aws

入门指南

Amazon Redshift

Amazon Redshift: 入门指南

Table of Contents

Amazon Redshift Serverless 数据仓库入门

如果您是首次接触 Amazon Redshift Serverless 的用户，我们建议您先阅读以下部分，以帮助您开始使用 Amazon Redshift Serverless。Amazon Redshift Serverless 的基本流程是创建无服务器资源，连接到 Amazon Redshift Serverless，加载示例数据，然后对数据运行查询。在本指南中，您可以选择从 Amazon Redshift Serverless 或 Amazon S3 存储桶加载示例数据。在 Amazon Redshift 文档中，会使用示例数据来演示功能。要开始使用 Amazon Redshift 预置数据仓库，请参阅 Amazon Redshift 预置数据仓库入门。

- the section called “注册 AWS”
- the section called “使用 Amazon Redshift Serverless 创建数据仓库”
- the section called “从 Amazon S3 加载数据”

注册 AWS

如果您还没有 AWS 账户，请先注册一个。如果您已有账户，则可以跳过此先决条件步骤，并使用您已有的账户。

1. 打开 https://portal.aws.amazon.com/billing/signup。
2. 按照屏幕上的说明操作。

   当您注册 AWS 账户时，系统会创建一个 AWS 账户根用户。根用户有权访问该账户中的所有 AWS 服务和资源。作为安全最佳实践，请为管理用户分配管理访问权限，并且只使用根用户执行需要根用户访问权限的任务。

使用 Amazon Redshift Serverless 创建数据仓库

首次登录 Amazon Redshift Serverless 控制台时，系统会提示您访问入门体验，您可以通过该体验创建和管理无服务器资源。在本指南中，您将使用 Amazon Redshift Serverless 的默认设置创建无服务器资源。

要更精细地控制您的设置，请选择 Customize settings（自定义设置）。

> **Note**
>
> Redshift Serverless 需要一个 Amazon VPC，且该 VPC 有三个子网位于三个不同的可用区中。Redshift Serverless 还需要至少 3 个可用 IP 地址。在继续操作之前，请确保您用于 Redshift Serverless 的 Amazon VPC 有三个子网位于三个不同的可用区中，并且至少有 3 个可用 IP 地址。有关在 Amazon VPC 中创建子网的更多信息，请参阅《Amazon Virtual Private Cloud 用户指南》中的创建子网。有关 Amazon VPC 中的 IP 地址的更多信息，请参阅为 VPC 和子网分配 IP 地址。

要使用原定设置进行配置，请执行以下操作：

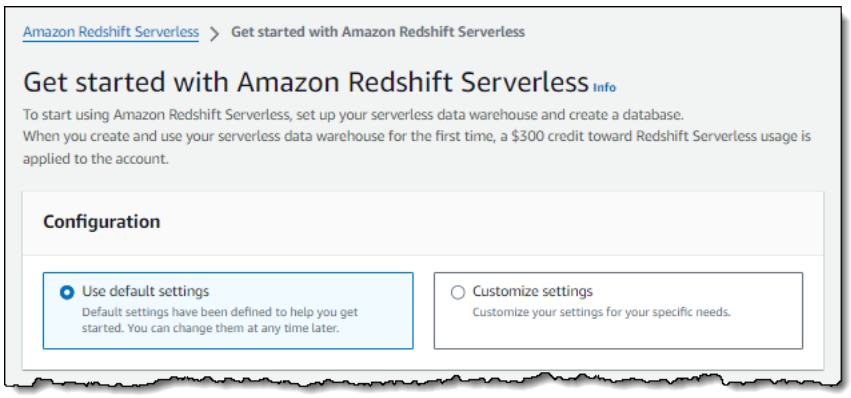
1. 登录到 AWS Management Console并打开 Amazon Redshift 控制台，网址：https://console.aws.amazon.com/redshiftv2/。

