SQL Reference

Amazon Kinesis Data Analytics SQL Reference

Amazon Kinesis Data Analytics SQL Reference: SQL Reference

Table of Contents

SQL Reference **1**
Streaming SQL Language Elements **2**
Identifiers 2
Data Types 3
Numeric Types and Precision 9
Streaming SQL Operators 11
IN Operator 12
EXISTS Operator 12
Scalar Operators 12
Arithmetic Operators 13
String Operators 14
Logical Operators 22
Expressions and Literals 29
Monotonic Expressions and Operators 32
Monotonic columns 33
Monotonic expressions 33
Rules for deducing monotonicity 34
Condition Clause 35
Temporal Predicates 36
Syntax 38
Example 39
Sample Use Case 40
Reserved Words and Keywords 40
Standard SQL Operators **48**
CREATE statements 48
CREATE STREAM 48
CREATE FUNCTION 50
CREATE PUMP 51
INSERT 53
Syntax 53
Pump Stream Insert 53
Query 54
Syntax 54
select 54

Streaming set operators ... 55
VALUES operator ... 56
SELECT statement ... 57
Syntax ... 57
The STREAM keyword and the principle of streaming SQL ... 57
SELECT ALL and SELECT DISTINCT ... 58
SELECT clause ... 60
FROM clause ... 63
JOIN clause ... 66
HAVING clause ... 83
GROUP BY clause ... 84
WHERE clause ... 87
WINDOW Clause (Sliding Windows) ... 87
ORDER BY clause ... 97
ROWTIME ... 100
Functions ... 103
Aggregate Functions ... 103
Streaming Aggregation and Rowtime Bounds ... 105
Aggregate Function List ... 105
Examples of Aggregate Queries on Streams (Streaming Aggregation) ... 107
Windowed Aggregation on Streams ... 110
AVG ... 116
COUNT ... 121
COUNT_DISTINCT_ITEMS_TUMBLING Function ... 125
EXP_AVG ... 128
FIRST_VALUE ... 129
LAST_VALUE ... 130
MAX ... 130
MIN ... 134
SUM ... 138
TOP_K_ITEMS_TUMBLING Function ... 142
Analytic Functions ... 145
Related Topics ... 146
Boolean Functions ... 146
ANY ... 146
EVERY ... 147

Conversion Functions ... 147
CAST ... 148
Date and Time Functions ... 170
Time Zones ... 171
Datetime Conversion Functions ... 171
Date, Timestamp, and Interval Operators ... 191
Date and Time Patterns ... 200
CURRENT_DATE ... 205
CURRENT_ROW_TIMESTAMP ... 205
CURRENT_TIME ... 206
CURRENT_TIMESTAMP ... 206
EXTRACT ... 207
LOCALTIME ... 208
LOCALTIMESTAMP ... 209
TSDIFF ... 209
Null Functions ... 210
COALESCE ... 210
NULLIF ... 211
Numeric Functions ... 211
ABS ... 212
CEIL / CEILING ... 213
EXP ... 214
FLOOR ... 215
LN ... 216
LOG10 ... 216
MOD ... 217
POWER ... 218
STEP ... 218
Log Parsing Functions ... 222
FAST_REGEX_LOG_PARSER ... 223
FIXED_COLUMN_LOG_PARSE ... 229
REGEX_LOG_PARSE ... 229
SYS_LOG_PARSE ... 233
VARIABLE_COLUMN_LOG_PARSE ... 233
W3C_LOG_PARSE ... 235
Sorting Functions ... 247

Group Rank ... 247
Statistical Variance and Deviation Functions ... 252
HOTSPOTS ... 253
RANDOM_CUT_FOREST ... 258
RANDOM_CUT_FOREST_WITH_EXPLANATION ... 265
STDDEV_POP ... 275
STDDEV_SAMP ... 279
VAR_POP ... 282
VAR_SAMP ... 286
Streaming SQL Functions ... 290
LAG ... 290
Monotonic Function ... 293
NTH_VALUE ... 294
String and Search Functions ... 294
CHAR_LENGTH / CHARACTER_LENGTH ... 295
INITCAP ... 296
LOWER ... 296
OVERLAY ... 297
POSITION ... 298
REGEX_REPLACE ... 299
SUBSTRING ... 301
TRIM ... 304
UPPER ... 305
Kinesis Data Analytics Developer Guide ... 307
Document History ... 308

Amazon Kinesis Data Analytics SQL Reference

> **Note**
>
> For new projects, we recommend that you use Kinesis Data Analytics Studio over Kinesis Data Analytics for SQL Applications. Kinesis Data Analytics Studio combines ease of use with advanced analytical capabilities, enabling you to build sophisticated stream processing applications in minutes.

The *Amazon Kinesis Data Analytics SQL Reference* describes the SQL language elements that are supported by Amazon Kinesis Data Analytics. The language is based on the SQL:2008 standard with some extensions to enable operations on streaming data.

For information about developing Kinesis Data Analytics applications, see the *Kinesis Data Analytics Developer Guide*.

This guide covers the following:

- Streaming SQL Language Elements – Data Types, Streaming SQL Operators, Functions.
- Standard SQL Operators – CREATE statements, SELECT statement.
- Operators for transforming and filtering incoming data – WHERE clause, JOIN clause, GROUP BY clause, WINDOW Clause (Sliding Windows).
- Logical Operators – AS, AND, OR, etc.

Streaming SQL Language Elements

The following topics discuss the language elements in Kinesis Data Analytics that underlie its syntax and operations:

Topics

- Identifiers
- Data Types
- Streaming SQL Operators
- Expressions and Literals
- Monotonic Expressions and Operators
- Condition Clause
- Temporal Predicates
- Reserved Words and Keywords

Identifiers

All identifiers may be up to 128 characters. Identifiers may be quoted (with case-sensitivity) by enclosing them in double-quote marks ("), or unquoted (with implicit uppercasing before both storage and lookup).

Unquoted identifiers must start with a letter or underscore, and be followed by letters, digits or underscores; letters are all converted to upper case.

Quoted identifiers can contain other punctuation too (in fact, any Unicode character except control characters: codes 0x0000 through 0x001F). You can include a double-quote in an identifier by escaping it with another double-quote.

In the following example, a stream is created with an unquoted identifier, which is converted to upper case before the stream definition is stored in the catalog. It can be referenced using its upper-case name, or by an unquoted identifier which is implicitly converted to upper case.

```
-- Create a stream. Stream name specified without quotes,
-- which defaults to uppercase.
CREATE OR REPLACE STREAM ExampleStream (col1 VARCHAR(4));

- example 1: OK, stream name interpreted as uppercase.
```

```
CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO ExampleStream
    SELECT * FROM SOURCE_SQL_STREAM_001;

- example 2: OK, stream name interpreted as uppercase.
CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO examplestream
    SELECT * FROM   customerdata;

- example 3: Ok.
CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO EXAMPLESTREAM
    SELECT * FROM   customerdata;

- example 2: Not found. Quoted names are case-sensitive.
CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "examplestream"
    SELECT * FROM   customerdata;
```

When objects are created in Amazon Kinesis Data Analytics, their names are implicitly quoted, so it is easy to create identifiers that contain lowercase characters, spaces, dashes, or other punctuation. If you reference those objects in SQL statements, you will need to quote their names.

Reserved Words and Keywords

Certain identifiers, called keywords, have special meaning if they occur in a particular place in a streaming SQL statement. A subset of these key words are called reserved words and may not be used as the name of an object, unless they are quoted. For more information, see Reserved Words and Keywords.

Data Types

The following table summarizes the data types supported by Amazon Kinesis Data Analytics.

SQL Data Type	JSON Data Type	Description	Notes
BIGINT	number	64-bit signed integer	
BINARY	BASE64-encoded string	Binary (non character) data	Substring works on BINARY. Concatena tion does not work on BINARY.

SQL Data Type	JSON Data Type	Description	Notes
BOOLEAN	boolean	TRUE, FALSE, or NULL	Evaluates to TRUE, FALSE, and UNKNOWN.
CHAR (n)	string	A character string of fixed length n. Also specifiable as CHARACTER	n must be greater than 0 and less than 65535.
DATE	string	A date is a calendar day (year/month/day).	Precision is day. Range runs from the largest value, approximately +229 (in years) to the smallest value, -229.
DECIMAL DEC NUMERIC	number	A fixed point, with up to 19 significant digits.	Can be specified with DECIMAL, DEC, or NUMERIC.
DOUBLE DOUBLE PRECISION	number	A 64-bit floating point number	64-bit approx value; -1.79E+308 to 1.79E+308. Follows the ISO DOUBLE PRECISION data type, 53 bits are used for the number's mantissa in scientific notation, representing 15 digits of precision and 8 bytes of storage.

SQL Data Type	JSON Data Type	Description	Notes
INTEGER INT	number		32-bit signed integer. Range is -2147483648 to 2147483647 [2**(31) to 2**(31)-1]
INTERVAL <timeunit> [TO <timeunit>]	string	Day-time intervals supported, year-month intervals not supported	Allowed in an expression in date arithmetic, but cannot be used as a datatype for a column in a table or stream.
<timeUnit>	string	The units of a INTERVAL value	Supported units are YEAR, MONTH, DAY, HOUR, MINUTE, and SECOND
SMALLINT	number	16-bit signed integer	Range is -32768 to 32767 [2**(15) to 2**(15)-1]
REAL	number	A 32-bit floating point number	Following the ISO REAL data type, 24 bits are used for the number's mantissa in scientific notation, representing 7 digits of precision and 4 bytes of storage. The minimum value is -3.40E+38; the maximum value is 3.40E+38.

SQL Data Type	JSON Data Type	Description	Notes
TIME	string	A TIME is a time in a day (hour:minute:second).	Its precision is milliseconds; its range is 00:00:00.000 to 23:59:59.999. Since the system clock runs in UTC, the timezone used for values stored in a TIME or TIMESTAMP column is not considered. for values stored in a TIME or TIMESTAMP column.

SQL Data Type	JSON Data Type	Description	Notes
TIMESTAMP	string	A TIMESTAMP is a combined DATE and TIME.	A TIMESTAMP value always has a precision of 1 millisecond. It has no particular timezone. Since the system clock runs in UTC, the timezone used for values stored in a TIME or TIMESTAMP column is not considered. Its range runs from the largest value, approximately +229 (in years) to the smallest value, -229. Each timestamp is stored as a signed 64-bit integer, with 0 representing the Unix epoch (Jan 1, 1970 00:00am). This means that the largest TIMESTAMP value represents approxima tely 300 million years after 1970, and the smallest value represents approximately 300 million years before 1970. Following the SQL standard, a TIMESTAMP value

SQL Data Type	JSON Data Type	Description	Notes
			has an undefined timezone.
TINYINT	number	8-bit signed integer	Range is -128 to 127,
VARBINARY (n)	BASE64-encoded string	Also specifiable as BINARY VARYING	n must be greater than 0 and less than 65535.
VARCHAR (n)	string	Also specifiable as CHARACTER VARYING	n must be greater than 0 and less than 65535.

Notes

Regarding characters:

- Amazon Kinesis Data Analytics supports only Java single-byte CHARACTER SETs.
- Implicit type conversion is not supported. That is, characters are mutually assignable if and only if they are taken from the same character repertoire and are values of the data types CHARACTER or CHARACTER VARYING.

Regarding numbers:

- Numbers are mutually comparable and mutually assignable if they are values of the data types NUMERIC, DECIMAL, INTEGER, BIGINT, SMALLINT, TINYINT, REAL, and DOUBLE PRECISION.

The following sets of data types are synonyms:

- DEC and DECIMAL
- DOUBLE PRECISION and DOUBLE
- CHARACTER and CHAR
- CHAR VARYING or CHARACTER VARYING and VARCHAR
- BINARY VARYING and VARBINARY
- INT and INTEGER

- Binary values (data types BINARY and BINARY VARYING) are always mutually comparable and are mutually assignable.

Regarding dates, times, and timestamps:

- Implicit type conversion is not supported (that is, datetime values are mutually assignable only if the source and target of the assignment are both of type DATE, or both of type TIME, or both of type TIMESTAMP).

- The Amazon Kinesis Data Analytics timezone is always UTC. The time functions, including the Amazon Kinesis Data Analytics extension CURRENT_ROW_TIMESTAMP, return time in UTC.

Numeric Types and Precision

For DECIMAL we support a maximum of 18 digits for precision and scale.

Precision specifies the maximum number of decimal digits that can be stored in the column, both to the right and to the left of the decimal point. You can specify precisions ranging from 1 digit to 18 digits or use the default precision of 18 digits.

Scale specifies the maximum number of digits that can be stored to the right of the decimal point. Scale must be less than or equal to the precision. You can specify a scale ranging from 0 digits to 18 digits, or use the default scale of 0 digits.

Rule for Divide

Let p1, s1 be the precision and scale of the first operand, such as DECIMAL (10,1).

Let p2, s2 be the precision and scale of the second operand, such as DECIMAL (10,3).

Let p, s be the precision and scale of the result.

Let d be the number of whole digits in the result. Then, the result type is a decimal as shown following:

`d = p1 - s1 + s2`	D = 10 - 1 + 3 Number of whole digits in result = 6

s <= MAX (6, s1 + p2 +1)	S <= MAX (6, 1 + 10 + 1) Scale of result = 14
p = d + s	Precision of result = 18

Precision and scale are capped at their maximum values (18, where scale cannot be larger than precision).

Precedence is first giving at least the scale of the first argument (s >= s1) followed by enough whole digits to represent the result without overflow

Rule for Multiply

Let p1, s1 be the precision and scale of the first operand DECIMAL (10,1).

Let p2, s2 be the precision and scale of the second operand DECIMAL (10,3).

Let p, s be the precision and scale of the result.

Then, the result type is a decimal as shown following:

p = p1 + p2	p = 10 + 10 Precision of result = 18
s = s1 + s2	s = 1 + 3 Scale of result = 4

Rule for Sum or Subtraction

Type-inference strategy whereby the result type of a call is the decimal sum of two exact numeric operands where at least one of the operands is a decimal.

Let p1, s1 be the precision and scale of the first operand DECIMAL (10,1).

Let p2, s2 be the precision and scale of the second operand DECIMAL (10,3).

Let p, s be the precision and scale of the result, as shown following:

s = max(s1, s2)	s = max (1,3) Scale of result = 3
p = max(p1 - s1, p2 - s2) + s + 1	p = max(10-1,10-3) + 3 + 1 Precision of result = 11

s and p are capped at their maximum values

Streaming SQL Operators

Subquery Operators

Operators are used in queries and subqueries to combine or test data for various properties, attributes, or relationships.

The available operators are described in the topics that follow, grouped into the following categories:

- Scalar Operators
  - Operator Types
  - Precedence

- Arithmetic Operators
- String Operators
  - (Concatenation)
  - LIKE patterns
  - SIMILAR TO patterns

- Date, Timestamp, and Interval Operators
- Logical Operators
  - 3-state boolean logic
  - Examples

IN Operator

As an operator in a condition test, IN tests a scalar or row value for membership in a list of values, a relational expression, or a subquery.

```
Examples:
1. --- IF column IN ('A','B','C')
2. --- IF (col1, col2) IN (
    select a, b from my_table
    )
```

Returns TRUE if the value being tested is found in the list, in the result of evaluating the relational expression, or in the rows returned by the subquery; returns FALSE otherwise.

> **Note**
>
> IN has a different meaning and use in CREATE FUNCTION.

EXISTS Operator

Tests whether a relational expression returns any rows; returns TRUE if any row is returned, FALSE otherwise.

Scalar Operators

Operator Types

The two general classes of scalar operators are:

- unary: A unary operator operates on only one operand. A unary operator typically appears with its operand in this format:

  ```
  operator operand
  ```

- binary: A binary operator operates on two operands. A binary operator appears with its operands in this format:

  ```
  operand1 operator operand2
  ```

A few operators that use a different format are noted specifically in the operand descriptions below.

If an operator is given a null operand, the result is almost always null (see the topic on logical operators for exceptions).

Precedence

Streaming SQL follows the usual precedence of operators:

1. Evaluate bracketed sub-expressions.
2. Evaluate unary operators (e.g., + or -, logical NOT).
3. Evaluate multiplication and divide (* and /).
4. Evaluate addition and subtraction (+ and -) and logical combination (AND and OR).

If one of the operands is NULL, the result is also NULL If the operands are of different but comparable types, the result will be of the type with the greatest precision. If the operands are of the same type, the result will be of the same type as the operands. For instance 5/2 = 2, not 2.5, as 5 and 2 are both integers.

Arithmetic Operators

Operator	Unary/Binary	Description
+	U	Identity
-	U	Negation
+	B	Addition
-	B	Subtraction
*	B	Multiplication
/	B	Division

Each of these operators works according to normal arithmetic behavior, with the following caveats:

1. If one of the operands is NULL, the result is also NULL
2. If the operands are of different but comparable types, the result will be of the type with the greatest precision.
3. If the operands are of the same type, the result will be of the same type as the operands. For instance 5/2 = 2, not 2.5, as 5 and 2 are both integers.

Examples

Operation	Result
1 + 1	2
2.0 + 2.0	4.0
3.0 + 2	5.0
5 / 2	2
5.0 / 2	2.500000000000
5*2+2	12

String Operators

You can use string operators for streaming SQL, including concatenation and string pattern comparison, to combine and compare strings.

Operator	Unary/Binary	Description	Notes
||	B	Concatenation	Also applies to binary types
LIKE	B	String pattern comparison	<string> LIKE <like pattern> [ESCAPE <escape character>]

Operator	Unary/Binary	Description	Notes
SIMILAR TO	B	String pattern comparison	<string> SIMILAR TO <similar to pattern> [ESCAPE <escape character>]

Concatenation

This operator is used to concatenate one or more strings as shown in the following table.

Operation	Result
'SQL'||'stream'	SQLstream
'SQL'||''||'stream'	SQLstream
'SQL'||'stream'||' Incorporated'	SQLstream Incorporated
<col1>||<col2>||<col3>||<col4>	<col1><col2><col3><col4>

LIKE patterns

LIKE compares a string to a string pattern. In the pattern, the characters _ (underscore) and % (percent) have special meaning.

Character in pattern	Effect
_	Matches any single character
%	Matches any substring, including the empty string
<any other character>	Matches only the exact same character

If either operand is NULL, the result of the LIKE operation is UNKNOWN.

To explicitly match a special character in the character string, you must specify an escape character using the ESCAPE clause. The escape character must then precede the special character in the pattern. The following table lists examples.

Operation	Result
'a' LIKE 'a'	TRUE
'a' LIKE 'A'	FALSE
'a' LIKE 'b'	FALSE
'ab' LIKE 'a_'	TRUE
'ab' LIKE 'a%'	TRUE
'ab' LIKE 'a_' ESCAPE '\'	FALSE
'ab' LIKE 'a\%' ESCAPE '\'	FALSE
'a_' LIKE 'a_' ESCAPE '\'	TRUE
'a%' LIKE 'a\%' ESCAPE '\'	TRUE
'a' LIKE 'a_'	FALSE
'a' LIKE 'a%'	TRUE
'abcd' LIKE 'a_'	FALSE
'abcd' LIKE 'a%'	TRUE
'' LIKE ''	TRUE
'1a' LIKE '_a'	TRUE
'123aXYZ' LIKE '%a%'	TRUE
'123aXYZ' LIKE '_%_a%_'	TRUE

SIMILAR TO patterns

SIMILAR TO compares a string to a pattern. It is much like the LIKE operator, but more powerful, as the patterns are regular expressions.

In the following SIMILAR TO table, *seq* means any sequence of characters explicitly specified, such as '13aq'. Non-alphanumeric characters intended for matching must be preceded by an escape character explicitly declared in the SIMILAR TO statement, such as '13aq\!' SIMILAR TO '13aq\!24br \!% ESCAPE '\' (This statement is TRUE).

When a range is indicated, as when a dash is used in a pattern, the current collating sequence is used. Typical ranges are 0-9 and a-z. PostgreSQL provides a typical discussion of pattern-matching, including ranges.

When a line requires multiple comparisons, the innermost pattern that can be matched will be matched first, then the "next-innermost," etc.

Expressions and matching operations that are enclosed within parentheses are evaluated before surrounding operations are applied, again by innermost-first precedence.

Delimiter	Character in pattern	Effect	Rule ID
parentheses ( )	(seq)	Groups the *seq* (used for defining precedence of pattern expressions)	1
brackets [ ]	[seq]	Matches any single character in the seq	2
caret or circumflex	[^seq]	Matches any single character not in the seq	3
	[seq ^ seq]	Matches any single character in seq and not in seq	4
dash	<character1>-<character2>	Specifies a range of characters between	5

Delimiter	Character in pattern	Effect	Rule ID
		character1 and character2 (using some known sequence like 1-9 or a-z)	
bar	[seq seq]	Matches either seq or seq	6
asterisk	seq*	Matches zero or more repetitions of seq	7
plus	seq+	Matches one or more repetitions of seq	8
braces	seq{<number>}	Matches exactly number repetitions of seq	9
	seq{<low number>,<high number>}	Matches low number or more repetitions of seq, to a maximum of high number	10
question-mark	seq?	Matches zero or one instances of seq	11
underscore	_	Matches any single character	12
percent	%	Matches any substring, including the empty string	13
character	<any other character>	Matches only the exact same character	14

Delimiter	Character in pattern	Effect	Rule ID
NULL	NULL	If either operand is NULL, the result of the SIMILAR TO operation is UNKNOWN.	15
Non-alphanumeric	Special characters	To explicitly match a special character in the character string, that special character must be preceded by an escape character defined using an ESCAPE clause specified at the end of the pattern.	16

The following table lists examples.

Operation	Result	Rule
'a' SIMILAR TO 'a'	TRUE	14
'a' SIMILAR TO 'A'	FALSE	14
'a' SIMILAR TO 'b'	FALSE	14
'ab' SIMILAR TO 'a_'	TRUE	12
'ab' SIMILAR TO 'a%'	TRUE	13
'a' SIMILAR TO 'a_'	FALSE	12 & 14
'a' SIMILAR TO 'a%'	TRUE	13

Operation	Result	Rule
'abcd' SIMILAR TO 'a_'	FALSE	12
'abcd' SIMILAR TO 'a%'	TRUE	13
'' SIMILAR TO ''	TRUE	14
'1a' SIMILAR TO '_a'	TRUE	12
'123aXYZ' SIMILAR TO ''	TRUE	14
'123aXYZ' SIMILAR TO '_%_a %_'	TRUE	13 & 12
'xy' SIMILAR TO '(xy)'	TRUE	1
'abd' SIMILAR TO '[ab][bcde]d'	TRUE	2
'bdd' SIMILAR TO '[ab][bcd e]d'	TRUE	2
'abd' SIMILAR TO '[ab]d'	FALSE	2
'cd' SIMILAR TO '[a-e]d'	TRUE	2
'cd' SIMILAR TO '[a-e^c]d'	FALSE	4
'cd' SIMILAR TO '[^(a-e)]d'	INVALID	
'yd' SIMILAR TO '[^(a-e)]d'	INVALID	
'amy' SIMILAR TO 'amyfred'	TRUE	6
'fred' SIMILAR TO 'amyfred'	TRUE	6
'mike' SIMILAR TO 'amyfred'	FALSE	6
'acd' SIMILAR TO 'ab*c+d'	TRUE	7 & 8
'accccd' SIMILAR TO 'ab*c+d'	TRUE	7 & 8

Operation	Result	Rule
'abd' SIMILAR TO 'ab*c+d'	FALSE	7 & 8
'aabc' SIMILAR TO 'ab*c+d'	FALSE	
'abb' SIMILAR TO 'a(b{3})'	FALSE	9
'abbb' SIMILAR TO 'a(b{3})'	TRUE	9
'abbbbb' SIMILAR TO 'a(b{3})'	FALSE	9
'abbbbb' SIMILAR TO 'ab{3,6}'	TRUE	10
'abbbbbbbb' SIMILAR TO 'ab{3,6}'	FALSE	10
'' SIMILAR TO 'ab?'	FALSE	11
'' SIMILAR TO '(ab)?'	TRUE	11
'a' SIMILAR TO 'ab?'	TRUE	11
'a' SIMILAR TO '(ab)?'	FALSE	11
'a' SIMILAR TO 'a(b?)'	TRUE	11
'ab' SIMILAR TO 'ab?'	TRUE	11
'ab' SIMILAR TO 'a(b?)'	TRUE	11
'abb' SIMILAR TO 'ab?'	FALSE	11
'ab' SIMILAR TO 'a_' ESCAPE '\'	FALSE	16
'ab' SIMILAR TO 'a\%' ESCAPE '\'	FALSE	16
'a_' SIMILAR TO 'a_' ESCAPE '\'	TRUE	16

Operation	Result	Rule
'a%' SIMILAR TO 'a\%' ESCAPE '\'	TRUE	16
'a(b{3})' SIMILAR TO 'a(b{3})'	FALSE	16
'a(b{3})' SIMILAR TO 'a\(b\{3\}\)' ESCAPE '\'	TRUE	16

Logical Operators

Logical operators let you establish conditions and test their results.

Operator	Unary/Binary	Description	Operands
NOT	U	Logical negation	Boolean
AND	B	Conjunction	Boolean
OR	B	Disjunction	Boolean
IS	B	Logical assertion	Boolean
IS NOT UNKNOWN	U	Negated unknown comparison: <expr> IS NOT UNKNOWN	Boolean
IS NULL	U	Null comparison: <expr> IS NULL	Any
IS NOT NULL	U	Negated null comparison: <expr> IS NOT NULL	Any

Operator	Unary/Binary	Description	Operands
=	B	Equality	Any
!=	B	Inequality	Any
<>	B	Inequality	Any
>	B	Greater than	Ordered types (Numeric, String, Date, Time)
>=	B	Greater than or equal to (not less than)	Ordered types
<	B	Less than	Ordered types
<=	B	Less than or equal to (not more than)	Ordered types
BETWEEN	Ternary	Range comparison: col1 BETWEEN expr1 AND expr2	Ordered types
IS DISTINCT FROM	B	Distinction	Any
IS NOT DISTINCT FROM	B	Negated distinction	Any

Three State Boolean Logic

SQL boolean values have three possible states rather than the usual two: TRUE, FALSE, and UNKNOWN, the last of which is equivalent to a boolean NULL. TRUE and FALSE operands generally function according to normal two-state boolean logic, but additional rules apply when pairing them with UNKNOWN operands, as the tables that follow will show.

> **Note**
>
> UNKOWN represents "maybe TRUE, maybe FALSE" or, to put it another way, "not definitely TRUE and not definitely FALSE." This understanding may help you clarify why some of the expressions in the tables evaluate as they do.

Negation (NOT)

Operation	Result
NOT TRUE	FALSE
NOT FALSE	TRUE
NOT UNKNOWN	UNKNOWN

Conjunction (AND)

Operation	Result
TRUE AND TRUE	TRUE
TRUE AND FALSE	FALSE
TRUE AND UNKNOWN	UNKNOWN
FALSE AND TRUE	FALSE
FALSE AND FALSE	FALSE
FALSE AND UNKNOWN	FALSE
UNKNOWN AND TRUE	UNKNOWN
UNKNOWN AND FALSE	FALSE
UNKNOWN AND UNKNOWN	UNKNOWN

Disjunction (OR)

Operation	Result
TRUE OR TRUE	TRUE
TRUE OR FALSE	TRUE
TRUE OR UNKNOWN	TRUE
FALSE OR TRUE	TRUE
FALSE OR FALSE	FALSE
FALSE OR UNKNOWN	UNKNOWN
UNKNOWN OR TRUE	TRUE
UNKNOWN OR FALSE	UNKNOWN
UNKNOWN OR UNKNOWN	UNKNOWN

Assertion (IS)

Operation	Result
TRUE IS TRUE	TRUE
TRUE IS FALSE	FALSE
TRUE IS UNKNOWN	FALSE
FALSE IS TRUE	FALSE
FALSE IS FALSE	TRUE
FALSE IS UNKNOWN	FALSE
UNKNOWN IS TRUE	FALSE
UNKNOWN IS FALSE	FALSE

Operation	Result
UNKNOWN IS UNKNOWN	TRUE

IS NOT UNKNOWN

Operation	Result
TRUE IS NOT UNKNOWN	TRUE
FALSE IS NOT UNKNOWN	TRUE
UNKNOWN IS NOT UNKNOWN	FALSE

IS NOT UNKNOWN is a special operator in and of itself. The expression "x IS NOT UNKNOWN" is equivalent to "(x IS TRUE) OR (x IS FALSE)", not "x IS (NOT UNKNOWN)". Thus, substituting in the table above:

x	Operation	Result		Result of substituting for x in "(x IS TRUE) OR (x IS FALSE)"
TRUE	TRUE IS NOT UNKNOWN	TRUE	becomes	"(TRUE IS TRUE) OR (TRUE IS FALSE)" -- hence TRUE
FALSE	FALSE IS NOT UNKNOWN	TRUE	becomes	"(FALSE IS TRUE) OR (FALSE IS FALSE)" -- hence TRUE
UNKNOWN	UNKNOWN IS NOT UNKNOWN	FALSE	becomes	"(UNKNOWN IS TRUE) OR (UNKNOWN IS

x	Operation	Result		Result of substituting for x in "(x IS TRUE) OR (x IS FALSE)"
				FALSE)" -- hence FALSE, since UNKNOWN is neither TRUE not FALSE

Since IS NOT UNKNOWN is a special operator, the operations above are not transitive around the word IS:

Operation	Result
NOT UNKNOWN IS TRUE	FALSE
NOT UNKNOWN IS FALSE	FALSE
NOT UNKNOWN IS UNKNOWN	TRUE

IS NULL and IS NOT NULL

Operation	Result
UNKNOWN IS NULL	TRUE
UNKNOWN IS NOT NULL	FALSE
NULL IS NULL	TRUE
NULL IS NOT NULL	FALSE

IS DISTINCT FROM and IS NOT DISTINCT FROM

Operation	Result
UNKNOWN IS DISTINCT FROM TRUE	TRUE
UNKNOWN IS DISTINCT FROM FALSE	TRUE
UNKNOWN IS DISTINCT FROM UNKNOWN	FALSE
UNKNOWN IS NOT DISTINCT FROM TRUE	FALSE
UNKNOWN IS NOT DISTINCT FROM FALSE	FALSE
UNKNOWN IS NOT DISTINCT FROM UNKNOWN	TRUE

Informally, "x IS DISTINCT FROM y" is similar to "x <> y", except that it is true even when either x or y (but not both) is NULL. DISTINCT FROM is the opposite of identical, whose usual meaning is that a value (true, false, or unknown) is identical to itself, and distinct from every other value. The IS and IS NOT operators treat UNKOWN in a special way, because it represents "maybe TRUE, maybe FALSE".

Other Logical Operators

For all other operators, passing a NULL or UNKNOWN operand will cause the result to be UNKNOWN (which is the same as NULL).

Examples

Operation	Result
TRUE AND CAST(NULL AS BOOLEAN)	UNKNOWN
FALSE AND CAST(NULL AS BOOLEAN)	FALSE
1 > 2	FALSE
1 < 2	TRUE

Operation	Result
'foo' = 'bar'	FALSE
'foo' <> 'bar'	TRUE
'foo' <= 'bar'	FALSE
'foo' <= 'bar'	TRUE
3 BETWEEN 1 AND 5	TRUE
1 BETWEEN 3 AND 5	FALSE
3 BETWEEN 3 AND 5	TRUE
5 BETWEEN 3 AND 5	TRUE
1 IS DISTINCT FROM 1.0	FALSE
CAST(NULL AS INTEGER) IS NOT DISTINCT FROM CAST (NULL AS INTEGER)	TRUE

Expressions and Literals

Value expressions

Value expressions are defined by the following syntax:

```
value-expression := <character-expression > | <number-expression> | <datetime-
expression> | <interval-expression> | <boolean-expression>
```

Character (string) expressions

Character expressions are defined by the following syntax:

```
character-expression := <character-literal>
                  | <character-expression> || <character-expression>
                  | <character-function> ( <parameters> )

 character-literal  := <quote> { <character> }* <quote>
```

```
 string-literal     := <quote> { <character> }* <quote>
 character-function   :=   CAST | COALESCE | CURRENT_PATH
                  | FIRST_VALUE  | INITCAP | LAST_VALUE
                  | LOWER | MAX | MIN | NULLIF
                  | OVERLAY | SUBSTRING| SYSTEM_USER
                  | TRIM | UPPER
                  | <user-defined-function>
```

Note that Amazon Kinesis Data Analytics streaming SQL supports unicode character literals, such as u&'foo'. As in the use of regular literals, you can escape single quotes in these, such as u&'can''t'. Unlike regular literals, you can have unicode escapes: e.g., u&'\0009' is a string consisting only of a tab character. You can escape a \ with another \, such as u&'back\\slash'. Amazon Kinesis Data Analytics also supports alternate escape characters, such as u&'!0009!!' uescape '!' is a tab character.

Numeric expressions

Numeric expressions are defined by the following syntax:

```
number-expression := <number-literal>
                  | <number-unary-oper> <number-expression>
                  | <number-expression> <number-operator> <number-expression>
                  | <number-function> [ ( <parameters> ) ]
 number-literal :=   <UNSIGNED_INTEGER_LITERAL> | <DECIMAL_NUMERIC_LITERAL>
                        | <APPROX_NUMERIC_LITERAL>
```

```
--Note: An <APPROX_NUMERIC_LITERAL> is a number in scientific notation, such as with
 an
--exponent, such as 1e2 or -1.5E-6.
number-unary-oper := + | -
 number-operator      :=   + | - | / | *

 number-function      :=   ABS | AVG | CAST | CEIL
                         | CEILING | CHAR_LENGTH
                         | CHARACTER_LENGTH | COALESCE
                         | COUNT | EXP | EXTRACT
                         | FIRST_VALUE
                         | FLOOR | LAST_VALUE
                         | LN | LOG10
                         | MAX | MIN  | MOD
                         | NULLIF
```

```
                          | POSITION | POWER
                          | SUM| <user-defined-function>
```

Date / Time expressions

Date / Time expressions are defined by the following syntax:

```
datetime-expression := <datetime-literal>
                          | <datetime-expression> [ + | - ] <number-expression>
                          | <datetime-function> [ ( <parameters> ) ]
 datetime-literal    :=  <left_brace> { <character-literal> } *  <right_brace>
                          |  <DATE> { <character-literal> } *
                          |  <TIME> { <character-literal> } *
                          |  <TIMESTAMP> { <character-literal> } *
 datetime-function   :=    CAST | CEIL | CEILING
                          | CURRENT_DATE | CURRENT_ROW_TIMESTAMP
                          | CURRENT_ROW_TIMESTAMP
                          | FIRST_VALUE| FLOOR
                          | LAST_VALUE | LOCALTIME
                          | LOCALTIMESTAMP | MAX | MIN
                          | NULLIF | ROWTIME
                          | <user-defined-function>
 <time unit>         :=  YEAR | MONTH | DAY | HOUR | MINUTE | SECOND
```

Interval Expression

Interval expressions are defined by the following syntax:

```
interval-expression := <interval-literal>
                          | <interval-function>
 interval-literal    :=    <INTERVAL> ( <MINUS> | <PLUS> ) <QUOTED_STRING>
 <IntervalQualifier>
 IntervalQualifier  :=  <YEAR> ( <UNSIGNED_INTEGER_LITERAL> )
                          | <YEAR> ( <UNSIGNED_INTEGER_LITERAL> ) <TO> <MONTH>
                          |  <MONTH> [ ( <UNSIGNED_INTEGER_LITERAL> ) ]
                          |  <DAY> [ (  <UNSIGNED_INTEGER_LITERAL> )  ]
                          |  <DAY> [ (  <UNSIGNED_INTEGER_LITERAL> )  ] <TO>
                                  { <HOUR> | <MINUTE> | <SECOND>
 [ ( <UNSIGNED_INTEGER_LITERAL> ) ] }
                          |  <HOUR> [ (  <UNSIGNED_INTEGER_LITERAL> ) ]
                          |  <HOUR> [ (  <UNSIGNED_INTEGER_LITERAL> ) ] <TO>
                                  { <MINUTE> | <SECOND>
 [ <UNSIGNED_INTEGER_LITERAL> ] }
```

```
                          |  <MINUTE> [ ( <UNSIGNED_INTEGER_LITERAL> )  ]
                          |  <MINUTE> [ ( <UNSIGNED_INTEGER_LITERAL> )  ] <TO>
                                    <SECOND> [ ( <UNSIGNED_INTEGER_LITERAL> ) ]
                          |  <SECOND> [ ( <UNSIGNED_INTEGER_LITERAL> )  ]
 interval-function   :=    ABS  | CAST | FIRST_VALUE
                          | LAST_VALUE | MAX | MIN
                          | NULLIF| <user-defined-function>
```

Boolean expression

Boolean expressions are defined by the following syntax:

```
boolean-expression := <boolean-literal>
                  | <boolean-expression> <boolean-operator> <boolean-expression>
                  | <boolean-unary-oper> <boolean-expression>
                        | <boolean-function> ( <parameters> )
                        | ( <boolean-expression> )
 boolean-literal    :=  TRUE | FALSE
 boolean-operator   :=  AND | OR
 boolean-unary-oper :=  NOT
 boolean-function   :=    CAST | FIRST_VALUE | LAST_VALUE
                        | NULLIF | <user-defined-function>
```

Monotonic Expressions and Operators

Since Amazon Kinesis Data Analytics queries operate on infinite streams of rows, some operations are only possible if something is known about those streams.

For example, given a stream of orders, it makes sense to ask for a stream summarizing orders by day and product (because day is increasing) but not to ask for a stream summarizing orders by product and shipping state. We can never complete the summary of, say Widget X to Oregon, because we never see the 'last' order of a Widget to Oregon.

This property, of a stream being sorted by a particular column or expression, is called monotonicity.

Some time-related definitions:

- **Monotonic**. An expression is monotonic if it is ascending or descending. An equivalent phrasing is "non-decreasing or non-increasing."
- **Ascending**. An expression e is ascending within a stream if the value of e for a given row is always greater than or equal to the value in the previous row.

- **Descending**. An expression e is descending within a stream if the value of e for a given row is always less than or equal to the value in the previous row.
- **Strictly Ascending**. An expression e is strictly ascending within a stream if for the value of e for a given row is always greater than the value in the previous row.
- **Strictly Descending**. An expression e is strictly descending within a stream if the value of e for a given row is always less than the value in the previous row.
- **Constant**. An expression e is constant within a stream if the value of e for a given row is always equal to the value in the previous row.

Note that by this definition, a constant expression is considered monotonic.

Monotonic columns

The ROWTIME system column is ascending. The ROWTIME column is not strictly ascending: it is acceptable for consecutive rows to have the same timestamp.

Amazon Kinesis Data Analytics prevents a client from inserting a row into a stream whose timestamp is less than the previous row it wrote into the stream. Amazon Kinesis Data Analytics also ensures that if multiple clients are inserting rows into the same stream, the rows are merged so that the ROWTIME column is ascending.

Clearly it would be useful to assert, for instance, that the orderId column is ascending; or that no orderId is ever more than 100 rows from sorted order. However, declared sort keys are not supported in the current release.

Monotonic expressions

Amazon Kinesis Data Analytics can deduce that an expression is monotonic if it knows that its arguments are monotonic. (See also the Monotonic Function.)

Another definition:

Functions or operators that are monotonic

A function or operator is monotonic if, when applied to a strictly increasing sequence of values, it yields a monotonic sequence of results.

For example, the FLOOR function, when applied to the ascending inputs {1.5, 3, 5, 5.8, 6.3}, yields {1, 3, 5, 5, 6}. Note that the input is strictly ascending, but the output is merely ascending (includes duplicate values).

Rules for deducing monotonicity

Amazon Kinesis Data Analytics requires that one or more grouping expressions are valid in order for a streaming GROUP BY statement to be valid. In other cases, Amazon Kinesis Data Analytics may be able to operate more efficiently if it knows about monotonicity; for example it may be able to remove entries from a table of windowed aggregate totals if it knows that a particular key will never be seen on the stream again.

In order to exploit monotonicity in this way, Amazon Kinesis Data Analytics uses a set of rules for deducing the monotonicity of an expression. Here are the rules for deducing monotonicity:

Expression	Monotonicity
c	Constant
FLOOR(m)	Same as m, but not strict
CEIL / CEILING(m)	Same as m, but not strict
CEIL / CEILING(m TO timeUnit)	Same as m, but not strict
FLOOR(m TO timeUnit)	Same as m, but not strict
SUBSTRING(m FROM 0 FOR c)	Same as m, but not strict
+ m	Same as m
- m	Reverse of m
m + c c + m	Same as m
m1 + m2	Same as m1, if m1 and m2 have same direction; otherwise not monotonic
c - m	Reverse of m
m * c	Same as m if c is positive;

Expression	Monotonicity
c * m	reverse of m is c is negative; constant (0) c is 0
c / m	Same as m if m is always positive or always negative, and c and m have same sign; reverse of m if m is always positive or always negative, and c and m have different sign; otherwise not monotonic
	Constant
LOCALTIME LOCALTIMESTAMP CURRENT_ROW_TIMESTAMP CURRENT_DATE	Ascending

Throughout the table, c is a constant, and m (also m1 and m2) is a monotonic expression.

Condition Clause

Referenced by:

- SELECT clauses: HAVING clause, WHERE clause, and JOIN clause. (See also the SELECT chart and its SELECT clause.)
- DELETE

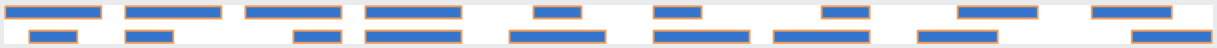
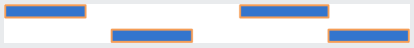
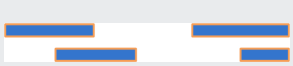
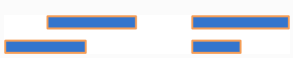
A condition is any value expression of type BOOLEAN, such as the following examples:

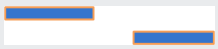
- 2<4
- TRUE
- FALSE
- expr_17 IS NULL

- NOT expr_19 IS NULL AND expr_23 < expr>29
- expr_17 IS NULL OR (NOT expr_19 IS NULL AND expr_23 < expr>29)

Temporal Predicates

The following table shows a graphic representation of temporal predicates supported by standard SQL and extensions to the SQL standard supported by Amazon Kinesis Data Analytics. It shows the relationships that each predicate covers. Each relationship is represented as an upper interval and a lower interval with the combined meaning *upperInterval predicate lowerInterval evaluates to TRUE*. The first 7 predicates are standard SQL. The last 10 predicates, shown in bold text, are Amazon Kinesis Data Analytics extensions to the SQL standard.

Pre	Covered Relationships
COI	
OVI	
EQU	
PRE	
SUC	
IMM LY PRE	
IMM LY SUC	
LEA	
LAC	

Pre	Covered Relationships
STF CO	
STF OV	
STF PRE	
STF SU(	
STF LEA	
STF LA(	
IMM LY LEA	
IMM LY LA(	

To enable concise expressions, Amazon Kinesis Data Analytics also supports the following extensions:

- Optional PERIOD keyword – The PERIOD keyword can be omitted.
- Compact chaining – If two of these predicates occur back to back, separated by an AND, the AND can be omitted provided that the right interval of the first predicate is identical to the left interval of the second predicate.
- TSDIFF – This function takes two TIMESTAMP arguments and returns their difference in milliseconds.

For example, you can write the following expression:

```
  PERIOD (s1,e1) PRECEDES PERIOD(s2,e2)
  AND PERIOD(s2, e2) PRECEDES PERIOD(s3,e3)
```

More concisely as follows:

```
(s1,e1) PRECEDES (s2,e2) PRECEDES PERIOD(s3,e3)
```

The following concise expression:

```
TSDIFF(s,e)
```

Means the following:

```
CAST((e - s) SECOND(10, 3) * 1000 AS BIGINT)
```

Finally, standard SQL allows the CONTAINS predicate to take a single TIMESTAMP as its right-hand argument. For example, the following expression:

```
PERIOD(s, e) CONTAINS t
```

Is equivalent to the following:

```
s <= t AND t < e
```

Syntax

Temporal predicates are integrated into a new BOOLEAN valued expression:

```
<period-expression> :=
  <left-period> <half-period-predicate> <right-period>

<half-period-predicate> :=
  <period-predicate> [ <left-period> <half-period-predicate> ]

<period-predicate> :=
   EQUALS
 | [ STRICTLY ] CONTAINS
 | [ STRICTLY ] OVERLAPS
```

```
 | [ STRICTLY | IMMEDIATELY ] PRECEDES
 | [ STRICTLY | IMMEDIATELY ] SUCCEEDS
 | [ STRICTLY | IMMEDIATELY ] LEADS
 | [ STRICTLY | IMMEDIATELY ] LAGS

<left-period> := <bounded-period>

<right-period> := <bounded-period> | <timestamp-expression>

<bounded-period> := [ PERIOD ] ( <start-time>, <end-time> )

<start-time> := <timestamp-expression>

<end-time> := <timestamp-expression>

<timestamp-expression> :=
  an expression which evaluates to a TIMESTAMP value

where <right-period> may evaluate to a <timestamp-expression> only if
the immediately preceding <period-predicate> is [ STRICTLY ] CONTAINS
```

This Boolean expression is supported by the following builtin function:

```
BIGINT tsdiff( startTime TIMESTAMP, endTime TIMESTAMP )
```

Returns the value of (endTime - startTime) in milliseconds.

```
```

Example

The following example code records an alarm if a window is open while the air conditioning is on:

```
create or replace pump alarmPump stopped as
  insert into alarmStream( houseID, roomID, alarmTime, alarmMessage )
    select stream w.houseID, w.roomID, current_timestamp,
                  'Window open while air conditioner is on.'
    from
        windowIsOpenEvents over (range interval '1' minute preceding) w
    join
        acIsOnEvents over (range interval '1' minute preceding) h
    on w.houseID = h.houseID
    where (h.startTime, h.endTime) overlaps (w.startTime, w.endTime);
```

Sample Use Case

The following query uses a temporal predicate to raise a fraud alarm when two people try to use the same credit card simultaneously at two different locations:

```
create pump creditCardFraudPump stopped as
 insert into alarmStream
  select stream
    current_timestamp, creditCardNumber, registerID1, registerID2
  from transactionsPerCreditCard
  where registerID1 <> registerID2
  and (startTime1, endTime1) overlaps (startTime2, endTime2)
;
```

The preceding code example uses an input stream with the following dataset:

```
(current_timestamp  TIMESTAMP,
  creditCardNumber  VARCHAR(16),
  registerID1       VARCHAR(16),
  registerID2       VARCHAR(16),
  startTime1        TIMESTAMP,
  endTime1          TIMESTAMP,
  startTime2        TIMESTAMP,
  endTime2          TIMESTAMP)
```

Reserved Words and Keywords

Reserved Words

The following is a list of reserved words in Amazon Kinesis Data Analytics applications as of version 5.0.1.

A		
ABS	ALL	ALLOCATE
ALLOW	ALTER	ANALYZE

AND	ANY	APPROXIMATE_ARRIVAL_TIME
ARE	ARRAY	AS
ASENSITIVE	ASYMMETRIC	AT
ATOMIC	AUTHORIZATION	AVG
	B	
BEGIN	BETWEEN	BIGINT
BINARY	BIT	BLOB
BOOLEAN	BOTH	BY
	C	
CALL	CALLED	CARDINALITY
CASCADED	CASE	CAST
CEIL	CEILING	CHAR
CHARACTER	CHARACTER_LENGTH	CHAR_LENGTH
CHECK	CHECKPOINT	CLOB
CLOSE	CLUSTERED	COALESCE
COLLATE	COLLECT	COLUMN
COMMIT	CONDITION	CONNECT
CONSTRAINT	CONVERT	CORR
CORRESPONDING	COUNT	COVAR_POP
COVAR_SAMP	CREATE	CROSS
CUBE	CUME_DIST	CURRENT

CURRENT_CATALOG	CURRENT_DATE	CURRENT_DEFAULT_TRANSFORM_GROUP
CURRENT_PATH	CURRENT_ROLE	CURRENT_SCHEMA
CURRENT_TIME	CURRENT_TIMESTAMP	CURRENT_TRANSFORM_GROUP_FOR_TYPE
CURRENT_USER	CURSOR	CYCLE
	D	
DATE	DAY	DEALLOCATE
DEC	DECIMAL	DECLARE
DEFAULT	DELETE	DENSE_RANK
DEREF	DESCRIBE	DETERMINISTIC
DISALLOW	DISCONNECT	DISTINCT
DOUBLE	DROP	DYNAMIC
	E	
EACH	ELEMENT	ELSE
END	END-EXEC	ESCAPE
EVERY	EXCEPT	EXEC
EXECUTE	EXISTS	EXP
EXPLAIN	EXP_AVG	EXTERNAL
EXTRACT		
	F	
FALSE	FETCH	FILTER

FIRST_VALUE	FLOAT	FLOOR
FOR	FOREIGN	FREE
FROM	FULL	FUNCTION
FUSION		
	G	
GET	GLOBAL	GRANT
GROUP	GROUPING	
	H	
HAVING	HOLD	HOUR
	I	
IDENTITY	IGNORE	IMPORT
IN	INDICATOR	INITCAP
INNER	INOUT	INSENSITIVE
INSERT	INT	INTEGER
INTERSECT	INTERSECTION	INTERVAL
INTO	IS	
	J	
JOIN		
	L	
LANGUAGE	LARGE	LAST_VALUE
LATERAL	LEADING	LEFT

LIKE	LIMIT	LN
LOCAL	LOCALTIME	LOCALTIMESTAMP
LOWER		
	M	
MATCH	MAX	MEMBER
MERGE	METHOD	MIN
MINUTE	MOD	MODIFIES
MODULE	MONTH	MULTISET
	N	
NATIONAL	NATURAL	NCHAR
NCLOB	NEW	NO
NODE	NONE	NORMALIZE
NOT	NTH_VALUE	NULL
NULLIF	NUMERIC	
	O	
OCTET_LENGTH	OF	OLD
ON	ONLY	OPEN
OR	ORDER	OUT
OUTER	OVER	OVERLAPS
OVERLAY		
	P	

PARAMETER	PARTITION	PARTITION_ID
PARTITION_KEY	PERCENTILE_CONT	PERCENTILE_DISC
PERCENT_RANK	POSITION	POWER
PRECISION	PREPARE	PRIMARY
PROCEDURE		
	R	
RANGE	RANK	READS
REAL	RECURSIVE	REF
REFERENCES	REFERENCING	REGR_AVGX
REGR_AVGY	REGR_COUNT	REGR_INTERCEPT
REGR_R2	REGR_SLOPE	REGR_SXX
REGR_SXY	RELEASE	RESPECT
RESULT	RETURN	RETURNS
REVOKE	RIGHT	ROLLBACK
ROLLUP	ROW	ROWS
ROWTIME	ROW_NUMBER	
	S	
SAVEPOINT	SCOPE	SCROLL
SEARCH	SECOND	SELECT
SENSITIVE	SEQUENCE_NUMBER	SESSION_USER
SET	SHARD_ID	SIMILAR

SMALLINT	SOME	SORT
SPECIFIC	SQL	SQLEXCEPTION
SQLSTATE	SQLWARNING	SQRT
START	STATIC	STDDEV
STDDEV_POP	STDDEV_SAMP	STOP
STREAM	SUBMULTISET	SUBSTRING
SUM	SYMMETRIC	SYSTEM
SYSTEM_USER		
	T	
TABLE	TABLESAMPLE	THEN
TIME	TIMESTAMP	TIMEZONE_HOUR
TIMEZONE_MINUTE	TINYINT	TO
TRAILING	TRANSLATE	TRANSLATION
TREAT	TRIGGER	TRIM
TRUE	TRUNCATE	
	U	
UESCAPE	UNION	UNIQUE
UNKNOWN	UNNEST	UPDATE
UPPER	USER	USING
	V	
VALUE	VALUES	VARBINARY

VARCHAR	VARYING	VAR_POP
VAR_SAMP		
	W	
WHEN	WHENEVER	WHERE
WIDTH_BUCKET	WINDOW	WITH
WITHIN	WITHOUT	
	Y	
YEAR		

Standard SQL Operators

The following topics discuss standard SQL operators:

Topics

- CREATE statements
- INSERT
- Query
- SELECT statement

CREATE statements

You can use the following CREATE statements with Amazon Kinesis Data Analytics:

- CREATE FUNCTION
- CREATE PUMP
- CREATE STREAM

CREATE STREAM

The CREATE STREAM statement creates a (local) stream. The name of the stream must be distinct from the name of any other stream in the same schema. It is good practice to include a description of the stream.

Like tables, streams have columns, and you specify the data types for these in the CREATE STREAM statement. These should map to the data source for which you are creating the stream. For column_name, any valid non-reserved SQL name is usable. Column values cannot be null.

- Specifying OR REPLACE re-creates the stream if it already exists, enabling a definition change for an existing object, implicitly dropping it without first needing to use a DRP command. Using CREATE OR REPLACE on a stream that already has data in flight kills the stream and loses all history.
- RENAME can be specified only if OR REPLACE has been specified.

- For the complete list of types and values in type_specification, such as TIMESTAMP, INTEGER, or varchar(2), see the topic Amazon Kinesis Data Analytics Data Types in the Amazon Kinesis Data Analytics SQL Reference Guide.
- For option_value, any string can be used.

Simple stream for unparsed log data

```
CREATE OR REPLACE STREAM logStream (
    source  VARCHAR(20),
    message VARCHAR(3072))
DESCRIPTION 'Head of webwatcher stream processing';
```

Stream capturing sensor data from Intelligent Travel System pipeline

```
CREATE OR REPLACE STREAM "LaneData" (
    -- ROWTIME is time at which sensor data collected
    LDS_ID  INTEGER,        -- loop-detector ID
    LNAME   VARCHAR(12),
    LNUM    VARCHAR(4),
    OCC     SMALLINT,
    VOL     SMALLINT,
    SPEED   DECIMAL(4,2)
) DESCRIPTION 'Conditioned LaneData for analysis queries';
```

Stream capturing order data from e-commerce pipeline

```
CREATE OR REPLACE STREAM "OrderData" (
    "key_order"    BIGINT NOT NULL,
    "key_user"     BIGINT,
    "country"      SMALLINT,
    "key_product"  INTEGER,
    "quantity"     SMALLINT,
    "eur"          DECIMAL(19,5),
    "usd"          DECIMAL(19,5)
) DESCRIPTION 'conditioned order data, ready for analysis';
```

CREATE FUNCTION

Amazon Kinesis Data Analytics provides a number of Functions, and also allows users to extend its capabilities by means of user-defined functions (UDFs). Amazon Kinesis Data Analytics supports UDFs defined in SQL only.

User-defined functions may be invoked using either the fully-qualified name or by the function name alone.

Values passed to (or returned from) a user-defined function or transformation must be exactly the same data types as the corresponding parameter definitions. In other words, implicit casting is not allowed in passing parameters to (or returning values from) a user-defined function.

User-Defined Function (UDF)

A user-defined function can implement complex calculations, taking zero or more scalar parameters and returning a scalar result. UDFs operate like built-in functions such as FLOOR() or LOWER(). For each occurrence of a user-defined function within a SQL statement, that UDF is called once per row with scalar parameters: constants or column values in that row.

Syntax

```
CREATE FUNCTION ''<function_name>'' ( ''<parameter_list>'' )
 RETURNS ''<data type>''
 LANGUAGE SQL
 [ SPECIFIC ''<specific_function_name>''  | [NOT] DETERMINISTIC ]
 CONTAINS SQL
 [ READS SQL DATA ]
 [ MODIFIES SQL DATA ]
 [ RETURNS NULL ON NULL INPUT | CALLED ON NULL INPUT ]
 RETURN ''<SQL-defined function body>''
```

SPECIFIC assigns a specific function name that is unique within the application. Note that the regular function name does not need to be unique (two or more functions may share the same name, as long as they are distinguishable by their parameter list).

DETERMINISTIC / NOT DETERMINISTIC indicates whether a function will always return the same result for a given set of parameter values. This may be used by your application for query optimization.

READS SQL DATA and MODIFIES SQL DATA indicate whether the function potentially reads or modifies SQL data, respectively. If a function attempts to read data from tables or streams without READS SQL DATA being specified, or insert to a stream or modify a table without MODIFIES SQL DATA being specified, an exception will be raised.

RETURNS NULL ON NULL INPUT and CALLED ON NULL INPUT indicate whether the function is defined as returning null if any of its parameters are null. If left unspecified, the default is CALLED ON NULL INPUT.

A SQL-defined function body consists only of a single RETURN statement.

Examples

```
CREATE FUNCTION get_fraction( degrees DOUBLE )
    RETURNS DOUBLE
    CONTAINS SQL
    RETURN degrees - FLOOR(degrees)
;
```

CREATE PUMP

A pump is an Amazon Kinesis Data Analytics Repository Object (an extension of the SQL standard) that provides a continuously running INSERT INTO stream SELECT ... FROM query functionality, thereby enabling the results of a query to be continuously entered into a named stream.

You need to specify a column list for both the query and the named stream (these imply a set of source-target pairs). The column lists need to match in terms of datatype, or the SQL validator will reject them. (These need not list all columns in the target stream; you can set up a pump for one column.)

For more information, see SELECT statement.

The following code first creates and sets a schema, then creates two streams in this schema:

- "OrderDataWithCreateTime" which will serve as the origin stream for the pump.
- "OrderData" which will serve as the destination stream for the pump.

```
CREATE OR REPLACE STREAM "OrderDataWithCreateTime" (
"key_order" VARCHAR(20),
```

```
"key_user" VARCHAR(20),
"key_billing_country" VARCHAR(20),
"key_product" VARCHAR(20),
"quantity" VARCHAR(20),
"eur" VARCHAR(20),
"usd" VARCHAR(20))
DESCRIPTION 'Creates origin stream for pump';

CREATE OR REPLACE STREAM "OrderData" (
"key_order" VARCHAR(20),
"key_user" VARCHAR(20),
"country" VARCHAR(20),
"key_product" VARCHAR(20),
"quantity" VARCHAR(20),
"eur" INTEGER,
"usd" INTEGER)
DESCRIPTION 'Creates destination stream for pump';
```

The following code uses these two streams to create a pump. Data is selected from "OrderDataWithCreateTime" and inserted into "OrderData".

```
CREATE OR REPLACE PUMP "200-ConditionedOrdersPump" AS
INSERT INTO "OrderData" (
"key_order", "key_user", "country",
"key_product", "quantity", "eur", "usd")
//note that this list matches that of the query
SELECT STREAM
"key_order", "key_user", "key_billing_country",
"key_product", "quantity", "eur", "usd"
//note that this list matches that of the insert statement
FROM "OrderDataWithCreateTime";
```

For more detail, see the topic In-Application Streams and Pumps in the *Amazon Managed Service for Apache Flink Developer Guide*.

Syntax

```
CREATE [ OR REPLACE ] PUMP <qualified-pump-name>
                      [ DESCRIPTION '<string-literal>' ] AS <streaming-insert>
```

where streaming-insert is an insert statement such as:

```
INSERT INTO ''stream-name'' SELECT "columns" FROM <source stream>
```

INSERT

INSERT is used to insert rows into a stream. It can also be used in a pump to insert the output of one stream into another.

Syntax

```
<insert statement> :=
  INSERT [ EXPEDITED ]
  INTO  <table-name > [ ( insert-column-specification ) ]
  <  query  >
<insert-column-specification> := < simple-identifier-list >
<simple-identifier-list> :=
   <simple-identifier> [ , < simple-identifier-list > ]
```

For a discussion of VALUES, see SELECT statement.

Pump Stream Insert

INSERT may also be specified as part of a CREATE PUMP statement.

```
CREATE PUMP "HighBidsPump" AS INSERT INTO "highBids" ( "ticker", "shares", "price")
SELECT  "ticker", "shares", "price"
FROM SALES.bids
WHERE "shares"*"price">100000
```

Here the results to be inserted into the "highBids" stream should come from a UNION ALL expression that evaluates to a stream. This will create a continuously running stream insert. Rowtimes of the rows inserted will be inherited from the rowtimes of the rows output from the select or UNION ALL. Again rows may be initially dropped if other inserters, ahead of this inserter, have inserted rows with rowtimes later than those initially prepared by this inserter, since the latter would then be out of time order. See the topic CREATE PUMP in this guide.

Query

Syntax

```
<query> :=
    <select>
  | <query> <set-operator> [ ALL ] <query>
  | VALUES <row-constructor> { , <row-constructor> }...
  | '(' <query> ')'
 <set-operator> :=
    EXCEPT
  | INTERSECT
  | UNION
 <row-constructor> :=
    [ ROW ] ( <expression> { , <expression> }... )
```

select

The select box in the chart above represents any SELECT command; that command is described in detail on its own page.

Set operators (EXCEPT, INTERSECT, UNION)

Set operators combine rows produced by queries using set operations:

- EXCEPT returns all rows that are in the first set but not in the second
- INTERSECT returns all rows that are in both first and second sets
- UNION returns all rows that are in either set

In all cases, the two sets must have the same number of columns, and the column types must be assignment-compatible. The column names of the resulting relation are the names of the columns of the first query.

With the ALL keyword, the operators use the semantics of a mathematical Multiset , meaning that duplicate rows are not eliminated. For example, if a particular row occurs 5 times in the first set and 2 times in the second set, then UNION ALL will emit the row 3 + 2 = 5 times.

ALL is not currently supported for EXCEPT or INTERSECT.

All operators are left-associative, and INTERSECT has higher precedence than EXCEPT or UNION, which have the same precedence. To override default precedence, you can use parentheses. For example:

```
SELECT * FROM a
UNION
SELECT * FROM b
INTERSECT
SELECT * FROM c
EXCEPT
SELECT * FROM d
EXCEPT
SELECT * FROM E
```

is equivalent to the fully-parenthesized query

```
( ( SELECT * FROM a
    UNION
    ( SELECT * FROM b
      INTERSECT
      SELECT * FROM c) )
  EXCEPT
  SELECT * FROM d )
EXCEPT
SELECT * FROM e
```

Streaming set operators

UNION ALL is the only set operator that can be applied to streams. Both sides of the operator must be streams; it is an error if one side is a stream and the other is a relation.

For example, the following query produces a stream of orders taken over the phone or via the web:

```
SELECT STREAM *
  FROM PhoneOrders
UNION ALL
SELECT STREAM *
  FROM WebOrders
```

Rowtime generation. The rowtime of a row emitted from streaming UNION ALL is the same as the timestamp of the input row.

Rowtime bounds: A rowtime bound is an assertion about the future contents of a stream. It states that the next row in the stream will have a ROWTIME no earlier than the value of the bound. For example, if a rowtime bound is 2018-12-0223:23:07, this tells the system that the next row will arrive no earlier than 2018-12-0223:23:07. Rowtime bounds are useful in managing gaps in data flow, such as those left overnight on a stock exchange.

Amazon Kinesis Data Analytics ensures that the ROWTIME column is ascending by merging the incoming rows on the basis of the time stamp. If the first set has rows that are timestamped 10:00 and 10:30, and the second set has only reached 10:15, Kinesis Data Analytics pauses the first set and waits for the second set to reach 10:30. In this case, it would be advantageous if the producer of the second set were to send a rowtime bound.

VALUES operator

The VALUES operator expresses a constant relation in a query. (See also the discussion of VALUES in the topic SELECT in this guide.)

VALUES can be used as a top-level query, as follows:

```
VALUES 1 + 2 > 3;
EXPR$0
======
FALSE
VALUES
    (42, 'Fred'),
    (34, 'Wilma');
EXPR$0 EXPR$1
====== ======
    42 Fred
    34 Wilma
```

Note that the system has generated arbitrary column names for anonymous expressions. You can assign column names by putting VALUES into a subquery and using an AS clause:

```
SELECT *
FROM (
    VALUES
        (42, 'Fred'),
        (34, 'Wilma')) AS t (age, name);
AGE NAME
```

```
=== =====
 42 Fred
 34 Wilma
```

SELECT statement

SELECT retrieves rows from streams. You can use SELECT as a top-level statement, or as part of a query involving set operations, or as part of another statement, including (for example) when passed as a query into a UDX. For examples, see the topics INSERT, IN, EXISTS, CREATE PUMP in this guide.

The subclauses of the SELECT statement are described in the topics SELECT clause, GROUP BY clause, Streaming GROUP BY, ORDER BY clause, HAVING clause, WINDOW Clause (Sliding Windows) and WHERE clause in this guide.

Syntax

```
<select> :=
   SELECT [ STREAM] [ DISTINCT | ALL ]
   <select-clause>
   FROM <from-clause>
   [ <where-clause> ]
   [ <group-by-clause> ]
   [ <having-clause> ]
   [ <window-clause> ]
   [ <order-by-clause> ]
```

The STREAM keyword and the principle of streaming SQL

The SQL query language was designed for querying stored relations, and producing finite relational results.

The foundation of streaming SQL is the STREAM keyword, which tells the system to compute the time differential of a relation. The time differential of a relation is the change of the relation with respect to time. A streaming query computes the change in a relation with respect to time, or the change in an expression computed from several relations.

To ask for the time-differential of a relation in Amazon Kinesis Data Analytics, we use the STREAM keyword:

```
SELECT STREAM * FROM Orders
```

If we start running that query at 10:00, it will produce rows at 10:15 and 10:25. At 10:30 the query is still running, waiting for future orders:

```
ROWTIME  orderId custName   product quantity
======== ======= ========== ======= ========
10:15:00     102 Ivy Black  Rice           6
10:25:00     103 John Wu    Apples         3
```

Here, the system is saying 'At 10:15:00 I executed the query SELECT * FROM Orders and found one row in the result that was not present at 10:14:59.999'. It generates the row with a value of 10:15:00 in the ROWTIME column because that is when the row appeared. This is the core idea of a stream: a relation that keeps updating over time.

You can apply this definition to more complicated queries. For example, the stream

```
SELECT STREAM * FROM Orders WHERE quantity > 5
```

has a row at 10:15 but no row at 10:25, because the relation

```
SELECT * FROM Orders WHERE quantity > 5
```

goes from empty to one row when order 102 is placed at 10:15, but is not affected when order 103 is placed at 10:25.

We can apply the same logic to queries involving any combination of SQL operators. Queries involving JOIN, GROUP BY, subqueries, set operations UNION, INTERSECT, EXCEPT, and even qualifiers such as IN and EXISTS, are well-defined when converted to streams. Queries combining streams and stored relations are also well-defined.

SELECT ALL and SELECT DISTINCT

If the ALL keyword is specified, the query does not eliminate duplicate rows. This is the default behavior if neither ALL nor DISTINCT is specified.

If the DISTINCT keyword is specified, a query eliminates rows that are duplicates according to the columns in the SELECT clause.

Note that for these purposes, the value NULL is considered equal to itself and not equal to any other value. These are the same semantics as for GROUP BY and the IS NOT DISTINCT FROM operator.

Streaming SELECT DISTINCT

SELECT DISTINCT can be used with streaming queries as long as there is a non-constant monotonic expression in the SELECT clause. (The rationale for the non-constant monotonic expression is the same as for streaming GROUP BY.) Amazon Kinesis Data Analytics emits rows for SELECT DISTINCT as soon as they are ready.

If ROWTIME is one of the columns in the SELECT clause, it is ignored for the purposes of duplicate-elimination. Duplicates are eliminated on the basis of the other columns in the SELECT clause.

For example:

```
SELECT STREAM DISTINCT ROWTIME, prodId, FLOOR(Orders.ROWTIME TO DAY)
FROM Orders
```

displays the set of unique products that are ordered in any given day.

If you are doing "GROUP BY floor(ROWTIME TO MINUTE)" and there are two rows in a given minute -- say 22:49:10 and 22:49:15 -- then the summary of those rows is going to come out timestamped 22:50:00. Why? Because that is the earliest time that row is complete.

Note: "GROUP BY ceil(ROWTIME TO MINUTE)" or "GROUP BY floor(ROWTIME TO MINUTE) - INTERVAL '1' DAY" would give identical behavior.

It is not the value of the grouping expression that determines row completion, it's when that expression changes value.

If you want the rowtimes of the output rows to be the time they are emitted, then in the following example you would need to change from form 1 to use form 2 instead:

```
(Form 1)
   select distinct floor(s.rowtime to hour), a,b,c
   from s
(Form 2)
   select min(s.rowtime) as rowtime, floor(s.rowtime to hour), a, b, c
   from s
   group by floor(s.rowtime to hour), a, b, c
```

SELECT clause

The <select-clause> uses the following items after the STREAM keyword:

```
 <select-list> :=
   <select-item> { , <select-item> }...
<select-item> :=
   <select-expression> [ [ AS ] <simple-identifier> ]
<simple-identifier> :=
   <identifier> | <quoted-identifier>
<select-expression> :=
   <identifier> . *  | *  | <expression>
```

Expressions

Each of these expressions may be:

- a scalar expression
- a call to an Aggregate Functions, if this is an aggregating query (see GROUP BY clause)
- a call to an Analytic Functions, if this is not an aggregating query
- the wildcard expression * expands to all columns of all relations in the FROM clause
- the wildcard expression alias.* expands to all columns of the relation named alias
- the ROWTIME
- a CASE expression

Each expression may be assigned an alias, using the AS column_name syntax. This is the name of the column in the result set of this query. If this query is in the FROM clause of an enclosing query, this will be the name that will be used to reference the column. The number of columns specified in the AS clause of a stream reference must match the number of columns defined in the original stream.

Amazon Kinesis Data Analytics has a few simple rules to derive the alias of an expression that does not have an alias. The default alias of a column expression is the name of the column: for example, EMPS.DEPTNO is aliased DEPTNO by default. Other expressions are given an alias like EXPR$0. You should not assume that the system will generate the same alias each time.

In a streaming query, aliasing a column AS ROWTIME has a special meaning: For more information, see ROWTIME.

> **Note**
>
> All streams have an implicit column called ROWTIME. This column may impact your use of the syntax 'AS t(c1, c2, ...)' that is now supported by SQL:2008. Previously in a FROM clause you could only write

```
SELECT ... FROM r1 AS t1 JOIN r2 as t2
```

but t1 and t2 would have the same columns as r1 and t2. The AS syntax enables you to rename r1's columns by writing the following:

```
SELECT ... FROM r1 AS t1(a, b, c)
```

(r1 must have precisely 3 columns for this syntax to work).

If r1 is a stream, then ROWTIME is implicitly included, but it doesn't count as a column. As a result, if a stream has 3 columns without including ROWTIME, you cannot rename ROWTIME by specifying 4 columns. For example, if the stream Bids has three columns, the following code is invalid.

```
SELECT STREAM * FROM Bids (a, b, c, d)
```

It is also invalid to rename another column ROWTIME, as in the following example.

```
SELECT STREAM * FROM Bids (ROWTIME, a, b)
```

because that would imply renaming another column to ROWTIME. For more information about expressions and literals, see Expressions and Literals.

CASE expression

The CASE expression enables you to specify a set of discrete test expressions and a specific return-value (expression) for each such test. Each test expression is specified in a WHEN clause; each return-value expression is specified in the corresponding THEN clause. Multiple such WHEN-THEN pairs can be specified.

If you specify a comparison-test-expression before the first WHEN clause, then each expression in a WHEN clause is compared to that comparison-test-expression. The first one to match the

comparison-test-expression causes the return-value from its corresponding THEN clause to be returned. If no WHEN clause expression matches the comparison-test-expression, the return-value is null unless an ELSE clause is specified, in which case the return-value in that ELSE clause is returned.

If you do not specify a comparison-test-expression before the first WHEN clause, then each expression in a WHEN clause is evaluated (left to right) and the first one to be true causes the return-value from its corresponding THEN clause to be returned. If no WHEN clause expression is true, the return-value is null unless an ELSE clause is specified, in which case the return-value in that ELSE clause is returned.

VALUES

VALUES uses expressions to calculate one or more row values, and is often used within a larger command. When creating more than one row, the VALUES clause must specify the same number of elements for every row. The resulting table-columns data-types are derived from the explicit or inferred types of the expressions appearing in that column. VALUES is allowed syntactically wherever SELECT is permitted. See also the discussion of VALUES as an operator, in the topic Query in this guide.

SYNTAX

```
VALUES ( expression [, ...] ) [, ...]
    [ ORDER BY sort_expression [ ASC | DESC | USING operator ] [, ...] ]
```

VALUES is a SQL operator, on a par with SELECT and UNION, enabling the following types of actions:

- You can write VALUES (1), (2) to return two rows each with a single anonymous column.
- You can write VALUES (1, 'a'), (2, 'b') to return two rows of two columns.
- You can name the columns using AS, as in the following example:

```
SELECT * FROM (VALUES (1, 'a'), (2, 'b')) AS t(x, y)
```

The most important use of VALUES is in an INSERT statement, to insert a single row:

```
  INSERT INTO emps (empno, name, deptno, gender)
    VALUES (107, 'Jane Costa', 22, 'F');
```

However, you can also insert multiple rows:

```
  INSERT INTO Trades (ticker, price, amount)
    VALUES ('MSFT', 30.5, 1000),
           ('ORCL', 20.25, 2000);
```

When you use VALUES in the FROM clause of a SELECT statement, the entire VALUES clause must be enclosed in parentheses, consistent with the fact that it operates as a query, not a table expression. For additional examples, see FROM clause.

> **Note**
>
> Using INSERT with streams engages some additional considerations as to rowtimes, pumps, and INSERT EXPEDITED. For more information, see INSERT.

FROM clause

The FROM clause is the source of rows for a query.

```
 <from-clause> :=
    FROM <table-reference> { , <table-reference> }...
 <table-reference> :=
    <table-name> [ <table-name> ] [ <correlation> ]
| <joined-table>
 <table-name> :=  <identifier>
 <table-over> :=  OVER <window-specification>
 <window-specification> :=
 (   <window-name>
| <query_partition_clause>
| ORDER BY <order_by_clause>
| <windowing_clause>
    )
 <windowing-clause> :=
    { ROWS | RANGE }
    { BETWEEN
    { UNBOUNDED PRECEDING
      | CURRENT ROW
      | <value-expression> { PRECEDING | FOLLOWING }
    }
       AND
    { UNBOUNDED FOLLOWING
```

```
      | CURRENT ROW
      | <value-expression> { PRECEDING | FOLLOWING }
   }
   | { UNBOUNDED { PRECEDING | FOLLOWING }
      | CURRENT ROW
      | <value-expression> { PRECEDING | FOLLOWING }
   }
   }
```

For charts on window-specification and windowing-clause, see the WINDOW Clause (Sliding Windows) under the Window statement.

```
<correlation> :=
   [ AS ] <correlation-name> [ '(' <column> { , <column> }... ')' ]
<joined-table> :=
   <table-reference> CROSS JOIN <table-reference>
 | <table-reference> NATURAL <join-type> JOIN <table-reference>
 | <table-reference> <join-type> JOIN <table-reference>
      [ USING '(' <column> { , <column>}... ')'
      | ON <condition>
      ]
<join-type> :=
   INNER
 | <outer-join-type> [ OUTER ]
<outer-join-type> :=
   LEFT
 | RIGHT
 | FULL
```

Relations

Several types of relation can appear in a FROM clause:

- A named relation (table, stream)
- A subquery enclosed in parentheses.
- A join combining two relations (see the topic JOIN in this guide).
- A transform expression.

Subqueries are described in more detail in the topic Query in this guide.

Here are some examples of subqueries:

```
// set operation as subquery
// (finds how many departments have no employees)
SELECT COUNT(*)
FROM (
  SELECT deptno FROM Dept
  EXCEPT
  SELECT deptno FROM Emp);
// table-constructor as a subquery,
// combined with a regular table in a join
SELECT *
FROM Dept AS d
  JOIN (VALUES ('Fred', 10), ('Bill', 20)) AS e (name, deptno)
  ON d.deptno = e.deptno;
```

Unlike subqueries in other parts of the SELECT statement, such as in the WHERE clause clause (WHERE Condition Clause), a subquery in the FROM clause cannot contain correlating variables. For example:

```
// Invalid query. Dept.deptno is an illegal reference to
// a column of another table in the enclosing FROM clause.
SELECT *
FROM Dept,
  (SELECT *
   FROM Emp
   WHERE Emp.deptno = Dept.Deptno)
```

FROM clause with multiple relations

If a FROM clause contains multiple, comma-separated relations, the query constructs the cartesian product of those relations; that is, it combines each row from each relation with each row from every other relation.

The comma in the FROM clause is therefore equivalent to the CROSS JOIN operator.

Correlation names

Each relation in the FROM clause can have a correlation name assigned using AS correlation-name. This name is an alternative name by which the relation can be referenced in expressions throughout the query. (Even though the relation may be a subquery or stream, it is conventionally called a 'table alias' to distinguish it from column aliases defined in the SELECT clause.)

Without an AS clause, a named relation's name becomes its default alias. (In streaming queries, the OVER clause does not prevent this default assignment from happening.)

An alias is necessary if a query uses the same named relation more than once, or if any of the relations are subqueries or table expressions.

For example, in the following query, the named relation EMPS is used twice; once with its default alias EMPS, and once with an assigned alias MANAGERS:

```
SELECT EMPS.NAME || ' is managed by ' || MANAGERS.NAME
FROM LOCALDB.Sales.EMPS,
  LOCALDB.Sales.EMPS AS MANAGERS
WHERE MANAGERS.EMPNO = EMPS.MGRNO
```

An alias can optionally be followed by a list of columns:

```
SELECT e.empname,
FROM LOCALDB.Sales.EMPS AS e(empname, empmgrno)
```

OVER clause

The OVER clause is only applicable for streaming joins. For more detail, see the topic JOIN clause in this guide.

JOIN clause

The JOIN clause in a SELECT statement combines columns from one or more streams or reference tables.

Topics

- Stream-to-Stream Joins
- Stream-to-Table Joins

Stream-to-Stream Joins

Amazon Kinesis Data Analytics supports joining an in-application stream with another in-application stream using SQL, bringing this important traditional database functionality into the streaming context.

This section describes the types of joins that Kinesis Data Analytics supports, including time-based and row-based window joins, and the details about streaming joins.

Join Types

There are five types of joins:

INNER JOIN (or just JOIN)	Returns all pairs of rows from the left and from the right for which the join condition evaluates to TRUE.
LEFT OUTER JOIN (or just LEFT JOIN)	As INNER JOIN, but rows from the left are kept even if they don't match any rows on the right. NULL values are generated on the right.
RIGHT OUTER JOIN (or just RIGHT JOIN)	As INNER JOIN, but rows from the right are kept even if they don't match any rows on the left. NULL values are generated on the left for these rows.
FULL OUTER JOIN (or just FULL JOIN)	As INNER JOIN, but rows from both sides are kept even if they don't match any rows on the other side. NULL values are generated on the other side for these rows.
CROSS JOIN	Returns the Cartesian product of the inputs: Every row from the left is paired with every row from the right.

Time-Based Window vs. Row-Based Window Joins

It isn't practical to join the entire history of the left stream to the entire history of the right. Therefore, you must restrict at least one stream to a time window by using an OVER clause. The OVER clause defines a window of rows that are to be considered for joining at a given time.

The window can be time-based or row-based:

- A time-based window uses the RANGE keyword. It defines the window as the set of rows whose ROWTIME column falls within a particular time interval of the query's current time.

For example, the following clause specifies that the window contains all rows whose ROWTIMEs are within the hour preceding the stream's current time:

```
OVER (RANGE INTERVAL '1' HOUR PRECEDING)
```

- A row-based window uses the ROWS keyword. It defines the window as a given count of rows before or after the row with the current time stamp.

  For example, the following clause specifies that only the latest 10 rows be included in the window:

  ```
  OVER (ROWS 10 PRECEDING)
  ```

> **Note**
>
> If no time window or row-based window is specified on the side of a join, then only the current row from that side participates in the join evaluation.

Examples of Stream-to-Stream Joins

The following examples demonstrate how an in-application stream-to-stream join works, when the results of the join are returned, and what the row times of the join results are.

Topics

- Example Dataset
- Example 1: Time Window on One Side of a JOIN (INNER JOIN)
- Example 2: Time Windows on Both Sides of a JOIN (INNER JOIN)
- Example 3: Time Window on One Side of a RIGHT JOIN (RIGHT OUTER JOIN)
- Example 4: Time Windows on Both Sides of a RIGHT JOIN (RIGHT OUTER JOIN)
- Example 5: Time Window on One Side of a LEFT JOIN (LEFT OUTER JOIN)
- Example 6: Time Windows on Both Sides of a LEFT JOIN (LEFT OUTER JOIN)
- Summary

Example Dataset

The examples in this section are based on the following datasets and stream definitions:

Sample of Orders Data

```
{
   "orderid":"101",
   "orders":"1"
}
```

Sample of Shipments Data

```
{
   "orderid":"101",
   "shipments":"2"
}
```

Creating the ORDERS_STREAM In-Application Stream

```
CREATE OR REPLACE STREAM "ORDERS_STREAM" ("orderid" int, "orderrowtime" timestamp);
CREATE OR REPLACE PUMP "ORDERS_STREAM_PUMP" AS INSERT INTO "ORDERS_STREAM"
SELECT STREAM "orderid", "ROWTIME"
FROM "SOURCE_SQL_STREAM_001" WHERE "orders" = 1;
```

Creating the SHIPMENTS_STREAM In-Application Stream

```
CREATE OR REPLACE STREAM "SHIPMENTS_STREAM" ("orderid" int, "shipmentrowtime"
 timestamp);
CREATE OR REPLACE PUMP "SHIPMENTS_STREAM_PUMP" AS INSERT INTO "SHIPMENTS_STREAM"
SELECT STREAM "orderid", "ROWTIME"
FROM "SOURCE_SQL_STREAM_001" WHERE "shipments" = 2;
```

Example 1: Time Window on One Side of a JOIN (INNER JOIN)

This example demonstrates a query that returns all orders with shipments that executed in the last minute.

Join Query

```
CREATE OR REPLACE STREAM "OUTPUT_STREAM" ("resultrowtime" timestamp, "orderid" int,
 "shipmenttime" timestamp, "ordertime" timestamp);
```

```
CREATE OR REPLACE PUMP "OUTPUT_STREAM_PUMP" AS
INSERT INTO "OUTPUT_STREAM"
  SELECT STREAM ROWTIME as "resultrowtime", s."orderid", s.rowtime as "shipmenttime",
 o.ROWTIME as "ordertime"
    FROM ORDERS_STREAM OVER (RANGE INTERVAL '1' MINUTE PRECEDING) AS o
      JOIN SHIPMENTS_STREAM AS s
        ON o."orderid" = s."orderid";
```

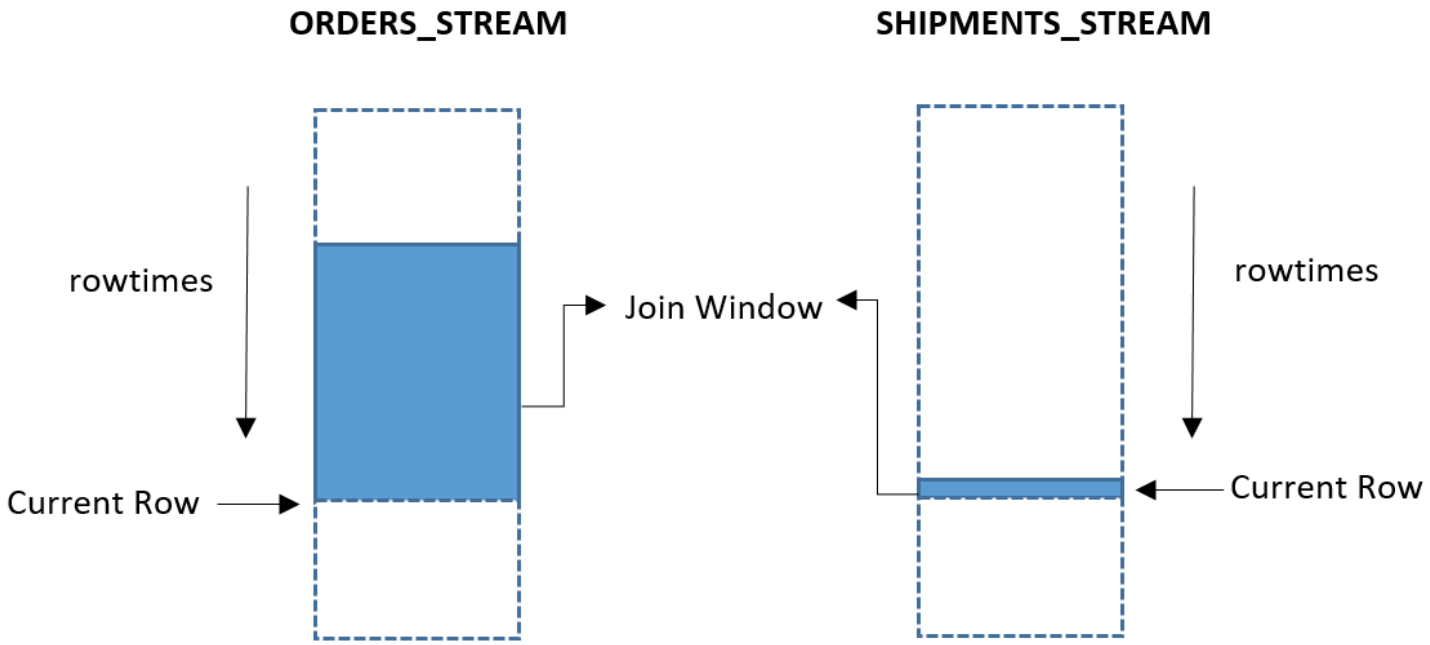
Query Results

ORDERS_STREAM		SHIPMENTS_STREAM		OUTPUT_STREAM			
ROWTIME	orderid	ROWTIME	orderid	resultrowtime	orderid	shipmenttime	OrderTime
10:00:00	101	10:00:00	100				
10:00:20	102						
10:00:30	103						
10:00:40	104						
		10:00:45	104				
10:00:45	100*			10:00:45	104	10:00:45	10:00:40
		10:00:50	105				

* - Record with orderid = 100 is a late event in the Orders stream.

Visual Representation of the Join

The following diagram represents a query that returns all orders with shipments that executed in the last minute.



Triggering of Results

The following describes the events that trigger results from the query.

- Because no time or row window is specified on the Shipments stream, only the current row of the Shipments stream participates in the join.
- Because the query on the Orders stream specifies a one-minute preceding window, the rows in the Orders stream with a ROWTIME in the last minute participate in the join.
- When the record in the Shipments stream arrived at 10:00:45 for orderid 104, the JOIN result was triggered because there is a match on orderid in the Orders stream in the preceding minute.
- The record in the Orders stream with orderid 100 arrived late, so the corresponding record in the Shipments stream was not the latest record. Because no window was specified on the Shipments stream, only the current row of the Shipments stream participates in the join. As a result, no records are returned by the JOIN statement for orderid 100. For information about including late rows in a JOIN statement, see Example 2.
- Because there is no matching record in the Shipments stream for orderid 105, no results are emitted, and the record is ignored.

ROWTIMES of Results

- The ROWTIME of the record in the output stream is the later of the ROWTIMEs of the rows that matched the join.

Example 2: Time Windows on Both Sides of a JOIN (INNER JOIN)

This example demonstrates a query that returns all orders that executed in the last minute, with shipments that executed in the last minute.

Join Query

```
CREATE OR REPLACE STREAM "OUTPUT_STREAM" ("resultrowtime" timestamp, "orderid" int,
 "shipmenttime" timestamp, "ordertime" timestamp);
CREATE OR REPLACE PUMP "OUTPUT_STREAM_PUMP" AS INSERT INTO "OUTPUT_STREAM"
  SELECT STREAM ROWTIME as "resultrowtime", s."orderid", s.rowtime as "shipmenttime",
 o.ROWTIME as "ordertime"
    FROM ORDERS_STREAM OVER (RANGE INTERVAL '1' MINUTE PRECEDING) AS o
      JOIN SHIPMENTS_STREAM OVER (RANGE INTERVAL '1' MINUTE PRECEDING) AS s
        ON o."orderid" = s."orderid";
```

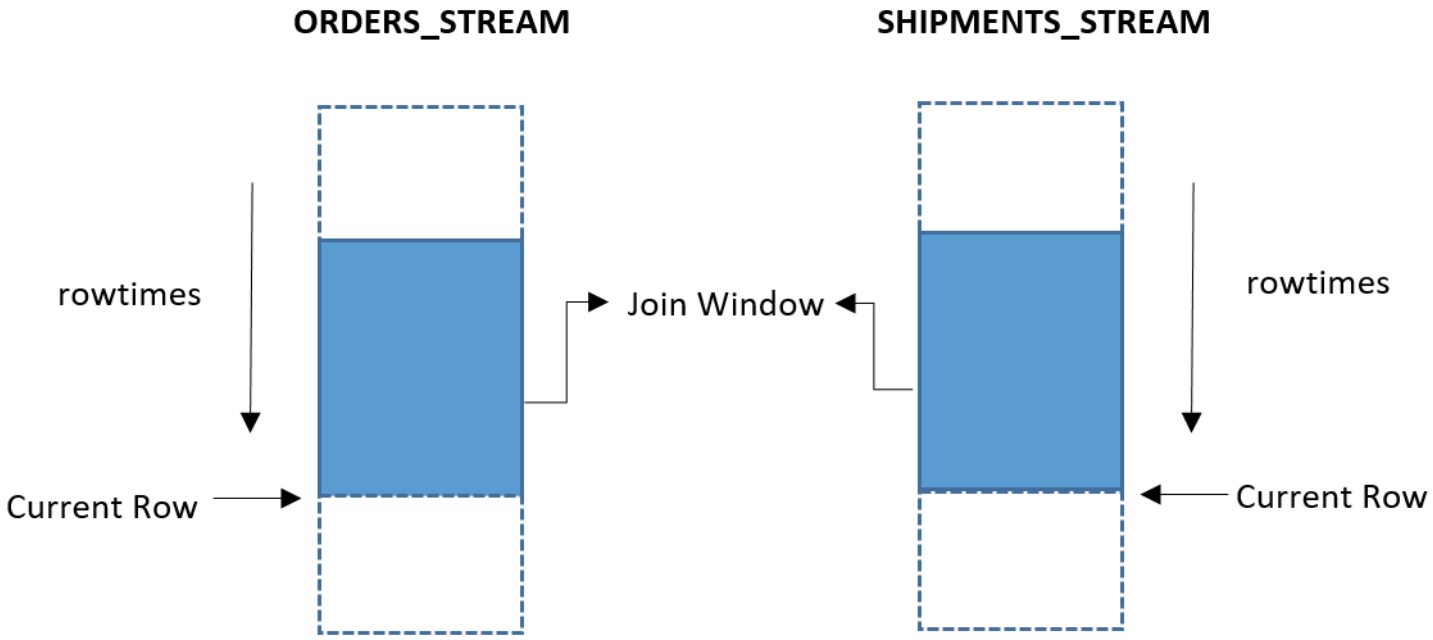
Query Results

ORDERS_STREAM		SHIPMENTS_STREAM		OUTPUT_STREAM			
ROWTIME	orderid	ROWTIME	orderid	resultrowtime	orderid	shipmenttime	OrderTime
10:00:00	101	10:00:00	100				
10:00:20	102						
10:00:30	103						
10:00:40	104						
		10:00:45	104				
10:00:45	100*			10:00:45	104	10:00:45	10:00:40
				10:00:45	100	10:00:00	10:00:45
		10:00:50	105				

* - Record with orderid = 100 is a late event in the Orders stream.

Visual Representation of the Join

The following diagram represents a query that returns all orders that executed in the last minute, with shipments that executed in the last minute.



Triggering of Results

The following describes the events that trigger results from the query.

- Windows are specified on both sides of the join. So all the rows in the minute preceding the current row of both the Orders stream and the Shipments stream participate in the join.
- When the record in the Shipments stream for orderid 104 arrived, the corresponding record in the Orders stream was within the one-minute window. So a record was returned to the Output stream.
- Even though the order event for orderid 100 arrived late in the Orders stream, the join result was returned. This is because the window in the Shipments stream includes the past minute of orders, which includes the corresponding record.
- Having a window on both sides of the join is helpful for including late-arriving records on either side of the join; for example, if an order or shipment record is received late or out of order.

ROWTIMEs of Results

- The ROWTIME of the record in the output stream is the later of the ROWTIMEs of the rows that matched the join.

Example 3: Time Window on One Side of a RIGHT JOIN (RIGHT OUTER JOIN)

This example demonstrates a query that returns all shipments that executed in the last minute, whether or not there are corresponding orders in the last minute.

Join Query

```
CREATE OR REPLACE STREAM "OUTPUT_STREAM" ("resultrowtime" timestamp, "orderid" int,
 "ordertime" timestamp);
CREATE OR REPLACE PUMP "OUTPUT_STREAM_PUMP" AS INSERT INTO "OUTPUT_STREAM"
  SELECT STREAM ROWTIME as "resultrowtime", s."orderid", o.ROWTIME as "ordertime"
    FROM ORDERS_STREAM OVER (RANGE INTERVAL '1' MINUTE PRECEDING) AS o
      RIGHT JOIN SHIPMENTS_STREAM AS s
        ON o."orderid" = s."orderid";
```

Query Results

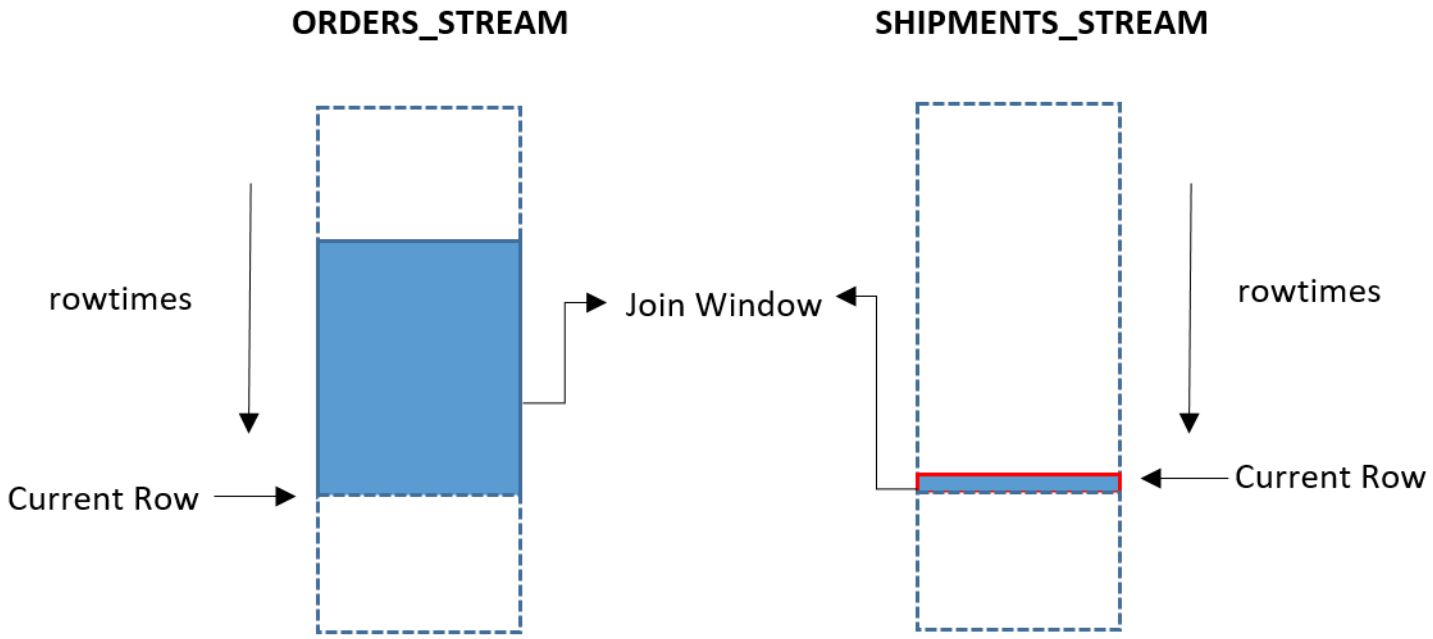
ORDERS_STREAM		SHIPMENTS_STREAM		OUTPUT_STREAM		
ROWTIME	orderid	ROWTIME	orderid	resultrowtime	orderid	OrderTime
10:00:00	101	10:00:00	100			
				10:00:00	100	null
10:00:20	102					
10:00:30	103					
10:00:40	104					
		10:00:45	104			
10:00:45	100*			10:00:45	104	10:00:40

ORDERS_STREAM		SHIPMENTS_STREAM		OUTPUT_STREAM		
		10:00:50	105			
				10:00:50	105	null

* - Record with orderid = 100 is a late event in the Orders stream.

Visual Representation of the Join

The following diagram represents a query that returns all shipments that executed in the last minute, whether or not there are corresponding orders in the last minute.



Triggering of Results

The following describes the events that trigger results from the query.

- When a record in the Shipments stream arrived for orderid 104, a result in the Output stream was emitted.
- As soon as the record in the Shipments stream arrived for orderid 105, a record was emitted in the Output stream. However, there is no matching record in the Orders stream, so the OrderTime value is null.

ROWTIMEs of Results

- The ROWTIME of the record in the output stream is the later of the ROWTIMEs of the rows that matched the join.
- Because the right side of the join (the Shipments stream) has no window, the ROWTIME of the result with an unmatched join is the ROWTIME of the unmatched row.

Example 4: Time Windows on Both Sides of a RIGHT JOIN (RIGHT OUTER JOIN)

This example demonstrates a query that returns all shipments that executed in the last minute, whether or not they have corresponding orders.

Join Query

```
CREATE OR REPLACE STREAM "OUTPUT_STREAM" ("resultrowtime" timestamp, "orderid" int,
 "shipmenttime" timestamp, "ordertime" timestamp);
CREATE OR REPLACE PUMP "OUTPUT_STREAM_PUMP" AS INSERT INTO "OUTPUT_STREAM"
  SELECT STREAM ROWTIME as "resultrowtime", s."orderid", s.ROWTIME as "shipmenttime",
 o.ROWTIME as "ordertime"
    FROM ORDERS_STREAM OVER (RANGE INTERVAL '1' MINUTE PRECEDING) AS o
      RIGHT JOIN SHIPMENTS_STREAM OVER (RANGE INTERVAL '1' MINUTE PRECEDING) AS s
        ON o."orderid" = s."orderid";
```

Query Results

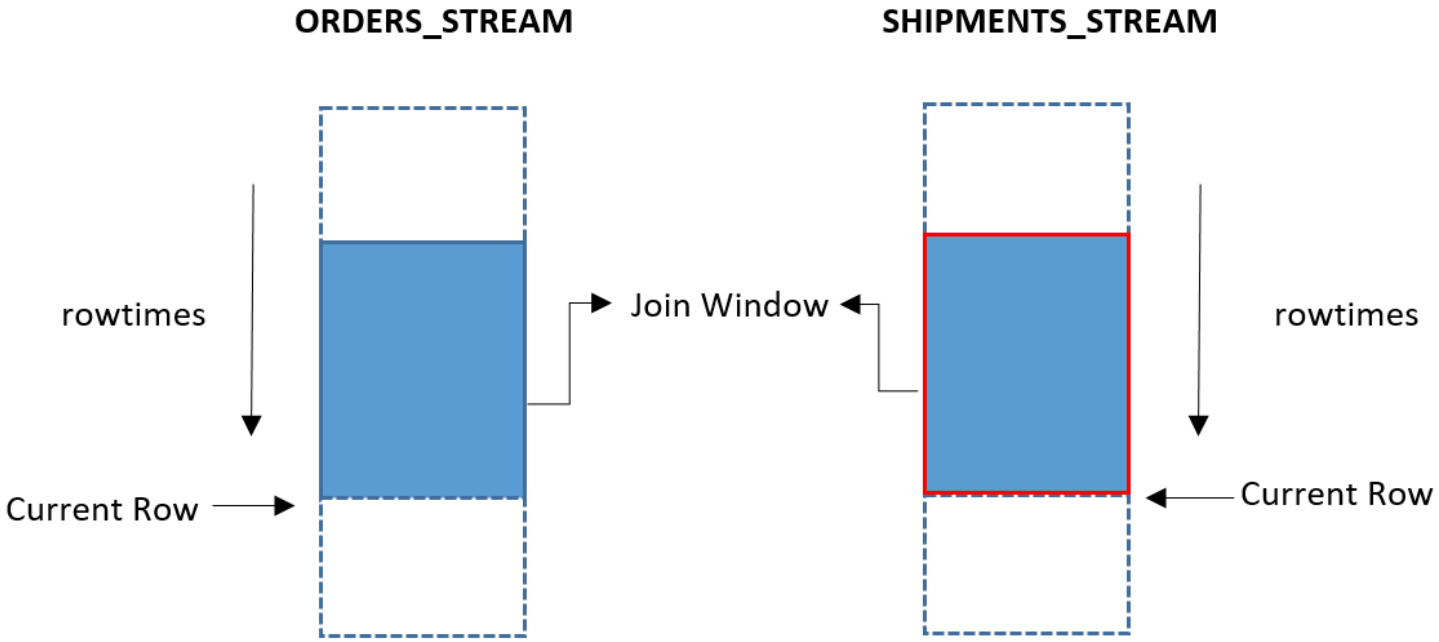
ORDERS_STREAM		SHIPMENTS_STREAM		OUTPUT_STREAM			
ROWTIME	orderid	ROWTIME	orderid	resultrowtime	orderid	shipmenttime	OrderTime
10:00:00	101	10:00:00	100				
10:00:20	102						
10:00:30	103						
10:00:40	104						

ORDERS_STREAM		SHIPMENTS_STREAM		OUTPUT_STREAM			
		10:00:45	104				
10:00:45	100*			10:00:45	104	10:00:40	10:00:45
				10:00:45	100	10:00:45	10:00:00
		10:00:50	105				
				10:01:50	105	10:00:50	null

* - Record with orderid = 100 is a late event in the Orders stream.

Visual Representation of the Join

The following diagram represents a query that returns all shipments that executed in the last minute, whether or not they have corresponding orders.



Triggering of Results

The following describes the events that trigger results from the query.

- When a record in the Shipments stream arrived for orderid 104, a result in the Output stream was emitted.
- Even though the order event for orderid 100 arrived late in the Orders stream, the join result is returned. This is because the window in the Shipments stream includes the past minute of orders, which includes the corresponding record.
- For the shipment for which the order is not found (for orderid 105), the result is not emitted to the Output stream until the end of the one-minute window on the Shipments stream.

ROWTIMEs of Results

- The ROWTIME of the record in the output stream is the later of the ROWTIMEs of the rows that matched the join.
- For shipment records with no matching order record, the ROWTIME of the result is the ROWTIME of the end of the window. This is because the right side of the join (from the Shipments stream) is now a one-minute window of events, and the service is waiting for the end of the window to determine whether any matching records arrive. When the window ends and no matching records are found, the result is emitted with a ROWTIME corresponding to the end of the window.

Example 5: Time Window on One Side of a LEFT JOIN (LEFT OUTER JOIN)

This example demonstrates a query that returns all orders that executed in the last minute, whether or not there are corresponding shipments in the last minute.

Join Query

```
CREATE OR REPLACE STREAM "OUTPUT_STREAM" ("resultrowtime" timestamp, "orderid" int,
 "ordertime" timestamp);
CREATE OR REPLACE PUMP "OUTPUT_STREAM_PUMP" AS INSERT INTO "OUTPUT_STREAM"
  SELECT STREAM ROWTIME as "resultrowtime", o."orderid", o.ROWTIME as "ordertime"
    FROM ORDERS_STREAM OVER (RANGE INTERVAL '1' MINUTE PRECEDING) AS o
      LEFT JOIN SHIPMENTS_STREAM AS s
        ON o."orderid" = s."orderid";
```

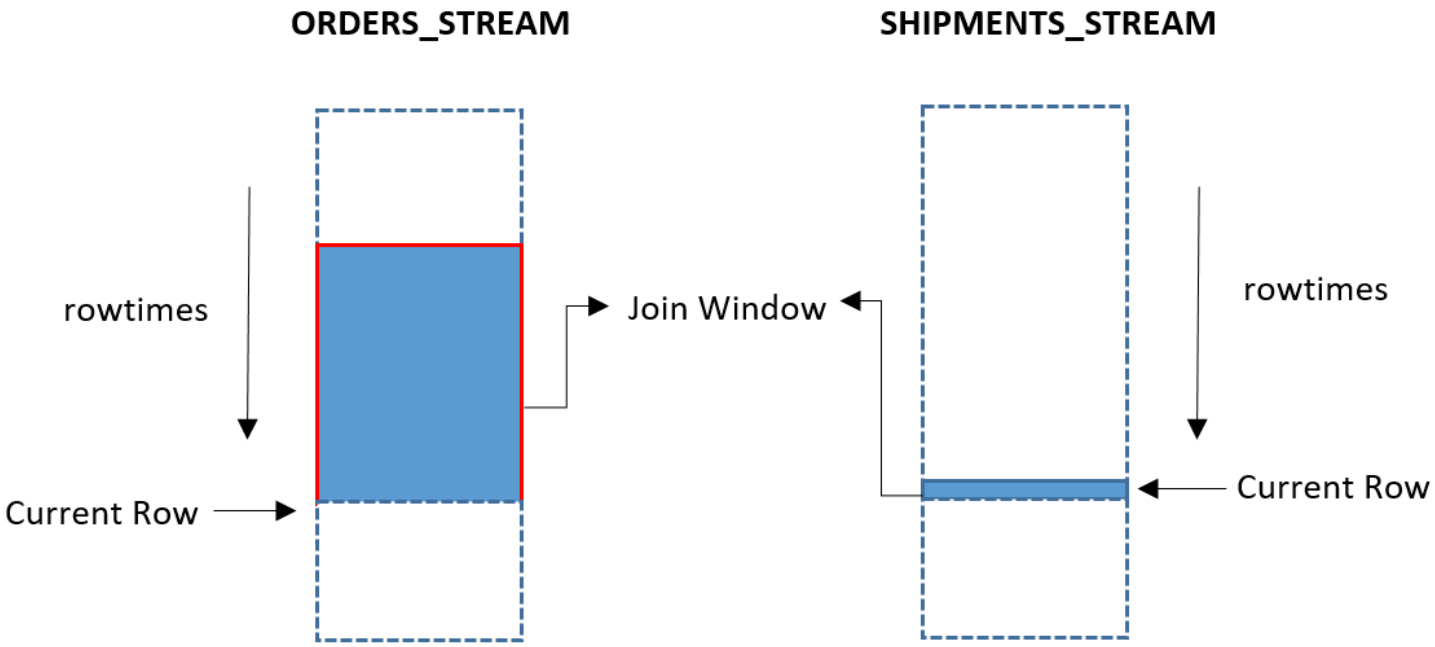
Query Results

ORDERS_STREAM		SHIPMENTS_STREAM		OUTPUT_STREAM		
ROWTIME	orderid	ROWTIME	orderid	resultrowtime	orderid	OrderTime
10:00:00	101	10:00:00	100			
10:00:20	102					
10:00:30	103					
10:00:40	104					
		10:00:45	104			
10:00:45	100*			10:00:45	104	10:00:40
		10:00:50	105			
				10:01:00	101	10:00:00
				10:01:20	102	10:00:20
				10:01:30	103	10:00:30
				10:01:40	104	10:00:40
				10:01:45	100	10:00:45

* - Record with orderid = 100 is a late event in the Orders stream.

Visual Representation of the Join

The following diagram represents a query that returns all orders that executed in the last minute, whether or not there are corresponding shipments in the last minute.



Triggering of Results

The following describes the events that trigger results from the query.

- When a record in the Shipments stream arrived for orderid 104, a result in the Output stream is emitted.
- For records in the Orders stream with no corresponding record in the Shipments stream, records are not emitted to the Output stream until the end of the one-minute window. This is because the service is waiting until the end of the window for matching records.

ROWTIMEs of Results

- The ROWTIME of the record in the output stream is the later of the ROWTIMEs of the rows that matched the join.
- For records in the Orders stream with no corresponding record in the Shipments stream, the ROWTIMEs of the results are the ROWTIMEs of the end of the current window.

Example 6: Time Windows on Both Sides of a LEFT JOIN (LEFT OUTER JOIN)

This example demonstrates a query that returns all orders that executed in the last minute, whether or not they have corresponding shipments.

Join Query

```
CREATE OR REPLACE STREAM "OUTPUT_STREAM" ("resultrowtime" timestamp, "orderid" int,
 "shipmenttime" timestamp, "ordertime" timestamp);
CREATE OR REPLACE PUMP "OUTPUT_STREAM_PUMP" AS INSERT INTO "OUTPUT_STREAM"
  SELECT STREAM ROWTIME as "resultrowtime", s."orderid", s.ROWTIME as "shipmenttime",
 o.ROWTIME as "ordertime"
    FROM ORDERS_STREAM OVER (RANGE INTERVAL '1' MINUTE PRECEDING) AS o
      LEFT JOIN SHIPMENTS_STREAM OVER (RANGE INTERVAL '1' MINUTE PRECEDING) AS s
        ON o."orderid" = s."orderid";
```

Query Results

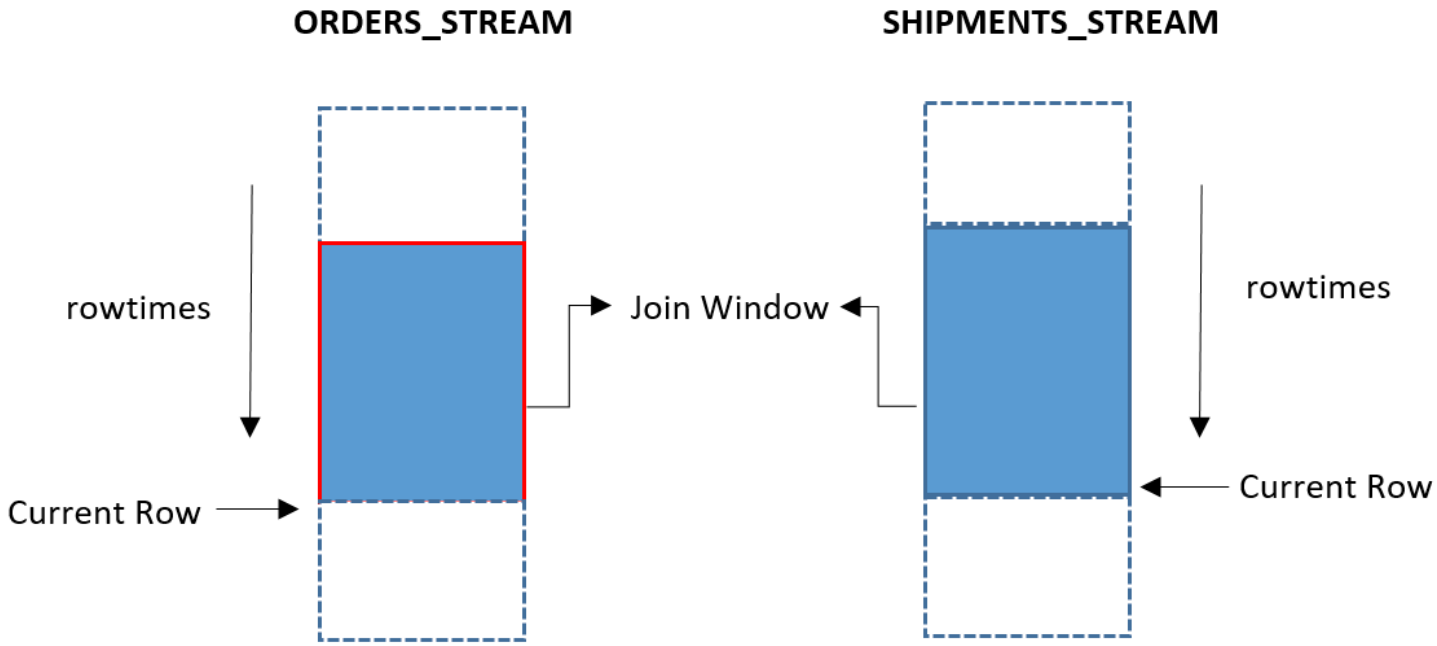
ORDERS_STREAM		SHIPMENTS_STREAM		OUTPUT_STREAM			
ROWTIME	orderid	ROWTIME	orderid	resultrowtime	orderid	shipmenttime	OrderTime
10:00:00	101	10:00:00	100				
10:00:20	102						
10:00:30	103						
10:00:40	104						
		10:00:45	104				
10:00:45	100*			10:00:45	104	10:00:40	10:00:45
		10:00:50	105	10:00:45	100	10:00:00	10:00:45
				10:01:00	101	null	10:00:00
				10:01:20	102	null	10:00:20
				10:01:30	103	null	10:00:30

ORDERS_STREAM		SHIPMENTS _STREAM		OUTPUT_STREAM			
				10:01:40	104	null	10:00:40
				10:01:45	100	null	10:00:45

* - Record with orderid = 100 is a late event in the Orders stream.

Visual Representation of the Join

The following diagram represents a query that returns all orders that executed in the last minute, whether or not they have corresponding shipments.



Triggering of Results

The following describes the events that trigger results from the query.

- When a record in the Shipments stream arrived for orderids 104 and 100, a result in the Output stream was emitted. This occurred even though the record in the Orders stream for orderid 100 arrived late.
- Records in the Orders stream with no corresponding record in the Shipments stream are emitted in the Output stream at the end of the one-minute window. This is because the service waits until the end of the window for corresponding records in the Shipments stream.

ROWTIMEs of Results

- The ROWTIME of the record in the Output stream is the later of the ROWTIMEs of the rows that matched the join.
- For records in the Orders stream with no corresponding record in the Shipments stream, the ROWTIMEs of the orders is the ROWTIME corresponding to the end of the window.

Summary

- Kinesis Data Analytics always returns rows from joins in ascending order of ROWTIME.
- For an inner join, the ROWTIME of an output row is the later of the ROWTIMEs of the two input rows. This is also true for an outer join for which matching input rows are found.
- For outer joins for which a match is not found, the ROWTIME of an output row is the later of the following two times:
  - The ROWTIME of the input row for which a match was not found.
  - The later bound of the window of the other input stream at the point any possible match could have been found.

Stream-to-Table Joins

If one of the relations is a stream and the other is a finite relation, it is referred to as a stream-table join. For each row in the stream, the query looks up the row or rows in the table that match the join condition.

For example, Orders is a stream and PriceList is a table. The effect of the join is to add price list information to the order.

For information about creating a reference data source and joining a stream to a reference table, see Example: Add Reference Data Source in the *Amazon Kinesis Data Analytics Developer Guide*.

HAVING clause

The HAVING clause in a SELECT specifies a condition to apply within a group or aggregate. In other words, HAVING filters rows after the aggregation of the GROUP BY clause has been applied. Since HAVING is evaluated after GROUP BY, it can only reference expressions constructed (or derivable) from grouping keys, aggregate expressions, and constants. (These are the same rules that apply to expressions in the SELECT clause of a GROUP BY query.) A HAVING clause must come after any

GROUP BY clause and before any ORDER BY clause. HAVING is like WHERE clause, but applies to groups. Results from a HAVING clause represent groupings or aggregations of original rows, whereas results from a WHERE clause are individual original rows.

In non-streaming applications, if there is no GROUP BY clause, GROUP BY () is assumed (though since there are no grouping expressions, expressions can consist only of constants and aggregate expressions). In streaming queries, HAVING cannot be used without a GROUP BY clause.

WHERE and HAVING can both appear in a single SELECT statement. The WHERE selects from the stream or table those individual rows that satisfy its condition (the WHERE-condition). The GROUP BY criteria apply only to the rows selected by the WHERE condition.

Such a grouping, for example "GROUP BY CustomerID", can be further qualified by a HAVING-condition, which then selects aggregations of rows satisfying its condition within the specified grouping. For example, "GROUP BY ClientID HAVING SUM(ShipmentValue) > 3600" would select only those clients whose various shipments that fit the WHERE criteria also had values that added up to exceed 3600.

See the WHERE clause syntax chart for the conditions, which applies to both HAVING and WHERE clauses.

The condition must be a Boolean predicate expression. The query returns only rows for which the predicate evaluates to TRUE.

The example below shows a streaming query that displays products for which there are more than $1000 of orders in the past hour.

```
SELECT STREAM "prodId"
FROM "Orders"
GROUP BY FLOOR("Orders".ROWTIME TO HOUR), "prodId"
HAVING SUM("quantity" * "price") > 1000;
```

GROUP BY clause

Syntax Chart for the GROUP BY Clause

(To see where this clause fits, see SELECT statement)

For example, GROUP BY <column name-or-expression>, where:

- the expression can be an aggregate; and,
- any column name used in the GROUP BY clause must also be in the SELECT statement.

Additionally, a column that is not named in or derivable from the GROUP BY clause cannot appear in the SELECT statement except within aggregations, such as SUM (allOrdersValue).

What derivable means is that a column specified in the GROUP BY clause enables access to the column you want to include in the SELECT clause. If a column is derivable, the SELECT statement can specify it even though it is not explicitly named in the GROUP BY clause.

Example: If the key to a table is in the GROUP BY clause, then any of that table's columns can appear in the select-list because, given that key, such columns are considered accessible.

The GROUP BY clause groups selected rows based on the value of the grouping expressions, returning a single summary row of information for each group of rows that have identical values in all columns.

Note that for these purposes, the value NULL is considered equal to itself and not equal to any other value. These are the same semantics as for the IS NOT DISTINCT FROM operator.

Streaming GROUP BY

GROUP BY can be used in a streaming query as long as one of the grouping expressions is a non-constant monotonic or time-based expression. This requirement is necessary in order for Amazon Kinesis Data Analytics to make progress, as explained below.

A monotonic expression is one that always moves in the same direction: it either ascends-or-stays-the-same, or it descends-or-stays the same; it doesn't reverse direction. It does not need to be strictly ascending or strictly descending, that is, every value always above the previous one or every value always below the previous one. A constant expression falls under the definition of monotonic -- it is technically both ascending and descending -- but is clearly unsuitable for these purposes. For more information about monotonicity, see Monotonic Expressions and Operators.

Consider the following query:

```
SELECT STREAM prodId, COUNT(*)
FROM Orders
GROUP BY prodId
```

The query is intended to compute the number of orders for each product, as a stream. However, since Orders is an infinite stream, Amazon Kinesis Data Analytics can never know that it has seen all orders for a given product, can never complete a particular row's total, and therefore can never output a row. Rather than allow a query that can never emit a row, the Amazon Kinesis Data Analytics validator rejects the query.

The syntax for streaming GROUP BY is as follows:

```
GROUP BY <monotonic or time-based expression> ,

<column name-or-expression, ...>
```

where any column name used in the GROUP BY clause needs to be in the SELECT statement; the expression can be an aggregate. Additionally, a column name that does not appear in the GROUP BY clause cannot appear in the SELECT statement except within aggregations, or if, as above, access to the column can be created from column that you specify in the GROUP BY clause.

For example, the following query, which computes the product counts per hour, uses the monotonic expression FLOOR(Orders.ROWTIME TO HOUR) is therefore valid:

```
SELECT STREAM FLOOR(Orders.ROWTIME TO HOUR) AS theHour, prodId, COUNT(*)
FROM Orders
GROUP BY FLOOR(Orders.ROWTIME TO HOUR), prodId
```

One of the expressions in the GROUP BY must be monotonic or time-based. For example GROUP BY FLOOR(S.ROWTIME) TO HOUR will yield one output row per hour for the previous hour's input rows. The GROUP BY can specify additional partitioning terms. For example, GROUP BY FLOOR(S.ROWTIME) TO HOUR, USERID will yield one output row per hour per USERID value. If you know for a fact that an expression is monotonic, you can declare it so by using the Monotonic Function. If the actual data are not monotonic, the resulting system behavior is indeterminate: results may not be as expected or desired.

See the topic Monotonic Function in this guide for more details.

Duplicate rowtimes can occur in a stream, and as long as the ROWTIME value is the same, the GROUP BY operation will keep accumulating rows. In order to emit a row, the ROWTIME value has to change at some point.

WHERE clause

The WHERE clause extracts records that meet a specified condition. The condition can be a numeric or string comparison, or use the BETWEEN, LIKE, OR IN operators: see Streaming SQL Operators. Conditions can be combined using logical operators such as AND, OR, and NOT.

The WHERE clause is like the HAVING clause clause. It applies to groups, that is, results from a WHERE clause are individual original rows, whereas results from a HAVING clause represent groupings or aggregations of original rows.

WHERE and HAVING can both appear in a single SELECT statement. The WHERE selects from the stream or table those individual rows that satisfy its condition (the WHERE-condition). The GROUP BY criteria apply only to the rows selected by the WHERE condition. Such a grouping, for example "GROUP BY CustomerID", can be further qualified by a HAVING-condition, which then selects aggregations of rows satisfying its condition within the specified grouping. For example, "GROUP BY ClientID HAVING SUM(ShipmentValue) > 3600" would select only those clients whose various shipments that fit the WHERE criteria also had values that added up to exceed 3600.

To see where this clause fits into the SELECT statement, see SELECT statement.

The condition must be a Boolean predicate expression. The query returns only rows for which the predicate evaluates to TRUE; if the condition evaluates to NULL, the row is not emitted.

The condition in the WHERE clause cannot contain windowed aggregation expressions, because if the where clause condition caused rows to be dropped, it would alter the contents of the window.

WHERE is also discussed in the topics JOIN clause and HAVING clause in this guide.

WINDOW Clause (Sliding Windows)

The `WINDOW` clause for a sliding windowed query specifies the rows over which analytic functions are computed across a group of rows in relation to the current row. These aggregate functions produce an output row aggregated by the keys in one or more columns for each input row. The `WINDOW` clause in a query specifies records in a stream partitioned by the time range interval or the number of rows, and an additional optional set of columns specified by the `PARTITION BY` clause. You can define named or inline window specifications that can be used in analytic functions and streaming `JOIN` clauses. For more information about analytic functions, see Analytic Functions.

Aggregate functions in a sliding window query are performed over each column specified in the `OVER` clause. The `OVER` clause can reference a named window specification or can be inline as part

of the SELECT statement for a pump. The following examples show how to use the OVER clause to reference a named window specification and inline in the SELECT statement.

Syntax

```
[WINDOW window_name AS
(
 {PARTITION BY partition_name
  RANGE INTERVAL 'interval' {SECOND | MINUTE | HOUR} PRECEDING |
  ROWS number PRECEDING
, …}
)
```

OVER Clause

The examples following show you how to use the OVER clause to reference a named window specification.

Example 1: OVER Referencing a Named Window Specification

The following example shows an aggregate function that references the window specification with the name W1. In this example, the average price is calculated over the set of records specified by the W1 window specification. To learn more about how to use the OVER clause with a window specification, see Examples, following.

```
AVG(price) OVER W1 AS avg_price
```

Example 2: OVER Referencing an Inline Window Specification

The following example shows an aggregate function that references an inline window specification. In this example, the average price is calculated over each input row with an inline window specification. To learn more about how to use the OVER clause with a window specification, see Examples, following.

```
AVG(price) OVER (
```

```
    PARTITION BY ticker_symbol
    RANGE INTERVAL '1' HOUR PRECEDING) AS avg_price
```

For more information about aggregate functions and the OVER clause, see Aggregate Functions.

Parameters

window-name

Specifies a unique name that can be referenced from OVER clauses or subsequent window definitions. The name is used in analytic functions and streaming `JOIN` clauses. For more information about analytic functions, see Analytic Functions.

AS

Defines the named window specification for the `WINDOW` clause.

PARTITION BY *partition-name*

Divides rows into groups that share the same values. After rows are partitioned, the window function computes all rows that fall into the same partition as the current row.

RANGE INTERVAL *'interval'* {SECOND | MINUTE | HOUR} PRECEDING

Specifies the window boundaries from the time range interval. The window function computes all rows that fall into the same time interval as the current row.

ROWS *number* PRECEDING

Specifies the window boundaries from the number of rows. The window function computes all rows that fall into the same number of rows.

Examples

Example Dataset

The examples following are based on the sample stock dataset that is part of Getting Started in the *Amazon Kinesis Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Analytics application that has the sample stock ticker input stream. To learn how to create an

Analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Analytics Developer Guide*. For additional samples, see Sliding Windows.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

Example 1: Time-Based Sliding Window That References a Named Window Specification

This example defines a named window specification with a partition boundary of one minute preceding the current row. The OVER clause of the SELECT statement for the pump references the named window specification.

```
WINDOW W1 AS (
    PARTITION BY ticker_symbol
    RANGE INTERVAL '1' MINUTE PRECEDING);
```

To run this example, create the stock sample application and run and save the SQL code following.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    min_price     DOUBLE,
    max_price     DOUBLE,
    avg_price     DOUBLE);
CREATE OR REPLACE PUMP "STREAM_PUMP" AS
    INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM ticker_symbol,
        MIN(price) OVER W1 AS min_price,
        MAX(price) OVER W1 AS max_price,
        AVG(price) OVER W1 AS avg_price
    FROM "SOURCE_SQL_STREAM_001"
    WINDOW W1 AS (
        PARTITION BY ticker_symbol
        RANGE INTERVAL '1' MINUTE PRECEDING);
```

The preceding example outputs a stream similar to the following.

ROWTIME	TICKER_SYMBOL	MIN_PRICE	MAX_PRICE	AVG_PRICE
2017-01-31 23:30:11.661	QXZ	215.0399932861328	215.0399932861328	215.0399932861328
2017-01-31 23:30:16.673	IOP	118.41999816894531	119.0999984741211	118.75999450683594
2017-01-31 23:30:16.673	AMZN	727.469970703125	727.469970703125	727.469970703125
2017-01-31 23:30:16.673	AMZN	713.0900268554688	727.469970703125	720.280029296875
2017-01-31 23:30:16.673	TBV	164.00999450683594	178.77999877929688	169.89332580566406

Example 2: Row-Based Sliding Window That References a Named Window Specification

This example defines a named window specification with a partition boundary of two rows preceding the current row and ten rows preceding the current row. The OVER clause of the SELECT statement for the pump references the named window specification.

```
 WINDOW
    last2rows AS (PARTITION BY ticker_symbol ROWS 2 PRECEDING),
    last10rows AS (PARTITION BY ticker_symbol ROWS 10 PRECEDING);
```

To run this example, create the stock sample application and run and save the SQL code following.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol      VARCHAR(4),
    price              DOUBLE,
    avg_last2rows      DOUBLE,
    avg_Last10rows     DOUBLE);
CREATE OR REPLACE PUMP "myPump" AS INSERT INTO "DESTINATION_SQL_STREAM"
SELECT STREAM ticker_symbol,
    price,
    AVG(price) OVER last2rows,
    AVG(price) OVER last10rows
FROM SOURCE_SQL_STREAM_001
WINDOW
    last2rows AS (PARTITION BY ticker_symbol ROWS 2 PRECEDING),
    last10rows AS (PARTITION BY ticker_symbol ROWS 10 PRECEDING);
```

The preceding example outputs a stream similar to the following.

ROWTIME	TICKER_SYMBOL	PRICE	AVERAGE_LAST2ROWS	AVERAGE_LAST10ROWS
2017-01-31 23:43:06.414	QXZ	75.33999633789062	114.7199935913086	100.65363311767578
2017-01-31 23:43:06.414	SLW	75.44000244140625	77.4233169555664	85.00454711914062
2017-01-31 23:43:06.414	SAC	41.709999084472656	42.95000076293945	44.18454360961914
2017-01-31 23:43:06.414	QWE	223.1999969482422	221.07000732421875	223.96910095214844
2017-01-31 23:43:06.414	WAS	14.039999961853027	13.993332862854004	13.739998817443848

Example 3: Time-Based Sliding Window with Inline Window Specification

This example defines an inline window specification with a partition boundary of one minute preceding the current row. The OVER clause of the SELECT statement for the pump uses the inline window specification.

To run this example, create the stock sample application and run and save the SQL code following.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    price         DOUBLE,
    avg_price     DOUBLE);
CREATE OR REPLACE PUMP "STREAM_PUMP" AS
    INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM ticker_symbol, price,
        AVG(Price) OVER (
            PARTITION BY ticker_symbol
            RANGE INTERVAL '1' HOUR PRECEDING) AS avg_price
    FROM "SOURCE_SQL_STREAM_001"
```

The preceding example outputs a stream similar to the following.

ROWTIME	TICKER_SYMBOL	PRICE	AVG_PRICE
2017-02-02 19:41:21.621	TBV	159.39999389648438	172.16751098632812
2017-02-02 19:41:21.621	JKL	15.300000190734863	15.295000076293945
2017-02-02 19:41:21.621	JKL	15.140000343322754	15.243332862854004
2017-02-02 19:41:21.621	QXZ	52.060001373291016	108.17571258544922
2017-02-02 19:41:21.621	WSB	8.430000305175781	8.569999694824219

Usage Notes

For the WINDOW clause and endpoints, Amazon Kinesis Analytics SQL follows SQL-2008 standards for windows over a range.

To include the endpoints of an hour, you can use the window syntax following.

```
WINDOW HOUR AS (RANGE INTERVAL '1' HOUR PRECEDING)
```

To not include the endpoints of the previous hour, you can use the window syntax following.

```
WINDOW HOUR AS (RANGE INTERVAL '59:59.999' MINUTE TO SECOND(3) PRECEDING);
```

For more information, see Allowed and Disallowed Window Specifications.

Related Topics

- Sliding Windows in the Kinesis Developer Guide
- Aggregate Functions
- SELECT statement
- CREATE STREAM statement
- CREATE PUMP statement

Allowed and Disallowed Window Specifications

Amazon Kinesis Data Analytics supports nearly all windows that end with the current row.

You cannot define an infinite window, a negative-sized window, or use negative integers in the window specification. Offset windows are currently unsupported.

- Infinite windows are windows with no bounds. Typically these point into the future, which for streams is infinite. For example "ROWS BETWEEN CURRENT ROW AND UNBOUNDED FOLLOWING" is not supported, because in a streaming context such a query would not produce

a result, since streams are continually expanding as new data arrives. All uses of UNBOUNDED FOLLOWING are unsupported.

- Negative windows . For example, "ROWS BETWEEN 0 PRECEDING AND 4 PRECEDING" is a window of negative size and is therefore illegal. Instead, you would use: "ROWS BETWEEN 4 PRECEDING AND 0 PRECEDING" in this case.

- Offset windows are windows that do not end with CURRENT ROW. These are not supported in the current release. For example, "ROWS BETWEEN UNBOUNDED PRECEDING AND 4 FOLLOWING" is not supported. (Window spans CURRENT ROW rather than starting or ending there.)

- Windows defined with negative integers. For example, "ROWS BETWEEN -4 PRECEDING AND CURRENT ROW" is invalid because negative integers are disallowed.

Also, the special case of ... 0 PRECEDING (and ... 0 FOLLOWING) cannot be used for windowed aggregation; instead, the synonym CURRENT ROW can be used.

For windowed aggregation, partitioned windows are allowed, but ORDER BY must not be present.

For windowed join, partitioned windows are NOT allowed, but ORDER BY can be present if it sorts by the ROWTIME column of one of the inputs.

Window examples

The following examples show a sample input data set, the definitions for several windows, and the contents of those windows at various times after 10:00, the time data starts to arrive for this example.

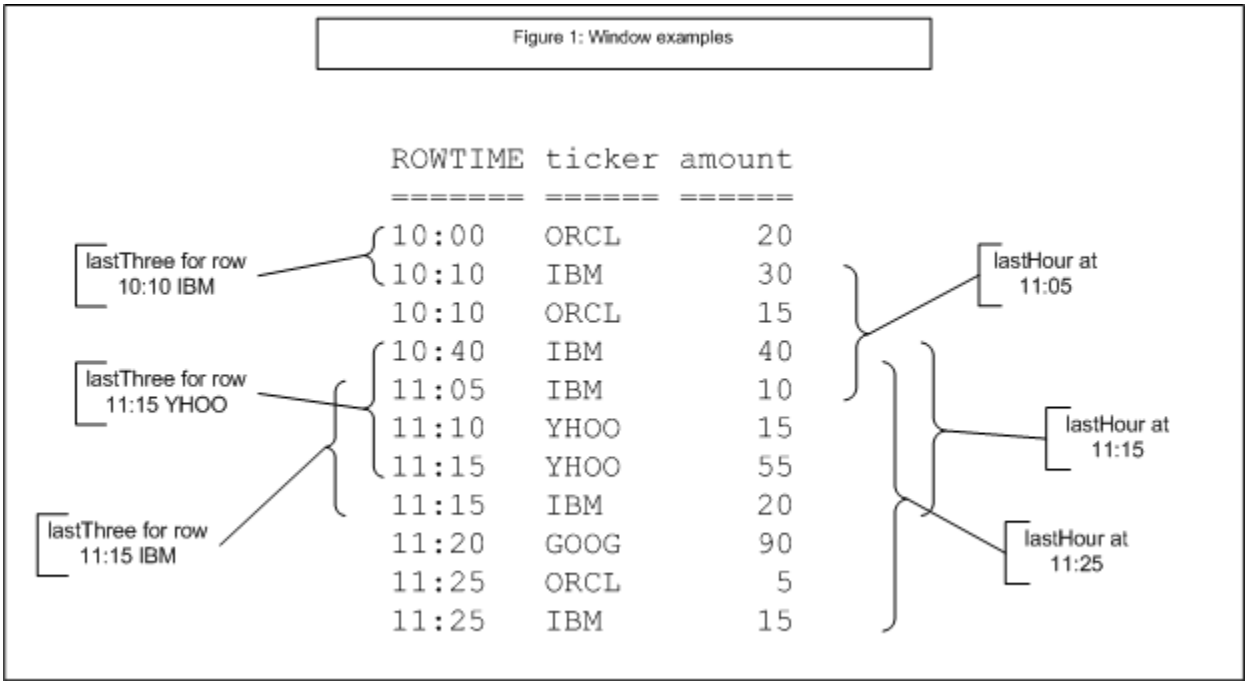
The windows are defined as follows:

```
SELECT STREAM
  ticker,
  sum(amount) OVER lastHour,
  count(*) OVER lastHour
  sum(amount) OVER lastThree
FROM Trades
WINDOW
  lastHour AS (RANGE INTERVAL '1' HOUR PRECEDING),
```

```
  lastThree AS (ROWS 3 PRECEDING),
  lastZeroRows AS (ROWS CURRENT ROW),
  lastZeroSeconds AS (RANGE CURRENT ROW),
  lastTwoSameTicker AS (PARTITION BY ticker ROWS 2 PRECEDING),
  lastHourSameTicker AS (PARTITION BY ticker RANGE INTERVAL '1' HOUR PRECEDING)
```

First Example: time-based windows versus row-based windows

As shown on the right side of the figure below, the time-based lastHour window contains varying numbers of rows, because window membership is defined by time range.



Examples of windows containing rows

The row-based lastThree window generally contains four rows: the three preceding and the current row. However for the row 10:10 IBM, it only contains two rows, because there is no data before 10:00.

A row-based window can contain several rows whose ROWTIME value is the same, though they arrive at different times (wall-clock times). The order of such a row in the row-based window depends on its arrival time; indeed, the row's arrival time can determine which window includes it.

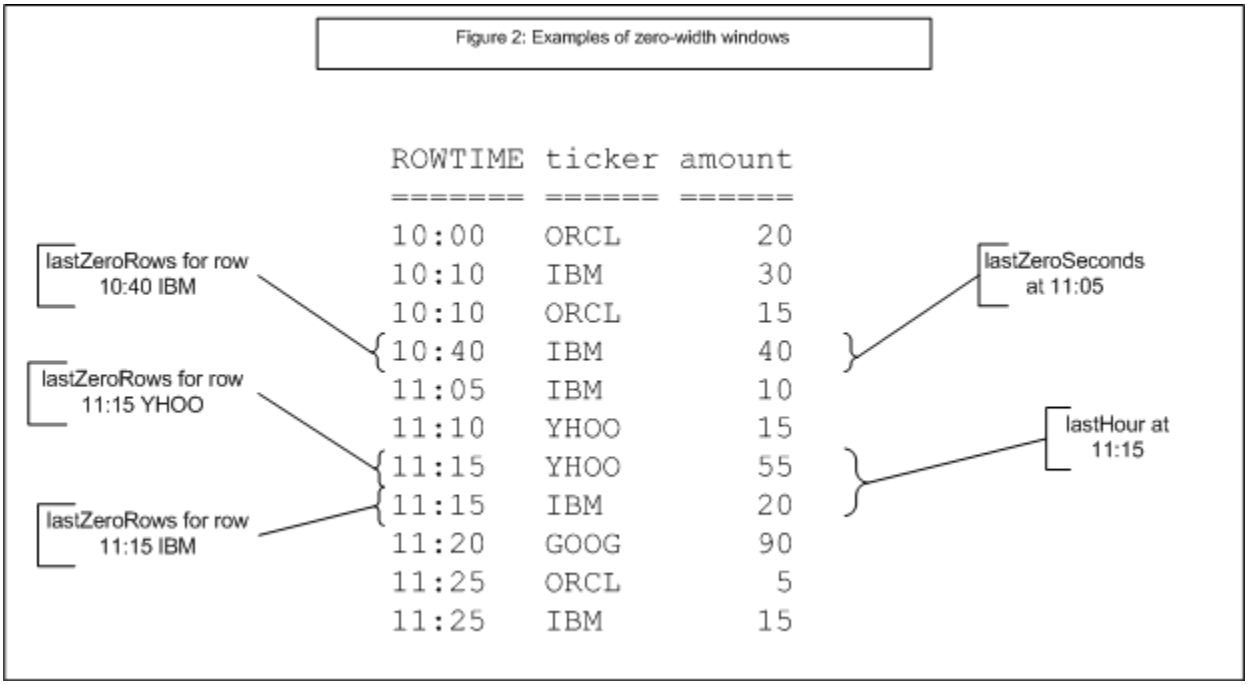
For example, the middle lastThree window in Figure 1 shows the arrival of a YHOO trade with ROWTIME 11:15 (and the last three trades before it). However, this window excludes the next trade, for IBM, whose ROWTIME is also 11:15 but which must have arrived later than the YHOO trade.

This 11:15 IBM trade is included in the 'next' window, as is the 11:15 YHOO trade, its immediate predecessor.

Second Example: zero width windows, row-based and time-based

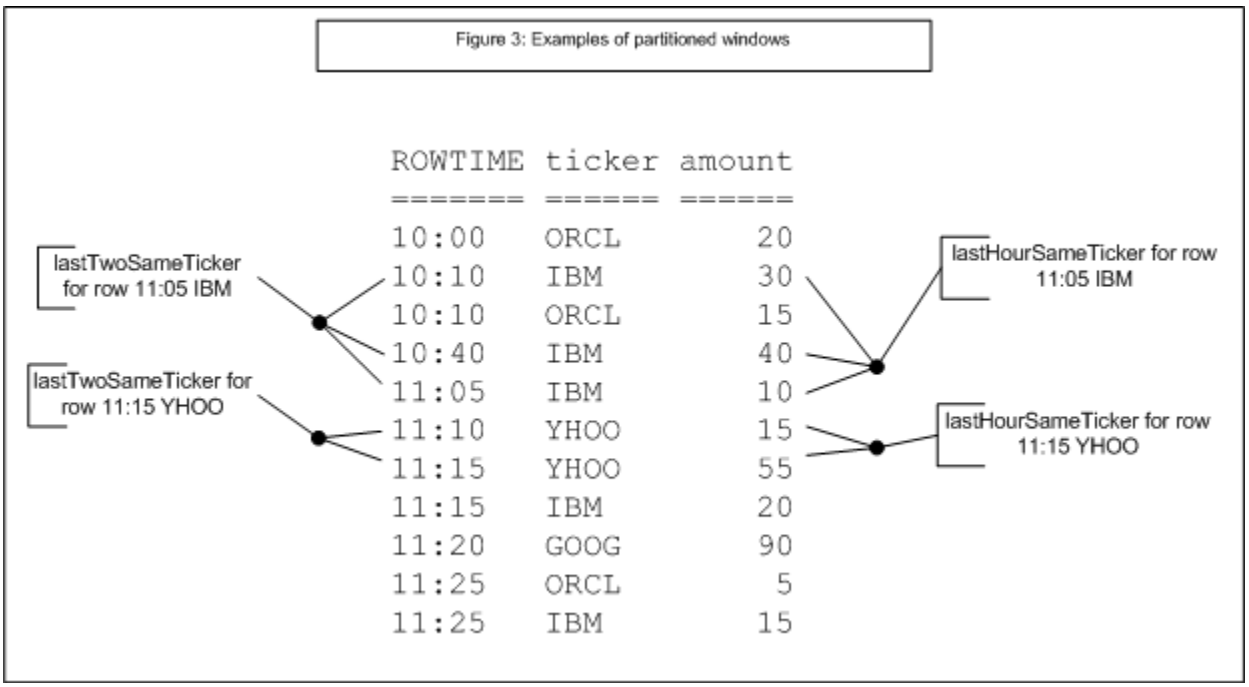
Figure 2: Examples of zero-width windows shows row-based and time-based windows of zero width. The row-based window lastZeroRows includes just the current row, and therefore always contains precisely one row. Note that ROWS CURRENT ROW is equivalent to ROWS 0 PRECEDING.

The time-based window lastZeroSeconds contains all rows with the same timestamp, of which there may be several. Note that RANGE CURRENT ROW is equivalent to RANGE INTERVAL '0' SECOND PRECEDING.



Third Example: Partitioning applied to row-based and time-based windows

Figure 3 shows windows that are similar to those in Figure 1 but with a PARTITION BY clause. For time-based window lastTwoSameTicker and the row-based window lastHourSameTicker, the window contains rows that meet the window criteria and have the same value of the ticker column. Note: Partitions are evaluated before windows.



ORDER BY clause

A streaming query can use ORDER BY if its leading expression is time-based and monotonic. For example, a streaming query whose leading expression is based on the ROWTIME column can use ORDER BY to do the following operations:

- Sort the results of a streaming GROUP BY.
- Sort a batch of rows arriving within a fixed time window of a stream.
- Perform streaming ORDER BY on windowed-joins.

The "time-based and monotonic" requirement on the leading expression means that the query

```
CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"
SELECT STREAM DISTINCT ticker FROM trades ORDER BY ticker
```

will fail, but the query

```
CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"
SELECT STREAM DISTINCT rowtime, ticker FROM trades ORDER BY ROWTIME, ticker
```

will succeed.

> **Note**
>
> The preceding examples use the DISTINCT clause to remove duplicate instances of the same ticker symbol from the result set, so that the results will be monotonic.

Streaming ORDER BY sorts rows using SQL-2008 compliant syntax for the ORDER BY clause. It can be combined with a UNION ALL statement, and can sort on expressions, such as:

```
CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"
SELECT STREAM x, y FROM t1
UNION ALL
SELECT STREAM a, b FROM t2 ORDER BY ROWTIME, MOD(x, 5)
```

The ORDER BY clause can specify ascending or descending sort order, and can use column ordinals, as well as ordinals specifying (referring to) the position of items in the select list.

> **Note**
>
> The UNION statement in the preceding query collects records from two separate streams for ordering.

Streaming ORDER BY SQL Declarations

The streaming ORDER BY clause includes the following functional attributes:

- Gathers rows until the monotonic expression in streaming ORDER BY clause does not change.
- Does not require streaming GROUP BY clause in the same statement.
- Can use any column with a basic SQL data type of TIMESTAMP, DATE, DECIMAL, INTEGER, FLOAT, CHAR, VARCHAR.
- Does not require that columns/expressions in the ORDER BY clause be present in the SELECT list of the statement.
- Applies all the standard SQL validation rules for ORDER BY clause.

The following query is an example of streaming ORDER BY:

```
CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"
SELECT STREAM state, city, SUM(amount)
FROM orders
GROUP BY FLOOR(ROWTIME TO HOUR), state, city
ORDER BY FLOOR(ROWTIME TO HOUR), state, SUM(amount)
```

T-sorting Stream Input

Amazon Kinesis Data Analytics real-time analytics use the fact that arriving data is ordered by ROWTIME. However, sometimes data arriving from multiple sources may not be time-synchronized.

While Amazon Kinesis Data Analytics can sort data from individual data sources that have been independently inserted into Amazon Kinesis Data Analytics application's native stream, in some cases such data may have already combined from multiple sources (such as for efficient consumption at an earlier stage in processing). At other times, high volume data sources could make direct insertion impossible.

In addition, an unreliable data source could block progress by forcing Amazon Kinesis Data Analytics application to wait indefinitely, unable to proceed until all connected data sources deliver. In this case, data from this source could be unsynchronized.

You can use the ORDER BY clause to resolve these issues. Amazon Kinesis Data Analytics uses a sliding time-based window of incoming rows to reorder those rows by ROWTIME.

Syntax

You specify the time-based parameter for sorting and the time-based window in which the streaming rows are to be time-sorted, using the following syntax:

```
ORDER BY <timestamp_expr> WITHIN
    <interval_literal>
```

Restrictions

The T-sort has the following restrictions:

- The datatype of the ORDER BY expression must be timestamp.
- The partially-ordered expression <timestamp_expr> must be present in the select list of the query with the alias ROWTIME.

- The leading expression of the ORDER BY clause must not contain the ROWTIME function and must not use the DESC keyword.
- The ROWTIME column needs to be fully qualified. For example:

- `ORDER BY FLOOR(ROWTIME TO MINUTE), ...` fails.
- `ORDER BY FLOOR(s.ROWTIME TO MINUTE), ...` works.

If any of these requirements are not met, the statement will fail with errors.

Additional notes:

- You cannot use incoming rowtimebounds. These are ignored by the system.
- If <timestamp_expr> evaluates to NULL, the corresponding row is discarded.

ROWTIME

ROWTIME is an operator and system column that returns the time at which a particular row of a stream was created.

It is used in four distinct ways:

- As an operator
- As a system column of a stream
- As a column alias, to override the timestamp on the current row
- As an ordinary column in a table

For more details, see the topics Timestamp, ROWTIME, and CURRENT_ROW_TIMESTAMP in this guide.

ROWTIME operator

When used in the SELECT clause of a streaming query, without being qualified by a preceding 'alias.' , ROWTIME is an operator that evaluates to the timestamp of the row that is just about to be generated.

Its type is always TIMESTAMP NOT NULL.

ROWTIME system column

Every stream has a ROWTIME column. To reference this column from within a query, qualify it with the stream name (or alias). For example, the following join query returns three timestamp columns: the system columns of its input streams, and the timestamp of the generated row.

```
SELECT STREAM
  o.ROWTIME AS leftRowtime,
  s.ROWTIME AS rightRowtime,
  ROWTIME AS joinRowtime
FROM Orders AS o
  JOIN Shipments OVER (RANGE INTERVAL '1' HOUR FOLLOWING) AS s
  ON o.orderId = s.orderId

leftRowtime         rightRowtime        joinRowtime
=================== =================== ===================
2008-02-20 10:15:00 2008-02-20 10:30:00 2008-02-20 10:15:00
2008-02-20 10:25:00 2008-02-20 11:15:00 2008-02-20 10:25:00
2008-02-20 10:25:30 2008-02-20 11:05:00 2008-02-20 10:25:30
```

As it happens, leftRowtime is always equal to joinRowtime, because the join is specified such that the output row timestamp is always equal to the ROWTIME column from the Orders stream.

It follows that every streaming query has a ROWTIME column. However, the ROWTIME column is not returned from a top-level JDBC query unless you explicitly include it in the SELECT clause. For example:

```
CREATE STREAM Orders(
  "orderId" INTEGER NOT NULL,
  "custId" INTEGER NOT NULL);
SELECT columnName
FROM ALL_STREAMS;

columnName
==========
orderId
custId

SELECT STREAM *
FROM Orders;
```

```
orderId custId
======= ======
    100    501
    101     22
    102    699

SELECT STREAM ROWTIME, *
FROM Orders;

ROWTIME             orderId custId
=================== ======= ======
2008-02-20 10:15:00     100    501
2008-02-20 10:25:00     101     22
2008-02-20 10:25:30     102    699
```

This is mainly to ensure compatibility with JDBC: the stream Orders declares two columns, so it makes sense that "SELECT STREAM *" should return two columns.

Functions

The topics in this section describe functions supported by streaming SQL.

Topics

- Aggregate Functions
- Analytic Functions
- Boolean Functions
- Conversion Functions
- Date and Time Functions
- Null Functions
- Numeric Functions
- Log Parsing Functions
- Sorting Functions
- Statistical Variance and Deviation Functions
- Streaming SQL Functions
- String and Search Functions

Aggregate Functions

Instead of returning a result calculated from a single row, an aggregate function returns a result calculated from aggregated data contained in a finite set of rows, or from information about a finite set of rows. An aggregate function may appear in any of the following:

- <selection list> portion of a SELECT clause
- ORDER BY clause
- HAVING clause

An aggregate function is different from Analytic Functions, which are always evaluated relative to a window that must be specified, and so they can't appear in a HAVING clause. Other differences are described in the table later in this topic.

Aggregate functions operate slightly differently in aggregate queries on tables than when you use them in aggregate queries on streams, as follows. If an aggregate query on tables contains a GROUP BY clause, the aggregate function returns one result per group in the set of input rows. Lacking an explicit GROUP BY clause is equivalent to GROUP BY (), and returns only one result for the entire set of input rows.

On streams, an aggregate query must contain an explicit GROUP BY clause on a monotonic expression based on rowtime. Without one, the sole group is the whole stream, which never ends, preventing any result from being reported. Adding a GROUP BY clause based on a monotonic expression breaks the stream into finite sets of rows, contiguous in time, and each such set can then be aggregated and reported.

Whenever a row arrives that changes the value of the monotonic grouping expression, a new group is started and the previous group is considered complete. Then, the Amazon Kinesis Data Analytics application outputs the value of the aggregate functions. Note that the GROUP BY clause may also include other non-monotonic expressions, in which case more than one result per set of rows may be produced.

Performing an aggregate query on streams is often referred to as streaming aggregation, as distinct from the windowed aggregation discussed in Analytic Functions and Windowed Aggregation on Streams. For more information about stream-to-stream joins, see JOIN clause.

If an input row contains a `null` in a column used as an input to a data analysis function, the data analysis function ignores the row (except for COUNT).

Differences Between Aggregate and Analytic Functions

Function Type	Outputs	Rows or Windows Used	Notes
Aggregate Functions	One output row per group of input rows.	All output columns are calculated over the same window or same group of rows.	COUNT DISTINCT is not allowed in streaming aggregati on. Statements of the following type are not allowed:

Function Type	Outputs	Rows or Windows Used	Notes
			SELECT COUNT(DISTINCT x) ... FROM ... GROUP BY ...
Analytic Functions	One output row for each input row.	Each output column may be calculated using a different window or partition.	COUNT DISTINCT can't be used as Analytic Functions or in windowed aggregation.

Streaming Aggregation and Rowtime Bounds

Normally, an aggregate query generates a result when a row arrives that changes the value of the monotonic expression in the GROUP BY. For example, if the query is grouped by FLOOR(rowtime TO MINUTE), and the rowtime of the current row is 9:59.30, then a new row with a rowtime of 10:00.00 will trigger the result.

Alternately, a rowtime bound can be used to advance the monotonic expression and enable the query to return a result. For example, if the query is grouped by FLOOR(rowtime TO MINUTE), and the rowtime of the current row is 9:59.30, then an incoming rowtime bound of 10:00.00 the query to return a result.

Aggregate Function List

Amazon Kinesis Data Analytics supports the following aggregate functions:

- AVG
- COUNT
- COUNT_DISTINCT_ITEMS_TUMBLING Function
- EXP_AVG
- FIRST_VALUE
- LAST_VALUE
- MAX

- MIN
- SUM
- TOP_K_ITEMS_TUMBLING Function

The following SQL uses the AVG aggregate function as part of a query to find the average age of all employees:

```
SELECT
    AVG(AGE) AS AVERAGE_AGE
FROM SALES.EMPS;
```

Result:

AVERAGE_AGE
38

To find the average age of employees in each department, we can add an explicit GROUP BY clause to the query:

```
SELECT
    DEPTNO,
    AVG(AGE) AS AVERAGE_AGE
FROM SALES.EMPS
GROUP BY DEPTNO;
```

Returns:

DEPTNO	AVERAGE_AGE
10	30
20	25
30	40

DEPTNO	AVERAGE_AGE
40	57

Examples of Aggregate Queries on Streams (Streaming Aggregation)

For this example, assume that the data in the following table is flowing through the stream called WEATHERSTREAM.

ROWTIME	CITY	TEMP
2018-11-01 01:00:00.0	Denver	29
2018-11-01 01:00:00.0	Anchorage	2
2018-11-01 06:00:00.0	Miami	65
2018-11-01 07:00:00.0	Denver	32
2018-11-01 09:00:00.0	Anchorage	9
2018-11-01 13:00:00.0	Denver	50
2018-11-01 17:00:00.0	Anchorage	10
2018-11-01 18:00:00.0	Miami	71
2018-11-01 19:00:00.0	Denver	43
2018-11-02 01:00:00.0	Anchorage	4
2018-11-02 01:00:00.0	Denver	39
2018-11-02 07:00:00.0	Denver	46
2018-11-02 09:00:00.0	Anchorage	3
2018-11-02 13:00:00.0	Denver	56
2018-11-02 17:00:00.0	Anchorage	2

ROWTIME	CITY	TEMP
2018-11-02 19:00:00.0	Denver	50
2018-11-03 01:00:00.0	Denver	36
2018-11-03 01:00:00.0	Anchorage	1

If you want to find the minimum and maximum temperature recorded anywhere each day (globally regardless of city), the minimum and maximum temperature can be calculated using the aggregate functions MIN and MAX respectively. To indicate that we want this information on a per-day basis (and to provide a monotonic expression as the argument of the GROUP BY clause), we use the FLOOR function to round each row's rowtime down to the nearest day:

```
SELECT STREAM

    FLOOR(WEATHERSTREAM.ROWTIME to DAY) AS FLOOR_DAY,
    MIN(TEMP) AS MIN_TEMP,
    MAX(TEMP) AS MAX_TEMP
FROM WEATHERSTREAM

GROUP BY FLOOR(WEATHERSTREAM.ROWTIME TO DAY);
```

The result of the aggregate query is shown in the following table.

FLOOR_DAY	MIN_TEMP	MAX_TEMP
2018-11-01 00:00:00.0	2	71
2018-11-02 00:00:00.0	2	56

There is no row for 2018-11-03, even though the example data does include temperature measurements on that day. This is because the rows for 2018-11-03 cannot be aggregated until all rows for that day are known to have arrived, and that will only happen when either a row with a rowtime of 2018-11-04 00:00:00.0 (or later) or a rowtime bound of 2018-11-04 00:00:00.0 (or later) arrives. If and when either did arrive, the next result would be as described in the following table.

FLOOR_DAY	MIN_TEMP	MAX_TEMP
2018-11-03 00:00:00.0	1	36

Let's say that instead of finding the global minimum and maximum temperatures each day, we want to find the minimum, maximum, and average temperature for each city each day. To do this, we use the SUM and COUNT aggregate functions to compute the average, and add CITY to the GROUP BY clause, as shown following:

```
SELECT STREAM
    FLOOR(WEATHERSTREAM.ROWTIME TO DAY) AS FLOOR_DAY,

    CITY,
    MIN(TEMP) AS MIN_TEMP,
    MAX(TEMP) AS MAX_TEMP,
    SUM(TEMP)/COUNT(TEMP) AS AVG_TEMP
FROM WEATHERSTREAM
GROUP BY FLOOR(WEATHERSTREAM.ROWTIME TO DAY), CITY;
```

The result of the aggregate query is shown in the following table.

FLOOR_DAY	CITY	MIN_TEMP	MAX_TEMP	AVG_TEMP
2018-11-01 00:00:00.0	Anchorage	2	10	7
2018-11-01 00:00:00.0	Denver	29	50	38
2018-11-01 00:00:00.0	Miami	65	71	68
2018-11-02 00:00:00.0	Anchorage	2	4	3
2018-11-02 00:00:00.0	Denver	39	56	47

In this case, the arrival of rows for a new day's temperature measurements triggers the aggregation of the previous day's data, grouped by CITY, which then results in one row being produced per city included in the day's measurements.

Here again, a rowtime bound 2018-11-04 00:00:00.0 could be used to prompt a result for 2018-11-03 prior to any actual measurements for 2018-11-04 coming in is shown in the following table.

FLOOR_DAY	CITY	MIN_TEMP	MAX_TEMP	AVG_TEMP
2018-11-03 00:00:00.0	Anchorage	1	1	1
2018-11-03 00:00:00.0	Denver	36	36	36

Windowed Aggregation on Streams

To illustrate how windowed aggregation works on Amazon Kinesis data streams, assume that the data in the following table is flowing through a stream called WEATHERSTREAM.

ROWTIME	CITY	TEMP
2018-11-01 01:00:00.0	Denver	29
2018-11-01 01:00:00.0	Anchorage	2
2018-11-01 06:00:00.0	Miami	65
2018-11-01 07:00:00.0	Denver	32
2018-11-01 09:00:00.0	Anchorage	9
2018-11-01 13:00:00.0	Denver	50
2018-11-01 17:00:00.0	Anchorage	10
2018-11-01 18:00:00.0	Miami	71

ROWTIME	CITY	TEMP
2018-11-01 19:00:00.0	Denver	43
2018-11-02 01:00:00.0	Anchorage	4
2018-11-02 01:00:00.0	Denver	39
2018-11-02 07:00:00.0	Denver	46
2018-11-02 09:00:00.0	Anchorage	3
2018-11-02 13:00:00.0	Denver	56
2018-11-02 17:00:00.0	Anchorage	2
2018-11-02 19:00:00.0	Denver	50
2018-11-03 01:00:00.0	Denver	36
2018-11-03 01:00:00.0	Anchorage	1

Suppose that you want to find the minimum and maximum temperature recorded in the 24-hour period prior to any given reading, globally, regardless of city. To do this, you define a window of `RANGE INTERVAL '1' DAY PRECEDING`, and use it in the `OVER` clause for the `MIN` and `MAX` analytic functions:

```
SELECT STREAM
       ROWTIME,
       MIN(TEMP) OVER W1 AS WMIN_TEMP,
       MAX(TEMP) OVER W1 AS WMAX_TEMP
FROM WEATHERSTREAM
WINDOW W1 AS (
   RANGE INTERVAL '1' DAY PRECEDING
);
```

Results

ROWTIME	WMIN_TEMP	WMAX_TEMP
2018-11-01 01:00:00.0	29	29
2018-11-01 01:00:00.0	2	29
2018-11-01 06:00:00.0	2	65
2018-11-01 07:00:00.0	2	65
2018-11-01 09:00:00.0	2	65
2018-11-01 13:00:00.0	2	65
2018-11-01 17:00:00.0	2	65
2018-11-01 18:00:00.0	2	71
2018-11-01 19:00:00.0	2	71
2018-11-02 01:00:00.0	2	71
2018-11-02 01:00:00.0	2	71
2018-11-02 07:00:00.0	4	71
2018-11-02 09:00:00.0	3	71
2018-11-02 13:00:00.0	3	71
2018-11-02 17:00:00.0	2	71
2018-11-02 19:00:00.0	2	56
2018-11-03 01:00:00.0	2	56
2018-11-03 01:00:00.0	1	56

Now, assume that you want to find the minimum, maximum, and average temperature recorded in the 24-hour period prior to any given reading, broken down by city. To do this, you add a `PARTITION BY` clause on `CITY` to the window specification, and add the `AVG` analytic function over the same window to the selection list:

```
SELECT STREAM
       ROWTIME,
       CITY,
       MIN(TEMP) over W1 AS WMIN_TEMP,
       MAX(TEMP) over W1 AS WMAX_TEMP,
       AVG(TEMP) over W1 AS WAVG_TEMP
FROM AGGTEST.WEATHERSTREAM
WINDOW W1 AS (
       PARTITION BY CITY
       RANGE INTERVAL '1' DAY PRECEDING
);
```

Results

ROWTIME	CITY	WMIN_TEMP	WMAX_TEMP	WAVG_TEMP
2018-11-01 01:00:00.0	Denver	29	29	29
2018-11-01 01:00:00.0	Anchorage	2	2	2
2018-11-01 06:00:00.0	Miami	65	65	65
2018-11-01 07:00:00.0	Denver	29	32	30
2018-11-01 09:00:00.0	Anchorage	2	9	5
2018-11-01 13:00:00.0	Denver	29	50	37

ROWTIME	CITY	WMIN_TEMP	WMAX_TEMP	WAVG_TEMP
2018-11-01 17:00:00.0	Anchorage	2	10	7
2018-11-01 18:00:00.0	Miami	65	71	68
2018-11-01 19:00:00.0	Denver	29	50	38
2018-11-02 01:00:00.0	Anchorage	2	10	6
2018-11-02 01:00:00.0	Denver	29	50	38
2018-11-02 07:00:00.0	Denver	32	50	42
2018-11-02 09:00:00.0	Anchorage	3	10	6
2018-11-02 13:00:00.0	Denver	39	56	46
2018-11-02 17:00:00.0	Anchorage	2	10	4
2018-11-02 19:00:00.0	Denver	39	56	46
2018-11-03 01:00:00.0	Denver	36	56	45
2018-11-03 01:00:00.0	Anchorage	1	4	2

Examples of Rowtime Bounds and Windowed Aggregation

This is an example of a windowed aggregate query:

```
SELECT STREAM ROWTIME, ticker, amount, SUM(amount)
   OVER (
       PARTITION BY ticker
       RANGE INTERVAL '1' HOUR PRECEDING)
AS hourlyVolume
FROM Trades
```

Because this is a query on a stream, rows pop out of this query as soon as they go in. For example, given the following inputs:

```
Trades: IBM 10 10 10:00:00
Trades: ORCL 20 10:10:00
Trades.bound: 10:15:00
Trades: ORCL 15 10:25:00
Trades: IBM 30 11:05:00
Trades.bound: 11:10:00
```

In this example, the output is as follows:

```
Trades: IBM 10 10 10:00:00
Trades: ORCL 20 20 10:10:00
Trades.bound: 10:15:00
Trades: ORCL 15 35 10:25:00
Trades: IBM 30 30 11:05:00
Trades.bound: 11:10:00
```

The rows still hang around behind the scenes for an hour, and thus the second ORCL row output has a total of 35; but the original IBM trade falls outside the "hour preceding" window, and so it is excluded from the IBM sum.

Example

Some business problems seem to need totals over the whole history of a stream, but this is usually not practical to compute. However, such business problems are often solvable by looking at the last day, the last hour, or the last N records. Sets of such records are called *windowed aggregates*.

They are easy to compute in a stream database, and can be expressed in ANSI (SQL:2008) standard SQL as follows:

```
SELECT STREAM ticker,
      avg(price) OVER lastHour AS avgPrice,
      max(price) OVER lastHour AS maxPrice
   FROM Bids
   WINDOW lastHour AS  (
      PARTITION BY ticker
      RANGE INTERVAL '1' HOUR PRECEDING)
```

Note

The `Interval_clause` must be of one of the following appropriate types:

- Integer literal with ROWS
- Numeric value for RANGE over a numeric column
- INTERVAL for a RANGE over a date/time/timestamp

AVG

Returns the average of a group of values from a windowed query. A windowed query is defined in terms of time or rows. For information about windowed queries, see Windowed Queries. To return an exponentially weighted average of a stream of value expressions selected in a specified time window, see EXP_AVG.

When you use AVG, be aware of the following:

- If you don't use the OVER clause, AVG is calculated as an aggregate function. In this case, the aggregate query must contain a GROUP BY clause on a monotonic expression based on ROWTIME that groups the stream into finite rows. Otherwise, the group is the infinite stream, and the query will never complete and no rows will be emitted. For more information, see Aggregate Functions.
- A windowed query that uses a GROUP BY clause processes rows in a tumbling window. For more information, see Tumbling Windows (Aggregations Using GROUP BY).
- If you use the OVER clause, AVG is calculated as an analytic function. For more information, see Analytic Functions.

- A windowed query that uses an OVER clause processes rows in a sliding window. For more information, see Sliding Windows

Syntax

Tumbling Windowed Query

```
AVG(number-expression) ... GROUP BY monotonic-expression | time-based-expression
```

Sliding Windowed Query

```
AVG([DISTINCT | ALL] number-expression) OVER window-specification
```

Parameters

DISTINCT

Performs the aggregate function only on each unique instance of a value.

ALL

Performs the aggregate function on all values. ALL is the default.

number-expression

Specifies the value expressions evaluated for each row in the aggregation.

OVER *window-specification*

Divides records in a stream partitioned by the time range interval or the number of rows. A window specification defines how records in the stream are partitioned by the time range interval or the number of rows.

GROUP BY *monotonic-expression* | *time-based-expression*

Groups records based on the value of the grouping expression, returning a single summary row for each group of rows that has identical values in all columns.

Examples

Example Dataset

The examples following are based on the sample stock dataset that is part of the Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Analytics application that has the sample stock ticker input stream. To learn how to create an Analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

Example 1: Return the Average of Values Using the GROUP BY Clause

In this example, the aggregate query has a `GROUP BY` clause on `ROWTIME` that groups the stream into finite rows. The `AVG` function is then calculated from the rows returned by the `GROUP BY` clause.

Using STEP (Recommended)

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    avg_price     DOUBLE);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS
  INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM
        ticker_symbol,
        AVG(price) AS avg_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol,
        STEP("SOURCE_SQL_STREAM_001".ROWTIME BY INTERVAL '60' SECOND);
```

Using FLOOR

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    avg_price     DOUBLE);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS
  INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM
        ticker_symbol,
        AVG(price) AS avg_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol,
        FLOOR("SOURCE_SQL_STREAM_001".ROWTIME TO MINUTE);
```

Results

The preceding examples output a stream similar to the following.

ROWTIME	TICKER_SYMBOL	AVG_PRICE
2017-02-16 21:46:00.0	NFS	85.84222412109375
2017-02-16 21:47:00.0	WAS	16.45800018310547
2017-02-16 21:47:00.0	PPL	23.33571434020996
2017-02-16 21:47:00.0	ALY	76.0999984741211

Example 2: Return the Average of Values Using the OVER Clause

In this example, the OVER clause divides records in a stream partitioned by the time range interval of '1' hour preceding. The AVG function is then calculated from the rows returned by the OVER clause.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    avg_price     DOUBLE);
CREATE OR REPLACE PUMP "STREAM_PUMP" AS
    INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM ticker_symbol,
        AVG(price) OVER (
```

```
            PARTITION BY ticker_symbol
            RANGE INTERVAL '1' HOUR PRECEDING) AS avg_price
    FROM "SOURCE_SQL_STREAM_001"
```

The preceding example outputs a stream similar to the following.

ROWTIME	TICKER_SYMBOL	AVG_PRICE
2017-02-16 21:57:25.509	AAPL	110.0180892944336
2017-02-16 21:57:25.509	TGT	54.96333312988281
2017-02-16 21:57:25.509	RFV	45.46332931518555
2017-02-16 21:57:25.509	SAC	17.972999572753906

Usage Notes

Amazon Kinesis Analytics doesn't support AVG applied to interval types. This functionality is a departure from the SQL:2008 standard.

When used as an analytic function, AVG returns null if the window being evaluated contains no rows, or if all rows contain null values. For more information, see Analytic Functions. AVG also returns null for a PARTITION BY clause for which the partition within the window matching the input row contains no rows or all rows are null. For more information about PARTITION BY, see WINDOW Clause (Sliding Windows).

AVG ignores null values from the set of values or a numeric expression. For example, each of the following return the value of 2:

- AVG(1, 2, 3) = 2
- AVG(1,null, 2, null, 3, null) = 2

Related Topics

- Windowed Queries
- EXP_AVG
- Aggregate Functions
- GROUP BY clause
- Analytic Functions

- Getting Started Exercise
- WINDOW Clause (Sliding Windows)

COUNT

Returns the number of qualifying rows of a group of values from a windowed query. A windowed query is defined in terms of time or rows. For information about windowed queries, see Windowed Queries.

When you use COUNT, be aware of the following:

- If you don't use the OVER clause, COUNT is calculated as an aggregate function. In this case, the aggregate query must contain a GROUP BY clause on a monotonic expression based on ROWTIME that groups the stream into finite rows. Otherwise, the group is the infinite stream, and the query will never complete and no rows will be emitted. For more information, see Aggregate Functions.
- A windowed query that uses a GROUP BY clause processes rows in a tumbling window. For more information, see Tumbling Windows (Aggregations Using GROUP BY).
- If you use the OVER clause, COUNT is calculated as an analytic function. For more information, see Analytic Functions.
- A windowed query that uses an OVER clause processes rows in a sliding window. For more information, see Sliding Windows

Syntax

Tumbling Windowed Query

```
COUNT(number-expression) ... GROUP BY monotonic-expression | time-based-expression
```

Sliding Windowed Query

```
COUNT(* | ALL number-expression) OVER window-specification
```

Parameters

*

Counts all rows.

ALL

Counts all rows. ALL is the default.

number-expression

Specifies the value expressions evaluated for each row in the aggregation.

OVER *window-specification*

Divides records in a stream partitioned by the time range interval or the number of rows. A window specification defines how records in the stream are partitioned by the time range interval or the number of rows.

GROUP BY *monotonic-expression* | *time-based-expression*

Groups records based on the value of the grouping expression returning a single summary row for each group of rows that has identical values in all columns.

Examples

Example Dataset

The examples following are based on the sample stock dataset that is part of Getting Started in the *Amazon Kinesis Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Analytics application that has the sample stock ticker input stream. To learn how to create an Analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

Example 1: Return the Number of Values Using the GROUP BY Clause

In this example, the aggregate query has a GROUP BY clause on ROWTIME that groups the stream into finite rows. The COUNT function is then calculated from the rows returned by the GROUP BY clause.

Using STEP (Recommended)

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    count_price     DOUBLE);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS
  INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM
        ticker_symbol,
        COUNT(Price) AS count_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol, STEP("SOURCE_SQL_STREAM_001".ROWTIME BY INTERVAL '60'
 SECOND);
```

Using FLOOR

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    count_price     DOUBLE);
-- CREATE OR REPLACE PUMP to insert into output
CREATE OR REPLACE PUMP "STREAM_PUMP" AS
  INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM
        ticker_symbol,
        COUNT(Price) AS count_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol, FLOOR("SOURCE_SQL_STREAM_001".ROWTIME TO MINUTE);
```

Results

The preceding examples output a stream similar to the following.

ROWTIME	TICKER_SYMBOL	COUNT_PRICE
2017-02-16 23:30:00.0	AAPL	9.0
2017-02-16 23:31:00.0	WSB	8.0
2017-02-16 23:31:00.0	AAPL	6.0
2017-02-16 23:31:00.0	UHN	9.0

Example 2: Return the Number of Values Using the OVER Clause

In this example, the OVER clause divides records in a stream partitioned by the time range interval of '1' hour preceding. The COUNT function is then calculated from the rows returned by the OVER clause.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    count_price     DOUBLE);
CREATE OR REPLACE PUMP "STREAM_PUMP" AS
    INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM ticker_symbol,
        COUNT(price) OVER (
            PARTITION BY ticker_symbol
            RANGE INTERVAL '1' HOUR PRECEDING) AS count_price
    FROM "SOURCE_SQL_STREAM_001"
```

The preceding example outputs a stream similar to the following.

ROWTIME	TICKER_SYMBOL	COUNT_PRICE
2017-02-16 23:36:16.729	WMT	3.0
2017-02-16 23:36:16.729	DFG	2.0
2017-02-16 23:36:16.729	RFV	4.0
2017-02-16 23:36:16.729	TGH	3.0

Usage Notes

Amazon Kinesis Analytics doesn't support the FILTER clause of the COUNT function or the use of COUNT DISTINCT in either aggregate functions or analytic functions. For more information on aggregate and analytic functions, see Aggregate Functions and Analytic Functions. This functionality is a departure from the SQL:2008 standard.

When used as an analytic function, COUNT returns zero if the window being evaluated contains no rows. For more information, see Analytic Functions. COUNT also returns zero for a PARTITION BY clause for which the partition within the window matching the input row contains no rows. For more information about PARTITION BY, see WINDOW Clause (Sliding Windows).

COUNT ignores null values from the set of values or a numeric expression. For example, each of the following return the value of 3:

- COUNT(1, 2, 3) = 3
- COUNT(1,null, 2, null, 3, null) = 3

Related Topics

- Windowed Queries
- Aggregate Functions
- GROUP BY clause
- Analytic Functions
- Getting Started Exercise
- WINDOW Clause (Sliding Windows)

COUNT_DISTINCT_ITEMS_TUMBLING Function

Returns a count of the number of distinct items in the specified in-application stream column over a tumbling window. The resulting count is approximate; the function uses the HyperLogLog algorithm.

For more information, see HyperLogLog.

When you use COUNT_DISTINCT_ITEMS_TUMBLING, be aware of the following:

- When there are less than or equal to 10,000 items in the window, the function returns an exact count.
- Getting an exact count of the number of distinct items can be inefficient and costly. Therefore, this function approximates the count. For example, if there are 100,000 distinct items, the algorithm may return 99,700. If cost and efficiency is not a consideration, you can write your own SELECT statement to get the exact count.

The following example demonstrates how to get an exact count of distinct rows for each ticker symbol in a five second tumbling window. The SELECT statement uses all of the columns (except ROWTIME) in determining the uniqueness.

```
CREATE OR REPLACE STREAM output_stream (ticker_symbol VARCHAR(4), unique_count
 BIGINT);


CREATE OR REPLACE PUMP stream_pump AS
INSERT INTO output_stream
SELECT STREAM TICKER_SYMBOL, COUNT(distinct_stream.price) AS unique_count
  FROM (
    SELECT STREAM DISTINCT rowtime as window_time,

      TICKER_SYMBOL,
      CHANGE,
      PRICE,
      STEP((SOURCE_SQL_STREAM_001.rowtime) BY INTERVAL '5' SECOND)

      FROM SOURCE_SQL_STREAM_001) as distinct_stream
  GROUP BY TICKER_SYMBOL,
      STEP((distinct_stream.window_time) BY INTERVAL '5' SECOND);
```

The function operates on a tumbling window. You specify the size of the tumbling window as a parameter.

Syntax

```
COUNT_DISTINCT_ITEMS_TUMBLING (

      in-application-streamPointer,
      'columnName',
      windowSize
   )
```

Parameters

The following sections describe the parameters.

in-application-streamPointer

Using this parameter, you provide a pointer to an in-application stream. You can set a pointer using the CURSOR function. For example, the following statement sets a pointer to `InputStream`.

```
CURSOR(SELECT STREAM * FROM InputStream)
```

columnName

Column name in your in-application stream that you want the function to use to count distinct values. Note the following about the column name:

- Must appear in single quotation marks ('). For example, `'column1'`.

windowSize

Size of the tumbling window in seconds. The size should be at least 1 second and should not exceed 1 hour = 3600 seconds.

Examples

Example Dataset

The examples following are based on the sample stock dataset that is part of Getting Started in the *Amazon Kinesis Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Analytics application that has the sample stock ticker input stream. To learn how to create an Analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

Example 1: Approximate the number of distinct values in a column

The following example demonstrates how to use the COUNT_DISTINCT_ITEMS_TUMBLING function to approximate the number of distinct TICKER_SYMBOL values in the current tumbling window of the in-application stream. For more information about tumbling windows, see Tumbling Windows.

```
CREATE OR REPLACE STREAM DESTINATION_SQL_STREAM (
    NUMBER_OF_DISTINCT_ITEMS BIGINT);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS
   INSERT INTO "DESTINATION_SQL_STREAM"
      SELECT STREAM *
      FROM TABLE(COUNT_DISTINCT_ITEMS_TUMBLING(
          CURSOR(SELECT STREAM * FROM "SOURCE_SQL_STREAM_001"),  -- pointer to the data
 stream
            'TICKER_SYMBOL',                                    -- name of column in
 single quotes
            60                                                  -- tumbling window
 size in seconds
      )
);
```

The preceding example outputs a stream similar to the following:

Filter by column name

ROWTIME	NUMBER_OF_DISTINCT_ITEMS
2017-03-16 21:22:57.49	47
2017-03-16 21:23:57.49	47
2017-03-16 21:24:57.49	47
2017-03-16 21:25:57.49	47

EXP_AVG

```
EXP_AVG ( expression, <time-interval> )
```

EXP_AVG returns an exponentially-weighted average (exponential moving average) of a stream of value expressions selected in a specified time window. EXP_AVG divides the specified window into intervals based on the value of <time-interval>. The values of the specified expression are weighted the most heavily for the most recent time-intervals and exponentially less heavily for earlier intervals.

Example

This example creates an exponentially-weighted average of the price of each stock ticker over a 30-second window such that the prices (for that ticker symbol) in the most recent 10-second subwindow carry double the weight of the prices in the middle 10-second subwindow and four times the weight of the prices in the oldest 10-second subwindow.

```
select stream t.rowtime, ticker, price,
exp_avg(price, INTERVAL '10' SECOND) over w as avgPrice
from t
window w as (partition by ticker range interval '30' second preceding);
```

In this example, 10 seconds is the half-life of the decay function, that is, the period over which the weights applied to the prices being averaged decrease by a factor of two. In other words, the older one will be given half as much weight as the newer one. It is specified as the time_interval in the call to EXP_AVG as interval '10' second .

FIRST_VALUE

```
FIRST_VALUE( <value-expression>) <null treatment> OVER <window-specification>
```

FIRST_VALUE returns the evaluation of the <value expression> from the first row that qualifies for the aggregate. FIRST_VALUE requires the OVER clause, and is considered an Analytic Functions. FIRST_VALUE has a null treatment option defined in the following table.

Null treatment option	Effect
FIRST_VALUE(x) IGNORE NULLS OVER <window-specification>	Returns first non null value of x in <window-specification>
FIRST_VALUE(x) RESPECT NULLS OVER <window-specification>	Returns first value, including null of x in <window-specification>

Null treatment option	Effect
FIRST_VALUE(x) OVER <window-specification>	Returns first value, including null of x in <window-specification>

LAST_VALUE

```
LAST_VALUE ( <value-expression> )  OVER <window-specification>
```

LAST_VALUE returns the evaluation of the <value expression> from the last row that qualifies for the aggregate.

Null Treatment Option	Effect
LAST_VALUE(x) IGNORE NULLS OVER <window-specification>	Returns last non null value of x in <window-specification>
LAST_VALUE(x) RESPECT NULLS OVER <window-specification>	Returns last value, including null of x in <window-specification>
LAST_VALUE(x) OVER <window-specification>	Returns last value, including null of x in <window-specification>

MAX

Returns the maximum value of a group of values from a windowed query. A windowed query is defined in terms of time or rows. For information about window queries, see Windowed Queries.

When you use MAX, be aware of the following:

- If you don't use the OVER clause, MAX is calculated as an aggregate function. In this case, the aggregate query must contain a GROUP BY clause on a monotonic expression based on ROWTIME that groups the stream into finite rows. Otherwise, the group is the infinite stream, and the query will never complete and no rows will be emitted. For more information, see Aggregate Functions.

- A windowed query that uses a GROUP BY clause processes rows in a tumbling window. For more information, see Tumbling Windows (Aggregations Using GROUP BY).
- If you use the OVER clause, MAX is calculated as an analytic function. For more information, see Analytic Functions.
- A windowed query that uses an OVER clause processes rows in a sliding window. For more information, see Sliding Windows

Syntax

Tumbling Windowed Query

```
MAX(number-expression) ... GROUP BY monotonic-expression | time-based-expression
```

Sliding Windowed Query

```
MAX(number-expression) OVER window-specification
```

Parameters

number-expression

Specifies the value expressions evaluated for each row in the aggregation.

OVER *window-specification*

Divides records in a stream partitioned by the time range interval or the number of rows. A window specification defines how records in the stream are partitioned by the time range interval or the number of rows.

GROUP BY *monotonic-expression* | *time-based-expression*

Groups records based on the value of the grouping expression returning a single summary row for each group of rows that has identical values in all columns.

Examples

Example Dataset

The examples following are based on the sample stock dataset that is part of Getting Started in the *Amazon Kinesis Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Analytics application that has the sample stock ticker input stream. To learn how to create an Analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

Example 1: Return the Maximum Value Using the GROUP BY Clause

In this example, the aggregate query has a GROUP BY clause on ROWTIME that groups the stream into finite rows. The MAX function is then calculated from the rows returned by the GROUP BY clause.

Using STEP (recommended)

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    max_price     DOUBLE);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS
  INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM
        ticker_symbol,
        MAX(Price) AS max_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol, STEP("SOURCE_SQL_STREAM_001".ROWTIME BY INTERVAL '60'
 SECOND);
```

Using FLOOR

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    max_price     DOUBLE);
-- CREATE OR REPLACE PUMP to insert into output
CREATE OR REPLACE PUMP "STREAM_PUMP" AS
  INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM
        ticker_symbol,
        MAX(Price) AS max_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol, FLOOR("SOURCE_SQL_STREAM_001".ROWTIME TO MINUTE);
```

Results

The preceding examples output a stream similar to the following.

ROWTIME	TICKER_SYMBOL	MAX_PRICE
2017-02-17 00:34:00.0	ASD	67.19999694824219
2017-02-17 00:35:00.0	BNM	179.3300018310547
2017-02-17 00:35:00.0	PPL	39.47999954223633
2017-02-17 00:35:00.0	QXZ	222.89999389648438

Example 2: Return the Maximum Value Using the OVER Clause

In this example, the OVER clause divides records in a stream partitioned by the time range interval of '1' hour preceding. The MAX function is then calculated from the rows returned by the OVER clause.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    max_price     DOUBLE);
CREATE OR REPLACE PUMP "STREAM_PUMP" AS
    INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM ticker_symbol,
        MAX(price) OVER (
            PARTITION BY ticker_symbol
```

```
            RANGE INTERVAL '1' HOUR PRECEDING) AS max_price
    FROM "SOURCE_SQL_STREAM_001"
```

The preceding example outputs a stream similar to the following.

ROWTIME	TICKER_SYMBOL	MAX_PRICE
2017-02-17 00:36:44.854	QAZ	198.2100067138672
2017-02-17 00:36:50.796	QXZ	232.49000549316406
2017-02-17 00:36:50.796	MJN	187.6999969482422
2017-02-17 00:36:50.796	WSB	101.0

Usage Notes

For string values, MAX is determined by which string is last in the collating sequence.

If MAX is used as an analytic function and the window being evaluated contains no rows, MAX returns null. For more information, see Analytic Functions.

Related Topics

- Windowed Queries
- Aggregate Functions
- GROUP BY clause
- Analytic Functions
- Getting Started Exercise
- WINDOW Clause (Sliding Windows)

MIN

Returns the minimum value of a group of values from a windowed query. A windowed query is defined in terms of time or rows. For information about windowed queries, see Windowed Queries.

When you use MIN, be aware of the following:

- If you don't use the `OVER` clause, `MIN` is calculated as an aggregate function. In this case, the aggregate query must contain a GROUP BY clause on a monotonic expression based on `ROWTIME` that groups the stream into finite rows. Otherwise, the group is the infinite stream, and the

query will never complete and no rows will be emitted. For more information, see Aggregate Functions.

- A windowed query that uses a GROUP BY clause processes rows in a tumbling window. For more information, see Tumbling Windows (Aggregations Using GROUP BY).
- If you use the OVER clause, MIN is calculated as an analytic function. For more information, see Analytic Functions.
- A windowed query that uses an OVER clause processes rows in a sliding window. For more information, see Sliding Windows

Syntax

Tumbling Windowed Query

```
MIN(number-expression) ... GROUP BY monotonic-expression | time-based-expression
```

Sliding Windowed Query

```
MIN(number-expression) OVER window-specification
```

Parameters

number-expression

Specifies the value expressions evaluated for each row in the aggregation.

OVER *window-specification*

Divides records in a stream partitioned by the time range interval or the number of rows. A window specification defines how records in the stream are partitioned by the time range interval or the number of rows.

GROUP BY *monotonic-expression* | *time-based-expression*

Groups records based on the value of the grouping expression returning a single summary row for each group of rows that has identical values in all columns.

Examples

Example Dataset

The examples following are based on the sample stock dataset that is part of the Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Analytics application that has the sample stock ticker input stream. To learn how to create an Analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

Example 1: Return the Minimum Value Using the GROUP BY Clause

In this example, the aggregate query has a `GROUP BY` clause on `ROWTIME` that groups the stream into finite rows. The `MIN` function is then calculated from the rows returned by the `GROUP BY` clause.

Using STEP (Recommended)

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    min_price     DOUBLE);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS
  INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM
        ticker_symbol,
        MIN(Price) AS min_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol, STEP("SOURCE_SQL_STREAM_001".ROWTIME BY INTERVAL '60'
 SECOND);
```

Using FLOOR

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    min_price     DOUBLE);
-- CREATE OR REPLACE PUMP to insert into output
CREATE OR REPLACE PUMP "STREAM_PUMP" AS
  INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM
        ticker_symbol,
        MIN(Price) AS min_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol, FLOOR("SOURCE_SQL_STREAM_001".ROWTIME TO MINUTE);
```

Results

The preceding examples output a stream similar to the following.

ROWTIME	TICKER_SYMBOL	MIN_PRICE
2017-02-17 22:45:00.0	QXZ	48.970001220703125
2017-02-17 22:46:00.0	WMT	62.540000915527344
2017-02-17 22:46:00.0	QWE	214.2100067138672
2017-02-17 22:46:00.0	CRM	28.260000228881836

Example 2: Return the Minimum Value Using the OVER Clause

In this example, the OVER clause divides records in a stream partitioned by the time range interval of '1' hour preceding. The MIN function is then calculated from the rows returned by the OVER clause.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    min_price     DOUBLE);
CREATE OR REPLACE PUMP "STREAM_PUMP" AS
    INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM ticker_symbol,
        MIN(price) OVER (
            PARTITION BY ticker_symbol
            RANGE INTERVAL '1' HOUR PRECEDING) AS min_price
```

```
    FROM "SOURCE_SQL_STREAM_001"
```

The preceding example outputs a stream similar to the following.

ROWTIME	TICKER_SYMBOL	MIN_PRICE
2017-02-17 22:49:51.147	NFS	84.94000244140625
2017-02-17 22:49:59.058	NFLX	112.37000274658203
2017-02-17 22:49:59.058	ASD	57.790000915527344
2017-02-17 22:49:59.058	DFG	147.3000030517578

Usage Notes

For string values, MIN is determined by which string is last in the collating sequence.

If MIN is used as an analytic function and the window being evaluated contains no rows, MIN returns null. For more information, see Analytic Functions.

Related Topics

- Windowed Queries
- Aggregate Functions
- GROUP BY clause
- Analytic Functions
- Getting Started Exercise
- WINDOW Clause (Sliding Windows)

SUM

Returns the sum of a group of values from a windowed query. A windowed query is defined in terms of time or rows. For information about windowed queries, see Windowed Queries.

When you use SUM, be aware of the following:

- If you don't use the OVER clause, SUM is calculated as an aggregate function. In this case, the aggregate query must contain a GROUP BY clause on a monotonic expression based on ROWTIME that groups the stream into finite rows. Otherwise, the group is the infinite stream, and the query will never complete and no rows will be emitted. For more information, see Aggregate Functions.

- A windowed query that uses a GROUP BY clause processes rows in a tumbling window. For more information, see Tumbling Windows (Aggregations Using GROUP BY).
- If you use the OVER clause, SUM is calculated as an analytic function. For more information, see Analytic Functions.
- A windowed query that uses an OVER clause processes rows in a sliding window. For more information, see Sliding Windows

Syntax

Tumbling Windowed Query

```
SUM(number-expression) ... GROUP BY monotonic-expression | time-based-expression
```

Sliding Windowed Query

```
SUM([DISTINCT | ALL] number-expression) OVER window-specification
```

Parameters

DISTINCT

Counts only distinct values.

ALL

Counts all rows. ALL is the default.

number-expression

Specifies the value expressions evaluated for each row in the aggregation.

OVER *window-specification*

Divides records in a stream partitioned by the time range interval or the number of rows. A window specification defines how records in the stream are partitioned by the time range interval or the number of rows.

GROUP BY *monotonic-expression* | *time-based-expression*

Groups records based on the value of the grouping expression returning a single summary row for each group of rows that has identical values in all columns.

Examples

Example Dataset

The examples following are based on the sample stock dataset that is part of the Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Analytics application that has the sample stock ticker input stream. To learn how to create an Analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

Example 1: Return the Sum of Values Using the GROUP BY Clause

In this example, the aggregate query has a `GROUP BY` clause on `ROWTIME` that groups the stream into finite rows. The `SUM` function is then calculated from the rows returned by the `GROUP BY` clause.

Using STEP (Recommended)

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    sum_price     DOUBLE);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS
  INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM
        ticker_symbol,
```

```
        SUM(price) AS sum_price
    FROM "SOURCE_SQL_STREAM_001"
     GROUP BY ticker_symbol, STEP("SOURCE_SQL_STREAM_001".ROWTIME BY INTERVAL '60'
 SECOND);
```

Using FLOOR

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    sum_price     DOUBLE);
-- CREATE OR REPLACE PUMP to insert into output
CREATE OR REPLACE PUMP "STREAM_PUMP" AS
  INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM
        ticker_symbol,
        SUM(price) AS sum_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol, FLOOR("SOURCE_SQL_STREAM_001".ROWTIME TO MINUTE);
```

Results

The preceding examples output a stream similar to the following.

ROWTIME	TICKER_SYMBOL	SUM_PRICE
2017-02-18 00:28:00.0	KIN	32.630001068115234
2017-02-18 00:29:00.0	VVS	50.409996032714844
2017-02-18 00:29:00.0	HJK	34.81999969482422
2017-02-18 00:29:00.0	KFU	356.6899719238281

Usage Notes

Amazon Kinesis Analytics doesn't support SUM applied to interval types. This functionality is a departure from the SQL:2008 standard.

SUM ignores null values from the set of values or a numeric expression. For example, each of the following return the value of 6:

- SUM(1, 2, 3) = 6
- SUM(1,null, 2, null, 3, null) = 6

Related Topics

- Windowed Queries
- Aggregate Functions
- GROUP BY clause
- Analytic Functions
- Getting Started Exercise
- WINDOW Clause (Sliding Windows)

TOP_K_ITEMS_TUMBLING Function

Returns the most frequently occurring values in the specified in-application stream column over a tumbling window. This can be used to find trending (most popular) values in a specified column.

For example, the Getting Started exercise uses a demo stream that provides continuous stock price updates (ticker_symbol, price, change, and other columns). Suppose you want to find the three most frequently traded stocks in each 1-minute tumbling window. You can use this function to find those ticker symbols.

When you use `TOP_K_ITEMS_TUMBLING`, be aware of the following:

- Counting each incoming record on your streaming source is not efficient, therefore the function approximates the most frequently occurring values. For example, when seeking the three most traded stocks, the function may return three of the five most traded stocks.

The function operates on a tumbling window. You specify the window size as a parameter.

For a sample application with step-by-step instructions, see Most Frequently Occurring Values.

Syntax

```
TOP_K_ITEMS_TUMBLING (
      in-application-streamPointer,
      'columnName',
      K,
      windowSize,
   )
```

Parameters

The following sections describe the parameters.

in-application-streamPointer

Pointer to an in-application stream. You can set a pointer using the CURSOR function. For example, the following statement sets a pointer to InputStream.

```
CURSOR(SELECT STREAM * FROM InputStream)
```

columnName

Column name in your in-application stream that you want to use to compute the topK values. Note the following about the column name:

> **Note**
>
> The column name must appear in single quotation marks ('). For example, `'column1'`.

K

Using this parameter, you specify how many of the most frequently occurring values from a specific column you want returned. The value K must be greater than or equal to one and cannot exceed 100,000.

windowSize

Size of the tumbling window in seconds. The size must be greater than or equal to one second and must not exceed 3600 seconds (one hour).

Examples

Example Dataset

The examples following are based on the sample stock dataset that is part of Getting Started in the *Amazon Kinesis Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Analytics application that has the sample stock ticker input stream. To learn how to create an Analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

Example 1: Return the Most Frequently Occurring Values

The following example retrieves the most frequently occuring values in the sample stream created in the Getting Started tutorial.

```
CREATE OR REPLACE STREAM DESTINATION_SQL_STREAM (
  "TICKER_SYMBOL" VARCHAR(4),
  "MOST_FREQUENT_VALUES" BIGINT
);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS
    INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM *
        FROM TABLE (TOP_K_ITEMS_TUMBLING(
            CURSOR(SELECT STREAM * FROM "SOURCE_SQL_STREAM_001"),
            'TICKER_SYMBOL',         -- name of column in single quotes
            5,                       -- number of the most frequently occurring values
            60                       -- tumbling window size in seconds
            )
        );
```

The preceding example outputs a stream similar to the following.

Filter by column name

ROWTIME	TICKER_SYMBOL	MOST_FREQUENT_VALUES
2017-04-12 17:14:45.305	QXZ	17
2017-04-12 17:15:45.305	QXZ	21
2017-04-12 17:15:45.305	AAPL	11
2017-04-12 17:15:45.305	DFT	8

Analytic Functions

An analytic function is one that returns a result calculated from data in (or about) a finite set of rows identified by a SELECT clause or in the ORDER BY clause.

The SELECT topic explains the order-by clause, showing the order-by chart, as well as the windowing clause (and window-specification chart). To see where an order-by clause is used in Select statements, see the Select chart in the SELECT topic of this guide.

1. Analytic functions must specify a window. Since there are a few restrictions on window specifications, and a few differences between specifying windows for windowed aggregation and windowed join, please see Allowed and Disallowed Window Specifications for explanations.
2. Analytic functions may only appear in the <selection list> portion of a SELECT clause or in the ORDER BY clause.

Other differences are described in the table later in this topic.

Performing queries using analytic functions is commonly referred to as windowed aggregation (discussed below), as distinct from Aggregate Functions.

Because of the presence of the window specification, queries that use analytic functions produce results in a different manner than do aggregate queries. For each row in the input set, the window specification identifies a different set of rows on which the analytic function operates. If the window specification also includes a PARTITION BY clause, then the only rows in the window that will be considered in producing a result will be those that share the same partition as the input row.

If an input row contains a null in a column used as an input to an analytic function, the analytic function ignores the row, except for COUNT, which does count rows with null values. In cases where the window (or in the case of a PARTITION BY, a partition within the window) contains no rows, an analytic function will return null. The exception to this is COUNT, which returns zero.

Differences Between Aggregate and Analytic Functions

Function Type	Outputs	Rows or Windows Used	Notes
Aggregate Functions	One output row per group of input rows.	All output columns are calculated over	COUNT DISTINCT is not allowed in Aggregate Functions

Function Type	Outputs	Rows or Windows Used	Notes
		the same window or same group of rows.	. Statements of the following type are not allowed: SELECT COUNT(DISTINCT x) ... FROM ... GROUP BY ...
Analytic Functions	One output row for each input row.	Each output column may be calculated using a different window or partition.	COUNT DISTINCT can't be used as analytic functions or in windowed aggregation.

Related Topics

- Windowed Aggregation on Streams
- SELECT statement
- SELECT clause

Boolean Functions

The topics in this section describe the boolean functions for Amazon Kinesis Data Analytics streaming SQL.

Topics

- ANY
- EVERY

ANY

```
ANY ( <boolean_expression> )
```

ANY returns true if the supplied boolean_expression is true in any of the selected rows. Returns false if the supplied boolean_expression is true in none of the selected rows.

Example

The following SQL snippet returns 'true' if the price for any ticker in the stream of trades is below 1. Returns 'false' if every price in the stream is 1 or greater.

```
SELECT STREAM ANY (price < 1) FROM trades
 GROUP BY (FLOOR trades.rowtime to hour)
```

EVERY

```
EVERY ( <boolean_expression> )
```

EVERY returns true if the supplied boolean_expression is true in all of the selected rows. Returns false if the supplied boolean_expression is false in any of the selected rows.

Example

The following SQL snippet returns 'true' if the price for every ticker in the stream of trades is below 1. Returns 'false' if any price is 1 or greater.

```
SELECT STREAM EVERY (price < 1) FROM trades
 GROUP BY (FLOOR trades.rowtime to hour)
```

Conversion Functions

The topics in this section describe the conversion functions for Amazon Kinesis Data Analytics streaming SQL.

For functions that convert to and from Datetime and Timestamp values, see Datetime Conversion Functions.

Topics

- CAST

CAST

CAST lets you convert one value expression or data type to another value expression or data type.

```
CAST ( <cast-operand> AS <cast-target> )
 <cast-operand> := <value-expression>
 <cast-target>  := <data-type>
```

Valid Conversions

Using CAST with source operands of the types listed in the first column below can create cast target types as listed in the second column, without restriction. Other target types are not supported.

Source Operand Types	Target Operand Types
Any numeric type (NUMERIC, DECIMAL, SMALLINT, INTEGER, BIGINT, REAL, DOUBLE)	VARCHAR, CHAR, or any numeric type (See Note A.)
VARCHAR, CHAR	All of the above, plus, DATE, TIME, TIMESTAMP , DAY-TIME INTERVAL, BOOLEAN
DATE	DATE, VARCHAR, CHAR, TIMESTAMP
TIME	TIME, VARCHAR, CHAR, TIMESTAMP
TIMESTAMP	TIME, VARCHAR, CHAR, TIMESTAMP, DATE
DAY-TIME INTERVAL	DAY-TIME INTERVAL, BIGINT, DECIMAL, CHAR, VARCHAR
BOOLEAN	VARCHAR, CHAR, BOOLEAN
BINARY, VARBINARY	BINARY, VARBINARY

Examples

2.1 DATE to CHAR/VARCHAR

```
+-------------+
```

```
|   EXPR$0    |
+-------------+
| 2008-08-23  |
+-------------+
1 row selected
```

(Note that if an inadequate output specification is supplied, no rows are selected:

```
values(cast(date'2008-08-23' as varchar(9)));
'EXPR$0'
No rows selected
```

(Because the date literal requires 10 characters)

In the next case, the date is blank-padded on the right (because of the semantics of the CHAR datatype):

```
+----------------------------+
|           EXPR$0           |
+----------------------------+
| 2008-08-23                 |
+----------------------------+
1 row selected
```

REAL to INTEGER

The real (NUMERIC or DECIMAL) is rounded by the cast:

```
+---------+
| EXPR$0  |
+---------+
| -2      |
+---------+
1 row selected
```

STRING to TIMESTAMP

There are two ways to convert a string to a timestamp. The first uses CAST, as shown in the next topic. The other uses Char To Timestamp(Sys).

Using CAST to Convert a String to a Timestamp

The example below illustrates this method for conversion:

```
'EXPR$0'
'2007-02-19 21:23:45'
1 row selected
```

If the input string lacks any one of the six fields (year, month, day, hours, minutes, seconds), or uses any delimiters different from those shown above, CAST will not return a value. (Fractional seconds are disallowed.)

If the input string is thus not in the appropriate format to be CAST, then to convert the string to a timestamp, you must use the CHAR_TO_TIMESTAMP method.

Using CHAR_TO_TIMESTAMP to convert a String to a Timestamp

When the input string is not in the appropriate format to be CAST, you can use the CHAR_TO_TIMESTAMP method. It has the additional advantage that you can specify which parts of the timestamp string you wish to use in subsequent processing, and create a TIMESTAMP value containing only those. To do so, you specify a template that identifies which parts you want, such as 'yyyy-MM' to use only the year and month parts.

The input-date-time string-to-be-converted can contain all or any parts of a full timestamp, that is, values for any or all of the standard elements ('yyyy-MM-dd hh:mm:ss'). If all these elements are present in your input string, and 'yyyy-MM-dd hh:mm:ss' is the template you supply, then the input-string elements are interpreted in that order as year, month, day, hour, minute, and seconds, such as in '2009-09-16 03:15:24'. The yyyy cannot be uppercase; the hh can be uppercase to mean using a 24-hour clock. For many examples of valid specifiers, see the table and examples later in this topic. For the full range of valid specifiers, see Class SimpleDateFormat on the Oracle website.

CHAR_TO_TIMESTAMP uses the template you specify as a parameter in the function call. The template causes the TIMESTAMP result to use only the parts of the input-date-time value that you specified in the template. Those fields in the resulting TIMESTAMP will then contain the corresponding data taken from your input-date-time string; fields not specified in your template will use default values (see below). The format of the template used by CHAR_TO_TIMESTAMP is defined by Class SimpleDateFormat, at which link all the specifiers are listed, some with examples. For more information, see Date and Time Patterns.

The function-call syntax is as follows:

```
CHAR_TO_TIMESTAMP('<format_string>','<input_date_time_string>')
```

Where <format_ string> is the template you specify for the parts of <date_time_string> you want, and <input_date_time_string> is the original string that is being converted to a TIMESTAMP result.

Each string must be enclosed in single quotes, and each element of the <input_date_time_string> must be in the range for its corresponding element in the template. Otherwise, no result is returned.

Example 1

- The input-string-element whose position corresponds with MM must be an integer from 1 to 12, because anything else does not represent a valid month.
- The input-string-element whose position corresponds with dd must be an integer from 1 to 31, because anything else does not represent a valid day.
- However, if MM is 2, dd cannot be 30 or 31, because February never has such days.

However, for months or days, the default starting value substituted for the omitted parts is 01.

For example, using '2009-09-16 03:15:24' as your input string, you can obtain a TIMESTAMP containing only the date, with zeros for the other fields such as hours, minutes, or seconds, by specifying

```
CHAR_TO_TIMESTAMP('yyyy-MM-dd','2009-09-16 03:15:24').
```

The result would be the TIMESTAMP 2009-09-16 00:00:00.

Example 2

- If the call had kept hours and minutes in the template while omitting months, days, and seconds, as illustrated in the following call --- --- CHAR_TO_TIMESTAMP('yyyy-hh-mm','2009-09-16 03:15:24') --- --- then the resulting TIMESTAMP would be 2009-01-01 03:15:00.

Template	Input String	Output TIMESTAMP	Notes
`'yyyy-MM-dd hh:mm:ss'`	`'2009-09-16 03:15:24'`	`'2009-09-16 03:15:24'`	Input string MUST use the form 'yyyy-

Template	Input String	Output TIMESTAMP	Notes
			MM-dd hh:mm:ss' or a subset or reorderin g thereof; using an input string like 'Wednesday, 16 September 2009 03:15:24' will NOT work, meaning that no output will result.
`'yyyy-mm'`	`'2012-02-08 07:23:19'`	`'2012-01-01 00:02:00'`	The template above specifies only year first and minutes second, so the second element in the input string ("02") is used as minutes. Default values are used for Month and Day ("01") and for hours and seconds ("00").

Template	Input String	Output TIMESTAMP	Notes
`'yyyy-ss-mm'`	`'2012-02-08 07:23:19'`	`'2012-01-01 00:08:02'`	The template above specifies only year, seconds, and minutes, in that order, so the second element in the input string ("02") is used as seconds and the third as minutes ("08"). Default values are used for Month and Day ("01") and for hours ("00").
`'MMM dd, yyyy'`	`'March 7, 2010'`	`'2010-03-07 00:00:00'`	MMM in the template above matches "March"; the template's 'comma space' matches the input string. --- --- If the template lacks the comma, so must the input string, or there is no output; --- --- If the input string lacks the comma, so must the template.

Template	Input String	Output TIMESTAMP	Notes
`'MMM dd,'`	`'March 7, 2010'`	`'1970-03-07 00:00:00'`	Note that the template above doesn't use a year specifier, causing the output TIMESTAMP to use the earliest year in this epoch, 1970.
`'MMM dd,y'`	`'March 7, 2010'`	`'2010-03-07 00:00:00'`	Using the template above, if the input string were 'March 7, 10', the output TIMESTAMP would be '0010-03-07 00:00:00'.
`'M-d'`	`'2-8'`	`'1970-02-08 00:00:00'`	Absent a yyyy specifier in the template, as above, the earliest year in this epoch (1970) is used. An input string of '2-8-2012' would give the same result; using '2012-2-8' would give no result because 2012 is not a valid month.

Template	Input String	Output TIMESTAMP	Notes
`'MM-dd-yyyy'`	`'06-23-2012 10:11:12'`	`'2012-06-23 00:00:00'`	Dashes as delimiters (as above) are fine, if template and input both use them in the same positions. Since the template omits hours, minutes, and seconds, zeroes are used in the output TIMESTAMP.
`'dd-MM-yy hh:mm:ss'`	`'23-06-11 10:11:12'`	`'2011-06-23 10:11:12'`	You can have the specifiers in any order as long as that order matches the meaning of the input string you supply, as above. The template and input string of the next example below have the same meaning (and the same output TIMESTAMP) as this example, but they specify months before days and seconds before hours.

Template	Input String	Output TIMESTAMP	Notes
`'MM-dd-yy ss:hh:mm'`	`'06-23-11 12:10:11'`	`'2011-06-23 10:11:12'`	In the template used above, the order of the month and day specifiers is reversed from the example just above, and the specifier for seconds is before hours instead of after minutes; but because the input string also puts months before days and seconds before hours, the meaning (and the output TIMESTAMP) is the same as the example ABOVE.
`'yy-dd-MM ss:hh:mm'`	`'06-23-11 12:10:11'`	`'2006-11-23 10:11:12'`	The template used above reverses (compared to the prior example above) the years and months specifiers, while the input string remains the same. In this case, the output TIMESTAMP uses the first element of the input string as the years, the second as the days, and the third as the months.

Template	Input String	Output TIMESTAMP	Notes
`'dd-MM-yy hh:mm'`	`'23-06-11 10:11:12'`	`'2011-06-23 10:11:00'`	With seconds omitted in the template, as above, the output TIMESTAMP uses 00 seconds. Any number of y specifiers produces the same result; but if the input string inadvertently uses a 1 instead of 11 for the year, as in '23-06-1 10:11:12', then the output TIMESTAMP becomes '0001-06-23 10:11:00'.

`'MM/dd/yy hh:mm:ss'`	`'12/19/11 10:11:12'`
	
	`'12/19/11 12:11:10'`
		`'2011-12-19 10:11:12'`
		
		`'2011-12-19 00:11:10'`
			Slashes as delimiters are fine, if template and input both use them in the same positions, as above; otherwise, no output.
			
			Using specifier hh, input times of 12:11:10 and 00:11:10 have the same meaning as a time in the morning.

Template	Input String	Output TIMESTAMP	Notes

`'MM/dd/yy HH:mm:ss'`	`'12/19/11 12:59:59'`
	`'12/19/11 21:08:07'`
		`'2011-12-19 12:59:59'`
		`'2011-12-19 21:08:07'`
			The input-string values '2011-12-19 00:11:12' or '2011-12-19 12:11:12' would fail with this template because '2011' is not a month, as required/expected by the template-string 'MM/dd/yy HH:mm:ss'.
			
			However, changing the template gives useful output:
			
			`values(cast(CHAR_TO_TIMESTAMP('y/MM/dd HH:mm:ss', '2011/12/19 00:11:12') as varchar(19)));`
			`'EXPR$0'`
			`'2011-12-19 00:11:12'`
			`1 row selected`
			
			`'12/19/11 00:11:12'` would fail with the above template (`'y/MM/dd'`), since 19 is not a valid month; supplying

Template	Input String	Output TIMESTAMP	Notes
			`'12/11/19 00:11:12'` works. `'2011-12-19 12:11:12'` would fail as input because dashes don't match the slashes in the template ; `'2011/12/19 12:11:12'` works. Note that for times after 12 noon, that is, for afternoon and evening times, the hours specifier must be HH instead of hh, and the input string must specify the afternoon or evening hour in 24-hour clock time, hours running from 00 to 23. --- --- Using specifier HH, input times of 12:11:10 and 00:11:10 have different meanings, the first as a time in the afternoon and the second as a time in the morning. --- --- Using the specifier hh, the

Template	Input String	Output TIMESTAMP	Notes
			times from 12:00 through 11:59:59 are morning times: --- --- Given the specifiers hh:mm:ss, the output TIMESTAMP will include '00:09:08' in the morning for both input string '12:09:08 ' and input string '00:09:08'; --- --- whereas --- --- Given the specifiers HH:mm:ss, the output TIMESTAMP for input string '00:09:08' in the morning will include '00:09:08' --- --- and the output TIMESTAMP for input string '12:09:08' in the afternoon will include '12:09:08'.

Further examples

The examples below illustrate using various templates with CHAR_TO_TIMESTAMP, including some common misunderstandings.

```
values (CHAR_TO_TIMESTAMP('yyyy-hh-mm','2009-09-16 03:15:24'));
```

```
'EXPR$0'
'2009-01-01 09:16:00'
1 row selected
```

Note that the fields in the input string above were used in the order given by the specifiers in the template, as defined by the dashes-as-delimiters in both template and input string: years first, then hours, then minutes. Since the specifiers for months and days are not present in the template, their values in the input string were ignored, with 01 substituted for both values in the output TIMESTAMP. The template specified hours and minutes as the second and third input values, so 09 became the hours and 16 became the minutes. No specifier was present for seconds, so 00 was used.

The years specifier can be alone or, after a delimiter matching the input string shows the end of the years specifier, with one of the hours:minutes:seconds specifiers:

```
values (CHAR_TO_TIMESTAMP('yyyy','2009-09-16 03:15:24') );
'EXPR$0'
'2009-01-01 00:00:00'
1 row selected
values (CHAR_TO_TIMESTAMP('yyyy hh','2009-09-16 03:15:24') );
'EXPR$0'
No rows selected
```

The template above fails because it has a space-as-delimiter before the "hh" rather than the dash delimiter used in the input string's date specification;

whereas the four templates below work because they use the same delimiter to separate the years specifier from the next specifier as is used in the input string's date specification (dash in the first case, space in the second, slash in the third, and dash in the fourth).

```
values (CHAR_TO_TIMESTAMP('yyyy-hh','2009-09-16 03:15:24') );
'EXPR$0'
'2009-01-01 09:00:00'
1 row selected
values (CHAR_TO_TIMESTAMP('yyyy hh','2009 09 16 03:15:24') );
'EXPR$0'
'2009-01-01 09:00:00'
1 row selected
values (CHAR_TO_TIMESTAMP('yyyy/hh','2009/09/16 03:15:24') );
'EXPR$0'
'2009-01-01 09:00:00'
```

```
1 row selected
values (CHAR_TO_TIMESTAMP('yyyy-mm','2009-09-16 03:15:24') );
'EXPR$0'
'2009-01-01 00:09:00'
1 row selected
```

However, if the template specifies months (MM), it cannot then specify hours, minutes, or seconds unless days are also specified:

Template specifying years and months only, thus omitting days/hours/minutes/seconds from the resulting TIMESTAMP:

```
values (CHAR_TO_TIMESTAMP('yyyy-MM','2009-09-16 03:15:24') );
'EXPR$0'
'2009-09-01 00:00:00'
1 row selected
```

The next two templates fail, lacking a 'days' specifier:

```
values (CHAR_TO_TIMESTAMP('yyyy-MM hh','2009-09-16 03:15:24') );
'EXPR$0'
No rows selected
values (CHAR_TO_TIMESTAMP('yyyy-MM hh:','2009-09-16 03:15:24') );
'EXPR$0'
No rows selected
```

The next three succeed, using a 'days' specifier:

```
 values (CHAR_TO_TIMESTAMP('yyyy-MM-dd hh','2009-09-16 03:15:24') );
'EXPR$0'
'2009-09-16 03:00:00'
1 row selected
```

The template above, 'yyyy-MM-dd hh', specifies only hours (hh) without minutes or seconds. Since hh is the 4th token/element of the template, its value is to be taken from the 4th token/element of the input string '2009-09-16 03:15:24' ; and that 4th element is 03, then used as the value output for hours. Since neither mm or ss is specified, the default or initial values defined as the starting point for mm and ss are used, which are zeroes.

```
values (CHAR_TO_TIMESTAMP('yyyy-MM-dd ss','2009-09-16 03:15:24') );
'EXPR$0'
```

```
'2009-09-16 00:00:03'
1 row selected
```

The template above, 'yyyy-MM-dd ss', specifies that the 4th token/element of the input string is to be used as seconds (ss). The 4th element of the input string '2009-09-16 03:15:24' is 03, which becomes the value output for seconds as specified in the template; and since neither hh nor mm is specified in the template, their default or initial values are used, which are zeroes.

```
values (CHAR_TO_TIMESTAMP('yyyy-MM-dd mm','2009-09-16 03:15:24') );
'EXPR$0'
'2009-09-16 00:03:00'
1 row selected
```

The template above, 'yyyy-MM-dd mm', specifies that the 4th token/element of the input string is to be used as minutes (mm). The 4th element of the input string '2009-09-16 03:15:24' is 03, which becomes the value output for minutes as specified in the template; and since neither hh nor ss is specified in the template, their default or initial values are used, which are zeroes.

Further failures, lacking a 'days' specifier:

```
values (CHAR_TO_TIMESTAMP('yyyy-MM- mm','2009-09-16 03:15:24') );
'EXPR$0'
No rows selected
values (CHAR_TO_TIMESTAMP('yyyy-MM   mm','2009-09-16 03:15:24') );
'EXPR$0'
No rows selected
values (CHAR_TO_TIMESTAMP('yyyy-MM   hh','2009-09-16 03:15:24') );
'EXPR$0'
No rows selected
```

About Delimiters and Values

Delimiters in the template must match those in the input string; values in the input string must be acceptable for the template specifiers to which they correspond.

As a general convention, a colon is used to separate hours from minutes, and minutes from seconds. Similarly, the general convention is to use a dash or slash to separate years from months and months from days. Any parallel usage seems to work, and the examples that follow illustrate this.

```
values (CHAR_TO_TIMESTAMP('MM/dd/yy hh:mm:ss','2009/09/16 03:15:24') );
```

```
'EXPR$0'
No rows selected
```

The example above fails because 2009 is not an acceptable value for months, which is the first specifier (MM) in the template.

```
values (CHAR_TO_TIMESTAMP('MM/dd/yy hh:mm:ss','09/16/11 03:15:24') );
'EXPR$0'
'2011-09-16 03:15:24'
1 row selected
```

The example above succeeds because the delimiters are parallel (slashes to slashes, colons to colons) and each value is acceptable for the corresponding specifier.

```
values (CHAR_TO_TIMESTAMP('MM/dd/yy hh/mm/ss','09/16/11 03/15/24') );
'EXPR$0'
'2011-09-16 03:15:24'
1 row selected
```

The example above succeeds because the delimiters are parallel (all slashes) and each value is acceptable for the corresponding specifier.

```
values (CHAR_TO_TIMESTAMP('MM/dd/yy hh-mm-ss','09/16/11 03-15-24') );
'EXPR$0'
'2011-09-16 03:15:24'
1 row selected
values (CHAR_TO_TIMESTAMP('yyyy|MM|dd hh|mm|ss','2009|09|16 03|15|24') );
'EXPR$0'
'2009-09-16 03:15:24'
1 row selected
values (CHAR_TO_TIMESTAMP('yyyy@MM@dd hh@mm@ss','2009@09@16 03@15@24') );
'EXPR$0'
'2009-09-16 03:15:24'
1 row selected
```

The examples above succeed because the delimiters are parallel and the values are acceptable per specifier.

In the following examples, note that omissions in the supplied string can cause the template value 'yyyy' to produce logical but unintended or unexpected results. The value given as the year in the resulting TIMESTAMP value derives directly from the first element in the supplied string.

```
 VALUES(CHAR_TO_TIMESTAMP('yyyy','09-16 03:15'));
'EXPR$0'
'0009-01-01 00:00:00'
1 row selected
VALUES(CHAR_TO_TIMESTAMP('yyyy','16 03:15'));
'EXPR$0'
'0016-01-01 00:00:00'
1 row selected
```

TIMESTAMP to STRING

```
 values(cast( TIMESTAMP '2007-02-19 21:25:35' AS VARCHAR(25)));
'EXPR$0'
'2007-02-19 21:25:35'
1 row selected
```

Note that CAST requires a TIMESTAMP-literal to have literally the full format of 'yyyy-mm-dd hh:mm:ss'. If any part of that full format is missing, the literal is rejected as illegal, as seen below:

```
 values( TIMESTAMP '2007-02-19 21:25');
Error: Illegal TIMESTAMP literal '2007-02-19 21:25':
                                  not in format 'yyyy-MM-dd HH:mm:ss' (state=,code=0)
 values( TIMESTAMP '2007-02-19 21:25:00');
'EXPR$0'
'2007-02-19 21:25:00'
1 row selected
```

Also, if an inadequate output specification is supplied, no rows are selected:

```
values(cast( TIMESTAMP '2007-02-19 21:25:35' AS VARCHAR(18)));
'EXPR$0'
No rows selected
(Because the timestamp literal requires 19 characters)
```

These restrictions apply similarly to CASTing to TIME or DATE types.

STRING to TIME

```
 values(cast(' 21:23:45.0' AS TIME));
```

```
'EXPR$0'
'21:23:45'
1 row selected
```

For more information, see Note A.

STRING to DATE

```
 values(cast('2007-02-19' AS DATE));
'EXPR$0'
'2007-02-19'
1 row selected
```

> **Note A**
>
> Note that CAST for strings requires that the string operand for casting to TIME or DATE have the exact form required to represent a TIME or DATE, respectively.

As shown below, the cast fails if:

- the string operand includes data extraneous to the targeted type, or
- the INTERVAL operand ('day hours:minutes:seconds.milliseconds') does not include necessary data, or
- the specified output field is too small to hold the conversion results.

```
 values(cast('2007-02-19 21:23:45.0' AS TIME));
'EXPR$0'
No rows selected
```

Fails because it includes date information not allowed as a TIME.

```
 values(cast('2007-02-19 21:23:45.0' AS DATE));
'EXPR$0'
No rows selected
```

Fails because it includes time information not allowed as a DATE.

```
 values(cast('2007-02-19 21' AS DATE));
'EXPR$0'
No rows selected
```

Fails because it includes time information not allowed as a DATE.

```
 values(cast('2009-02-28' AS DATE));
'EXPR$0'
'2009-02-28'
1 row selected
```

Succeeds because it includes a correct representation of date string.

```
 values(CAST (cast('2007-02-19 21:23:45.0' AS TIMESTAMP) AS DATE));
'EXPR$0'
'2007-02-19'
1 row selected
```

Succeeds because it correctly converts string to TIMESTAMP before casting to DATE.

```
 values(cast('21:23' AS TIME));
'EXPR$0'
No rows selected
```

Fails because it lacks time information (seconds) required for a TIME.

(Specifying fractional seconds is allowed but not required.)

```
 values(cast('21:23:34:11' AS TIME));
'EXPR$0'
No rows selected
```

Fails because it includes incorrect representation of fractional seconds.

```
 values(cast('21:23:34.11' AS TIME));
'EXPR$0'
'21:23:34'
```

```
1 row selected
```

Succeeds because it includes correct representation of fractional seconds.

```
 values(cast('21:23:34' AS TIME));
'EXPR$0'
'21:23:34'
1 row selected
```

This example succeeds because it includes correct representation of seconds without fractions of a second.

INTERVAL to exact numerics

CAST for intervals requires that the INTERVAL operand have only one field in it, such as MINUTE, HOUR, SECOND.

If the INTERVAL operand has more than one field, such as MINUTE TO SECOND, the cast fails, as shown below:

```
 values ( cast (INTERVAL '120' MINUTE(3) as decimal(4,2)));
+---------+
| EXPR$0  |
+---------+
+---------+
No rows selected

 values ( cast (INTERVAL '120' MINUTE(3) as decimal(4)));
+---------+
| EXPR$0  |
+---------+
| 120     |
+---------+
1 row selected

 values ( cast (INTERVAL '120' MINUTE(3) as decimal(3)));
+---------+
| EXPR$0  |
+---------+
| 120     |
+---------+
1 row selected
```

```
 values ( cast (INTERVAL '120' MINUTE(3) as decimal(2)));
+---------+
| EXPR$0  |
+---------+
+---------+
No rows selected

 values cast(interval '1.1' second(1,1) as decimal(2,1));
+---------+
| EXPR$0  |
+---------+
| 1.1     |
+---------+
1 row selected

 values cast(interval '1.1' second(1,1) as decimal(1,1));
+---------+
| EXPR$0  |
+---------+
+---------+
No rows selected
```

For year, decimal fractions are disallowed as input and as output.

```
values cast(interval '1.1' year (1,1) as decimal(2,1));
Error: org.eigenbase.sql.parser.SqlParseException: Encountered "," at line 1, column
 35.
Was expecting:
    ")" ... (state=,code=0)
values cast(interval '1.1' year (1) as decimal(2,1));
Error: From line 1, column 13 to line 1, column 35:
              Illegal interval literal format '1.1' for INTERVAL YEAR(1)
 (state=,code=0)
values cast(interval '1.' year (1) as decimal(2,1));
Error: From line 1, column 13 to line 1, column 34:
              Illegal interval literal format '1.' for INTERVAL YEAR(1) (state=,code=0)
values cast(interval '1' year (1) as decimal(2,1));
+---------+
| EXPR$0  |
+---------+
| 1.0     |
+---------+
1 row selected
```

For additional examples, see SQL Operators: Further examples.

Limitations

Amazon Kinesis Data Analytics does not support directly casting numeric values to interval values. This is a departure from the SQL:2008 standard. The recommended way to convert a numeric to an interval is to multiply the numeric value against a specific interval value. For example, to convert the integer time_in_millis to a day-time interval:

```
time_in_millis * INTERVAL '0 00:00:00.001' DAY TO SECOND
```

For example:

```
 values cast( 5000 * (INTERVAL '0 00:00:00.001' DAY TO SECOND) as varchar(11));
'EXPR$0'
'5000'
1 row selected
```

Date and Time Functions

The following built-in functions relate to dates and time.

Topics

- Time Zones
- Datetime Conversion Functions
- Date, Timestamp, and Interval Operators
- Date and Time Patterns
- CURRENT_DATE
- CURRENT_ROW_TIMESTAMP
- CURRENT_TIME
- CURRENT_TIMESTAMP
- EXTRACT
- LOCALTIME
- LOCALTIMESTAMP

- TSDIFF

Of these, the SQL extension CURRENT_ROW_TIMESTAMP is the most useful for a streaming context, because it gives you information about the times of streaming data as it emerges, not just when the query is run. This is a key difference between a streaming query and a traditional RDMS query: streaming queries remain "open," producing more data, so the timestamp for when the query was run does not offer good information.

LOCALTIMESTAMP, LOCALTIME, CURRENT_DATE, and CURRENT_TIMESTAMP all produce results which are set to values at the time the query first executes. Only CURRENT_ROW_TIMESTAMP generates a row with a unique timestamp (date and time) for each row.

A query run with LOCALTIMESTAMP (or CURRENT_TIMESTAMP or CURRENT_TIME) as one of the columns puts into all output rows the time the query is first run. If that column instead contains CURRENT_ROW_TIMESTAMP, each output row gets a newly-calculated value of TIME representing when that row was output.

To return a part (such as the day of the month) from a Datetime value, use EXTRACT

Time Zones

Amazon Kinesis Data Analytics runs in UTC. As a result, all time functions return time in UTC.

Datetime Conversion Functions

You specify date and time formats using patterned letters. Date and time pattern strings use unquoted letters from 'A' to 'Z' and from 'a' to 'z', with each letter representing a formatting element.

For more information, see Class SimpleDateFormat on the Oracle website.

> **Note**
>
> If you include other characters, they will be incorporated into the output string during formatting or compared to the input string during parsing.

The pattern letters in the following table are defined (all other characters from 'A' to 'Z' and from 'a' to 'z' are reserved).

Letter	Date or Time Component	Presentation	Examples
y	Year	Year	yyyy; yy 2018;18
Y	Week year	Year	YYYY; YY 2009; 09
M	Month in year	Month	MMM;MM;MM July; Jul; 07
w	Week in year	Number	ww; 27
W	Week in month	Number	W 2
D	Day in year	Number	DDD 321
d	Day in month	Number	dd 10
F	Day of week in month	Number	F 2
E	Day name in week	Text	Tuesday; Tue
u	Day number of week (1 = Monday, ..., 7 = Sunday)	Number	1
a	Am/pm marker	Text	PM
H	Hour in day (0-23)	Number	0
k	Hour in day (1-24)	Number	24
K	Hour in am/pm (0-11)	Number	0
h	Hour in am/pm (1-12)	Number	12
m	Minute in hour	Number	30
s	Second in minute	Number	55

Letter	Date or Time Component	Presentation	Examples
S	Millisecond	Number	978
z	Time zone	General time zone	Pacific Standard Time; PST; GMT-08:00
Z	Time zone	RFC 822 time zone	-0800
X	Time zone	ISO 8601 time zone	-08; -0800; -08:00

You determine the exact presentation by repeating pattern letters, along the lines of YYYY.

Text

If the number of repeated pattern letters is 4 or more, the full form is used; otherwise a short or abbreviated form is used if available. For parsing, both forms are accepted, independent of the number of pattern letters.

Number

For formatting, the number of pattern letters is the minimum number of digits, and shorter numbers are zero-padded to this amount. For parsing, the number of pattern letters is ignored unless it's needed to separate two adjacent fields.

Year

If the formatter's Calendar is the Gregorian calendar, the following rules are applied.

- For formatting, if the number of pattern letters is 2, the year is truncated to 2 digits; otherwise it is interpreted as a number.
- For parsing, if the number of pattern letters is more than 2, the year is interpreted literally, regardless of the number of digits. So using the pattern "MM/dd/yyyy", "01/11/12" parses to Jan 11, 12 A.D.

For parsing with the abbreviated year pattern ("y" or "yy"), SimpleDateFormat must interpret the abbreviated year relative to some century. It does this by adjusting dates to be within 80 years before and 20 years after the time the SimpleDateFormat instance is created. For example,

using a pattern of "MM/dd/yy" and a SimpleDateFormat instance created on Jan 1, 2018, the string "01/11/12" would be interpreted as Jan 11, 2012 while the string "05/04/64" would be interpreted as May 4, 1964. During parsing, only strings consisting of exactly two digits, as defined by Character.isDigit(char), will be parsed into the default century. Any other numeric string, such as a one digit string, a three or more digit string, or a two digit string that isn't all digits (for example, "-1"), is interpreted literally. So "01/02/3" or "01/02/003" are parsed, using the same pattern, as Jan 2, 3 AD. Likewise, "01/02/-3" is parsed as Jan 2, 4 BC.

Otherwise, calendar system specific forms are applied. For both formatting and parsing, if the number of pattern letters is 4 or more, a calendar specific long form is used. Otherwise, a calendar specific short or abbreviated form is used.

Char To Timestamp(Sys)

The Char to Timestamp function is one of the most frequently-used system functions, because it lets you create a timestamp out of any correctly formatted input string. Using this function, you can specify which parts of the timestamp string you wish to use in subsequent processing, and create a TIMESTAMP value containing only those. To do so, you specify a template that identifies the parts of the timestamp you want. For example, to use only year and month, you would specify 'yyyy-MM'.

The input date-time string can contain any parts of a full timestamp ('yyyy-MM-dd hh:mm:ss'). If all these elements are present in your input string, and 'yyyy-MM-dd hh:mm:ss' is the template you supply, then the input-string elements are interpreted in that order as year, month, day, hour, minute, and seconds, such as in '2009-09-16 03:15:24'. The yyyy cannot be uppercase; the hh can be uppercase to mean using a 24-hour clock.

For the full range of valid specifiers, see Class SimpleDateFormat on the Oracle website.

CHAR_TO_TIMESTAMP uses the template you specify as a parameter in the function call. The template causes the TIMESTAMP result to use only the parts of the input-date-time value that you specified in the template. Those fields in the resulting TIMESTAMP contain the corresponding data taken from your input-date-time string. Fields not specified in your template will use default values (see below). The format of the template used by CHAR_TO_TIMESTAMP is defined by the Class SimpleDateFormat on the Oracle website. For more information, see Date and Time Patterns.

The function-call syntax is as follows:

```
CHAR_TO_TIMESTAMP('<format_string>','<input_date_time_string>')
```

Where <format_ string> is the template you specify for the parts of <date_time_string> you want, and <input_date_time_string> is the original string that is being converted to a TIMESTAMP result.

Note that each string must be enclosed in single quotes and each element of the <input_date_time_string> must be in the range for its corresponding element in the template, otherwise no result is returned.

For example, the input-string-element whose position corresponds with MM must be an integer from 1 to 12, because anything else does not represent a valid month. Similarly, the input-string-element whose position corresponds with dd must be an integer from 1 to 31, because anything else does not represent a valid day. (However, if MM is 2, dd cannot be 30 or 31, because February never has such days.)

For hours, minutes, or seconds, the default starting value is zero, so when those specifiers are omitted from the template, zeroes are substituted. For months or days, the default starting value substituted for the omitted parts is 01.

For example, using '2009-09-16 03:15:24' as your input string, you can obtain a TIMESTAMP containing only the date, with zeros for the other fields such as hours, minutes, or seconds.

```
CHAR_TO_TIMESTAMP('yyyy-MM-dd','2009-09-16 03:15:24').
```

The result would is TIMESTAMP 2009-09-16 00:00:00.

If the call had kept hours and minutes in the template while omitting months, days, and seconds, as illustrated in the following call.

```
--- --- CHAR_TO_TIMESTAMP('yyyy-hh-mm','2009-09-16 03:15:24')
```

Then, the resulting TIMESTAMP would be 2009-01-01 03:15:00.

Template Strings to Create Specific Output Timestamps shows further illustrative examples of templates and input strings used to create the indicated output TIMESTAMPs.

> **Note**
>
> Input string MUST use the form 'yyyy-MM-dd hh:mm:ss' or a subset or reordering thereof. As a result, using an input string like 'Wednesday, 16 September 2009 03:15:24' will NOT work, meaning that no output will result.

About Delimiters and Values

Delimiters in the template must match those in the input string and values in the input string must be acceptable for the template specifiers to which they correspond.

As a general convention, a colon is used to separate hours from minutes, and minutes from seconds. Similarly, the general convention is to use a dash or slash to separate years from months and months from days.

For example, the following template has values that line up correctly with the input string.

```
values (CHAR_TO_TIMESTAMP('MM/dd/yy hh:mm:ss','09/16/11 03:15:24') );
'EXPR$0'
'2011-09-16 03:15:24'
1 row selected
```

If values in the input string are not acceptable for the template specifiers to which they correspond, the result fails, as in the following example.

```
values (CHAR_TO_TIMESTAMP('MM/dd/yy hh:mm:ss','2009/09/16 03:15:24') );
'EXPR$0'
No rows selected
```

This example returns no rows because 2009 is not an acceptable value for months, which is the first specifier (MM) in the template.

Omissions in the supplied string can cause the template value 'yyyy' to produce logical but unintended or unexpected results. The following examples each return an erroneous year, but one that derives directly from the first element in the supplied string.

```
 VALUES(CHAR_TO_TIMESTAMP('yyyy','09-16 03:15'));
'EXPR$0'
'0009-01-01 00:00:00'
1 row selected
VALUES(CHAR_TO_TIMESTAMP('yyyy','16 03:15'));
'EXPR$0'
'0016-01-01 00:00:00'
1 row selected
```

Examples Using Templates to Create TIMESTAMPS

The order of the template must match the input string. That means that you cannot specify "hh" after "yyyy" and expect the method to find the hour automatically. For example, the following template specifies years first, then hours, then minutes, and returns an erroneous result.

```
 values (CHAR_TO_TIMESTAMP('yyyy-hh-mm','2009-09-16 03:15:24'));
'EXPR$0'
'2009-01-01 09:16:00'
1 row selected
```

Since the specifiers for months and days are not present in the template, their values in the input string were ignored, with 01 substituted for both values in the output TIMESTAMP. The template specified hours and minutes as the second and third input values, so 09 became the hours and 16 became the minutes. No specifier was present for seconds, so 00 was used.

The years specifier can be alone or after a delimiter matching the input string shows the end of the years specifier, with one of the hours:minutes:seconds specifiers.

```
values (CHAR_TO_TIMESTAMP('yyyy','2009-09-16 03:15:24') );
'EXPR$0'
'2009-01-01 00:00:00'
1 row selected
```

In contrast, the template below fails because it has a space-as-delimiter before the "hh" rather than the dash delimiter used in the input string's date specification.

```
  values (CHAR_TO_TIMESTAMP('yyyy hh','2009-09-16 03:15:24') );
  'EXPR$0'
  No rows selected
```

The four templates below work because they use the same delimiter to separate the years specifier from the next specifier as is used in the input string's date specification (dash in the first case, space in the second, slash in the third, and dash in the fourth).

```
values (CHAR_TO_TIMESTAMP('yyyy-hh','2009-09-16 03:15:24') );
'EXPR$0'
'2009-01-01 09:00:00'
1 row selected
values (CHAR_TO_TIMESTAMP('yyyy hh','2009 09 16 03:15:24') );
'EXPR$0'
```

```
'2009-01-01 09:00:00'
1 row selected
values (CHAR_TO_TIMESTAMP('yyyy/hh','2009/09/16 03:15:24') );
'EXPR$0'
'2009-01-01 09:00:00'
1 row selected
values (CHAR_TO_TIMESTAMP('yyyy-mm','2009-09-16 03:15:24') );
'EXPR$0'
'2009-01-01 00:09:00'
1 row selected
```

However, if the template specifies months (MM), it cannot then specify hours, minutes, or seconds unless days are also specified.

Template Strings to Create Specific Output Timestamps

Template	Input String	Output TIMESTAMP	Notes
`'yyyy-MM-dd hh:mm:ss'`	`'2009-09-16 03:15:24'`	`'2009-09-16 03:15:24'`	
`'yyyy-mm'`	`'2011-02-08 07:23:19'`	`'2011-01-01 00:02:00'`	The template above specifies only year first and minutes second, so the second element in the input string ("02") is used as minutes. Default values are used for Month and Day ("01") and for hours and seconds ("00").
`'MMM dd, yyyy'`	`'March 7, 2010'`	`'2010-03-07 00:00:00'`	MMM in the template above matches "March"; the template's 'comma space' matches the input string.

Template	Input String	Output TIMESTAMP	Notes
			If the template lacks the comma, so must the input string, or there is no output; If the input string lacks the comma, so must the template.
`'MMM dd,'`	`'March 7, 2010'`	`'1970-03-07 00:00:00'`	Note that the template above doesn't use a year specifier, causing the output TIMESTAMP to use the earliest year in this epoch, 1970.
`'MMM dd,y'`	`'March 7, 2010'`	`'2010-03-07 00:00:00'`	Using the template above, if the input string were 'March 7, 10', the output TIMESTAMP would be '0010-03-07 00:00:00'.

Template	Input String	Output TIMESTAMP	Notes
`'M-d'`	`'2-8'`	`'1970-02-08 00:00:00'`	Absent a yyyy specifier in the template, as above, the earliest year in this epoch (1970) is used. An input string of '2–8–2011' would give the same result; using '2011–2–8' would give no result because 2011 is not a valid month.
`'MM-dd-yyyy'`	`'06-23-2011 10:11:12'`	`'2011-06-23 00:00:00'`	Dashes as delimiters (as above) are fine, if template and input both use them in the same positions. Since the template omits hours, minutes, and seconds, zeroes are used in the output TIMESTAMP.

Template	Input String	Output TIMESTAMP	Notes
`'dd-MM-yy hh:mm:ss'`	`'23-06-11 10:11:12'`	`'2011-06-23 10:11:12'`	You can have the specifiers in any order as long as that order matches the meaning of the input string you supply. The template and input string of the next example below have the same meaning (and the same output TIMESTAMP) as this example, but they specify months before days and seconds before hours.

Template	Input String	Output TIMESTAMP	Notes
`'MM-dd-yy ss:hh:mm'`	`'06-23-11 12:10:11'`	`'2011-06-23 10:11:12'`	In the template used above, the order of the month and day specifiers is reversed from the example just above, and the specifier for seconds is before hours instead of after minutes; but because the input string also puts months before days and seconds before hours, the meaning (and the output TIMESTAMP) is the same as the example ABOVE.
`'yy-dd-MM ss:hh:mm'`	`'06-23-11 12:10:11'`	`'2006-11-23 10:11:12'`	The template used above reverses (compared to the prior example above) the years and months specifiers, while the input string remains the same. In this case, the output TIMESTAMP uses the first element of the input string as the years, the second as the days, and the third as the months.

Template	Input String	Output TIMESTAMP	Notes
`'dd-MM-yy hh:mm'`	`'23-06-11 10:11:12'`	`'2011-06-23 10:11:00'`	With seconds omitted in the template, as above, the output TIMESTAMP uses 00 seconds. Any number of y specifiers produces the same result; but if the input string inadvertently uses a 1 instead of 11 for the year, as in '23-06-1 10:11:12', then the output TIMESTAMP becomes '0001-06-23 10:11:00'.

`'MM/dd/yy hh:mm:ss'`	`'12/19/11 10:11:12'`
	
	`'12/19/11 12:11:12'`
		`'2011-12-19 10:11:12'`
		
		`'12/19/11 00:11:12'`
			Slashes as delimiters are fine, if template and input both use them in the same positions, as above. Using specifier hh, input times of 12:11:10 and 00:11:10 have the same meaning as a time in the morning.

Template	Input String	Output TIMESTAMP	Notes

`'MM/dd/yy HH:mm:ss'`	`'12/19/11 12:59:59'`
	
	`'12/19/11 21:08:07'`
	
	`'2011-12-19 00:11:12'`
	
	`'2011-12-19 12:11:12'`
		`'2011-12-19 12:59:59'`
		
		`'2011-12-19 21:08:07'`
			The input-string values `'2011-12-19 00:11:12'` or `'2011-12-19 12:11:12'` would fail with this template because `'2011'` is not a month, as required/expected by the template-string `'MM/dd/yy HH:mm:ss'` .
			
			However, changing the template gives useful output:
			
			`values(cast(CHAR_TO_TIMESTAMP('y/MM/dd HH:mm:ss', '2011/12/19 00:11:12') as varchar(19))); 'EXPR$0' '2011-12-19 00:11:12'`
			
			1 row selected
			
			`'12/19/11 00:11:12'` would fail with the above template (`'y/MM/dd'`), since 19 is

Template	Input String	Output TIMESTAMP	Notes
			not a valid month; supplying '12/11/19 00:11:12' works. '2011-12-19 12:11:12' would fail as input because dashes don't match the slashes in the template, '2011/12/19 12:11:12' works. Note that for times after 12 noon (that is, for afternoon and evening times), the hours specifier must be HH instead of hh, and the input string must specify the afternoon or evening hour in 24-hour clock time, hours running from 00 to 23. Using specifier HH, input times of 12:11:10 and 00:11:10 have different meanings, the first as a time in the afternoon and the second as a time in the morning.

Template	Input String	Output TIMESTAMP	Notes
			Using the specifier hh, the times from 12:00 through 11:59:59 are morning times: • Given the specifier s hh:mm:ss, the output TIMESTAMP will include `'00:09:08'` in the morning for both input string '12:09:08' and input string `'00:09:08'` ; whereas, • Given the specifier s HH:mm:ss, the output TIMESTAMP for input string `'00:09:08'` in the morning will include `'00:09:08'` and the output TIMESTAMP for input string `'12:09:08'` in the afternoon will include `'12:09:08'` .

CHAR_TO_DATE

Converts a string to a date, according to the specified format string.

```
CHAR_TO_DATE(format,dateString);
```

CHAR_TO_TIME

Converts a string to a date, according to the specified format string

```
CHAR_TO_TIME(format,dateString);
```

DATE_TO_CHAR

The DATE_TO_CHAR converts a date to a string.

```
DATE_TO_CHAR(format,d);
```

Where d is a date that will be converted to a string.

TIME_TO_CHAR

Uses a format string to format a time. Returns the formatted time or portion of a time as a string.

```
TIME_TO_CHAR(format,time);
```

TIMESTAMP_TO_CHAR

Uses a format string to format a timestamp as char. Returns the timestamp as a string.

```
TIMESTAMP_TO_CHAR(format,ts);
```

Where ts is timestamp.

> **Note**
>
> If the input is `null`, the output will be the string "`null`".

TO_TIMESTAMP

Converts a Unix timestamp to a SQL timestamp in 'YYYY-MM-DD HH:MM:SS' format.

Syntax

```
TO_TIMESTAMP(unixEpoch)
```

Parameters

unixEpoch

A Unix timestamp in the format milliseconds since '1970-01-01 00:00:00' UTC, expressed as a BIGINT.

Example

Example Dataset

The examples following are based on the sample stock dataset that is part of Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*.

> **Note**
>
> The sample dataset has been modified to include a Unix timestamp value (CHANGE_TIME).

To run each example, you need an Amazon Kinesis Analytics application that has the input stream for the sample stock ticker. To learn how to create an Analytics application and configure the input stream for the sample stock ticker, see Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
change_time     BIGINT,    --The UNIX timestamp value
```

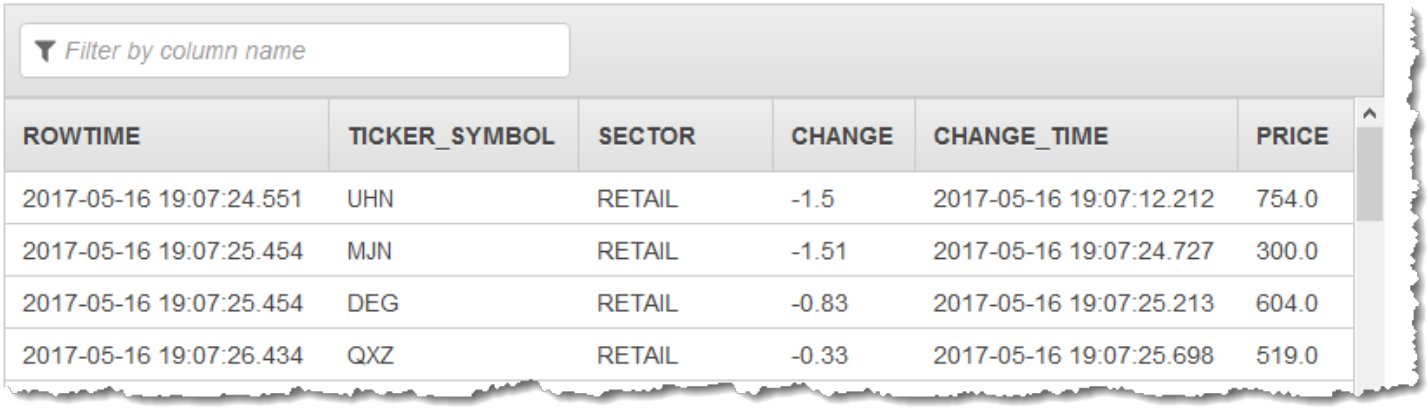
```
price           REAL)
```

Example 1: Convert a Unix Timestamp to a SQL Timestamp

In this example, the `change_time` value in the source stream is converted to a SQL TIMESTAMP value in the in-application stream.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    sector VARCHAR(64),
    change REAL,
    change_time TIMESTAMP,
    price REAL);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM   TICKER_SYMBOL,
                SECTOR,
                CHANGE,
                TO_TIMESTAMP(CHANGE_TIME),
                PRICE

FROM "SOURCE_SQL_STREAM_001"
```

The preceding example outputs a stream similar to the following.



Notes

TO_TIMESTAMP is not part of the SQL:2008 standard. It is an Amazon Kinesis Data Analytics streaming SQL extension.

UNIX_TIMESTAMP

Converts a SQL timestamp to a Unix timestamp that is expressed in milliseconds since '1970-01-01 00:00:00' UTC and that is a BIGINT.

Syntax

```
UNIX_TIMESTAMP(timeStampExpr)
```

Parameters

timeStampExpr

A SQL TIMESTAMP value.

Example

Example Dataset

The examples following are based on the sample stock dataset that is part of Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*.

> **Note**
>
> The sample dataset has been modified to include a Timestamp value (CHANGE_TIME).

To run each example, you need an Amazon Kinesis Analytics application that has the input stream for the sample stock ticker. To learn how to create an Analytics application and configure the input stream for the sample stock ticker, see Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

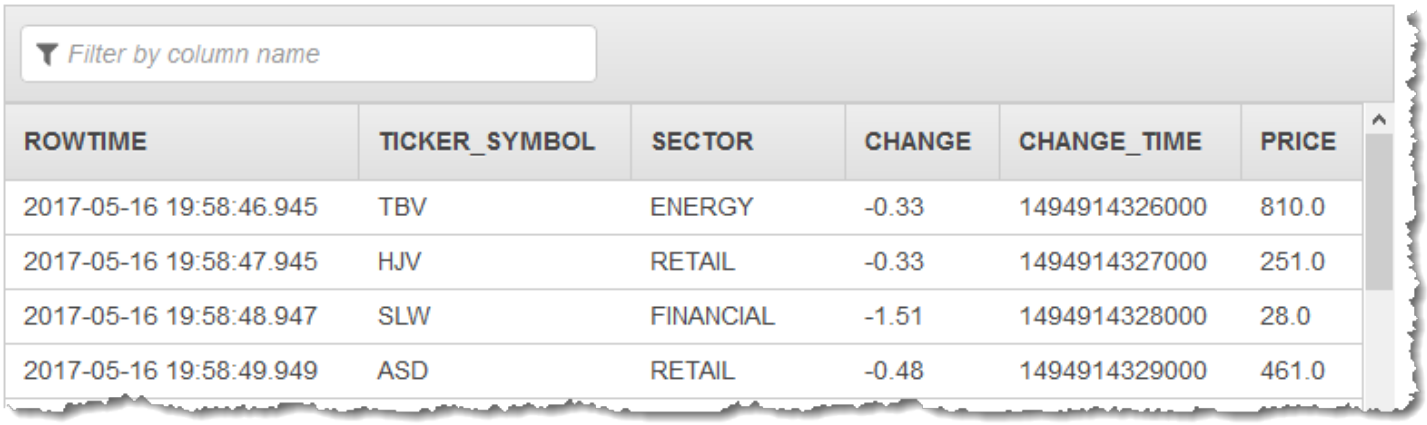
```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
change_time     TIMESTAMP,    --The timestamp value to convert
price           REAL)
```

Example 1: Convert a Timestamp to a UNIX Timestamp

In this example, the `change_time` value in the source stream is converted to a TIMESTAMP value in the in-application stream.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
        ticker_symbol VARCHAR(4),
        SECTOR VARCHAR(16),
        CHANGE REAL,
        CHANGE_TIME BIGINT,
        PRICE REAL);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM   TICKER_SYMBOL,
                SECTOR,
                CHANGE,
                UNIX_TIMESTAMP(CHANGE_TIME),
                PRICE
FROM "SOURCE_SQL_STREAM_001"
```

The preceding example outputs a stream similar to the following.



ROWTIME	TICKER_SYMBOL	SECTOR	CHANGE	CHANGE_TIME	PRICE
2017-05-16 19:58:46.945	TBV	ENERGY	-0.33	1494914326000	810.0
2017-05-16 19:58:47.945	HJV	RETAIL	-0.33	1494914327000	251.0
2017-05-16 19:58:48.947	SLW	FINANCIAL	-1.51	1494914328000	28.0
2017-05-16 19:58:49.949	ASD	RETAIL	-0.48	1494914329000	461.0

Notes

UNIX_TIMESTAMP is not part of the SQL:2008 standard. It is an Amazon Kinesis Data Analytics streaming SQL extension.

Date, Timestamp, and Interval Operators

The arithmetic operators +, -, *, and / are binary operators.

Operator	Description	Notes
+	Addition	interval + interval = interval interval + datetime = datetime datetime + interval = datetime
-	Subtraction	interval - interval = interval datetime - interval = datetime (<datetime> - <datetime>) Date, Timestamp, and Interval Operators <interval qualifier> = interval
*	Multiplication	interval * numeric = interval numeric * interval = interval
/	Division	interval / numeric = interval

Examples

Example	Operation	Result
1	INTERVAL '1' DAY + INTERVAL '3' DAY	INTERVAL '4' DAY
2	INTERVAL '1' DAY + INTERVAL '3 4' DAY TO HOUR	INTERVAL '+4 04' DAY TO HOUR
3	INTERVAL '1' DAY - INTERVAL '3 4' DAY TO HOUR	INTERVAL '-2 04' DAY TO HOUR

Example	Operation	Result
4	INTERVAL '1' YEAR + INTERVAL '3-4' YEAR TO MONTH	INTERVAL '+4-04' YEAR TO MONTH
5	2 * INTERVAL '3 4' DAY TO HOUR	INTERVAL '6 8' DAY TO HOUR
6	INTERVAL '3 4' DAY TO HOUR / 2	INTERVAL ' 1 14' DAY TO HOUR

In the example 3, '3 4 DAY means 3 days and 4 hours, so the result in that row means 24 hours minus 76 hours, resulting in minus 52 hours, which is a negative 2 days and 4 hours.

Example 4 uses TO MONTH rather than TO HOUR, so the INTERVAL specified as '3-4' means 3 years and 4 months, or 40 months.

In example 6, the "/2" applies to the INTERVAL '3 4', which is 76 hours, half of which is 38, or 1 day and 14 hours.

Further Examples of Interval Operations

Streaming SQL also supports subtracting two datetimes, giving an interval. You specify what kind of interval you want for the result, as shown following:

```
(<datetime> - <datetime>) <interval qualifier>
```

The following examples show operations that can be useful in Amazon Kinesis Data Analytics applications.

Example 1 – Time Difference (as minutes to the nearest second or as seconds)

```
 values cast ((time  '12:03:34' - time '11:57:23') minute to second as varchar(8));
+---------+
 EXPR$0
+---------+
 +6:11
+---------+
1 row selected
```

```
.............. 6 minutes, 11 seconds
or
 values cast ((time  '12:03:34' - time '11:57:23') second as varchar(8));
+---------+
 EXPR$0
+---------+
 +371
+---------+
1 row selected
```

Example 2 – Time Difference (as minutes only)

```
values cast ((time  '12:03:34' - time '11:57:23') minute as varchar(8));
+---------+
 EXPR$0
+---------+
 +6
+---------+
1 row selected
.............. 6 minutes; seconds ignored.
values cast ((time  '12:03:23' - time '11:57:23') minute as varchar(8));
+---------+
 EXPR$0
+---------+
 +6
+---------+
1 row selected
.............. 6 minutes
```

Example 3 – Time-to-Timestamp Difference (as days to the nearest second) Invalid

```
values cast ((time '12:03:34'-timestamp '2004-04-29 11:57:23') day to second as
 varchar(8));
Error: From line 1, column 14 to line 1, column 79: Parameters must be of the same type
```

Example 4 – Timestamp difference (as days to the nearest second)

```
values cast ((timestamp  '2004-05-01 12:03:34' - timestamp '2004-04-29 11:57:23') day
 to
                second as varchar(8));
+-----------+
  EXPR$0
```

```
+-----------+
 +2 00:06
+-----------+
1 row selected
.............. 2 days, 6 minutes
.............. Although "second" was specified above, the varchar(8) happens to allow
only room enough to show only the minutes, not the seconds.
The example below expands to varchar(11), showing the full result:
values cast ((timestamp  '2004-05-01 12:03:34' - timestamp '2004-04-29 11:57:23') day
 to
                second as varchar(11));
+--------------+
    EXPR$0
+--------------+
 +2 00:06:11
+--------------+
1 row selected
.............. 2 days, 6 minutes, 11 seconds
```

Example 5 – Timestamp Difference (as days to the nearest second)

```
values cast ((timestamp  '2004-05-01 1:03:34' - timestamp '2004-04-29 11:57:23') day to
                second as varchar(11));
+--------------+
    EXPR$0
+--------------+
 +1 13:06:11
+--------------+
1 row selected
.............. 1 day, 13 hours, 6 minutes, 11 seconds
values cast ((timestamp  '2004-05-01 13:03:34' - timestamp '2004-04-29 11:57:23') day
 to
                second as varchar(11));
+--------------+
    EXPR$0
+--------------+
 +2 01:06:11
+--------------+
1 row selected
.............. 2 days, 1 hour, 6 minutes, 11 seconds
```

Example 6 – Timestamp Difference (as days)

```
values cast ((timestamp  '2004-05-01 12:03:34' - timestamp '2004-04-29 11:57:23') day
               as varchar(8));
+---------+
 EXPR$0
+---------+
 +2
+---------+
1 row selected
.............. 2 days
```

Example 7 – Time Difference (as days)

```
values cast ((date '2004-12-02 ' - date '2003-12-01 ') day  as varchar(8));
Error: Illegal DATE literal '2004-12-02 ': not in format 'yyyy-MM-dd'
.............. Both date literals end with a space;  disallowed.
values cast ((date '2004-12-02' - date '2003-12-01 ') day  as varchar(8));
Error: Illegal DATE literal '2003-12-01 ': not in format 'yyyy-MM-dd'
.............. Second date literal still ends with a space;  disallowed.
values cast ((date '2004-12-02' - date '2003-12-01') day  as varchar(8));
+---------+
 EXPR$0
+---------+
 +367
+---------+
1 row selected
.............. 367 days
```

Example 8 – Not Supported (Simple Difference of Dates)

If you don't specify "day" as the intended unit, as shown following, the subtraction is not supported.

```
    values cast ((date '2004-12-02' - date '2003-12-01') as varchar(8));
    Error: From line 1, column 15 to line 1, column 51:
           Cannot apply '-' to arguments of type '<DATE> - <DATE>'.
    Supported form(s): '<NUMERIC> - <NUMERIC>'
                       '<DATETIME_INTERVAL> - <DATETIME_INTERVAL>'
                       '<DATETIME> - <DATETIME_INTERVAL>'
```

Why Use "as varchar" in Conversion Examples?

The reason for using the "values cast (<expression> AS varchar(N))" syntax in the examples above is that while the SQLline client used above (with Amazon Kinesis Data Analytics running) does return an interval, JDBC does not support returning that result so as to display it. Therefore, that "values" syntax is used to see/show it.

If you close the Amazon Kinesis Data Analytics (with a !kill command) or if you don't start it before running SQLline, then you can run the sqllineEngine (rather than the sqllineClient) from the bin subdirectory of your Amazon Kinesis Data Analytics home, which can show your results without the Amazon Kinesis Data Analytics application or JDBC:

Rules for Specifying Intervals

A Day-Time Interval Literal is a string that denotes a single interval value: for example '10' SECONDS. Note it has two parts: the value (which must always be in single-quotes) and the qualifier (here, SECONDS), which give the units for the value.

The qualifier takes the following form:

```
DAY  HOUR  MINUTE  SECOND [TO HOUR  MINUTE  SECOND]
```

> **Note**
>
> YEAR TO MONTH intervals require a dash separating the values, whereas DAY TO HOUR intervals use a space to separate the values, as seen in the 2nd, 3rd, 5th, and 6th examples in that topic.

In addition, the leading term has to be of greater significance than the optional trailing term, so this means you can only specify:

```
DAY
HOUR
MINUTE
SECOND
DAY TO HOUR
DAY TO MINUTE
DAY TO SECOND
HOUR TO MINUTE
HOUR TO SECOND
```

```
  MINUTE TO SECOND
```

The easiest way to understand these may be to translate X TO Y as "Xs to the nearest Y". Hence, DAY TO HOUR is "days to the nearest hour".

When DAY, HOUR, or MINUTE is the leading term, you can specify a precision, e.g., DAY(3) TO HOUR, indicating the number of digits the associated field in the value can have. The maximum precision is 10, and the default is 2. You can't specify precision for HOUR, OR MINUTE in the trailing term - they are always of precision 2. So for example, HOUR(3) TO MINUTE is legal, HOUR TO MINUTE(3) is not.

SECOND can also take a precision, but the way it is specified differs depending on whether it is the leading or trailing field.

- If SECOND is the leading field, you can specify the digits before and after the decimal point. For example, SECOND(3,3) would allow you to specify up to 999.999 seconds. The default is (2,3), which is actually a deviation from the SQL:2008 spec (it should be (2,6), but we only have millisecond precision).
- If SECOND is the trailing field, you can only specify precision for the fractional seconds, that is, the part shown after the seconds' decimal point below. For example, SECOND(3) would indicate milliseconds. The default is 3 digits after the decimal point, but as above this is a deviation from the standard of 6.

As for the value, it takes the general form of:

```
  [+-]'[+-]DD HH:MM:SS.SSS'
```

Where DD are digits indicating days, HH hours, MM minutes, and SS.SSS is seconds (adjust the number of digits appropriately if precision is explicitly specified).

Not all values have to include all fields—you can trim from both front or back, but not from in the middle. So you could make it 'DD HH' or 'MM:SS.SSS', but not 'DD MM'.

However you write it, though, the value must match the qualifier, as shown following:

```
INTERVAL '25 3' DAY to HOUR ------> legal
INTERVAL '3:45:04.0' DAY TO HOUR --> illegal
```

As stated in the SQL spec, if the precision is not explicitly specified, it is implied to be 2. Thus:

- INTERVAL '120' MINUTE is an illegal interval. The legal form for the desired interval is INTERVAL '120' MINUTE(2)

  and

- INTERVAL '120' SECOND is not legal. The legal form for the desired interval is INTERVAL '120' SECOND(3).

```
values INTERVAL '120' MINUTE(2);
Error: From line 1, column 8 to line 1, column 31:
                    Interval field value 120 exceeds precision of MINUTE(2) field
values INTERVAL '120' MINUTE(3);
Conversion not supported
```

Also, if HOUR, MINUTE, or SECOND are not the leading field, they must fall in the following ranges (taken from Table 6 in topic 4.6.3 of the SQL:2008 foundation spec), as shown following:

```
HOUR: 0-23
MINUTE: 0-59
SECOND: 0-59.999
```

Year-month intervals are similar, except that the qualifiers are as shown following:

```
YEAR
MONTH
YEAR TO MONTH
```

Precision can be specified just as with DAY and HOUR, and the max of 10 and default of 2 is the same.

The value format for year-month is: 'YY-MM'. If MONTH is the trailing field, it must fall in the range 0-11.

```
<interval qualifier> := <start field> TO <end field>  <single datetime field>

<start field> := <non-second primary datetime field> [ <left paren> <interval leading
 field precision> <right paren> ]

<end field> := <non-second primary datetime field>  SECOND [ <left paren> <interval
 fractional seconds precision> <right paren> ]
```

```
<single datetime field> := <non-second primary datetime field> [ <left paren> <interval
 leading field precision> <right paren> ]
   SECOND [ <left paren> <interval leading field precision>
          [ <comma> <interval fractional seconds precision> ] <right paren> ]
<primary datetime field> := <non-second primary datetime field>       SECOND
<non-second primary datetime field> := YEAR  MONTH  DAY  HOUR      MINUTE
<interval fractional seconds precision> := <unsigned integer>
<interval leading field precision> := <unsigned integer>
```

Date and Time Patterns

Date and time formats are specified by date and time pattern strings. In these pattern strings, unquoted letters from A to Z and from a to z represent components of a data or time value. If a letter or text string is enclosed within a pair of single quotes, that letter or text is not interpreted but rather used as is, as are all other characters in the pattern string. During printing, that letter or text is copied as is to the output string; during parsing, they are matched against the input string. "'" represents a single quote.

The following pattern letters are defined for the indicated Date or Time Component. All other characters from 'A' to 'Z' and from 'a' to 'z' are reserved. For an alphabetic ordering of the pattern letters, see Date and Time Pattern Letters in Alphabetic Order.

Date or Time Component	Pattern Letter	Presentation as text or number	Examples
Era designator	G	Text	AD
Year	y	Year	1996; 96
Month in year	M	Month	July; Jul; 07
Week in year	w	Number	27
Week in month	W	Number	2
Day in year	D	Number	189
Day in month	d	Number	10
Day of week in month	F	Number	2

Date or Time Component	Pattern Letter	Presentation as text or number	Examples
Day in week	E	Text	EE=Tu; EEE=Tue; EEEE=Tuesday
Am/pm marker	a	Text	PM
Hour in day (0-23)	H	Number	0
Hour in day (1-24)	k	Number	24
Hour in am/pm (0-11)	K	Number	0
Hour in am/pm (1-12)	h	Number	12
Minute in hour	m	Number	30
Second in minute	s	Number	55
Millisecond	S	Number	978
Time zone	z	General	Pacific Standard Time; PST; GMT-08:00
Time zone	Z	RFC	-0800

Pattern letters are usually repeated, as their number determines the exact presentation:

Text

For formatting, if the number of pattern letters is 4 or more, the full form is used; otherwise a short or abbreviated form is used if available. For parsing, both forms are accepted, independent of the number of pattern letters.

Number

For formatting, the number of pattern letters is the minimum number of digits, and shorter numbers are zero-padded to this amount. For parsing, the number of pattern letters is ignored unless it's needed to separate two adjacent fields.

Year

Time zones are interpreted as text if they have names. For time zones representing a GMT offset value, the following syntax is used:

```
GMTOffsetTimeZone:
GMT Sign Hours : Minutes
Sign: one of
+ -
Hours:
Digit
Digit Digit
Minutes:
Digit Digit
Digit: one of
0 1 2 3 4 5 6 7 8 9
```

Hours must be between 0 and 23, and Minutes must be between 00 and 59. The format is locale independent and digits must be taken from the Basic Latin block of the Unicode standard.

For parsing, RFC 822 time zones are also accepted.

RFC 822 time zone

For formatting, the RFC 822 4-digit time zone format is used:

```
RFC822TimeZone:
Sign TwoDigitHours Minutes
TwoDigitHours:
Digit Digit
```

TwoDigitHours must be between 00 and 23. Other definitions are as for general time zones.

For parsing, general time zones are also accepted.

SimpleDateFormat also supports "localized date and time pattern" strings. In these strings, the pattern letters described above may be replaced with other, locale dependent, pattern letters. SimpleDateFormat does not deal with the localization of text other than the pattern letters; that's up to the client of the class.

Examples

The following examples show how date and time patterns are interpreted in the U.S. locale. The given date and time are 2001-07-04 12:08:56 local time in the U.S. Pacific time zone.

Date and Time Pattern	Result
"yyyy.MM.dd G 'at' HH:mm:ss z"	2001.07.04 AD at 12:08:56 PDT
"EEE, MMM d, ''yy"	Wed, Jul 4, '01
"h:mm a"	12:08 PM
"hh 'o''clock' a, zzzz"	12 o'clock PM, Pacific Daylight Time
"K:mm a, z"	0:08 PM, PDT
"yyyyy.MMMMM.dd GGG hh:mm aaa"	02001.July.04 AD 12:08 PM
"EEE, d MMM yyyy HH:mm:ss Z"	Wed, 4 Jul 2001 12:08:56 -0700
"yyMMddHHmmssZ"	010704120856-0700
"yyyy-MM-dd'T'HH:mm:ss.SSSZ"	2001-07-04T12:08:56.235-0700

Date and Time Pattern Letters in Alphabetic Order

The same pattern letters shown at first, above, in Date or Time Component order are shown below in alphabetic order for easy reference.

Pattern Letter	Date or Time Component	Presentation as text or number	Examples
a	Am/pm marker	Text	PM

Pattern Letter	Date or Time Component	Presentation as text or number	Examples
D	Day in year	Number	189
d	Day in month	Number	10
E	Day in week	Text	EE=Tu; EEE=Tue; EEEE=Tuesday
F	Day of week in month	Number	2
G	Era designator	Text	AD
H	Hour in day (0-23)	Number	0
h	Hour in am/pm (1-12)	Number	12
k	Hour in day (1-24)	Number	24
K	Hour in am/pm (0-11)	Number	0
M	Month in year	Month	July; Jul; 07
m	Minute in hour	Number	30
s	Second in minute	Number	55
S	Millisecond	Number	978
w	Week in year	Number	27
W	Week in month	Number	2
y	Year	Year	1996; 96
z	Time zone	General	Pacific Standard Time; PST; GMT-08:00

Pattern Letter	Date or Time Component	Presentation as text or number	Examples
Z	Time zone	RFC	-0800

CURRENT_DATE

Returns the current Amazon Kinesis Data Analytics system date when the query executes as YYYY-MM-DD when the query executes.

For more information, see CURRENT_TIME, CURRENT_TIMESTAMP, LOCALTIMESTAMP, LOCALTIME, and CURRENT_ROW_TIMESTAMP.

Example

```
+---------------+
| CURRENT_DATE  |
+---------------+
| 2008-08-27    |
+---------------+
```

CURRENT_ROW_TIMESTAMP

CURRENT_ROW_TIMESTAMP is an Amazon Kinesis Data Analytics extension to the SQL:2008 specification. This function returns the current timestamp as defined by the environment on which the Amazon Kinesis Data Analytics application is running. CURRENT_ROW_TIMESTAMP is always returned as UTC, not the local timezone.

CURRENT_ROW_TIMESTAMP is similar to LOCALTIMESTAMP, but returns a new timestamp for each row in a stream.

A query run with LOCALTIMESTAMP (or CURRENT_TIMESTAMP or CURRENT_TIME) as one of the columns puts into all output rows the time the query is first run.

If that column instead contains CURRENT_ROW_TIMESTAMP, each output row gets a newly-calculated value of TIME representing when that row was output.

> **Note**
>
> CURRENT_ROW_TIMESTAMP is not defined in the SQL:2008 specification; it is an Amazon Kinesis Data Analytics extension.

For more information, see CURRENT_TIME, CURRENT_DATE, CURRENT_TIMESTAMP, LOCALTIMESTAMP, LOCALTIME, and CURRENT_ROW_TIMESTAMP.

CURRENT_TIME

Returns the current Amazon Kinesis Data Analytics system time when the query executes. Time is in UTC, not the local time zone.

For more information, see CURRENT_TIMESTAMP, LOCALTIMESTAMP, LOCALTIME, CURRENT_ROW_TIMESTAMP, and CURRENT_DATE.

Example

```
+---------------+
| CURRENT_TIME  |
+---------------+
| 20:52:05      |
```

CURRENT_TIMESTAMP

Returns the current database system timestamp (as defined on the environment on which Amazon Kinesis Data Analytics is running) as a datetime value.

For more information, see CURRENT_TIME, CURRENT_DATE, LOCALTIME, LOCALTIMESTAMP, and CURRENT_ROW_TIMESTAMP.

Example

```
+--------------------+
| CURRENT_TIMESTAMP  |
+--------------------+
| 20:52:05           |
```

```
+--------------------+
```

EXTRACT

```
EXTRACT(YEAR|MONTH|DAY|HOUR|MINUTE|SECOND FROM <datetime expression>|<interval
 expression>)
```

The EXTRACT function extracts one field from a DATE, TIME, TIMESTAMP or INTERVAL expression. Returns BIGINT for all fields other than SECOND. For SECOND it returns DECIMAL(5,3) and includes milliseconds.

Syntax

Examples

Function	Result
`EXTRACT(DAY FROM INTERVAL '2 3:4:5.678 ' DAY TO SECOND)`	2
`EXTRACT(HOUR FROM INTERVAL '2 3:4:5.678' DAY TO SECOND)`	3
`EXTRACT(MINUTE FROM INTERVAL '2 3:4:5.678' DAY TO SECOND)`	4
`EXTRACT(SECOND FROM INTERVAL '2 3:4:5.678' DAY TO SECOND)`	5.678
`EXTRACT(MINUTE FROM CURRENT_R OW_TIMESTAMP) where CURRENT_ROW_TIMESTAMP is 2016-09-23 04:29:26.234`	29
`EXTRACT (HOUR FROM CURRENT_ROW_TIMEST AMP)`	4

Function	Result
where CURRENT_ROW_TIMESTAMP is 2016-09-23 04:29:26.234	

Use in Function

EXTRACT can be used for conditioning data, as in the following function which returns a 30 minute floor when CURRENT_ROW_TIMESTAMP is input for p_time.

```
CREATE or replace FUNCTION FLOOR30MIN( p_time TIMESTAMP )
RETURNS  TIMESTAMP
CONTAINS SQL
RETURNS NULL ON NULL INPUT
RETURN  floor(p_time to HOUR) + (( EXTRACT (  MINUTE FROM p_time  ) / 30)* INTERVAL
 '30' MINUTE ) ;
```

You would implement this function using code along the following lines:

```
SELECT stream FLOOR30MIN( CURRENT_ROW_TIMESTAMP ) as ROWTIME , * from "MyStream" ) over
 (range current row ) as r
```

> **Note**
>
> The code above assumes that you have previously created a stream called "MyStream."

LOCALTIME

Returns the current time when the query executes as defined by the environment on which Amazon Kinesis Data Analytics is running. LOCALTIME is always returned as UTC (GMT), not the local timezone.

For more information, see CURRENT_TIME, CURRENT_DATE, CURRENT_TIMESTAMP, LOCALTIMESTAMP, and CURRENT_ROW_TIMESTAMP.

Example

```
VALUES localtime;
```

```
+------------+
| LOCALTIME  |
+------------+
| 01:11:15   |
+------------+
1 row selected (1.558 seconds)
```

Limitations

Amazon Kinesis Data Analytics does not support the optional <time precision> parameter specified in SQL:2008. This is a departure from the SQL:2008 standard.

LOCALTIMESTAMP

Returns the current timestamp as defined by the environment on Amazon Kinesis Data Analytics application is running. Time is always returned as UTC (GMT), not the local timezone.

For more information, see CURRENT_TIME, CURRENT_DATE, CURRENT_TIMESTAMP, LOCALTIME, and CURRENT_ROW_TIMESTAMP.

Example

```
values localtimestamp;
+-------------------------+
|      LOCALTIMESTAMP     |
+-------------------------+
| 2008-08-27 01:13:42.206  |
+-------------------------+
1 row selected (1.133 seconds)
```

Limitations

Amazon Kinesis Data Analytics does not support the optional <timestamp precision> parameter specified in SQL:2008. This is a departure from the SQL:2008 standard.

TSDIFF

Returns NULL if any of the arguments is null.

Otherwise returns the difference between the two timestamps in milliseconds.

Syntax

```
TSDIFF(startTime, endTime)
```

Parameters

startTime

A Unix timestamp in the format milliseconds since '1970-01-01 00:00:00' UTC, expressed as a BIGINT.

endTime

A Unix timestamp in the format milliseconds since '1970-01-01 00:00:00' UTC, expressed as a BIGINT.

Null Functions

The topics in this section describe the null functions for Amazon Kinesis Data Analytics streaming SQL.

Topics

- COALESCE
- NULLIF

COALESCE

```
COALESCE (
      <value-expression>
      {,<value-expression>}... )
```

The COALESCE function takes a list of expressions (all of which must be of the same type) and returns the first non-null argument from the list. If all of the expressions are null, COALESCE returns null.

Examples

Expression	Result
COALESCE('chair')	chair
COALESCE('chair', null, 'sofa')	chair
COALESCE(null, null, 'sofa')	sofa
COALESCE(null, 2, 5)	2

NULLIF

```
NULLIF ( <value-expression>, <value-expression> )
```

Returns null if the two input arguments are equal, otherwise returns the first value. Both arguments must be of comparable type, or an exception is raised.

Examples

Function	Result
NULLIF(4,2)	4
NULLIF(4,4)	<null>
NULLIF('amy','fred')	amy
NULLIF('amy', cast(null as varchar(3)))	amy
NULLIF(cast(null as varchar(3)),'fred')	<null>

Numeric Functions

The topics in this section describe the numeric functions for Amazon Kinesis Data Analytics streaming SQL.

Topics

- ABS
- CEIL / CEILING
- EXP
- FLOOR
- LN
- LOG10
- MOD
- POWER
- STEP

ABS

Returns the absolute value of the input argument. Returns `null` if the input argument is null.

```
ABS ( <numeric-expression>  <interval-expression>
      )
```

Examples

Function	Result
ABS(2.0)	2.0
ABS(-1.0)	1.0
ABS(0)	0
ABS(-3 * 3)	9
ABS(INTERVAL '-3 4:20' DAY TO MINUTE)	INTERVAL '3 4:20' DAY TO MINUTE

If you use `cast as` VARCHAR in SQLline to show the output, the value is returned as `+3 04:20`.

```
values(cast(ABS(INTERVAL '-3 4:20' DAY TO MINUTE) AS VARCHAR(8)));
 +-----------+
```

```
  EXPR$0
+-----------+
 +3 04:20
+-----------+
1 row selected
```

CEIL / CEILING

```
CEIL | CEILING ( <number-expression> )
CEIL | CEILING ( <datetime-expression> TO <time-unit> )
CEIL | CEILING ( <number-expression> )
CEIL | CEILING ( <datetime-expression> TO <[[time-unit> )
```

When called with a numeric argument, CEILING returns the smallest integer equal to or larger than the input argument.

When called with a date, time, or timestamp expression, CEILING returns the smallest value equal to or larger than the input, subject to the precision specified by the <time unit>.

Returns null if any input argument is null.

Examples

Function	Result
CEIL(2.0)	2
CEIL(-1.0)	-1
CEIL(5.2)	6
CEILING(-3.3)	-3
CEILING(-3 * 3.1)	-9
CEILING(TIMESTAMP '2004-09-30 13:48:23' TO HOUR)	TIMESTAMP '2004-09-30 14:00:00'
CEILING(TIMESTAMP '2004-09-30 13:48:23' TO MINUTE)	TIMESTAMP '2004-09-30 13:49:00'

Function	Result
CEILING(TIMESTAMP '2004-09-30 13:48:23' TO DAY)	TIMESTAMP '2004-10-01 00:00:00.0'
CEILING(TIMESTAMP '2004-09-30 13:48:23' TO YEAR)	TIMESTAMP '2005-01-01 00:00:00.0'

Notes

- CEIL and CEILING are synonyms for this function provided by the SQL:2008 standard.
- CEIL(<datetime value expression> TO <time unit>) is an Amazon Kinesis Data Analytics extension.
- For more information, see FLOOR.

EXP

```
EXP ( <number-expression> )
```

Returns the value of e (approximately 2.7182818284590455) raised to the power of the input argument. Returns null if the input argument is null.

Examples

Function	Result
EXP(1)	2.7182818284590455
EXP(0)	1.0
EXP(-1)	0.36787944117144233
EXP(10)	22026.465794806718
EXP(2.5)	12.182493960703473

FLOOR

```
FLOOR ( <time-unit> )
```

When called with a numeric argument, FLOOR returns the largest integer equal to or smaller than the input argument.

When called with a date, time, or timestamp expression, FLOOR returns the largest value equal to or smaller than the input, subject to the precision specified by <time unit>.

FLOOR returns null if any input argument is null.

Examples

Function	Result
FLOOR(2.0)	2
FLOOR(-1.0)	-1
FLOOR(5.2)	5
FLOOR(-3.3)	-4
FLOOR(-3 * 3.1)	-10
FLOOR(TIMESTAMP '2004-09-30 13:48:23' TO HOUR)	TIMESTAMP '2004-09-30 13:00:00'
FLOOR(TIMESTAMP '2004-09-30 13:48:23' TO MINUTE)	TIMESTAMP '2004-09-30 13:48:00'
FLOOR(TIMESTAMP '2004-09-30 13:48:23' TO DAY)	TIMESTAMP '2004-09-30 00:00:00.0'
FLOOR(TIMESTAMP '2004-09-30 13:48:23' TO YEAR)	TIMESTAMP '2004-01-01 00:00:00.0'

Notes

> **Note**
>
> FLOOR (<datetime expression> TO <timeunit>) is an Amazon Kinesis Data Analytics extension.
> The STEP function is similar to FLOOR but can round values down to arbitrary intervals, such as 30 seconds. For more information, see STEP.

LN

```
LN ( <number-expression> )
```

Returns the natural log (that is, the log with respect to base e) of the input argument. If the argument is negative or 0, an exception is raised. Returns null if the input argument is null.

For more information, see LOG10 and EXP.

Examples

Function	Result
LN(1)	0.0
LN(10)	2.302585092994046
LN(2.5)	0.9162907318741551

LOG10

```
LOG10 ( <number-expression> )
```

Returns the base 10 logarithm of the input argument. If the argument is negative or 0, an exception is raised. Returns null if the input argument is null.

Examples

Function	Result
LOG10(1)	0.0
LOG10(100)	2.0
log10(cast('23' as decimal))	1.3617278360175928

> **Note**
>
> LOG10 is not a SQL:2008 standard function; it is an Amazon Kinesis Data Analytics extension to the standard.

MOD

```
MOD ( <dividend>, <divisor> )
 <dividend> := <integer-expression>
 <divisor>  := <integer-expression>
```

Returns the remainder when the first argument (the dividend is divided by the second numeric argument (the divisor). If the divisor is zero, a divide by zero error is raised.

Examples

Function	Result
MOD(4,2)	0
MOD(5,3)	2
MOD(-4,3)	-1
MOD(5,12)	5

Limitations

The Amazon Kinesis Data Analytics MOD function only supports arguments of scale 0 (integers). This is a departure from the SQL:2008 standard, which supports any numeric argument. Other numeric arguments can be CAST to an integer, of course.

POWER

```
POWER ( <base>, <exponent> )
<base> := <number-expression>
<exponent> := <number-expression>
```

Returns the value of the first argument (the base) raised to the power of the second argument (the exponent). Returns null if either the base or the exponent is null, and raises an exception if the base is zero and the exponent is negative, or if the base is negative and the exponent is not a whole number.

Examples

Function	Result
POWER(3,2)	9
POWER(-2,3)	-8
POWER(4,-2)	1/16 ..or.. 0.0625
POWER(10.1,2.5)	324.19285157140644

STEP

```
STEP ( <time-unit> BY INTERVAL '<integer-literal>' <interval-literal> )
STEP ( <integer-expression> BY <integer-literal> )
```

STEP rounds down the input value (<time-unit> or <integer-expression>) to the nearest multiple of <integer-literal>.

The STEP function works on datetime data types or integer types. STEP is a scalar function that performs an operation similar to FLOOR. However, by using STEP you can specify an arbitrary time or integer interval for rounding down the first argument.

STEP returns null if any input argument is null.

STEP with an Integer Argument

When called with an integer argument, STEP returns the largest integer multiple of the <interval-literal> argument equal to or smaller than the <integer-expression> argument. For example, `STEP(23 BY 5)` returns 20, because 20 is the greatest multiple of 5 that is less than 23.

`STEP ( <integer-expression > BY <integer-literal> )` is equivalent to the following.

```
( <integer-expression> / <integer-literal> ) * <integer-literal>
```

Examples

In the following examples, the return value is the largest multiple of <integer-literal> that is equal to or less than <integer-expression>.

Function	Result
STEP(23 BY 5)	20
STEP(30 BY 10)	30

STEP with a Date Type Argument

When called with a date, time, or timestamp argument, STEP returns the largest value equal to or smaller than the input, subject to the precision specified by <time unit>.

`STEP(<datetimeExpression> BY <intervalLiteral>)` is equivalent to the following.

```
(<datetimeExpression> - timestamp '1970-01-01 00:00:00')  /  <intervalLiteral> )  *
 <intervalLiteral> + timestamp '1970-01-01 00:00:00'
```

<intervalLiteral> can be one of the following:

- YEAR

- MONTH
- DAY
- HOUR
- MINUTE
- SECOND

Examples

In the following examples, the return value is the latest multiple of <integer-literal> of the unit specified by <intervalLiteral> that is equal to or earlier than <datetime-expression>.

Function	Result
STEP(CAST('2004-09-30 13:48:23' as TIMESTAMP) BY INTERVAL '10' SECOND)	'2004-09-30 13:48:20'
STEP(CAST('2004-09-30 13:48:23' as TIMESTAMP) BY INTERVAL '2' HOUR)	'2004-09-30 12:00:00'
STEP(CAST('2004-09-30 13:48:23' as TIMESTAMP) BY INTERVAL '5' MINUTE)	'2004-09-30 13:45:00'
STEP(CAST('2004-09-27 13:48:23' as TIMESTAMP) BY INTERVAL '5' DAY)	'2004-09-25 00:00:00.0'
STEP(CAST('2004-09-30 13:48:23' as TIMESTAMP) BY INTERVAL '1' YEAR)	'2004-01-01 00:00:00.0'

STEP in a GROUP BY clause (tumbling window)

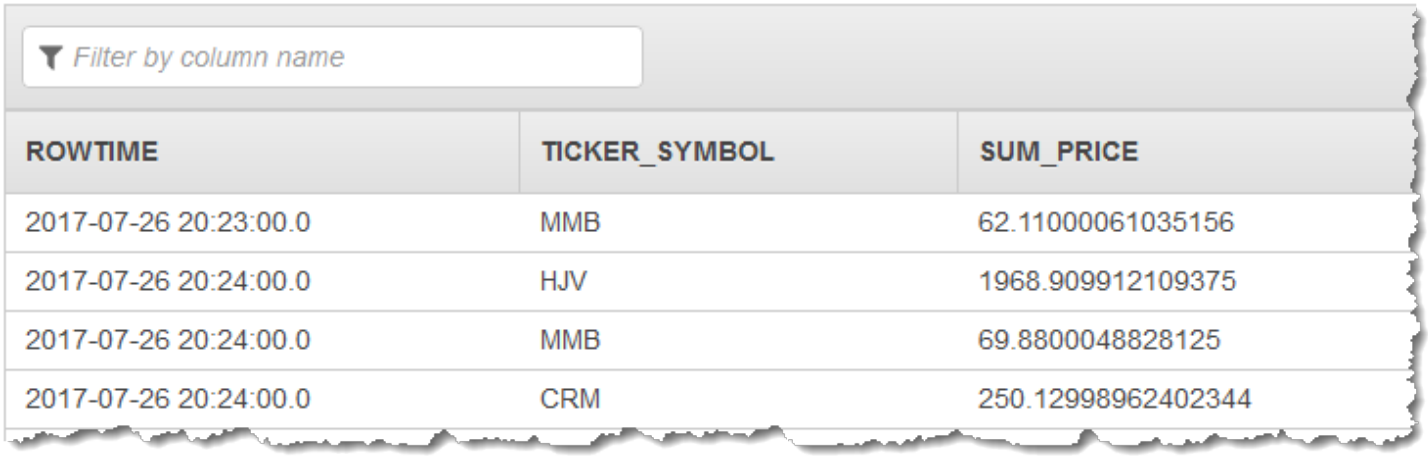
In this example, an aggregate query has a GROUP BY clause with STEP applied to ROWTIME that groups the stream into finite rows.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    sum_price     DOUBLE);
```

```
-- CREATE OR REPLACE PUMP to insert into output
CREATE OR REPLACE PUMP "STREAM_PUMP" AS
  INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM
        ticker_symbol,
        SUM(price) AS sum_price
    FROM "SOURCE_SQL_STREAM_001"
     GROUP BY ticker_symbol, STEP("SOURCE_SQL_STREAM_001".ROWTIME BY INTERVAL '60'
 SECOND);
```

Results

The preceding example outputs a stream similar to the following.



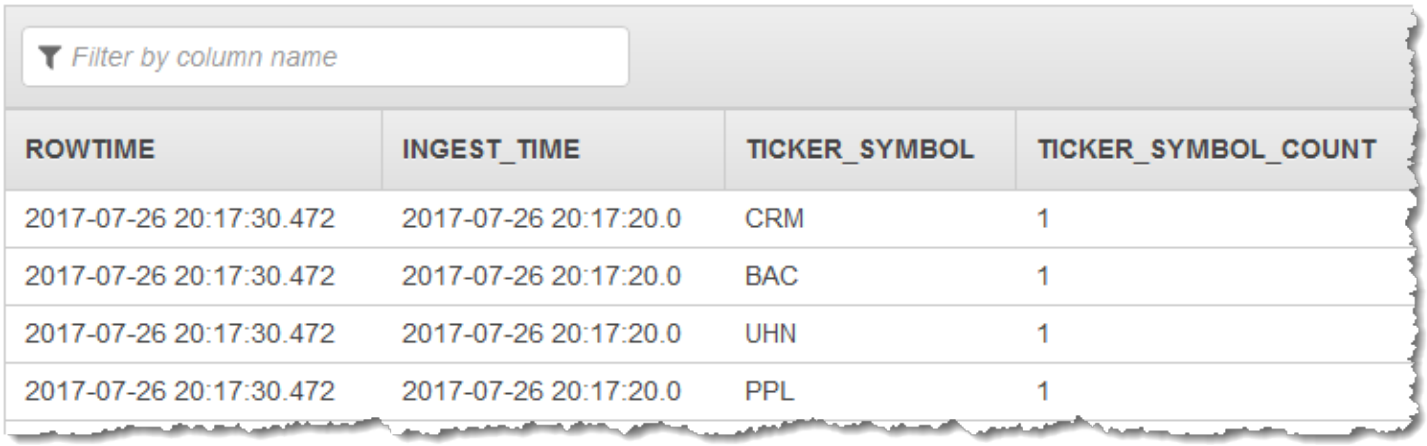
STEP in an OVER clause (sliding window)

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ingest_time TIMESTAMP,
    ticker_symbol VARCHAR(4),
    ticker_symbol_count integer);

--Create pump data into output
CREATE OR REPLACE PUMP "STREAM_PUMP" AS
INSERT INTO "DESTINATION_SQL_STREAM"
-- select the ingest time used in the GROUP BY clause
SELECT STREAM STEP(source_sql_stream_001.approximate_arrival_time BY INTERVAL '10'
 SECOND) as ingest_time,
    ticker_symbol,
   count(*) over w1 as ticker_symbol_count
FROM source_sql_stream_001
```

```
WINDOW w1 AS (
    PARTITION BY ticker_symbol,
   -- aggregate records based upon ingest time
       STEP(source_sql_stream_001.approximate_arrival_time BY INTERVAL '10' SECOND)
   -- use process time as a trigger, which can be different time window as the
 aggregate
   RANGE INTERVAL '10' SECOND PRECEDING);
```

Results

The preceding example outputs a stream similar to the following.



Filter by column name

ROWTIME	INGEST_TIME	TICKER_SYMBOL	TICKER_SYMBOL_COUNT
2017-07-26 20:17:30.472	2017-07-26 20:17:20.0	CRM	1
2017-07-26 20:17:30.472	2017-07-26 20:17:20.0	BAC	1
2017-07-26 20:17:30.472	2017-07-26 20:17:20.0	UHN	1
2017-07-26 20:17:30.472	2017-07-26 20:17:20.0	PPL	1

Notes

STEP (<datetime expression> BY <literal expression>) is an Amazon Kinesis Data Analytics extension.

You can use STEP to aggregate results using tumbling windows. For more information on tumbling windows, see Tumbling Window Concepts.

Log Parsing Functions

Amazon Kinesis Data Analytics features the following functions for log parsing:

- FAST_REGEX_LOG_PARSER works similarly to the regex parser, but takes several "shortcuts" to ensure faster results. For example, the fast regex parser stops at the first match it finds (known as "lazy" semantics.)
- FIXED_COLUMN_LOG_PARSE parses fixed-width fields and automatically converts them to the given SQL types.

- REGEX_LOG_PARSE uses the default Java regular expression parser. For more information about this parser, see Pattern in the Java Platform documentation on the Oracle website.
- SYS_LOG_PARSE processes entries commonly found in UNIX/Linux system logs.
- VARIABLE_COLUMN_LOG_PARSE splits an input string (its first argument, <character-expression>) into fields separated by a delimiter character or delimiter string.
- W3C_LOG_PARSE processes entries in W3C-predefined-format logs.

FAST_REGEX_LOG_PARSER

```
FAST_REGEX_LOG_PARSE('input_string', 'fast_regex_pattern')
```

The FAST_REGEX_LOG_PARSE works by first decomposing the regular expression into a series of regular expressions, one for each expression inside a group and one for each expression outside a group. Any fixed length portions at the start of any expressions are moved to the end of the previous expression. If any expression is entirely fixed length, it is merged with the previous expression. The series of expressions is then evaluated using lazy semantics with no backtracking. (In regular expression parsing parlance, "lazy" means don't parse more than you need to at each step. "Greedy" means parse as much as you can at each step.)

The columns returned will be COLUMN1 through COLUMNn, where n is the number of groups in the regular expression. The columns will be of type varchar(1024). See sample usage below at First FRLP Example and at Further FRLP Examples.

FAST_REGEX_LOG_PARSER (FRLP)

FAST_REGEX_LOG_PARSER uses a lazy search - it stops at the first match. By contrast, the REGEX_LOG_PARSE is greedy unless possessive quantifiers are used.

FAST_REGEX_LOG_PARSE scans the supplied input string for all the characters specified by the Fast Regex pattern.

- All characters in that input string must be accounted for by the characters and scan groups defined in the Fast Regex pattern. Scan groups define the fields-or-columns resulting when a scan is successful.
- If all characters in the input_string are accounted for when the Fast Regex pattern is applied, then FRLP creates an output field (column) from each parenthetical expression in that Fast Regex pattern, in left-to-right order. The first (leftmost) parenthetical expression creates the

first output field, the next (second) parenthetical expression creates the second output field, up through the last parenthetical expression creating the last output field.

- If the input_string contains any characters not accounted for (matched) by applying Fast Regex pattern, then FRLP returns no fields at all.

Character Class Symbols for Fast Regex

Fast Regex uses a different set of character class symbols from the regular regex parser:

Symbol or Construct	Meaning
-	Character range, including endpoints
[charclasses]	Character class
[^ charclasses]	Negated character class
|	Union
&	Intersection
?	Zero or one occurrence
*	Zero or more occurrences
+	One or more occurrences
{n}	n occurrences
{n,}	n or more occurrences
{n,m}	n to m occurrences, including both
.	Any single character
#	The empty language
@	Any string
"<Unicode string without double-quotes>"	A string)

Symbol or Construct	Meaning
( )	The empty string)
(unionexp)	Precedence override
< <identifier> >	Named pattern
<n-m>	Numerical interval
charexp:=<Unicode character>	A single non-reserved character
\ <Unicode character>	A single character)

We support the following POSIX standard identifiers as named patterns:

<Digit> - "[0-9]"

<Upper> - "[A-Z]"

<Lower> - "[a-z]"

<ASCII> - "[\u0000-\u007F]"

<Alpha> - "<Lower>|<Upper>"

<Alnum> - "<Alpha>|<Digit>"

<Punct> - "[!\"#$%&'()*+,-./:;<=>?@[\\\]^_`{|}~]"

<Blank> - "[\t]"

<Space> - "[\t\n\f\r\u000B]"

<Cntrl> - "[\u0000-\u001F\u007F]"

<XDigit> - "0-9a-fA-F"

<Print> - "<Alnum>|<Punct>"

<Graph> - "<Print>"

First FRLP Example

This first example uses the Fast Regex pattern '(.*)_(._.*)_.*'

```
select t.r."COLUMN1", t.r."COLUMN2" from
. . . . . .> (values (FAST_REGEX_LOG_PARSE('Mary_had_a_little_lamb',
 '(.*)_(._.*)_.*'))) t(r);
+------------------------+-----------------------+
|         COLUMN1        |         COLUMN2       |
+------------------------+-----------------------+
| Mary_had               |     a_little_lamb     |
+------------------------+-----------------------+
1 row selected
```

1. The scan of input_string ('Mary_had_a_little_lamb') begins with the 1st group defined in Fast Regex pattern: (.*), which means "find any character 0 or more times."

   '**(.*)**_(._.*)_.*'

2. This group specification, defining the first column to be parsed, asks the Fast Regex Log Parser to accept input string characters starting from the input string's first character until it finds the next group in the Fast Regex Pattern or the next literal character or string that is not inside a group (not in parentheses). In this example, the next literal character after the first group is an underscore:

   '(.*)_(._.*)_.*'

3. The parser scans each character in the input string until it finds the next specification in the Fast Regex pattern: an underscore:

   '(.*)_**(._.*)**_.*'

4. Group-2 thus begins with "a_l". Next, the parser needs to determine the end of this group, using the remaining specification in the pattern:

   '(.*)_(._.*)**_.***'

> **Note**
>
> Character-strings or literals specified in the pattern but not inside a group must be found in the input string but will not be included in any output field.

If the Fast Regex pattern had omitted the final asterisk, no results would be obtained.

Further FRLP Examples

The next example uses a "+", which means repeat the last expression 1 or more times ("*" means 0 or more times).

Example A

In this case, the longest prefix is the first underscore. The first field/column group will match on "Mary" and the second will not match.

```
select t.r."COLUMN1", t.r."COLUMN2" from
      . . . . . . . . . . . . .> (values
 (FAST_REGEX_LOG_PARSE('Mary_had_a_little_lamb',
        '(.*)_+(._.*)'))) t(r);
      +----------+----------+
      | COLUMN1  | COLUMN2  |
      +----------+----------+
      +----------+----------+
      No rows selected
```

The preceding example returns no fields because the "+" required there be at least one more underscore-in-a-row; and the input_string does not have that.

Example B

In the following case, the '+' is superfluous because of the lazy semantics:

```
select t.r."COLUMN1", t.r."COLUMN2" from
      . . . . . . . . . . . . .> (values
 (FAST_REGEX_LOG_PARSE('Mary____had_a_little_lamb',
        '(.*)_+(.*)'))) t(r);
      +-------------------------+-------------------------+
      |         COLUMN1         |         COLUMN2         |
      +-------------------------+-------------------------+
      | Mary                    |    had_a_little_lamb    |
      +-------------------------+-------------------------+
      1 row selected
```

The preceding example succeeds in returning two fields because after finding the multiple underscores required by the "_+" specification, the group-2 specification (.*) accepts all remaining characters in the .input_string. Underscores do not appear trailing "Mary" nor leading "had" because the "_+" specification is not enclosed in parentheses.

As mentioned in the introduction, "lazy" in regular expression parsing parlance means don't parse more than you need to at each step; "Greedy" means parse as much as you can at each step.

The first case in this topic, A, fails because when it gets to the first underscore, the regex processor has no way of knowing without backtracking that it can't use the underscore to match "_+", and FRLP doesn't backtrack, whereas REGEX_LOG_PARSE does.

The search directly above, B, gets turned into three searches:

```
(.*)_
_*(._
.*)
```

Notice that the second field group gets split between the second and third searches, also that "_+" is considered the same as "__*" (that is, it considers "underscore repeat-underscore-1-or-more-times" the same as "underscore underscore repeat-underscore-0-or-more-times".)

Case A demonstrates the main difference between REGEX_LOG_PARSE and FAST_REGEX_LOG_PARSE, because the search in A would work under REGEX_LOG_PARSE because that function would use backtracking.

Example C

In the following example, the plus is not superfluous, because the "<Alpha> (any alphabetic char) is fixed length thus will be used as a delimiter for the " +" search.

```
select t.r."COLUMN1", t.r."COLUMN2" from
. . . . . .> (values (FAST_REGEX_LOG_PARSE('Mary____had_a_little_lamb',
 '(.*)_+(<Alpha>.*)'))) t(r);
+---------------------------+---------------------------+
|          COLUMN1          |          COLUMN2          |
+---------------------------+---------------------------+
| Mary                      | had_a_little_lamb         |
+---------------------------+---------------------------+
1 row selected
```

```
'(.*) +(<Alpha>.*)' gets converted into three regular expressions:
'.* '
' *<Alpha>'
'.*$'
```

Each is matched in turn using lazy semantics.

The columns returned will be COLUMN1 through COLUMNn, where n is the number of groups in the regular expression. The columns will be of type varchar(1024).

FIXED_COLUMN_LOG_PARSE

Parses fixed-width fields and automatically converts them to the given SQL types.

```
FIXED_COLUMN_LOG_PARSE ( <string value expression>, <column description string
 expression> )
 <column description string expression> := '<column description> [,...]'
 <column description> :=
   <identifier> TYPE <data type> [ NOT NULL ]
   START <numeric value expression> [FOR <numeric constant expression>]
```

Starting position of column is 0. Column specifications for types DATE,TIME and TIMESTAMP support a format parameter allowing the user to specify exact time component layout. The parser uses the Java class java.lang.SimpleDateFormat to parse the strings for types DATE, TIME and TIMESTAMP. The Date and Time Patterns topic gives a full description and examples of timestamp format strings. The following is an example of a column definition with a format string:

```
"name" TYPE TIMESTAMP 'dd/MMM/yyyy:HH:mm:ss'
```

Related Topics

REGEX_LOG_PARSE

REGEX_LOG_PARSE

```
REGEX_LOG_PARSE (<character-expression>,<regex-pattern>,<columns>)<regex-pattern> :=
 <character-expression>[OBJECT] <columns> := <columnname> [ <datatype> ] {,
 <columnname> <datatype> }*
```

Parses a character string based on Java Regular Expression patterns as defined in java.util.regex.pattern.

Columns are based on match groups defined in the regex-pattern. Each group defines a column, and the groups are processed from left to right. Failure to match produces a NULL value result: If the regular expression does not match the string passed as the first parameter, NULL is returned.

The columns returned will be COLUMN1 through COLUMNn, where n is the number of groups in the regular expression. The columns will be of type varchar(1024).

Examples

Example Dataset

The examples following are based on the sample stock dataset that is part of the Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Analytics application that has the sample stock ticker input stream. To learn how to create an Analytics application and configure the sample stock ticker input stream, see the Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

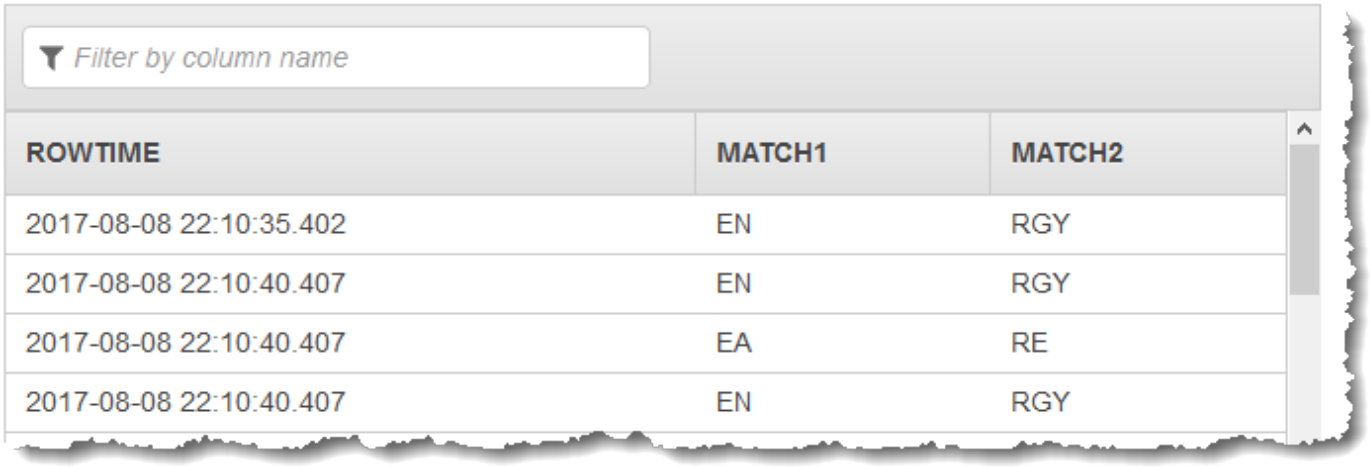
Example 1: Return results from two capture groups

The following code example searches the contents of the `sector` field for a letter E and the character that follows it, and then searches for a letter R, and returns it and all characters following it:

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (match1 VARCHAR(1024), match2
 VARCHAR(1024));

CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM T.REC.COLUMN1, T.REC.COLUMN2
```

```
    FROM
         (SELECT STREAM SECTOR,
             REGEX_LOG_PARSE(SECTOR, '.*([E].).*([R].*)') AS REC
             FROM SOURCE_SQL_STREAM_001) AS T;
```

The preceding code example produces results similar to the following:

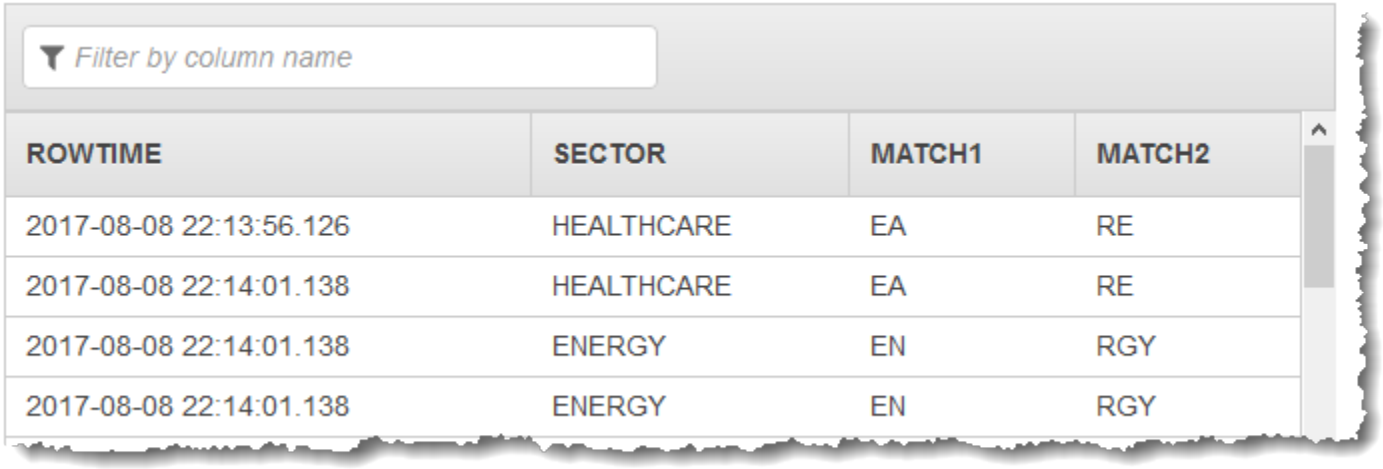


Example 2: Return a stream field and results from two capture groups

The following code example returns the `sector` field, and searches the contents of the `sector` field for a letter E and returns it and the character that follows it, and then searches for a letter R, and returns it and all characters following it:

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (sector VARCHAR(24), match1
 VARCHAR(24), match2 VARCHAR(24));

CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM T.SECTOR, T.REC.COLUMN1, T.REC.COLUMN2
    FROM
         (SELECT STREAM SECTOR,
             REGEX_LOG_PARSE(SECTOR, '.*([E].).*([R].*)') AS REC
             FROM SOURCE_SQL_STREAM_001) AS T;
```

The preceding code example produces results similar to the following:



For more information, see FAST_REGEX_LOG_PARSER.

Quick Regex Reference

For full details on Regex, see java.util.regex.pattern

[xyz] Find single character of: x, y or z

[^abc] Find any single character except: x, y, or z

[r-z] Find any single character between r-z

[r-zR-Z] Find any single character between r-z or R-Z

^ Start of line

$ End of line

\A Start of string

\z End of string

. Any single character

\s Find any whitespace character

\S Find any non-whitespace character

\w Find any word character (letter, number, underscore)

\W Find any non-word character

\b Find any word boundary

(...) Capture everything enclosed

(x|y) Find x or y (also works with symbols such as \d or \s)

x? Find zero or one of x (also works with symbols such as \d or \s)

x* Find zero or more of x (also works with symbols such as \d or \s)

x+ Find one or more of x (also works with symbols such as \d or \s)

\d Find any digit \D Find any non-digit	x{3} Find exactly 3 of x (also works with symbols such as \d or \s) x{3,} Find 3 or more of x (also works with symbols such as \d or \s) x{3,6} Find between 3 and 6 of x (also works with symbols such as \d or \s)

SYS_LOG_PARSE

Parses the standard syslog format:

```
Mon DD HH:MM:SS server message
```

SYS_LOG_PARSE processes entries commonly found in UNIX/Linux system logs. System log entries start with a timestamp and are followed with a free form text field. SYS_LOG_PARSE output consists of two columns. The first column is named "COLUMN1" and is SQL data type TIMESTAMP. The second column is named "COLUMN2" and is SQL type VARCHAR().

Note

For more information about SYSLOG, see IETF RFC3164. For more information about date-time patterns and matching, see Date and Time Patterns.

VARIABLE_COLUMN_LOG_PARSE

```
VARIABLE_COLUMN_LOG_PARSE(
 <character-expression>, <columns>, <delimiter-string>
 [ , <escape-string>, <quote-string> ] )
 <columns> := <number of columns> | <list of columns>
 <number of columns> := <numeric value expression>
 <list of columns> := '<column description>[, ...]'
 <column description> := <identifier> TYPE <data type> [ NOT NULL ]
 <delimiter string> := <character-expression>
 <escape-string> := <character-expression>
 <quote-string> := '<begin quote character> [ <end quote character> ]'
```

VARIABLE_COLUMN_LOG_PARSE splits an input string (its first argument, <character-expression>) into fields separated by a delimiter character or delimiter string. Thus it handles comma-separated values or tab-separated values. It can be combined with FIXED_COLUMN_LOG_PARSE to handle something like maillog, where some fields are fixed-length and others are variable-length.

> **Note**
>
> Parsing of binary files is not supported.

The arguments <escape-string> and <quote-string> are optional. Specifying an <escape-string> allows the value of a field to contain an embedded delimiter. As a simple example, if the <delimiter-string> specified a comma, and the <escape-string> specified a backslash, then an input of "a,b' would be split into two fields "a" and "b", but an input of "a\,b" would result in a single field "a,b".

Since Amazon Kinesis Data Analytics supports Expressions and Literals, a tab can also be a delimiter, specified using a unicode escape, e.g., u&'\0009', which is a string consisting only of a tab character.

Specifying a <quote-string> is another way to hide an embedded delimiter. The <quote-string> should be a one or two character expression: the first is used as the <begin quote character> character; the second, if present, is used as the <end quote character> character. If only one character is supplied, it is used as both to begin and to end quoted strings. When the input includes a quoted string, that is, a string enclosed in the characters specified as <quote-string>, then that string appears in one field, even if it contains a delimiter.

Note that the <begin quote character> and <end quote character> are single characters and can be different. The <begin quote character> can be used to start and end the quoted string, or the <begin quote character> can start the quoted string and the <end quote character> used to end that quoted string.

When a list of columns <list of columns> is supplied as the second parameter <columns>, the column specifications (<column description>) for types DATE, TIME, and TIMESTAMP support a format parameter allowing the user to specify exact time component layout. The parser uses the Java class java.lang.SimpleDateFormat to parse the strings for those types. Date and Time Patterns gives a full description of timestamp format strings, with examples. The following is an example of a column definition with a format string:

```
    "name" TYPE TIMESTAMP 'dd/MMM/yyyy:HH:mm:ss'
```

By default, the output columns are named COLUMN1, COLUMN2, COLUMN3, etc., each of SQL data type VARCHAR(1024).

W3C_LOG_PARSE

```
W3C_LOG_PARSE( <character-expression>, <format-string> )
<format-string> := '<predefined-format> | <custom-format>'
<predefined format> :=
        COMMON
      | COMMON WITH VHOST
      | NCSA EXTENDED
      | REFERER
      | AGENT
      | IIS
<custom-format> := [an Apache log format specifier]
```

W3C Predefined Formats

Specifying the following W3C-predefined-format names summarizes using the format specifiers indicated, as shown in the following statement:

```
select stream W3C_LOG_PARSE(message, 'COMMON') r  from w3ccommon t;
```

Format Name	W3C Name	Format Specifiers
COMMON	Common Log Format (CLF)	%h %l %u %t "%r" %>s %b
COMMON WITH VHOST	Common Log Format with Virtual Host	%v %h %l %u %t "%r" %>s %b
NCSA EXTENDED	NCSA extended/combined log format	%h %l %u %t "%r" %>s %b "%[Referer]i" "%[User-agent]i"
REFERER	Referer log format	%[Referer]i ---> %U

Format Name	W3C Name	Format Specifiers
AGENT	Agent (Browser) log format	%[User-agent]i

W3C Format Specifiers

The format specifiers are listed below. W3C_LOG_PARSE automatically detects these specifiers and output records with one column for each specifier. The column's type is automatically chosen based on the possible outputs of the specifier. For example, %b represents the number of bytes sent in processing an HTTP request, so the column type is numeric. For %B, however, zero bytes is represented by a dash - forcing the column type to be text. Note A explains what the "..." and "<" or ">" markings shown in the specifier table mean.

The following table lists W3C format specifiers alphabetically by command.

Format Specifier	Explanation
%	The percent sign (Apache 2.0.44 and later)
%...a	Remote IP-address
%...A	Local IP-address
%...B	Size of response in bytes, excluding HTTP headers.
%...b	Size of response in bytes, excluding HTTP headers, in CLF format, which means that when no bytes are sent, uses a '-' rather than a 0.
%...[Customerdata]C	The contents of cookie Customerdata in the request sent to the server.
%...D	The time taken to serve the request, in microseconds.

Format Specifier	Explanation
%...[CUSTOMERDATA]e	The contents of the environment variable CUSTOMERDATA
%...f	Filename
%...h	Remote host
%...H	The request protocol
%...[Customerdata]i	The contents of Customerdata: header line(s) in the request sent to the server.
%...l	Remote logname (from identd, if supplied)
%...m	The request method
%...[Customerdata]n	The contents of note Customerdata from another module.
%...[Customerdata]o	The contents of Customerdata: header line(s) in the reply.
%...p	The canonical port of the server serving the request
%...P	The process ID of the child that serviced the request.
%...[format]P	The process ID or thread id of the child that serviced the request. Valid formats are pid and tid. (Apache 2.0.46 and later)
%...q	The query string (prepended with a ? if a query string exists, otherwise an empty string)
%...r	First line of request

Format Specifier	Explanation
%...s	Status. For requests that got internally redirected, this is the status of the *original* request --- %...>s for the last.
%...t	Time, in common log format time format (standard English format)
%...[format]t	The time, in the form given by format, which should be in strimmer(3) format. (potentially localized)
%...T	The time taken to serve the request, in seconds.
%...u	Remote user (from auth; may be bogus if return status (%s) is 401)
%...U	The URL path requested, not including any query string.
%...v	The canonical ServerName of the server serving the request.
%...V	The server name according to the UseCanoni calName setting.

Format Specifier	Explanation
%...X	Connection status when response is completed X = connection aborted before the response completed. + = connection may be kept alive after the response is sent. - = connection will be closed after the response is sent. (The %..X directive was %...c in late versions of Apache 1.3, but this conflicted with the historical ssl %...[var]c syntax.)
:%...I:	Bytes received, including request and headers, cannot be zero. You need to enable mod_logio to use this.
:%...O:	Bytes sent, including headers, cannot be zero. You need to enable mod_logio to use this.

Note

Some W3C format specifiers are shown as containing a "..." indication or a "<" or ">", which are optional controls on suppressing or redirecting the output of that specifier. The "..." can either be empty (as in the COMMON specification "\%h %u %r \%s %b") or it can indicate conditions for including the item. The conditions are a list of HTTP status codes, possibly preceded by "!", and if the specified condition is not met, then the column or field returned shows "-".
For example, as described in the Apache documentation, specifying "%400,501[User-agent]i" will log the User-agent only on 400 errors and 501 errors (Bad Request, Not

Implemented). Similarly, "%!200,304,302[Referer]i" will log the Referer: on all requests that fail to return some sort of normal status.
The modifiers "<" and ">" can be used to choose whether the original or final (respectively) request should be consulted when a request has been internally redirected. By default, the % directives %s, %U, %T, %D, and %r look at the original request while all others look at the final request. So for example, %>s can be used to record the final status of the request and %<u can be used to record the original authenticated user on a request that is internally redirected to an unauthenticated resource.
For security reasons, starting with Apache 2.0.46, non-printable and other special characters are escaped mostly by using \xhh sequences, where hh stands for the hexadecimal representation of the raw byte. Exceptions from this rule are " and \ which are escaped by prepending a backslash, and all white space characters which are written in their C-style notation (\n, \t etc). In httpd 2.0 versions prior to 2.0.46, no escaping was performed on the strings from %...r, %...i and %...o, so great care was needed when dealing with raw log files, since clients could have inserted control characters into the log.
Also, in httpd 2.0, the B format strings represent simply the size in bytes of the HTTP response (which will differ, for instance, if the connection is aborted, or if SSL is used). For the actual number of bytes sent over the network to the client, use the %O format provided by mod_logio.

W3C Format Specifiers by Function or Category

The categories are bytes sent, connection status, content of environmental variable, filename, host, IP, notes, protocol, query string, replies, requests, and time. For the markings "..." or "<" or "<", see the previous note.

Function or Category	W3C Format Specifiers
Bytes sent, excluding HTTP headers	
with a "0" when no bytes are sent	%...B
with a "-" (CLF format) when no bytes are sent	%...b
Bytes received, including request and headers, cannot be zero	:% ... I:

Function or Category	W3C Format Specifiers
Must enable mod_logio to use this.	
Bytes sent, including headers, cannot be zero Must enable mod_logio to use this.	:%... O:
Connection status when response is completed	
Connection aborted before the response completed	X
Connection may be kept alive after the response is sent	+
Connection will be closed after the response is sent	-

> **Note**
>
> The %..X directive was %...c in late versions of Apache 1.3, but this conflicted with the historical ssl %...[var]c syntax.

Environment variable CUSTOMERDATA	
contents	%...[CUSTOMERDATA]e
Filename	%...f
Host (remote)	%...h
Protocol	%...H
IP addresses	
Remote	%...a
Local	%...A
Notes	

Function or Category	W3C Format Specifiers
Contents of note Customerdata from another module	%...[Customerdata]n
Protocol (request)	%...H
Query string **Note** If query exists, prepended with a ? If not, the empty string.	%...q
Replies	
Contents of Customerdata (header lines in the reply)	%...[Customerdata]o

The W3C format specifiers for the response and time categories are listed following table.

Function or Category	W3C Format Specifiers
Requests	
Canonical port of the server serving the request	%...p
Contents of cookie Customerdata in the request sent to server	%... [Customerdata]C
Contents of BAR:header line(s)	%... [BAR]i
First line sent:	%...r
Microseconds taken to serve a request	%...D
Protocol	%...H

Function or Category	W3C Format Specifiers
Process ID of the child that serviced the request	%...P
Process ID or thread id of the child that serviced the request. Valid formats are pid and tid. (Apache 2.0.46 and later)	%...[format]P
Remote logname (from identd, if supplied)	%...l
Remote user: (from auth; may be bogus if return status (%s) is 401)	%...u
Server (canonical ServerName) serving the request	%...v
Server name by the UseCanonicalName setting	%...V
Request method	%...m
Return status	%s
Seconds taken to serve the request	%...T
Status of the *original* request that was internally redirected	%...s
Status of the last request	%...>s
URL path requested, not including any query string	%...U
Time	
Common log format time format (standard English format)	%...t

Function or Category	W3C Format Specifiers
Time in strftime(3) format, potentially localized	%...[format]t
Seconds taken to serve the request	%...T

W3C Examples

W3C_LOG_PARSE supports access to logs generated by W3C-compliant applications like the Apache web server, producing output rows with one column for each specifier. The data types are derived from the log entry description specifiers listed in the Apache mod_log_config specification.

Example 1

The input in this example is taken from an Apache log file and is representative of the COMMON log format.

Input

```
(192.168.254.30 - John [24/May/2004:22:01:02 -0700]
                    "GET /icons/apache_pb.gif HTTP/1.1" 304 0),
(192.168.254.30 - Jane [24/May/2004:22:01:02 -0700]
                    "GET /icons/small/dir.gif HTTP/1.1" 304 0);
```

DDL

```
CREATE OR REPLACE PUMP weblog AS
        SELECT STREAM
            l.r.COLUMN1,
            l.r.COLUMN2,
            l.r.COLUMN3,
            l.r.COLUMN4,
            l.r.COLUMN5,
            l.r.COLUMN6,
            l.r.COLUMN7
        FROM (SELECT STREAM W3C_LOG_PARSE(message, 'COMMON')
              FROM "weblog_read) AS l(r);
```

Output

```
192.168.254.30 -  John  [24/May/2004:22:01:02 -0700] GET /icons/apache_pb.gif HTTP/1.1
 304  0
192.168.254.30 -  Jane  [24/May/2004:22:01:02 -0700] GET /icons/small/dir.gif HTTP/1.1
 304  0
```

The specification of COMMON in the FROM clause means the Common Log Format (CLF), which uses the specifiers %h %l %u %t "%r" %>s %b.

The W3C-predefined formats shows the COMMON and other predefined specifier sets.

The specification of COMMON in the FROM clause means the Common Log Format (CLF), which uses the specifiers %h %l %u %t "%r" %>s %b.

The table below, Specifiers used by the Common Log Format, describes the specifiers used by COMMON in the FROM clause.

Specifiers Used by the Common Log Format

Output Column	Format Specifier	Returns
COLUMN1	%h	The IP address of the remote host
COLUMN2	%l	The remote logname
COLUMN3	%u	The remote user
COLUMN4	%t	The time
COLUMN5	"%r"	The first line of the request
COLUMN6	%>s	The status: For internally redirected requests, the status of the *original* request --- %...>s for the last.

Output Column	Format Specifier	Returns
COLUMN7	%b	The number of bytes sent, excluding HTTP headers

Example 2

The DDL in this example shows how to rename output columns and filter out unneeded columns.

DDL

```
CREATE OR REPLACE VIEW "Schema1".weblogreduced AS
    SELECT STREAM CAST(s.COLUMN3 AS VARCHAR(5)) AS LOG_USER,
    CAST(s.COLUMN1 AS VARCHAR(15)) AS ADDRESS,
    CAST(s.COLUMN4 AS VARCHAR(30)) as TIME_DATES
    FROM "Schema1".weblog s;
```

Output

```
+----------+-----------------+---------------------------------+
| LOG_USER |     ADDRESS     |           TIME_DATES            |
|          |                 |                                 |
+----------+-----------------+---------------------------------+
| Jane     | 192.168.254.30  | [24/May/2004:22:01:02 -0700]    |
|          |                 |                                 |
| John     | 192.168.254.30  | [24/May/2004:22:01:02 -0700]    |
+----------+-----------------+---------------------------------+
```

W3C Customized Formats

The same results would be created by naming the specifiers directly rather than using the "COMMON" name, as shown following:

```
CREATE OR REPLACE FOREIGN STREAM schema1.weblog
    SERVER logfile_server
    OPTIONS (LOG_PATH '/path/to/logfile',
             ENCODING 'UTF-8',
             SLEEP_INTERVAL '10000',
             MAX_UNCHANGED_STATS '10',
             PARSER 'W3C',
```

```
                PARSER_FORMAT '%h %l %u %t \"%r\" %>s %b');
    or
     CREATE FOREIGN STREAM "Schema1".weblog_read
     SERVER "logfile_server"
     OPTIONS (log_path '/path/to/logfile',
     encoding 'UTF-8',
     sleep_interval '10000',
     max_unchanged_stats '10');
     CREATE OR REPLACE VIEW "Schema1".weblog AS
        SELECT STREAM
            l.r.COLUMN1,
            l.r.COLUMN2,
            l.r.COLUMN3,
            l.r.COLUMN4,
            l.r.COLUMN5,
            l.r.COLUMN6
        FROM (SELECT STREAM W3C_LOG_PARSE(message, '%h %l %u %t \"%r\" %>s %b')
              FROM "Schema1".weblog_read) AS l(r);
```

Note

If you change %t to [%t], the date column contains the following:

```
        24/May/2004:22:01:02 -0700
```

(instead of [24/May/2004:22:01:02 -0700])

Sorting Functions

The topics in this section describe the sorting functions for Amazon Kinesis Data Analytics streaming SQL.

Topics

- Group Rank

Group Rank

This function applies a RANK( ) function to logical groups of rows and optionally delivers the group in sorted order.

Applications of `group_rank` include the following:

- To sort results of a streaming `GROUP BY`.
- To determine a relationship within the results of a group.

Group Rank can do the following actions:

- Apply rank to a specified input column.
- Supply either sorted or unsorted output.
- Enable the user to specify a period of inactivity for data flush.

SQL Declarations

The functional attributes and DDL are described in the sections that follow.

Functional Attributes for Group_Rank

This function acts as follows:

- Gathers rows until either a rowtime change is detected or a specified idle-time limit is exceeded.
- Accepts any streaming rowset.
- Uses any column with a basic SQL data type of `INTEGER`, `CHAR`, `VARCHAR` as the column by which to do the ranking.
- Orders the output rows either in the order received or in ascending or descending order of values in the selected column.

DDL for Group_Rank

```
group_rank(c cursor, rankByColumnName VARCHAR(128),
  rankOutColumnName VARCHAR(128), sortOrder VARCHAR(10), outputOrder VARCHAR(10),
  maxIdle INTEGER, outputMax INTEGER)
returns table(c.*, "groupRank" INTEGER)
```

The parameters to the function are listed in the following table.

Parameter	Description
`c`	CURSOR to Streaming Result Set
`rankByColumnName`	String naming the column to use for ranking the group.
`rankOutColumnName`	String naming the column to use for returning the rank. This string must match the name of the `groupRank` column in the `RETURNS` clause of the `CREATE FUNCTION` statement.
`sortOrder`	Controls ordering of rows for rank assignment. Valid values are as follows: • 'asc' - Ascending based on the rank. • 'desc' - Descending based on the rank.
`outputOrder`	Controls ordering of output. Valid values are as follows: • 'asc' - Ascending based on the rank. • 'desc' - Descending based on the rank.
`maxIdle`	Time limit in milliseconds for holding a group for ranking. When `maxIdle` expires, the current group is released to the stream. A value of zero indicates no idle timeout.
`outputMax`	Maximum number of rows the function outputs in a given group. A value of 0 indicates no limit.

Example

Example Dataset

The following example is based on the sample stock dataset that is part of the Getting Started Exercise in the *Amazon Kinesis Data Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Data Analytics application that has the sample stock ticker input stream. To learn how to create an analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Data Analytics Developer Guide*.

The sample stock dataset has the following schema:

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

Example 1: Sort the Results of a GROUP BY Clause

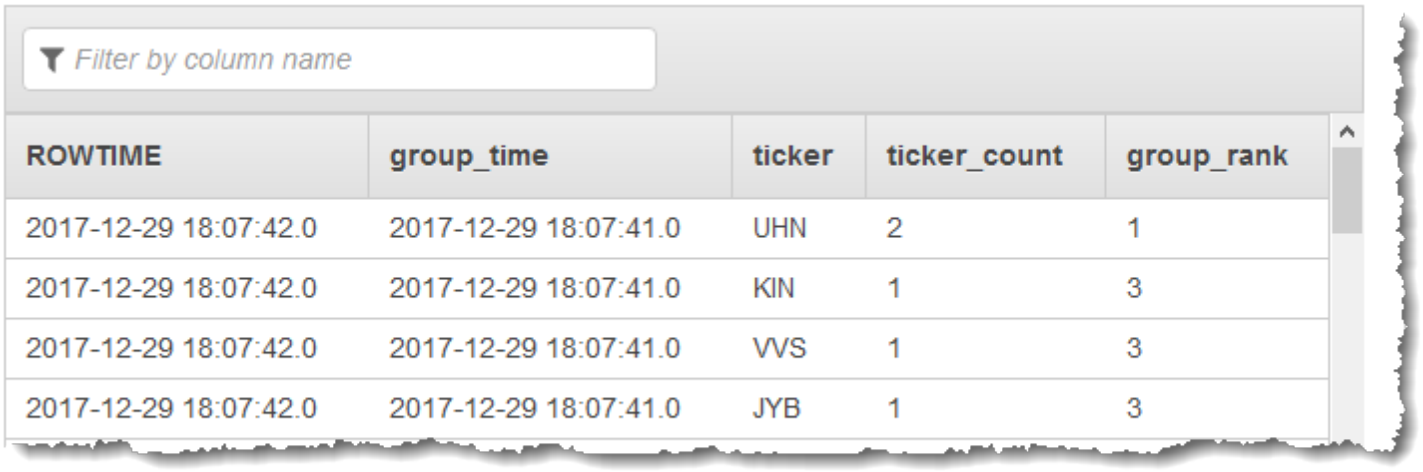
In this example, the aggregate query has a GROUP BY clause on ROWTIME that groups the stream into finite rows. The GROUP_RANK function then sorts the rows returned by the GROUP BY clause.

```
CREATE OR REPLACE STREAM "ticker_grouped" (
    "group_time" TIMESTAMP,
    "ticker" VARCHAR(65520),
    "ticker_count" INTEGER);

CREATE OR REPLACE STREAM "destination_sql_stream" (
    "group_time" TIMESTAMP,
    "ticker" VARCHAR(65520),
    "ticker_count" INTEGER,
    "group_rank" INTEGER);

CREATE OR REPLACE PUMP "ticker_pump" AS
    INSERT INTO "ticker_grouped"
    SELECT STREAM
        FLOOR(SOURCE_SQL_STREAM_001.ROWTIME TO SECOND),
        "TICKER_SYMBOL",
        COUNT(TICKER_SYMBOL)
    FROM SOURCE_SQL_STREAM_001
```

```
    GROUP BY FLOOR(SOURCE_SQL_STREAM_001.ROWTIME TO SECOND), TICKER_SYMBOL;

CREATE OR REPLACE PUMP DESTINATION_SQL_STREAM_PUMP AS
    INSERT INTO "destination_sql_stream"
    SELECT STREAM
        "group_time",
        "ticker",
        "ticker_count",
        "groupRank"
    FROM TABLE(
        GROUP_RANK(
            CURSOR(SELECT STREAM * FROM "ticker_grouped"),
            'ticker_count',
            'groupRank',
            'desc',
            'asc',
            5,
            0));
```

Results

The preceding examples output a stream similar to the following.



Operational Overview

Rows are buffered from the input cursor for each group (that is, rows with the same rowtimes). Ranking of the rows is done either after the arrival of a row with a different rowtime (or when the idle timeout occurs). Rows continue to be read while ranking is performed on the group of rows with the same rowtime.

The `outputMax` parameter specifies the maximum number of rows to be returned for each group after ranks are assigned.

By default, `group_rank` supports column pass through, as the example illustrates by using `c.*` as the standard shortcut directing pass through of all input columns in the order presented. You can, instead, name a subset using the notation "`c.columName`", allowing you to reorder the columns. However, using specific column names ties the UDX to a specific input set, whereas using the `c.*` notation allows the UDX to handle any input set.

The `rankOutColumnName` parameter specifies the output column used to return ranks. This column name must match the column name specified in the `RETURNS` clause of the `CREATE FUNCTION` statement.

Statistical Variance and Deviation Functions

Each of these functions takes a set of numbers, ignores nulls, and can be used as either an aggregate function or an analytical function. For more information, see Aggregate Functions and Analytic Functions.

The relationships among these functions are described in the following table.

Function purpose	Function name	Formula	Comments
Hotspots	HOTSPOTS (expr)	Detects hotspots of frequently occurring data in the data stream.	
Random Cut Forest	RANDOM_CUT_FOREST (expr)	Detects anomalies in the data stream.	
Random Cut Forest with Explanation	RANDOM_CUT_FOREST_WITH_EXPLANATION (expr)	Detects anomalies in the data stream, and returns an attribution score based on how anomalous the data in each column is.	

Function purpose	Function name	Formula	Comments
Population variance	VAR_POP(expr)	(SUM(expr*expr) - SUM(expr)*SUM(expr) / COUNT(expr)) / COUNT(expr)	Applied to an empty set, it returns null.
Population standard deviation	STDDEV_POP	Square root of the population variance (VAR_POP).	When VAR_POP returns null, STDDEV_POP returns null.
Sample variance	VAR_SAMP	(SUM(expr*expr) - SUM(expr)*SUM(expr) / COUNT(expr)) / (COUNT(expr)−1)	Applied to an empty set, it returns null. Applied to an input set of one element, VAR_SAMP returns null.
Sample standard deviation	STDDEV_SAMP (expr)	Square root of the sample variance (VAR_SAMP).	Applied to only 1 row of input data, STDDEV_SAMP returns null.

HOTSPOTS

Detects *hotspots*, or regions of activity that is significantly higher than the norm, in your data stream. A hotspot is defined as a small region of space that is relatively dense with data points.

Using the HOTSPOTS function, you can use a simple SQL function to identify relatively dense regions in your data without having to explicitly build and train complicated machine learning models. You can then identify subsections of your data that need attention so that you can take immediate action.

For example, hotspots in your data might indicate a collection of overheated servers in a data center, a high concentration of vehicles indicating a traffic bottleneck, ride share rides in a certain

area indicating a high-traffic event, or increased sales of products in a category that share similar features.

> **Note**
>
> The ability of the HOTSPOTS function to detect frequent data points is application-dependent. To cast your business problem so that it can be solved with this function requires domain expertise. For example, you might want to determine which combination of columns in your input stream to pass to the function, and how to normalize the data if necessary.

The algorithm accepts the DOUBLE, INTEGER, FLOAT, TINYINT, SMALLINT, REAL, and BIGINT data types. DECIMAL is not a supported type. Use DOUBLE instead.

> **Note**
>
> The HOTSPOT function does not return the records that make up a hotspot. You can use the ROWTIME column to determine which records belong to a given hotspot.

Syntax

```
HOTSPOTS (inputStream,
  windowSize,
  scanRadius,
  minimumNumberOfPointsInAHotspot)
```

Parameters

The following sections describe HOTSPOT function parameters.

inputStream

Pointer to your input stream. You set a pointer using the CURSOR function. For example, the following statements set a pointer to InputStream:

```
--Select all columns from input stream
CURSOR(SELECT STREAM * FROM InputStream)
--Select specific columns from input stream
CURSOR(SELECT STREAM PRICE, CHANGE FROM InputStream)
-- Normalize the column X value.
CURSOR(SELECT STREAM IntegerColumnX / 100, IntegerColumnY FROM InputStream)
-- Combine columns before passing to the function.
CURSOR(SELECT STREAM IntegerColumnX - IntegerColumnY FROM InputStream)
```

Note

Only numeric columns from the input stream will be analyzed for hotspots. The HOTSPOTS function ignores other columns included in the cursor.

windowSize

Specifies the number of records that are considered for each timestep by the sliding window over the stream.

You can set this value between 100 and 1000 inclusive.

By increasing the window size, you can get a better estimate of hotspot position and density (relevance), but this also increases the running time.

scanRadius

Specifies the typical distance between a hotspot point and its nearest neighbors.

This parameter is analogous to the **ε** value in the DBSCAN algorithm.

Set this parameter to a value that is smaller than the typical distance between points that are not in a hotspot, but large enough so that points in a hotspot have neighbors within this distance.

You can set this value to any double value greater than zero. The lower the value for `scanRadius`, the more similar any two records belonging to the same hotspot are. However, low values for `scanRadius` also increase running time. Lower values of `scanRadius` result in hotspots that are smaller but more numerous.

minimumNumberOfPointsInAHotspot

Specifies the number of records that are required for the records to form a hotspot.

> **Note**
>
> This parameter should be set in consideration with windowSize. It is best to think of `minimumNumberOfPointsInAHotspot` as some fraction of `windowSize`. The exact fraction is discoverable through experimentation.

You can set this value between 2 and your configured value for window size inclusive. Choose a value that best models the problem you are solving in light of your chosen value for window size.

Output

The output of the HOTSPOTS function is a table object that has the same schema as the input, with the following additional column:

HOTSPOT_RESULTS

A JSON string describing all the hotspots found around the record. The function returns all potential hotspots; you can filter out hotspots below a certain `density` threshold in your application. The field has the following nodes, with values for each of the input columns:

- `density`: The number of records in the hotspot divided by the hotspot size. You can use this value to determine the relative relevance of the hotspot.
- `maxValues`: The maximum values for the records in the hotspot for each data column.
- `minValues`: The minimum values for the records in the hotspot for each data column.

Data type: VARCHAR.

> **Note**
>
> The trends that machine learning functions use to determine analysis scores are infrequently reset when the Kinesis Data Analytics service performs service maintenance. You might unexpectedly see analysis scores of 0 after service maintenance occurs. We recommend you set up filters or other mechanisms to treat these values appropriately as they occur.

Example

The following example executes the HOTSPOTS function on the demo stream, which contains random data without meaningful hotspots. For an example that executes the HOTSPOTS function on a custom data stream with meaningful data hotspots, see Example: Detect Hotspots.

Example Dataset

The example following is based on the sample stock dataset that is part of the Getting Started Exercise in the *Amazon Kinesis Data Analytics Developer Guide*. To run the example, you need an Kinesis Data Analytics application that has the sample stock ticker input stream. To learn how to create a Kinesis Data Analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Data Analytics Developer Guide*.

The sample stock dataset has the following schema:

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

Example 1: Return Hotspots on the Sample Data Stream

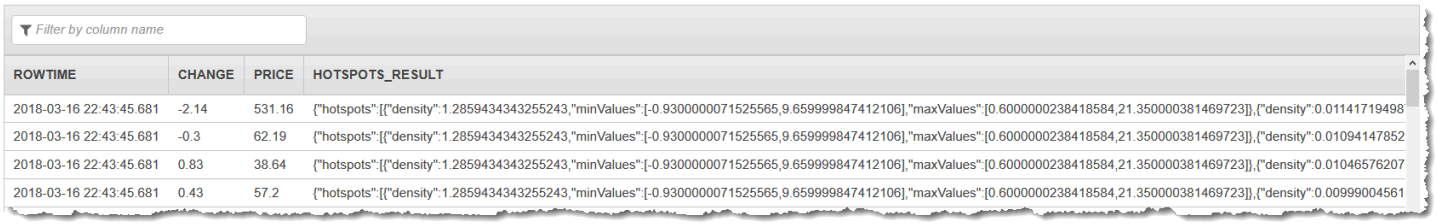
In this example, a destination stream is created for the output of the HOTSPOTS function. A pump is then created that runs the HOTSPOTS function on the specified values in the sample data stream.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM"(
    CHANGE REAL,
    PRICE REAL,
    HOTSPOTS_RESULT VARCHAR(10000));

CREATE OR REPLACE PUMP "STREAM_PUMP" AS
    INSERT INTO "DESTINATION_SQL_STREAM"
        SELECT
            "CHANGE",
            "PRICE",
            "HOTSPOTS_RESULT"
```

```
        FROM TABLE (
            HOTSPOTS(
                CURSOR(SELECT STREAM "CHANGE", "PRICE" FROM "SOURCE_SQL_STREAM_001"),
                100,
                0.013,
                20)
            );
```

Results

This example outputs a stream similar to the following.



Filter by column name

ROWTIME	CHANGE	PRICE	HOTSPOTS_RESULT
2018-03-16 22:43:45.681	-2.14	531.16	{"hotspots":[{"density":1.2859434343255243,"minValues":[-0.9300000071525565,9.659999847412106],"maxValues":[0.6000000238418584,21.350000381469723]},{"density":0.01141719498
2018-03-16 22:43:45.681	-0.3	62.19	{"hotspots":[{"density":1.2859434343255243,"minValues":[-0.9300000071525565,9.659999847412106],"maxValues":[0.6000000238418584,21.350000381469723]},{"density":0.01094147852
2018-03-16 22:43:45.681	0.83	38.64	{"hotspots":[{"density":1.2859434343255243,"minValues":[-0.9300000071525565,9.659999847412106],"maxValues":[0.6000000238418584,21.350000381469723]},{"density":0.01046576207
2018-03-16 22:43:45.681	0.43	57.2	{"hotspots":[{"density":1.2859434343255243,"minValues":[-0.9300000071525565,9.659999847412106],"maxValues":[0.6000000238418584,21.350000381469723]},{"density":0.00999004561

RANDOM_CUT_FOREST

Detects anomalies in your data stream. A record is an anomaly if it is distant from other records. To detect anomalies in individual record columns, see RANDOM_CUT_FOREST_WITH_EXPLANATION.

> **Note**
>
> The RANDOM_CUT_FOREST function's ability to detect anomalies is application-dependent. To cast your business problem so that it can be solved with this function requires domain expertise. For example, determining which combination of columns in your input stream to pass to the function and potentially normalize the data. For more information, see inputStream.

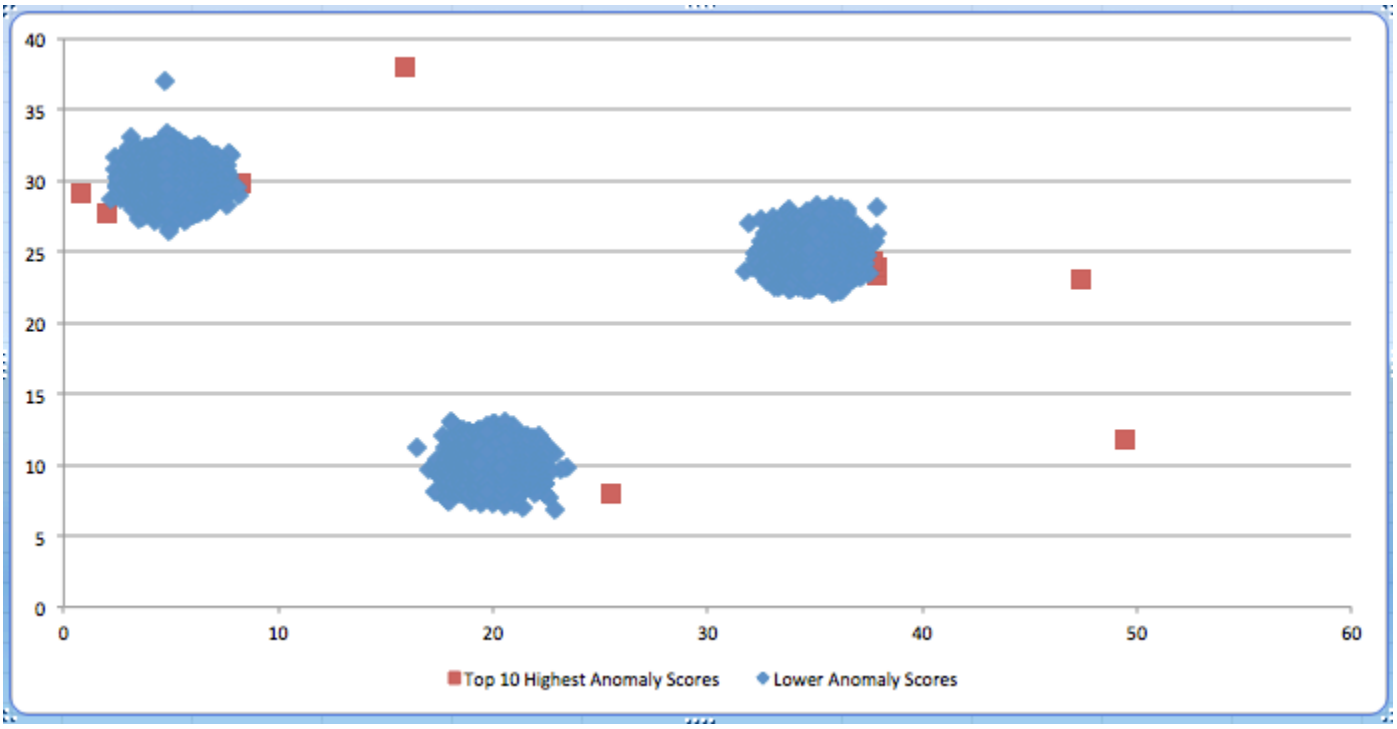
A stream record can have non-numeric columns, but the function uses only numeric columns to assign an anomaly score. A record can have one or more numeric columns. The algorithm uses all of the numeric data in computing an anomaly score. If a record has ***n*** numeric columns, the underlying algorithm assumes each record is a point in ***n***-dimensional space. A point in ***n***-dimensional space that is distant from other points receives a higher anomaly score.

The algorithm starts developing the machine learning model using current records in the stream when you start the application. The algorithm does not use older records in the stream for machine learning, nor does it use statistics from previous executions of the application.

The algorithm accepts the `DOUBLE`, `INTEGER`, `FLOAT`, `TINYINT`, `SMALLINT`, `REAL`, and `BIGINT` data types.

> **Note**
>
> DECIMAL is not a supported type. Use DOUBLE instead.

The following is an example of anomaly detection. The diagram shows three clusters and a few anomalies randomly interjected. The red squares show the records that received the highest anomaly score according to the `RANDOM_CUT_FOREST` function. The blue diamonds represent the remaining records. Note how the highest scoring records tend to be outside the clusters.



For a sample application with step-by-step instructions, see Detect Anomalies.

Syntax

```
RANDOM_CUT_FOREST (inputStream,
                   numberOfTrees,
                   subSampleSize,
                   timeDecay,
                   shingleSize)
```

Parameters

The following sections describe the parameters.

inputStream

Pointer to your input stream. You set a pointer using the CURSOR function. For example, the following statements sets a pointer to `InputStream`.

```
CURSOR(SELECT STREAM * FROM InputStream)
CURSOR(SELECT STREAM IntegerColumnX, IntegerColumnY FROM InputStream)
-- Perhaps normalize the column X value.
CURSOR(SELECT STREAM IntegerColumnX / 100, IntegerColumnY FROM InputStream)
-- Combine columns before passing to the function.
CURSOR(SELECT STREAM IntegerColumnX - IntegerColumnY FROM InputStream)
```

The CURSOR function is the only required parameter for the RANDOM_CUT_FOREST function. The function assumes the following default values for the other parameters:

numberOfTrees = 100

subSampleSize = 256

timeDecay = 100,000

shingleSize = 1

When using this function, your input stream can have up to 30 numeric columns.

numberOfTrees

Using this parameter, you specify the number of random cut trees in the forest.

> **Note**
>
> By default, the algorithm constructs a number of trees, each constructed using a given number of sample records (see `subSampleSize` later in this list) from the input stream. The algorithm uses each tree to assign an anomaly score. The average of all these scores is the final anomaly score.

The default value for `numberOfTrees` is 100. You can set this value between 1 and 1,000 (inclusive). By increasing the number of trees in the forest, you can get a better estimate of the anomaly score, but this also increases the running time.

subSampleSize

Using this parameter, you can specify the size of the random sample that you want the algorithm to use when constructing each tree. Each tree in the forest is constructed with a (different) random sample of records. The algorithm uses each tree to assign an anomaly score. When the sample reaches `subSampleSize` records, records are removed randomly, with older records having a higher probability of removal than newer records.

The default value for `subSampleSize` is 256. You can set this value between 10 and 1,000 (inclusive).

Note that the `subSampleSize` must be less than the `timeDecay` parameter (which is set to 100,000 by default). Increasing the sample size provides each tree a larger view of the data, but also increases the running time.

> **Note**
>
> The algorithm returns zero for the first `subSampleSize` records while the machine learning model is trained.

timeDecay

The `timeDecay` parameter allows you to specify how much of the recent past to consider when computing an anomaly score. This is because data streams naturally evolve over time. For example, an eCommerce website's revenue may continuously increase, or global temperatures may rise over

time. In such situations, we want an anomaly to be flagged relative to recent data, as opposed to data from the distant past.

The default value is 100,000 records (or 100,000 shingles if shingling is used, as described in the following section). You can set this value between 1 and the maximum integer (that is, 2147483647). The algorithm exponentially decays the importance of older data.

If you choose the `timeDecay` default of 100,000, the anomaly detection algorithm does the following:

- Uses only the most recent 100,000 records in the calculations (and ignores older records).
- Within the most recent 100,000 records, assigns exponentially more weight to recent records and less to older records in anomaly detection calculations.

If you don't want to use the default value, you can calculate the number of records to use in the algorithm. To do this, multiply the number of expected records per day by the number of days you want the algorithm to consider. For example, if you expect 1,000 records per day, and you want to analyze 7 days of records, set this parameter to 7,000 (1,000 * 7).

The `timeDecay` parameter determines the maximum quantity of recent records kept in the working set of the anomaly detection algorithm. Smaller `timeDecay` values are desirable if the data is changing rapidly. The best `timeDecay` value is application-dependent.

shingleSize

The explanation given here is for a one-dimensional stream (that is, a stream with one numeric column), but it can also be used for multi-dimensional streams.

A shingle is a consecutive sequence of the most recent records. For example, a `shingleSize` of 10 at time *t* corresponds to a vector of the last 10 records received up to and including time *t*. The algorithm treats this sequence as a vector over the last `shingleSize` number of records.

If data is arriving uniformly in time, a shingle of size 10 at time *t* corresponds to the data received at time *t-9, t-8,…,t*. At time *t+1*, the shingle slides over one unit and consists of data from time *t-8,t-7, …, t, t+1*. These shingled records gathered over time correspond to a collection of 10-dimensional vectors over which the anomaly detection algorithm runs.

The intuition is that a shingle captures the shape of the recent past. Your data may have a typical shape. For example, if your data is collected hourly, a shingle of size 24 may capture the daily rhythm of your data.

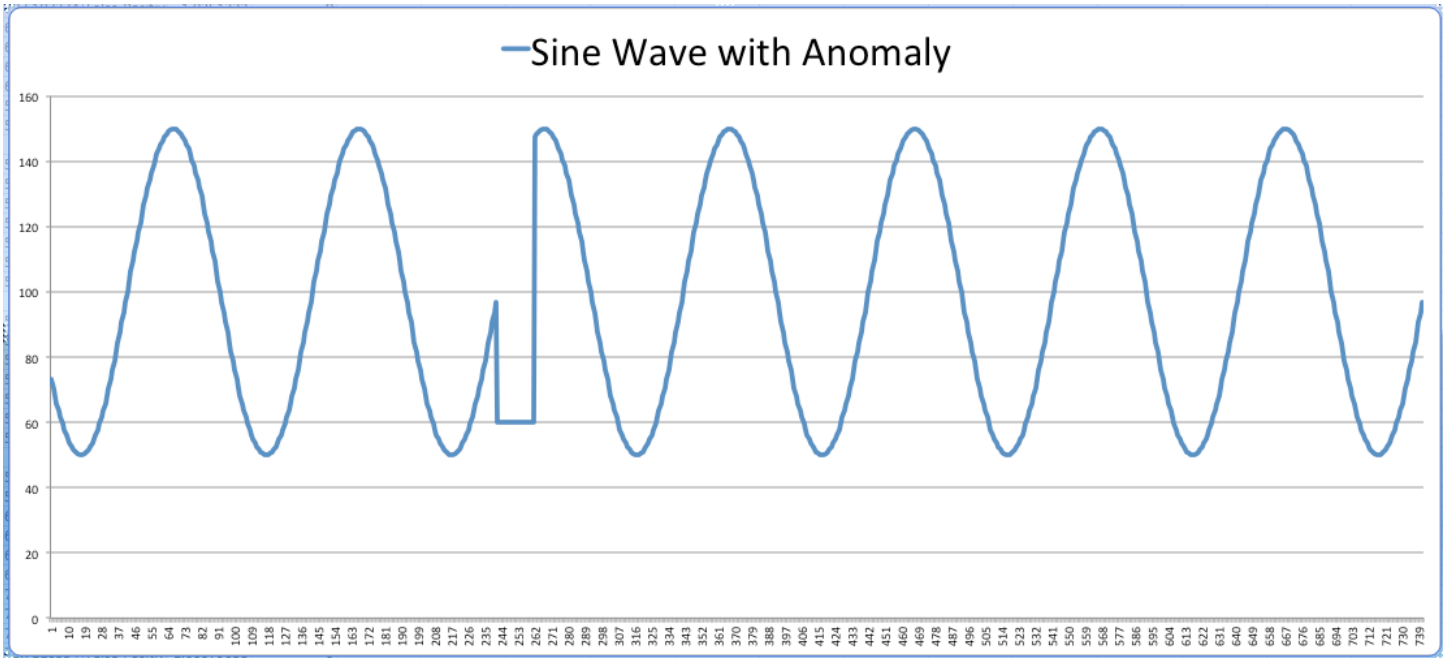
The default `shingleSize` is one record (because shingle size is data dependent). You can set this value between 1 and 30 (inclusive).

Note the following about setting the `shingleSize`:

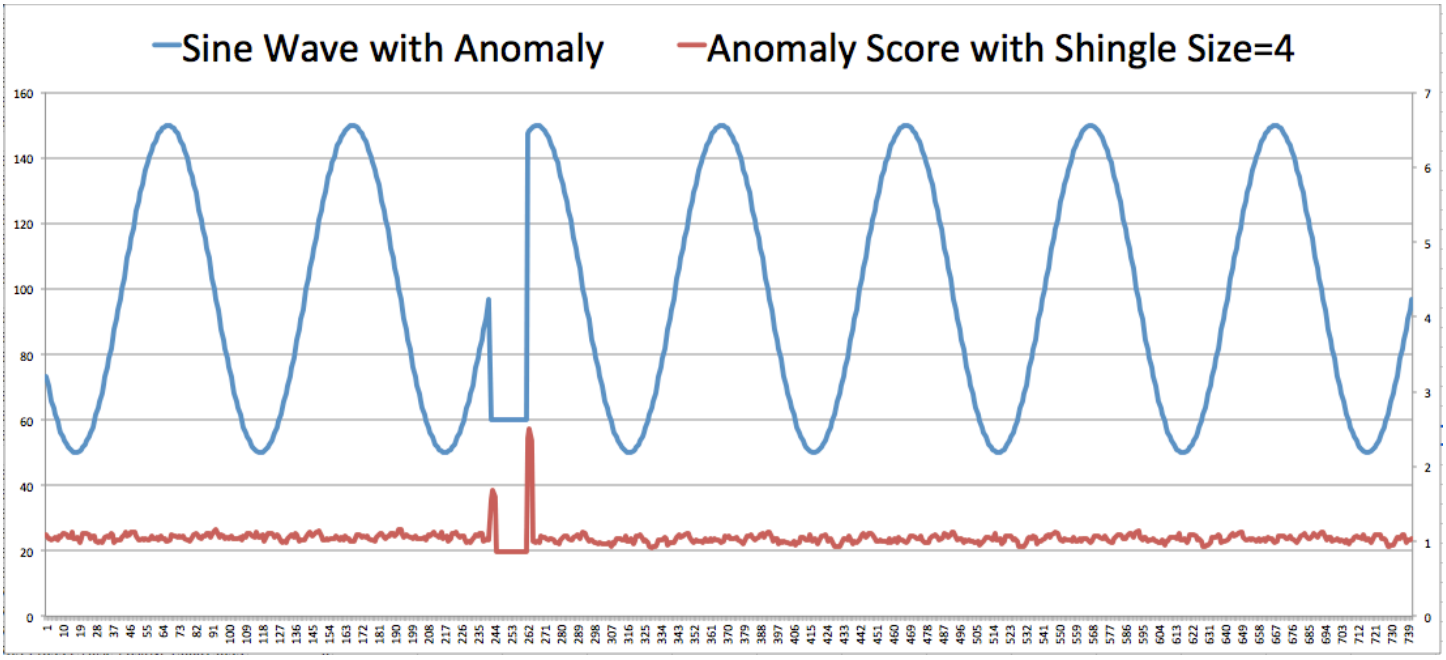
- If you set the `shingleSize` too small, the algorithm will be more susceptible to minor fluctuations in the data, leading to high-anomaly scores for records that are not anomalous.
- If you set the `shingleSize` too large, it may take more time to detect anomalous records because there are more records in the shingle that are not anomalous. It may also take more time to determine that the anomaly has ended.
- Identifying the right shingle size is application-dependent. Experiment with different shingle sizes to ascertain the effect.

The following example illustrates how you can catch anomalies when you monitor the records with the highest anomaly score. In this particular example, the two highest anomaly scores also signal the beginning and end of an artificially injected anomaly.

Consider this stylized one-dimensional stream represented as a sine wave, intended to capture a circadian rhythm. This curve illustrates the typical number of orders that an eCommerce site receives per hour, the number of users logged into a server, the number of ad clicks received per hour, etc. A severe dip of 20 consecutive records is artificially injected in the middle of the plot.



We ran the RANDOM_CUT_FOREST function with a shingle size of four records. The result is shown below. The red line shows the anomaly score. Note that the beginning and the end of the anomaly received high scores.



When you use this function, we recommend that you investigate the highest scoring points as potential anomalies.

Note

The trends that machine learning functions use to determine analysis scores are infrequently reset when the Kinesis Data Analytics service performs service maintenance. You might unexpectedly see analysis scores of 0 after service maintenance occurs. We recommend you set up filters or other mechanisms to treat these values appropriately as they occur.

For more information, see the Robust Random Cut Forest Based Anomaly Detection On Streams white paper at the Journal of Machine Learning Research website.

RANDOM_CUT_FOREST_WITH_EXPLANATION

Computes an anomaly score and explains it for each record in your data stream. The anomaly score for a record indicates how different it is from the trends that have recently been observed for your stream. The function also returns an attribution score for each column in a record, based on how anomalous the data in that column is. For each record, the sum of the attribution scores of all columns is equal to the anomaly score.

You also have the option of getting information about the direction in which a given column is anomalous (whether it's high or low relative to the recently observed data trends for that column in the stream).

For example, in e-commerce applications, you might want to know when there's a change in the recently observed pattern of transactions. You also might want to know how much of the change is due to a change in the number of purchases made per hour, and how much is due to a change in the number of carts abandoned per hour—information that is represented by the attribution scores. You also might want to look at directionality to know whether you're being notified of the change due to an increase or a decrease in each of those values.

Note

The `RANDOM_CUT_FOREST_WITH_EXPLANATION` function's ability to detect anomalies is application-dependent. Casting your business problem so that it can be solved with this function requires domain expertise. For example, you may need to determine which

combination of columns in your input stream to pass to the function, and you may potentially benefit from normalizing the data. For more information, see inputStream.

A stream record can have non-numeric columns, but the function uses only numeric columns to assign an anomaly score. A record can have one or more numeric columns. The algorithm uses all the numeric data in computing an anomaly score.

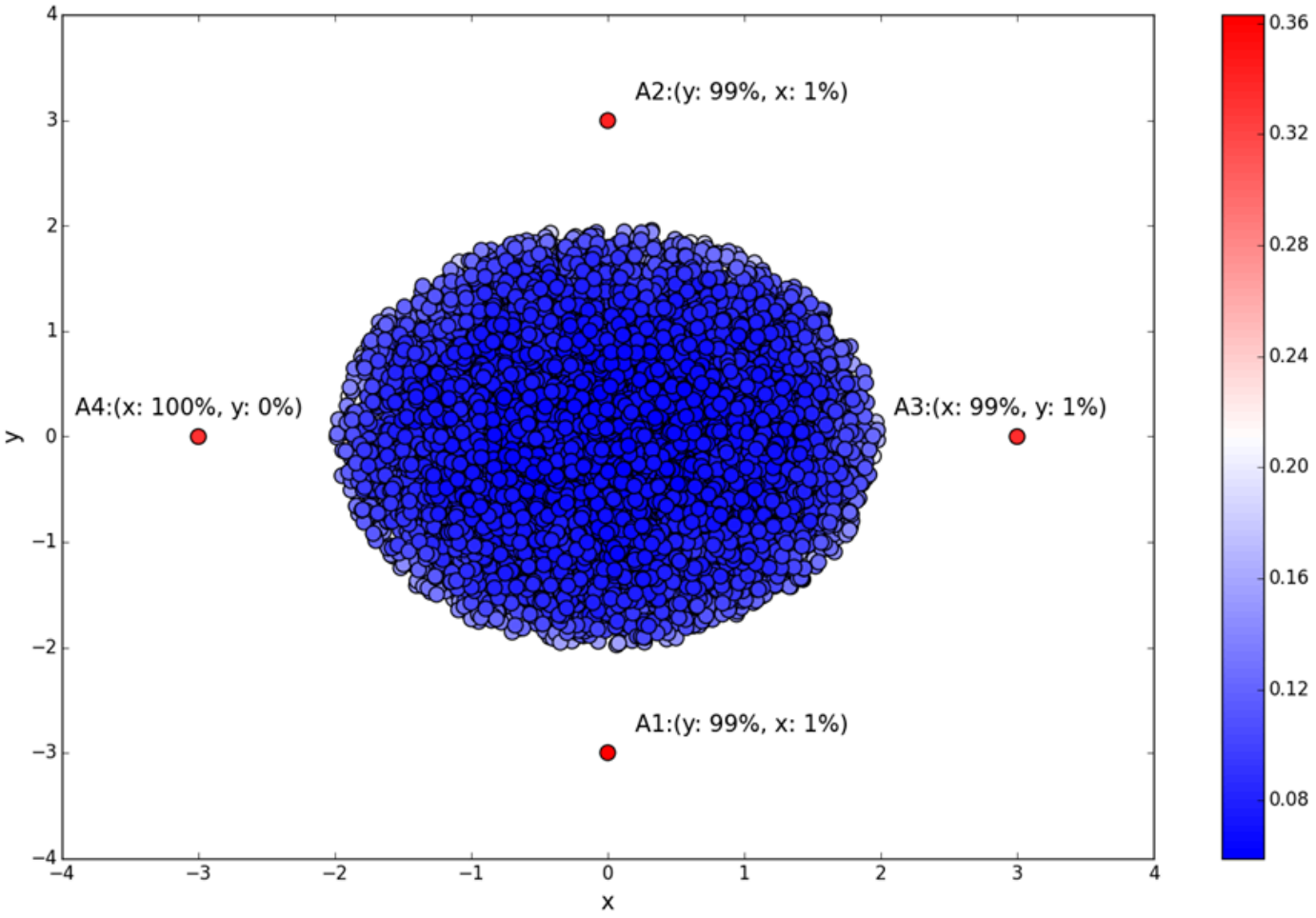
The algorithm starts developing the machine learning model using current records in the stream when you start the application. The algorithm does not use older records in the stream for machine learning, nor does it use statistics from previous executions of the application.

The algorithm accepts the `DOUBLE`, `INTEGER`, `FLOAT`, `TINYINT`, `SMALLINT`, `REAL`, and `BIGINT` data types.

Note

`DECIMAL` is not a supported type. Use `DOUBLE` instead.

The following is a simple visual example of anomaly detection with different attribution scores in two-dimensional space. The diagram shows a cluster of blue data points and four outliers shown as red points. The red points have similar anomaly scores, but these four points are anomalous for different reasons. For points A1 and A2, most of the anomaly is attributable to their outlying y-values. In the case of A3 and A4, you can attribute most of the anomaly to their outlying x-values. Directionality is LOW for the y-value of A1, HIGH for the y-value of A2, HIGH for the x-value of A3, and LOW for the x-value of A4.



Syntax

```
RANDOM_CUT_FOREST_WITH_EXPLANATION (inputStream,
                    numberOfTrees,
                    subSampleSize,
                    timeDecay,
                    shingleSize,
                    withDirectionality
)
```

Parameters

The following sections describe the parameters of the RANDOM_CUT_FOREST_WITH_EXPLANATION function.

inputStream

Pointer to your input stream. You set a pointer using the CURSOR function. For example, the following statements set a pointer to `InputStream`.

```
CURSOR(SELECT STREAM * FROM InputStream)
CURSOR(SELECT STREAM IntegerColumnX, IntegerColumnY FROM InputStream)
-- Perhaps normalize the column X value.
CURSOR(SELECT STREAM IntegerColumnX / 100, IntegerColumnY FROM InputStream)
-- Combine columns before passing to the function.
CURSOR(SELECT STREAM IntegerColumnX - IntegerColumnY FROM InputStream)
```

The CURSOR function is the only required parameter for the `RANDOM_CUT_FOREST_WITH_EXPLANATION` function. The function assumes the following default values for the other parameters:

`numberOfTrees` = 100

`subSampleSize` = 256

`timeDecay` = 100,000

`shingleSize` = 1

`withDirectionality` = FALSE

When you use this function, your input stream can have up to 30 numeric columns.

numberOfTrees

Using this parameter, you specify the number of random cut trees in the forest.

> **Note**
>
> By default, the algorithm constructs a number of trees, each constructed using a given number of sample records (see `subSampleSize` later in this list) from the input stream. The algorithm uses each tree to assign an anomaly score. The average of all these scores is the final anomaly score.

The default value for `numberOfTrees` is 100. You can set this value between 1 and 1,000 (inclusive). By increasing the number of trees in the forest, you can get a better estimate of the anomaly and attribution scores, but this also increases the running time.

subSampleSize

Using this parameter, you can specify the size of the random sample that you want the algorithm to use when constructing each tree. Each tree in the forest is constructed with a (different) random sample of records. The algorithm uses each tree to assign an anomaly score. When the sample reaches `subSampleSize` records, records are removed randomly, with older records having a higher probability of removal than newer records.

The default value for `subSampleSize` is 256. You can set this value between 10 and 1,000 (inclusive).

The `subSampleSize` must be less than the `timeDecay` parameter (which is set to 100,000 by default). Increasing the sample size provides each tree a larger view of the data, but it also increases the running time.

> **Note**
>
> The algorithm returns zero for the first `subSampleSize` records while the machine learning model is trained.

timeDecay

You can use the `timeDecay` parameter to specify how much of the recent past to consider when computing an anomaly score. Data streams naturally evolve over time. For example, an e-commerce website's revenue might continuously increase, or global temperatures might rise over time. In such situations, you want an anomaly to be flagged relative to recent data, as opposed to data from the distant past.

The default value is 100,000 records (or 100,000 shingles if shingling is used, as described in the following section). You can set this value between 1 and the maximum integer (that is, 2147483647). The algorithm exponentially decays the importance of older data.

If you choose the `timeDecay` default of 100,000, the anomaly detection algorithm does the following:

- Uses only the most recent 100,000 records in the calculations (and ignores older records).
- Within the most recent 100,000 records, assigns exponentially more weight to recent records and less to older records in anomaly detection calculations.

The `timeDecay` parameter determines the maximum quantity of recent records kept in the working set of the anomaly detection algorithm. Smaller `timeDecay` values are desirable if the data is changing rapidly. The best `timeDecay` value is application-dependent.

shingleSize

The explanation given here applies to a one-dimensional stream (that is, a stream with one numeric column), but shingling can also be used for multi-dimensional streams.

A shingle is a consecutive sequence of the most recent records. For example, a `shingleSize` of 10 at time *t* corresponds to a vector of the last 10 records received up to and including time *t*. The algorithm treats this sequence as a vector over the last `shingleSize` number of records.

If data is arriving uniformly in time, a shingle of size 10 at time *t* corresponds to the data received at time *t-9, t-8,…,t*. At time *t+1*, the shingle slides over one unit and consists of data from time *t-8,t-7, …, t, t+1*. These shingled records gathered over time correspond to a collection of 10-dimensional vectors over which the anomaly detection algorithm runs.

The intuition is that a shingle captures the shape of the recent past. Your data might have a typical shape. For example, if your data is collected hourly, a shingle of size 24 might capture the daily rhythm of your data.

The default `shingleSize` is one record (because shingle size is data-dependent). You can set this value between 1 and 30 (inclusive).

Note the following about setting the `shingleSize`:

- If you set the `shingleSize` too small, the algorithm is more susceptible to minor fluctuations in the data, leading to high anomaly scores for records that are not anomalous.
- If you set the `shingleSize` too large, it might take more time to detect anomalous records because there are more records in the shingle that are not anomalous. It also might take more time to determine that the anomaly has ended.
- Identifying the right shingle size is application-dependent. Experiment with different shingle sizes to determine the effects.

withDirectionality

A Boolean parameter that defaults to `false`. When set to `true`, it tells you the direction in which each individual dimension makes a contribution to the anomaly score. It also provides the strength of the recommendation for that directionality.

Results

The function returns an anomaly score of 0 or more and an explanation in JSON format.

The anomaly score starts out at 0 for all the records in the stream while the algorithm goes through the learning phase. You then start to see positive values for the anomaly score. Not all positive anomaly scores are significant; only the highest ones are. To get a better understanding of the results, look at the explanation.

The explanation provides the following values for each column in the record:

- **Attribution score:** A nonnegative number that indicates how much this column has contributed to the anomaly score of the record. In other words, it indicates how different the value of this column is from what's expected based on the recently observed trend. The sum of the attribution scores of all columns for the record is equal to the anomaly score.
- **Strength:** A nonnegative number representing the strength of the directional recommendation. A high value for strength indicates a high confidence in the directionality that is returned by the function. During the learning phase, the strength is 0.
- **Directionality:** This is either HIGH if the value of the column is above the recently observed trend or LOW if it's below the trend. During the learning phase, this defaults to LOW.

> **Note**
>
> The trends that machine learning functions use to determine analysis scores are infrequently reset when the Kinesis Data Analytics service performs service maintenance. You might unexpectedly see analysis scores of 0 after service maintenance occurs. We recommend you set up filters or other mechanisms to treat these values appropriately as they occur.

Examples

Stock Ticker Data Example

This example is based on the sample stock dataset that is part of the Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*. To run the example, you need a Kinesis Data Analytics application that has the sample stock ticker input stream. To learn how to create a Kinesis Data Analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the following schema:

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

In this example, the application calculates an anomaly score for the record and an attribution score for the PRICE and CHANGE columns, which are the only numeric columns in the input stream.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (anomaly REAL, ANOMALY_EXPLANATION
 VARCHAR(20480));
CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"
SELECT "ANOMALY_SCORE", "ANOMALY_EXPLANATION" FROM TABLE
 (RANDOM_CUT_FOREST_WITH_EXPLANATION(CURSOR(SELECT STREAM * FROM
 "SOURCE_SQL_STREAM_001"), 100, 256, 100000, 1, true)) WHERE ANOMALY_SCORE > 0
```

The preceding example outputs a stream similar to the following.

Filter by column name

ROWTIME	ANOMALY	ANOMALY_EXPLANATION
2017-10-23 20:09:21.466	0.57327205	{"CHANGE":{"DIRECTION":"LOW","STRENGTH":"0.4824","ATTRIBUTION_SCORE":"0.3780"},"PRICE":{"DIRECTION":"HIGH","STRENGTH":"0.0675","ATTRIBUTION_SCORE":"0.1953"}}
2017-10-23 20:09:21.466	0.73723656	{"CHANGE":{"DIRECTION":"HIGH","STRENGTH":"0.3178","ATTRIBUTION_SCORE":"0.3757"},"PRICE":{"DIRECTION":"HIGH","STRENGTH":"0.1695","ATTRIBUTION_SCORE":"0.3616"}}
2017-10-23 20:09:21.466	0.6443901	{"CHANGE":{"DIRECTION":"HIGH","STRENGTH":"0.1731","ATTRIBUTION_SCORE":"0.3396"},"PRICE":{"DIRECTION":"LOW","STRENGTH":"0.2367","ATTRIBUTION_SCORE":"0.3048"}}
2017-10-23 20:09:21.466	0.55428815	{"CHANGE":{"DIRECTION":"HIGH","STRENGTH":"0.0888","ATTRIBUTION_SCORE":"0.2327"},"PRICE":{"DIRECTION":"LOW","STRENGTH":"0.1295","ATTRIBUTION_SCORE":"0.3216"}}
2017-10-23 20:09:21.466	0.5416738	{"CHANGE":{"DIRECTION":"HIGH","STRENGTH":"0.0316","ATTRIBUTION_SCORE":"0.2805"},"PRICE":{"DIRECTION":"HIGH","STRENGTH":"0.0208","ATTRIBUTION_SCORE":"0.2612"}}
2017-10-23 20:09:21.466	0.67421293	{"CHANGE":{"DIRECTION":"HIGH","STRENGTH":"0.3364","ATTRIBUTION_SCORE":"0.3058"},"PRICE":{"DIRECTION":"LOW","STRENGTH":"0.2090","ATTRIBUTION_SCORE":"0.3684"}}
2017-10-23 20:09:21.466	0.6528528	{"CHANGE":{"DIRECTION":"HIGH","STRENGTH":"0.2020","ATTRIBUTION_SCORE":"0.3746"},"PRICE":{"DIRECTION":"HIGH","STRENGTH":"0.0107","ATTRIBUTION_SCORE":"0.2783"}}
2017-10-23 20:09:21.466	0.83734107	{"CHANGE":{"DIRECTION":"HIGH","STRENGTH":"0.0177","ATTRIBUTION_SCORE":"0.2954"},"PRICE":{"DIRECTION":"LOW","STRENGTH":"0.0204","ATTRIBUTION_SCORE":"0.5419"}}
2017-10-23 20:09:21.466	0.9346224	{"CHANGE":{"DIRECTION":"HIGH","STRENGTH":"0.2364","ATTRIBUTION_SCORE":"0.2299"},"PRICE":{"DIRECTION":"LOW","STRENGTH":"0.3877","ATTRIBUTION_SCORE":"0.7047"}}
2017-10-23 20:09:21.466	0.6726167	{"CHANGE":{"DIRECTION":"HIGH","STRENGTH":"0.0397","ATTRIBUTION_SCORE":"0.3335"},"PRICE":{"DIRECTION":"HIGH","STRENGTH":"0.0631","ATTRIBUTION_SCORE":"0.3392"}}
2017-10-23 20:09:21.466	0.6636287	{"CHANGE":{"DIRECTION":"LOW","STRENGTH":"0.0571","ATTRIBUTION_SCORE":"0.3147"},"PRICE":{"DIRECTION":"HIGH","STRENGTH":"0.1943","ATTRIBUTION_SCORE":"0.3489"}}
2017-10-23 20:09:21.466	0.7353514	{"CHANGE":{"DIRECTION":"HIGH","STRENGTH":"0.1851","ATTRIBUTION_SCORE":"0.1992"},"PRICE":{"DIRECTION":"HIGH","STRENGTH":"0.0564","ATTRIBUTION_SCORE":"0.5361"}}

Network and CPU Utilization Example

This theoretical example shows two sets of data that follow an oscillating pattern. In the following graph, they're represented by the red and blue curves at the top. The red curve shows network utilization over time, and the blue curve shows idle CPU over time for the same computer system. The two signals, which are out of phase with each other, are regular most of the time. But they both also show occasional anomalies, which appear as irregularities in the graph. The following is an explanation of what the curves in the graph represent from the top curve to the bottom curve.

- The top curve, which is red, represents network utilization over time. It follows a cyclical pattern and is regular most of the time, except for two anomalous periods, each representing a drop in utilization. The first anomalous period occurs between time values 500 and 1,000. The second anomalous period occurs between time values 1,500 and 2,000.
- The second curve from the top (blue in color) is idle CPU over time. It follows a cyclical pattern and is regular most of the time, with the exception of two anomalous periods. The first anomalous period occurs between time values 1,000 and 1,500 and shows a drop in idle CPU time. The second anomalous period occurs between time values 1,500 and 2,000 and shows an increase in idle CPU time.
- The third curve from the top shows the anomaly score. At the beginning, there's a learning phase during which the anomaly score is 0. After the learning phase, there's steady noise in the curve, but the anomalies stand out.

  The first anomaly, which is marked in red on the black anomaly score curve, is more attributable to the network utilization data. The second anomaly, marked in blue, is more attributable to the CPU data. The red and blue markings are provided in this graph as a visual aide. They aren't

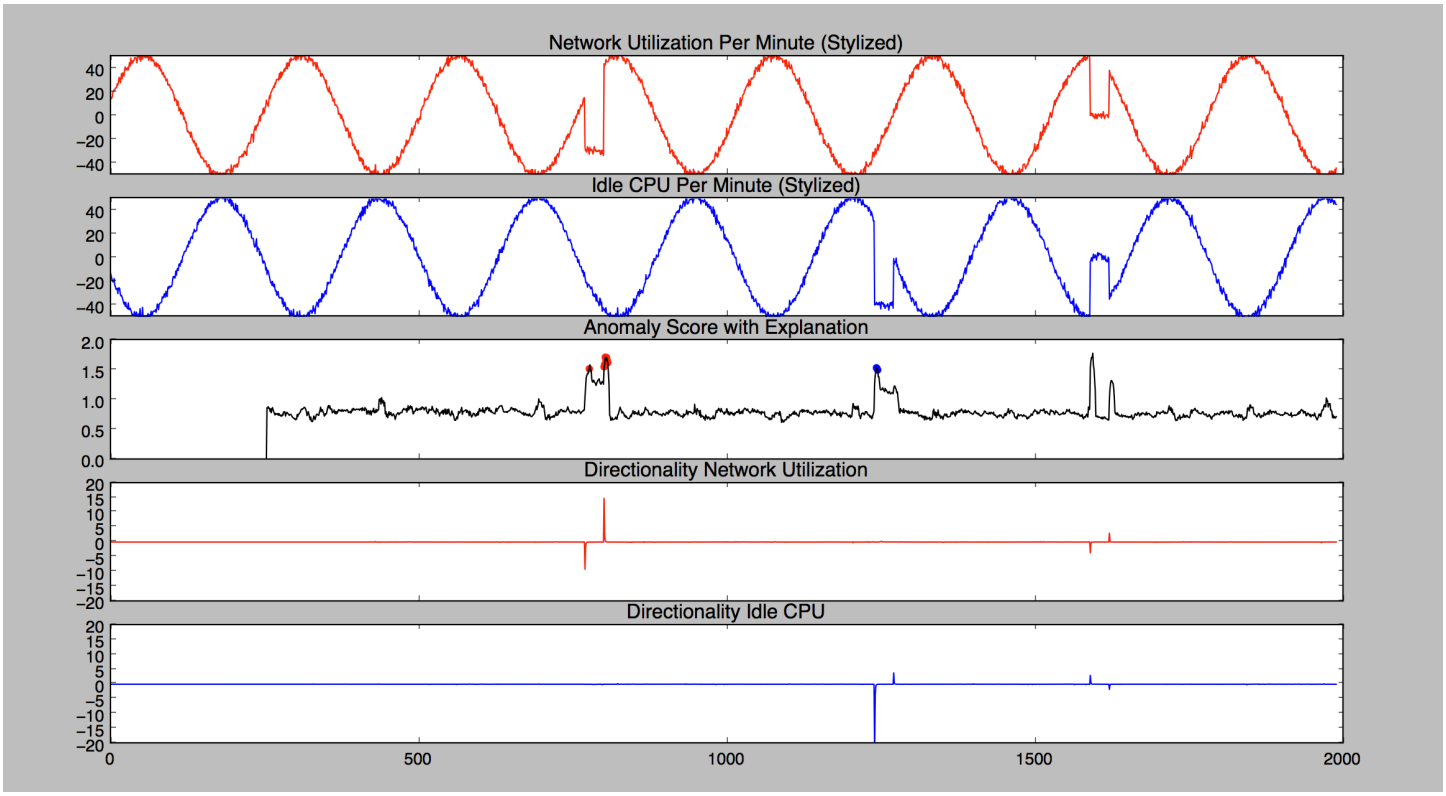
produced by the RANDOM_CUT_FOREST_WITH_EXPLANATION function. Here's how we obtained these red and blue markings:

- After running the function, we selected the top 20 anomaly score values.
- From this set of top 20 anomaly score values, we selected those values for which the network utilization had an attribution greater than or equal to 1.5 times the attribution for CPU. We colored the points in this new set with red markers in the graph.
- We colored with blue markers the points for which the CPU attribution score was greater than or equal to 1.5 times the attribution for network utilization.

- The second curve from the bottom is a graphical representation of directionality for the network utilization signal. We obtained this curve by running the function, multiplying the strength by -1 for LOW directionality and by +1 for HIGH directionality, and plotting the results against time.

  When there's a drop in the cyclical pattern of network utilization, there’s a corresponding negative spike in directionality. When network utilization shows an increase back to the regular pattern, directionality shows a positive spike corresponding to that increase. Later on, there's another negative spike, followed closely by another positive spike. Together they represent the second anomaly seen in the network utilization curve.
- The bottom curve is a graphical representation of directionality for the CPU signal. We obtained it by multiplying the strength by -1 for LOW directionality and by +1 for HIGH directionality, and then plotting the results against time.

  With the first anomaly in the idle CPU curve, this directionality curve shows a negative spike followed immediately by a smaller, positive spike. The second anomaly in the idle CPU curve produces a positive spike followed by a negative spike in directionality.



Blood Pressure Example

For a more detailed example, with code that detects and explains anomalies in blood pressure readings, see Example: Detecting Data Anomalies and Getting an Explanation.

STDDEV_POP

Returns the square root of the VAR_POP population variance for <number expression>, evaluated for each row remaining in the group.

When you use STDDEV_POP, be aware of the following:

- When the input set has no non-null data, STDDEV_POP returns NULL.
- If you don't use the OVER clause, STDDEV_POP is calculated as an aggregate function. In this case, the aggregate query must contain a GROUP BY clause on a monotonic expression based on ROWTIME that groups the stream into finite rows. Otherwise, the group is the infinite stream, and the query will never complete and no rows will be emitted. For more information, see Aggregate Functions.
- A windowed query that uses a GROUP BY clause processes rows in a tumbling window. For more information, see Tumbling Windows (Aggregations Using GROUP BY).

- If you use the OVER clause, STDDEV_POP is calculated as an analytic function. For more information, see Analytic Functions.
- A windowed query that uses an OVER clause processes rows in a sliding window. For more information, see Sliding Windows

Syntax

```
STDDEV_POP ( [DISTINCT | ALL] number-expression )
```

Parameters

ALL

Includes duplicate values in the input set. ALL is the default.

DISTINCT

Excludes duplicate values in the input set.

Examples

Example Dataset

The examples following are based on the sample stock dataset that is part of the Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Analytics application that has the sample stock ticker input stream. To learn how to create an Analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

Example 1: Determine the standard deviation of the population in a column in a tumbling window query

The following example demonstrates how to use the STDDEV_POP function to determine the standard deviation of the values in a tumbling window of the PRICE column of the example dataset. DISTINCT is not specified, so duplicate values are included in the calculation.

Using STEP (Recommended)

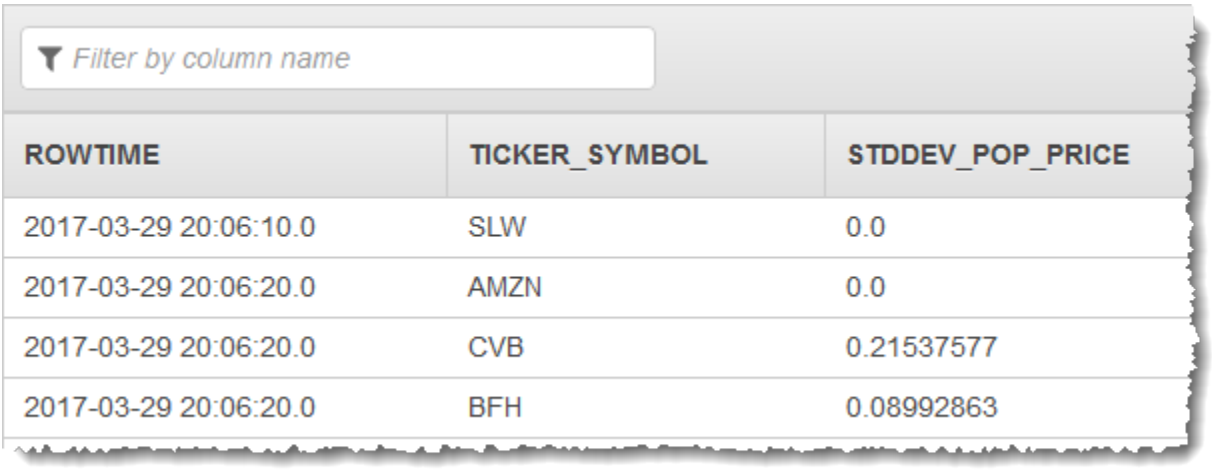
```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (ticker_symbol VARCHAR(4),
 stddev_pop_price REAL);

CREATE OR REPLACE  PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM ticker_symbol, STDDEV_POP(price) AS stddev_pop_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol, STEP(("SOURCE_SQL_STREAM_001".ROWTIME) BY INTERVAL '60'
 SECOND);
```

Using FLOOR

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (ticker_symbol VARCHAR(4),
 stddev_pop_price REAL);

CREATE OR REPLACE  PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM ticker_symbol, STDDEV_POP(price) AS stddev_pop_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol, FLOOR(("SOURCE_SQL_STREAM_001".ROWTIME - TIMESTAMP
 '1970-01-01 00:00:00') SECOND / 10 TO SECOND);
```

Results

The preceding examples output a stream similar to the following:

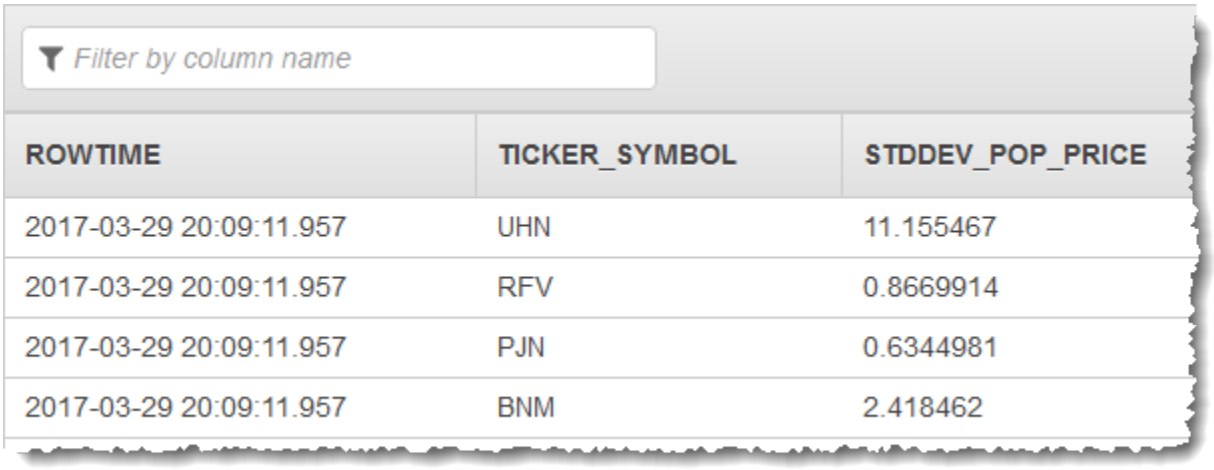


Example 2: Determine the standard deviation of the population of the values in a column in a sliding window query

The following example demonstrates how to use the STDDEV_POP function to determine the standard deviation of the values in a sliding window of the PRICE column of the example dataset. DISTINCT is not specified, so duplicate values are included in the calculation.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (ticker_symbol VARCHAR(4),
 stddev_pop_price REAL);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM ticker_symbol, STDDEV_POP(price) OVER TEN_SECOND_SLIDING_WINDOW AS
 stddev_pop_price
FROM "SOURCE_SQL_STREAM_001"

WINDOW TEN_SECOND_SLIDING_WINDOW AS (
  PARTITION BY ticker_symbol
  RANGE INTERVAL '10' SECOND PRECEDING);
```

The preceding example outputs a stream similar to the following:


Filter by column name

ROWTIME	TICKER_SYMBOL	STDDEV_POP_PRICE
2017-03-29 20:09:11.957	UHN	11.155467
2017-03-29 20:09:11.957	RFV	0.8669914
2017-03-29 20:09:11.957	PJN	0.6344981
2017-03-29 20:09:11.957	BNM	2.418462

See Also

- Sample standard deviation: STDDEV_SAMP
- Sample variance: VAR_SAMP
- Population variance: VAR_POP

STDDEV_SAMP

Returns the statistical standard deviation of all values in <number-expression>, evaluated for each row remaining in the group and defined as the square root of the VAR_SAMP.

When you use STDDEV_SAMP, be aware of the following:

- When the input set has no non-null data, STDDEV_SAMP returns NULL.
- If you don't use the OVER clause, STDDEV_SAMP is calculated as an aggregate function. In this case, the aggregate query must contain a GROUP BY clause on a monotonic expression based on ROWTIME that groups the stream into finite rows. Otherwise, the group is the infinite stream, and the query will never complete and no rows will be emitted. For more information, see Aggregate Functions.
- A windowed query that uses a GROUP BY clause processes rows in a tumbling window. For more information, see Tumbling Windows (Aggregations Using GROUP BY).
- If you use the OVER clause, STDDEV_SAMP is calculated as an analytic function. For more information, see Analytic Functions.
- A windowed query that uses an OVER clause processes rows in a sliding window. For more information, see Sliding Windows

- STD_DEV is an alias of STDDEV_SAMP.

Syntax

```
STDDEV_SAMP ( [DISTINCT | ALL] number-expression )
```

Parameters

ALL

Includes duplicate values in the input set. ALL is the default.

DISTINCT

Excludes duplicate values in the input set.

Examples

Example Dataset

The examples following are based on the sample stock dataset that is part of the Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Analytics application that has the sample stock ticker input stream. To learn how to create an Analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

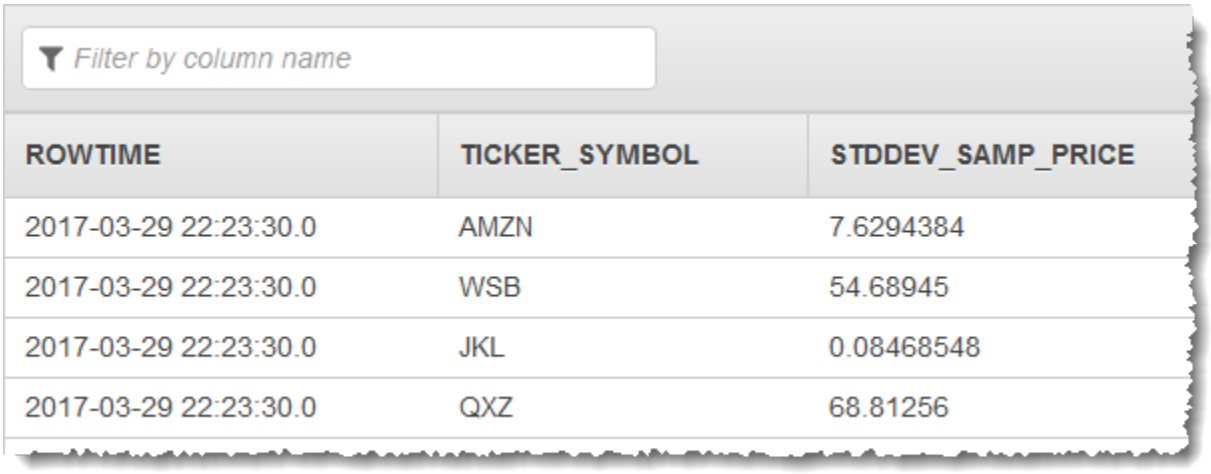
Example 1: Determine the statistical standard deviation of the values in a column in a tumbling window query

The following example demonstrates how to use the STDDEV_SAMP function to determine the standard deviation of the values in a tumbling window of the PRICE column of the example dataset. DISTINCT is not specified, so duplicate values are included in the calculation.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (ticker_symbol VARCHAR(4),
 stddev_samp_price REAL);

CREATE OR REPLACE  PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM ticker_symbol, STDDEV_SAMP(price) AS stddev_samp_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol, FLOOR(("SOURCE_SQL_STREAM_001".ROWTIME - TIMESTAMP
 '1970-01-01 00:00:00') SECOND / 10 TO SECOND);
```

Results

The preceding examples output a stream similar to the following:



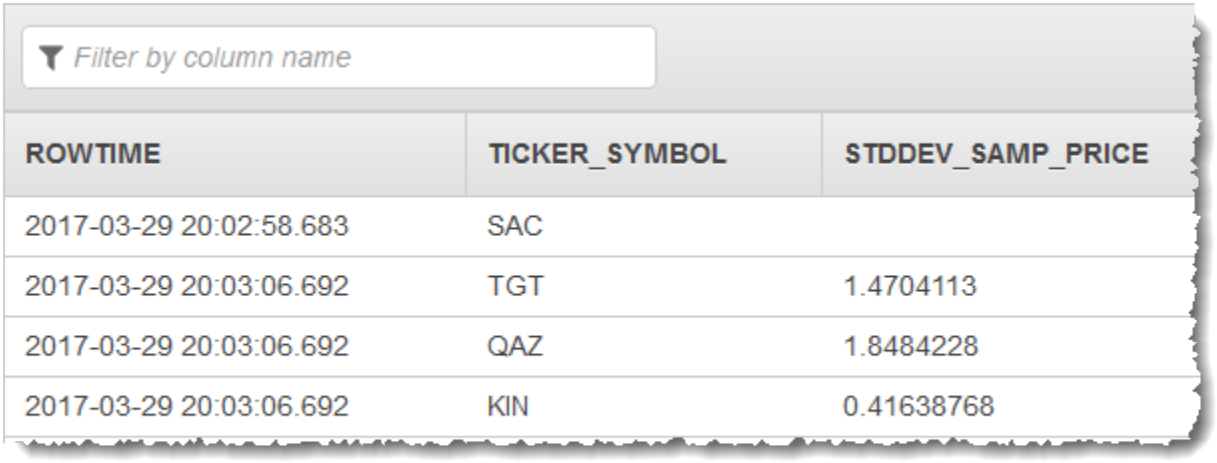
Example 2: Determine the statistical standard deviation of the values in a columm in a sliding window query

The following example demonstrates how to use the STDDEV_SAMP function to determine the standard deviation of the values in a sliding window of the PRICE column of the example dataset. DISTINCT is not specified, so duplicate values are included in the calculation.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (ticker_symbol VARCHAR(4),
 stddev_samp_price REAL);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM ticker_symbol, STDDEV_SAMP(price) OVER TEN_SECOND_SLIDING_WINDOW AS
 stddev_samp_price
FROM "SOURCE_SQL_STREAM_001"
```

```
WINDOW TEN_SECOND_SLIDING_WINDOW AS (
  PARTITION BY ticker_symbol
  RANGE INTERVAL '10' SECOND PRECEDING);
```

The preceding example outputs a stream similar to the following:



See Also

- Population standard deviation: STDDEV_POP
- Sample variance: VAR_SAMP
- Population variance: VAR_POP

VAR_POP

Returns the population variance of a non-null set of numbers (nulls being ignored)

VAR_POP uses the following calculation:

- (SUM(expr*expr) - SUM(expr)*SUM(expr) / COUNT(expr)) / COUNT(expr)

In other words, for a given set of non-null values, using S1 as the sum of the values and S2 as the sum of the squares of the values, VAR_POP returns the result (S2-S1*S1/N)/N.

When you use VAR_POP, be aware of the following:

- When the input set has no non-null data, or when applied to an empty set, VAR_POP returns NULL.
- If you don't use the OVER clause, VAR_POP is calculated as an aggregate function. In this case, the aggregate query must contain a GROUP BY clause on a monotonic expression based on ROWTIME that groups the stream into finite rows. Otherwise, the group is the infinite stream, and the query will never complete and no rows will be emitted. For more information, see Aggregate Functions.
- A windowed query that uses a GROUP BY clause processes rows in a tumbling window. For more information, see Tumbling Windows (Aggregations Using GROUP BY).
- If you use the OVER clause, VAR_POP is calculated as an analytic function. For more information, see Analytic Functions.
- A windowed query that uses an OVER clause processes rows in a sliding window. For more information, see Sliding Windows

Syntax

```
VAR_POP ( [DISTINCT | ALL] number-expression )
```

Parameters

ALL

Includes duplicate values in the input set. ALL is the default.

DISTINCT

Excludes duplicate values in the input set.

Examples

Example Dataset

The examples following are based on the sample stock dataset that is part of the Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Analytics application that has the sample stock ticker input stream. To learn how to create an Analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

Example 1: Determine the population variance in a column in a tumbling window query

The following example demonstrates how to use the VARPOP function to determine the population variance of the values in a tumbling window of the PRICE column of the example dataset. DISTINCT is not specified, so duplicate values are included in the calculation.

Using STEP (Recommended)

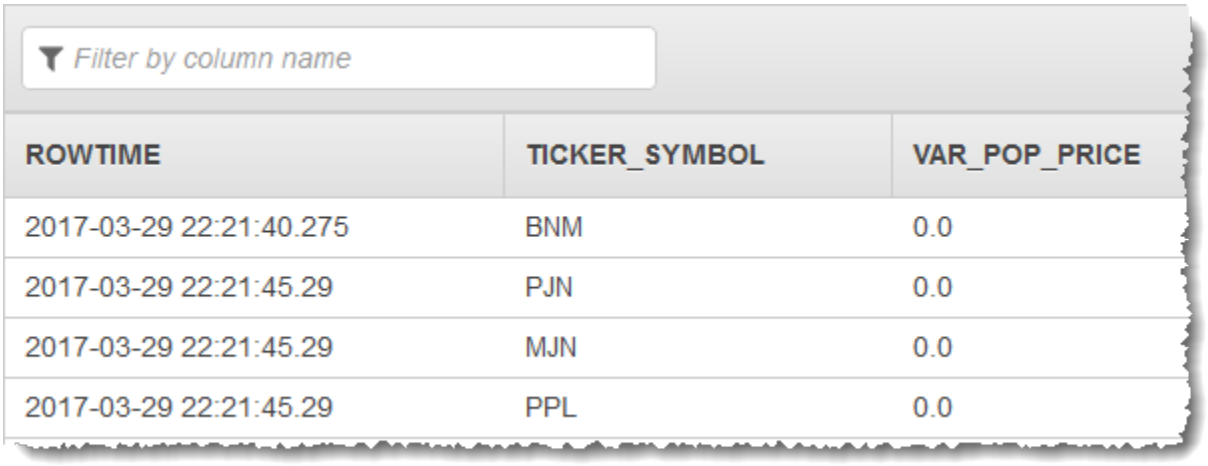
```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (ticker_symbol VARCHAR(4),
 var_pop_price REAL);

CREATE OR REPLACE  PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM ticker_symbol, VAR_POP(price) AS var_pop_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol, STEP(("SOURCE_SQL_STREAM_001".ROWTIME) BY INTERVAL '60'
 SECOND);
```

Using FLOOR

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (ticker_symbol VARCHAR(4),
 var_pop_price REAL);

CREATE OR REPLACE  PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM ticker_symbol, VAR_POP(price) AS var_pop_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol, FLOOR(("SOURCE_SQL_STREAM_001".ROWTIME - TIMESTAMP
 '1970-01-01 00:00:00') SECOND / 10 TO SECOND);
```

Results

The preceding examples output a stream similar to the following:

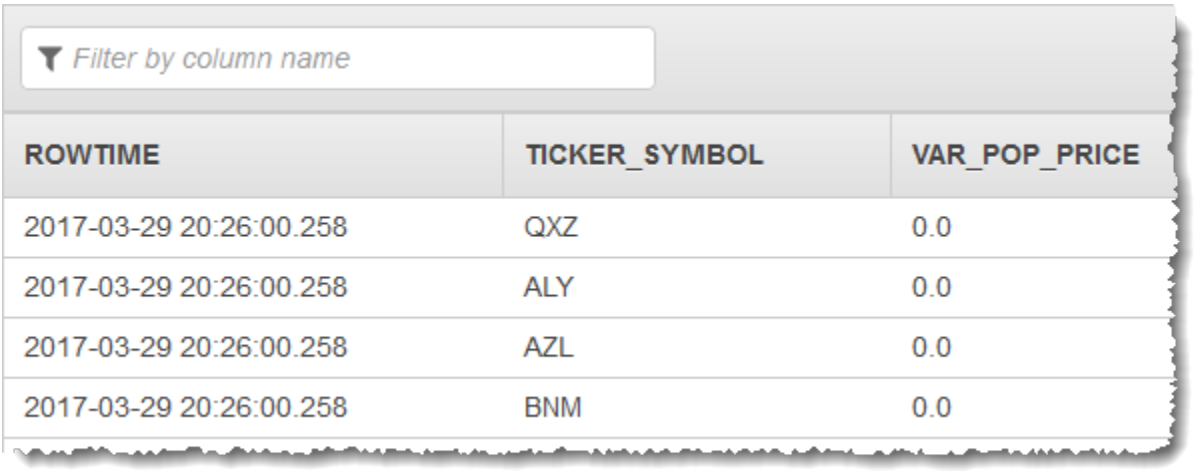


Example 2: Determine the population variance of the values in a columm in a sliding window query

The following example demonstrates how to use the VARPOP function to determine the population variance of the values in a sliding window of the PRICE column of the example dataset. DISTINCT is not specified, so duplicate values are included in the calculation.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (ticker_symbol VARCHAR(4),
 var_pop_price REAL);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM ticker_symbol, VAR_POP(price) OVER TEN_SECOND_SLIDING_WINDOW AS
 var_pop_price
FROM "SOURCE_SQL_STREAM_001"

WINDOW TEN_SECOND_SLIDING_WINDOW AS (
  PARTITION BY ticker_symbol
  RANGE INTERVAL '10' SECOND PRECEDING);
```

The preceding example outputs a stream similar to the following:



See Also

- Population standard deviation: STDDEV_POP
- Sample standard deviation: STDDEV_SAMP
- Sample variance: VAR_SAMP

VAR_SAMP

Returns the sample variance of a non-null set of numbers (nulls being ignored).

VAR_SAMP uses the following calculation:

- (SUM(expr*expr) - SUM(expr)*SUM(expr) / COUNT(expr)) / (COUNT(expr)-1)

In other words, for a given set of non-null values, using S1 as the sum of the values and S2 as the sum of the squares of the values, `VAR_SAMP` returns the result (S2-S1*S1/N)/(N-1).

When you use `VAR_SAMP`, be aware of the following:

- When the input set has no non-null data, `VAR_SAMP` returns `NULL`. Given an input set of null or one element, `VAR_SAMP` returns `null`.
- If you don't use the `OVER` clause, `VAR_SAMP` is calculated as an aggregate function. In this case, the aggregate query must contain a GROUP BY clause on a monotonic expression based on `ROWTIME` that groups the stream into finite rows. Otherwise, the group is the infinite stream, and the query will never complete and no rows will be emitted. For more information, see Aggregate Functions.

- A windowed query that uses a GROUP BY clause processes rows in a tumbling window. For more information, see Tumbling Windows (Aggregations Using GROUP BY).
- If you use the OVER clause, VAR_SAMP is calculated as an analytic function. For more information, see Analytic Functions.
- A windowed query that uses an OVER clause processes rows in a sliding window. For more information, see Sliding Windows

Syntax

```
VAR_SAMP ( [DISTINCT | ALL] number-expression )
```

Parameters

ALL

Includes duplicate values in the input set. ALL is the default.

DISTINCT

Excludes duplicate values in the input set.

Examples

Example Dataset

The examples following are based on the sample stock dataset that is part of the Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Analytics application that has the sample stock ticker input stream. To learn how to create an Analytics application and configure the sample stock ticker input stream, see Getting Started in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

Example 1: Determine the sample variance in a column in a tumbling window query

The following example demonstrates how to use the VAR_SAMP function to determine the sample variance of the values in a tumbling window of the PRICE column of the example dataset. DISTINCT is not specified, so duplicate values are included in the calculation.

Using STEP (Recommended)

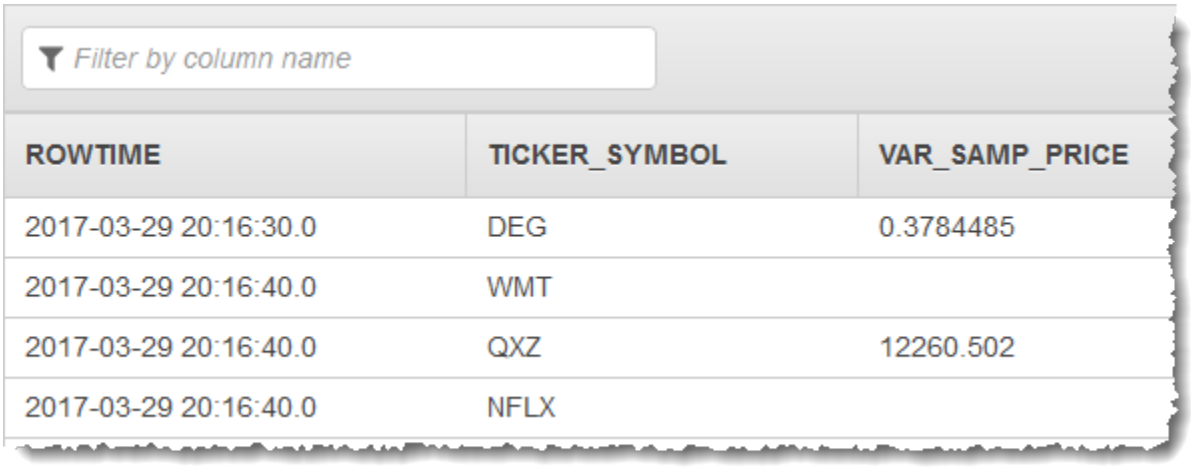
```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (ticker_symbol VARCHAR(4),
 var_samp_price REAL);

CREATE OR REPLACE  PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM ticker_symbol, VAR_SAMP(price) AS var_samp_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol, STEP(("SOURCE_SQL_STREAM_001".ROWTIME) BY INTERVAL '60'
 SECOND);
```

Using FLOOR

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (ticker_symbol VARCHAR(4),
 var_samp_price REAL);

CREATE OR REPLACE  PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM ticker_symbol, VAR_SAMP(price) AS var_samp_price
    FROM "SOURCE_SQL_STREAM_001"
    GROUP BY ticker_symbol, FLOOR(("SOURCE_SQL_STREAM_001".ROWTIME - TIMESTAMP
 '1970-01-01 00:00:00') SECOND / 10 TO SECOND);
```

Results

The preceding examples output a stream similar to the following:



Filter by column name

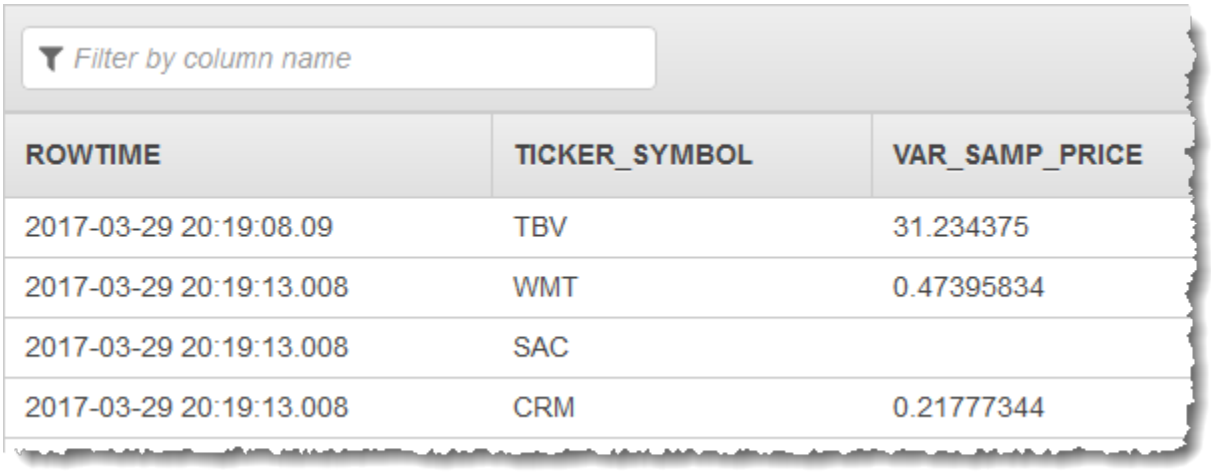
ROWTIME	TICKER_SYMBOL	VAR_SAMP_PRICE
2017-03-29 20:16:30.0	DEG	0.3784485
2017-03-29 20:16:40.0	WMT	
2017-03-29 20:16:40.0	QXZ	12260.502
2017-03-29 20:16:40.0	NFLX	

Example 2: Determine the sample variance of the values in a column in a sliding window query

The following example demonstrates how to use the VAR_SAMP function to determine the sample variance of the values in a sliding window of the PRICE column of the example dataset. DISTINCT is not specified, so duplicate values are included in the calculation.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (ticker_symbol VARCHAR(4),
 var_samp_price REAL);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM ticker_symbol, VAR_SAMP(price) OVER TEN_SECOND_SLIDING_WINDOW AS
 var_samp_price
FROM "SOURCE_SQL_STREAM_001"

WINDOW TEN_SECOND_SLIDING_WINDOW AS (
  PARTITION BY ticker_symbol
  RANGE INTERVAL '10' SECOND PRECEDING);
```

The preceding example outputs a stream similar to the following:



See Also

- Population standard deviation: STDDEV_POP
- Sample standard deviation: STDDEV_SAMP
- Population variance:VAR_POP

Streaming SQL Functions

The topics in this section describe the streaming functions for Amazon Kinesis Data Analytics streaming SQL.

Topics

- LAG
- Monotonic Function
- NTH_VALUE

LAG

LAG returns the evaluation of the expression (such as the name of a column) for the record that is N records before the current record in a given window. Both offset and default are evaluated with respect to the current record. If there is no such record, LAG instead returns a specified default expression. LAG returns a value of the same type as the expression.

Syntax

```
LAG(expr [ , N [ , defaultExpr]]) [ IGNORE NULLS | RESPECT NULLS ] OVER [ window-
definition ]
```

Parameters

expr

An expression that is evaluated on a record.

N

The number of records before the current record to query. The default is 1.

defaultExpr

An expression of the same type as *expr* that is returned if the record queried (*n* before the current record) falls outside the window. If not specified, *null* is returned for values that fall outside the window.

> **Note**
>
> The *defaultExpr* expression doesn't replace actual *null* values returned from the source stream.

IGNORE NULLS

A clause that specifies that null values are not counted when determining the offset. For example, suppose that `LAG(expr, 1)` is queried, and the previous record has a null value for *expr*. Then the second record previous is queried, and so on.

RESPECT NULLS

A clause that specifies that null values are counted when determining the offset. This behavior is the default.

OVER *window-specification*

A clause that divides records in a stream partitioned by the time range interval or the number of records. A window specification defines how records in the stream are partitioned, whether by the time range interval or the number of records.

Example

Example Dataset

The examples following are based on the sample stock dataset that is part of Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*. To run each example, you need an Amazon Kinesis Analytics application that has the input stream for the sample stock ticker. To learn how to create an Analytics application and configure the input stream for the sample stock ticker, see Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

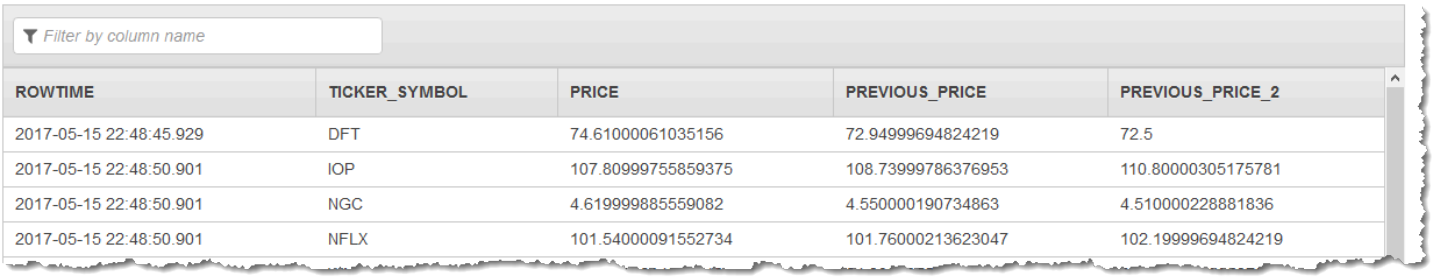
Example 1: Return Values from Previous Records in an OVER Clause

In this example, the OVER clause divides records in a stream partitioned by the time range interval of '1' minute preceding. The LAG function then retrieves price values from the two previous records that contain the given ticker symbol, skipping records if `price` is null.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
    ticker_symbol VARCHAR(4),
    price     DOUBLE,
    previous_price DOUBLE,
    previous_price_2 DOUBLE);
CREATE OR REPLACE PUMP "STREAM_PUMP" AS
    INSERT INTO "DESTINATION_SQL_STREAM"
    SELECT STREAM ticker_symbol,
          price,
          LAG(price, 1, 0) IGNORE NULLS OVER (
             PARTITION BY ticker_symbol
             RANGE INTERVAL '1' MINUTE PRECEDING),
          LAG(price, 2, 0) IGNORE NULLS OVER (
```

```
            PARTITION BY ticker_symbol
            RANGE INTERVAL '1' MINUTE PRECEDING)
    FROM "SOURCE_SQL_STREAM_001"
```

The preceding example outputs a stream similar to the following.



Filter by column name

ROWTIME	TICKER_SYMBOL	PRICE	PREVIOUS_PRICE	PREVIOUS_PRICE_2
2017-05-15 22:48:45.929	DFT	74.61000061035156	72.94999694824219	72.5
2017-05-15 22:48:50.901	IOP	107.80999755859375	108.73999786376953	110.80000305175781
2017-05-15 22:48:50.901	NGC	4.619999885559082	4.550000190734863	4.510000228881836
2017-05-15 22:48:50.901	NFLX	101.54000091552734	101.76000213623047	102.19999694824219

Notes

LAG is not part of the SQL:2008 standard. It is an Amazon Kinesis Data Analytics streaming SQL extension.

Monotonic Function

```
MONOTONIC(<expression>)
```

Streaming GROUP BY requires that at least one of the grouped expressions be monotonic and non-constant. The only column known in advance to be monotonic is ROWTIME. For more information, see Monotonic Expressions and Operators.

The MONOTONIC function allows you to declare that a given expression is monotonic, enabling a streaming GROUP BY to use that expression as a key.

The MONOTONIC function evaluates its argument and returns the result (as the same type as its argument).

By enclosing an expression in MONOTONIC, you are asserting that values of that expression are either non-increasing or non-decreasing and never change direction. For example, if you have a stream LINEITEMS consisting of the line items of orders, and you wrote MONOTONIC(orderId), you are asserting that line items are consecutive in the stream. It would be OK if there were line items for order 1000, followed by line items for order 1001, followed by line items for order 1005. It would be illegal if there were then a line item for order 1001 (that is, the line item sequence became 1000, 1001, 1005, 1001). Similarly, a line item sequence of 987, 974, 823 would be legal, but the following line item sequences would be illegal:

- 987, 974, 823, 973
- 987, 974, 823, 1056

An expression declared monotonic can decrease, or even have arbitrary order.

Note that the definition of MONOTONIC is precisely what is needed for GROUP BY to make progress.

If an expression declared monotonic is not monotonic (that is, if the assertion is not valid for the actual data) then Amazon Kinesis Data Analytics behavior is unspecified.

In other words, if you are certain that an expression is monotonic, you can use this MONOTONIC function to enable Amazon Kinesis Data Analytics to treat the expression as monotonic.

However, if you are mistaken and the values resulting from evaluating the expression change from ascending to descending or from descending to ascending, unexpected results may arise. Amazon Kinesis Data Analytics streaming SQL will take you at your word and operate on your assurance that the expression is monotonic. But if in fact it is not monotonic, the resulting Amazon Kinesis Data Analytics behavior cannot be determined in advance, and so results may not be as expected or desired.

NTH_VALUE

```
NTH_VALUE(x, n) [ <from first or last> ] [ <null treatment> ] over w
```

Where:

<null treatment> := RESPECT NULLS | IGNORE NULL

<from first or last> := FROM FIRST | FROM LAST

NTH_VALUE returns the nth value of x from the first or last value in the window. Default is first. If <null treatment> is set to IGNORE NULLS, then function will skip over nulls while counting.

If there aren't enough rows in the window to reach nth value, the function returns NULL.

String and Search Functions

The topics in this section describe the string and search functions for Amazon Kinesis Data Analytics streaming SQL.

Topics

- CHAR_LENGTH / CHARACTER_LENGTH
- INITCAP
- LOWER
- OVERLAY
- POSITION
- REGEX_REPLACE
- SUBSTRING
- TRIM
- UPPER

CHAR_LENGTH / CHARACTER_LENGTH

```
CHAR_LENGTH | CHARACTER_LENGTH ( <character-expression> )
```

Returns the length in characters of the string passed as the input argument. Returns null if input argument is null.

Examples

`CHAR_LENGTH('one')`	3
`CHAR_LENGTH('')`	0
`CHARACTER_LENGTH('fred')`	4
`CHARACTER_LENGTH( cast (null as varchar(16) )`	null
`CHARACTER_LENGTH( cast ('fred' as char(16) )`	16

Limitations

Amazon Kinesis Data Analytics streaming SQL does not support the optional USING CHARACTERS | OCTETS clause. This is a departure from the SQL:2008 standard.

INITCAP

```
INITCAP ( <character-expression> )
```

Returns a converted version of the input string such that the first character of each space-delimited word is upper-cased, and all other characters are lower-cased.

Examples

Function	Result
INITCAP('Each FIRST lEtTeR is cAPITALIZED')	Each First Letter Is Capitalized

> **Note**
>
> The INITCAP function is not part of the SQL:2008 standard. It is an Amazon Kinesis Data Analytics extension.

LOWER

```
LOWER ( <character-expression> )
```

Converts a string to all lower-case characters. Returns null if input argument is null, and the empty string if the input argument is an empty string.

Examples

Function	Result
LOWER('abcDEFghi123')	abcdefghi123

OVERLAY

```
OVERLAY ( <original-string>
          PLACING <replacement-string>
          FROM <start-position>
          [ FOR <string-length> ]
        )
<original-string> := <character-expression>
<replacement-string> := <character-expression>
<start-position> := <integer-expression>
<string-length> := <integer-expression>
```

The OVERLAY function is used to replace a portion of the first string argument (the original string) with the second string argument (the replacement string).

The start position indicates the character position in the original string where the replacement string should be overlaid. The optional string length parameter determines how many characters of the original string to replace (if not specified, it defaults to the length of the replacement string). If there are more characters in the replacement string than are left in the original string, the remaining characters are simply appended.

If the start position is greater than the length of the original string, the replacement string is simply appended. If the start position is less than 1, then (1 - start position) characters of the replacement string is prepended to the result, and the rest overlaid on the original (see examples below).

If the string length is less than zero, an exception is raised.

If any of the input arguments are null, the result is null.

Examples

Function	Result
OVERLAY ('12345' PLACING 'foo' FROM 1)	foo45
OVERLAY ('12345' PLACING 'foo' FROM 0)	foo345
OVERLAY ('12345' PLACING 'foo' FROM -2)	foo12345

Function	Result
OVERLAY ('12345' PLACING 'foo' FROM 4)	123foo
OVERLAY ('12345' PLACING 'foo' FROM 17)	12345foo
OVERLAY ('12345' PLACING 'foo' FROM 2 FOR 0)	1foo2345
OVERLAY ('12345' PLACING 'foo' FROM 2 FOR 2)	1foo45
OVERLAY ('12345' PLACING 'foo' FROM 2 FOR 9)	1foo

Limitations

Amazon Kinesis Data Analytics does not support the optional USING CHARACTERS | OCTETS clause defined in SQL:2008; USING CHARACTERS is simply assumed. Strict SQL:2008 also requires that a start position less than 1 return a null result, rather than the behavior described above. These are departures from the standard.

POSITION

```
POSITION ( <search-string> IN <source-string> )
search-string := <character-expression>
source-string := <character-expression>
```

The POSITION function searches for the first input argument (the search string) within the second input argument (the source string).

If the search string is found within the source string, POSITION returns the character position of the first instance of the search string (subsequent instances are ignored). If the search string is the empty string, POSITION returns 1.

If the search string is not found, POSITION returns 0.

If either the search string or the source string is null, POSITION returns null.

Examples

Function	Result
POSITION ('findme' IN '1234findmeXXX')	5
POSITION ('findme' IN '1234not-hereXXX')	0
POSITION ('1' IN '1234567')	1
POSITION ('7' IN '1234567')	7
POSITION ('' IN '1234567')	1

Limitations

Amazon Kinesis Data Analytics streaming SQL does not support the optional USING CHARACTERS | OCTETS clause defined in SQL:2008; USING CHARACTERS is simply assumed. This is a departure from the standard.

REGEX_REPLACE

REGEX_REPLACE replaces a substring with an alternative substring. It returns the value of the following Java expression.

```
java.lang.String.replaceAll(regex, replacement)
```

Syntax

```
REGEX_REPLACE(original VARCHAR(65535), regex VARCHAR(65535), replacement
 VARCHAR(65535), startPosition int, occurence int)

RETURNS VARCHAR(65535)
```

Parameters

original

The string on which to execute the regex operation.

regex

The regular expression to match. If the encoding for *regex* doesn't match the encoding for *original*, an error is written to the error stream.

replacement

The string to replace *regex* matches in the *original* string. If the encoding for *replacement* doesn't match the encoding for *original* or *regex*, an error is written to the error stream.

startPosition

The first character in the *original* string to search. If *startPosition* is less than 1, an error is written to the error stream. If *startPosition* is greater than the length of *original*, then *original* is returned.

occurence

The occurrence of the string that matches the *regex* expression to replace. If *occurence* is 0, all substrings matching *regex* are replaced. If *occurence* is less than 0, an error is written to the error stream.

Example

Example Dataset

The examples following are based on the sample stock dataset that is part of Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*.

To run each example, you need an Amazon Kinesis Analytics application that has the input stream for the sample stock ticker. To learn how to create an Analytics application and configure the input stream for the sample stock ticker, see Getting Started Exercise in the *Amazon Kinesis Analytics Developer Guide*.

The sample stock dataset has the schema following.

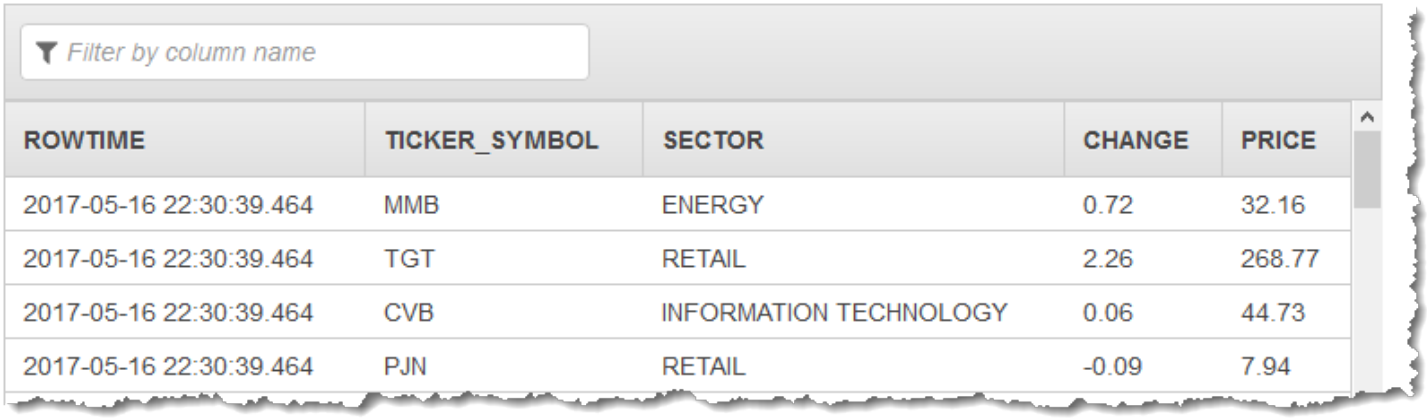
```
(ticker_symbol  VARCHAR(4),
sector          VARCHAR(16),
change          REAL,
price           REAL)
```

Example 1: Replace All String Values in a Source String with a New Value

In this example, all character strings in the `sector` field are replaced if they match a regular expression.

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
        ticker_symbol VARCHAR(4),
        SECTOR VARCHAR(24),
        CHANGE REAL,
        PRICE REAL);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM   TICKER_SYMBOL,
                REGEX_REPLACE(SECTOR, 'TECHNOLOGY', 'INFORMATION TECHNOLOGY', 1, 0);
                CHANGE,
                PRICE
FROM "SOURCE_SQL_STREAM_001"
```

The preceding example outputs a stream similar to the following.



Notes

REGEX_REPLACE is not part of the SQL:2008 standard. It is an Amazon Kinesis Data Analytics streaming SQL extension.

REGEX_REPLACE returns `null` if any parameters are `null`.

SUBSTRING

```
SUBSTRING ( <source-string> FROM <start-position> [ FOR <string-length> ] )
SUBSTRING ( <source-string>, <start-position> [ , <string-length> ] )
```

```
SUBSTRING ( <source-string> SIMILAR <pattern> ESCAPE <escape-char> )
<source-string> := <character-expression>
<start-position> := <integer-expression>
<string-length> := <integer-expression>
<regex-expression> := <character-expression>
<pattern> := <character-expression>
<escape-char> := <character-expression>
```

SUBSTRING extracts a portion of the source string specified in the first argument. Extraction starts at the value of *start-position* or the first expression matching the value of *regex-expression*.

If a value is specified for *string-length*, only that number of characters is returned. If there aren't that many characters left in the string, only the characters that are left are returned. If *string-length* is not specified, the string length defaults to the remaining length of the input string.

If the start position is less than 1, then the start position is interpreted as if it is 1 and the string length is reduced by (1–start position). For examples, see following. If the start position is greater than the number of characters in the string, or the length parameter is 0, the result is an empty string.

Parameters

source-string

The string to search for positional or regular-expression matches.

start-position

The first character of *source-string* to return. If *start-position* is greater than the length of *source-string*, SUBSTRING returns null.

string-length

The number of characters from *source-string* to return.

regex-expression

A pattern of characters to match and return from *source-string*. Only the first match is returned.

pattern

A three-part pattern of characters that consists of the following:

- The string to be found before the returned substring

- The returned substring
- The string to be found after the returned substring

The parts are delimited by a double quotation mark (") and a specified escape character. For more information, see Similar...Escape Samples following.

Examples

FROM/ FOR

Function	Result
SUBSTRING('123456789' FROM 3 FOR 4)	3456
SUBSTRING('123456789' FROM 17 FOR 4)	<empty string>
SUBSTRING('123456789' FROM -1 FOR 4)	12
SUBSTRING('123456789' FROM 6 FOR 0)	<empty string>
SUBSTRING('123456789' FROM 8 FOR 4)	89

FROM Regex

Function	Result
SUBSTRING('TECHNOLOGY' FROM 'L[A-Z]*')	LOGY
SUBSTRING('TECHNOLOGY' FROM 'FOO')	null
SUBSTRING('TECHNOLOGY' FROM 'O[A-Z]')	OL

Numeric

Function	Result
SUBSTRING('123456789', 3, 4)	3456

Function	Result
SUBSTRING('123456789', 7, 4)	789
SUBSTRING('123456789', 10, 4)	null

Similar...Escape

Function	Result
SUBSTRING('123456789' SIMILAR '23#"456#"78' ESCAPE '#')	456
SUBSTRING('TECHNOLOGY' SIMILAR 'TECH%"NOLO%"GY' ESCAPE '%')	NOLO

Notes

- Amazon Kinesis Data Analytics streaming SQL doesn't support the optional 'USING CHARACTERS | OCTETS' clause defined in SQL:2008. USING CHARACTERS is simply assumed.
- The second and third forms of the SUBSTRING function listed preceding (using a regular expression, and using commas rather than FROM...FOR) are not part of the SQL:2008 standard. They are part of the streaming SQL extension to Amazon Kinesis Data Analytics.

TRIM

```
TRIM ( [ [ <trim-specification> ] [ <trim-character> ] FROM ] <trim-source> )
 <trim-specification> := LEADING | TRAILING | BOTH
 <trim-character> := <character-expression>
 <trim-source> := <character-expression>
```

TRIM removes instances of the specified trim-character from the beginning and/or end of the trim-source string as dictated by the trim-specification (that is, LEADING, TRAILING, or BOTH). If LEADING is specified, only repetitions of the trim character at the beginning of the source string are removed. If TRAILING is specified, only repetitions of the trim character at the end of the source

string are removed. If BOTH is specified, or the trim specifier is left out entirely, then repetitions are removed from both the beginning and end of the source string.

If the trim-character is not explicitly specified, it defaults to the space character (' '). Only one trim character is allowed; specifying an empty string or a string longer than one character results in an exception.

If either input is null, null is returned.

Examples

Function	Result
`TRIM(' Trim front and back ')`	`'Trim front and back'`
`TRIM (BOTH FROM ' Trim front and back ')`	`'Trim front and back'`
`TRIM (BOTH ' ' FROM ' Trim front and back ')`	`'Trim front and back'`
`TRIM (LEADING 'x' FROM 'xxxTrim frontxxx')`	`'Trim frontxxx'`
`TRIM (TRAILING 'x' FROM 'xxxTrimx Backxxx')`	`'xxxTrimxBack'`
`TRIM (BOTH 'y' FROM 'xxxNo y to trimxxx')`	`'xxxNo y to trimxxx'`

UPPER

```
< UPPER ( <character-expression> )
```

Converts a string to all upper-case characters. Returns null if the input argument is null, and the empty string if the input argument is an empty string.

Examples

Function	Result
UPPER('abcDEFghi123')	ABCDEFGHI123

Kinesis Data Analytics Developer Guide

For information about developing Kinesis Data Analytics applications, see the Kinesis Data Analytics Developer Guide.

Document History

The following table describes important changes to the documentation since the last release of the Amazon Kinesis Data Analytics SQL Reference.

- **API version: 2015-08-14**
- **Latest documentation update:** March 19, 2018

Change	Description	Date
New HOTSPOTS function	Locate and return informati on about relatively dense regions in your data. For more information, see HOTSPOTS.	March 19, 2018
Stream-to-stream JOIN examples	Examples for JOIN clause queries. For more informati on, see JOIN clause.	February 28, 2018
New TSDIFF function	Obtain the difference between two time stamps. For more information, see TSDIFF.	December 11, 2017
New RANDOM_CUT_FOREST_ WITH_EXPLANATION function	Get an explanation of what fields contribute to an anomaly score in a data stream. For more information, see RANDOM_CUT_FOREST_ WITH_EXPLANATION.	November 2, 2017
New REGEX_LOG_PARSE function	Get regex matches from a source string in columnar format. For more information, see REGEX_LOG_PARSE.	August 21, 2017

Change	Description	Date
Table of contents reorganization	Topic categories are now more intuitive.	August 18, 2017
New SQL functions	Addition of STEP, LAG, TO_TIMESTAMP, UNIX_TIMESTAMP, REGEX_REPLACE, and addition of regex support to SUBSTRING	August 3, 2017
New guide	This is the first release of the *Amazon Kinesis Data Analytics SQL Reference* guide.	August 11, 2016