   选择试用 Redshift Serverless 免费试用版。

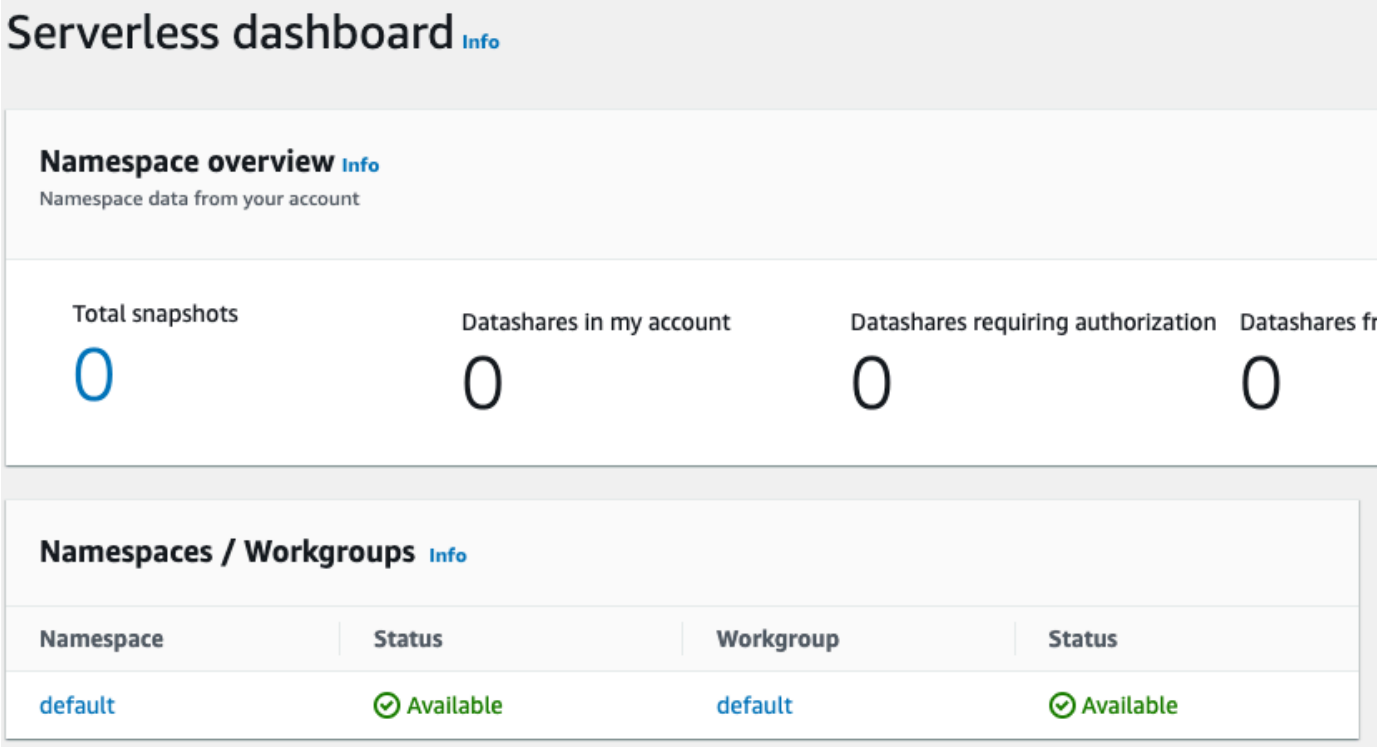
2. 在 Configuration（配置）下，选择 Use default settings（使用原定设置）。Amazon Redshift Serverless 将创建一个默认命名空间，其中包含与该命名空间关联的默认工作组。选择 Save configuration。

   > **Note**
   >
   > 命名空间是数据库对象和用户的集合。命名空间将您在 Redshift Serverless 中使用的所有资源组合在一起，例如架构、表、用户、数据共享和快照。
   > 工作组是计算资源的集合。工作组存放 Redshift Serverless 运行计算任务所用的计算资源。

   以下屏幕截图显示 Amazon Redshift Serverless 的默认设置。



3. 设置完成后，选择 Continue（继续）以转到 Serverless dashboard（Serverless 控制面板）。您可以看到无服务器工作组和命名空间可用。



> **Note**
>
> 如果 Redshift Serverless 未能成功创建工作组，您可以执行以下操作：
>
> - 解决 Redshift Serverless 报告的任何错误，例如 Amazon VPC 中的子网过少。
> - 通过在 Redshift Serverless 控制面板中选择 default-namespace，然后选择操作、删除命名空间，来删除命名空间。删除命名空间需要几分钟时间。
> - 再次打开 Redshift Serverless 控制台时，会出现欢迎屏幕。

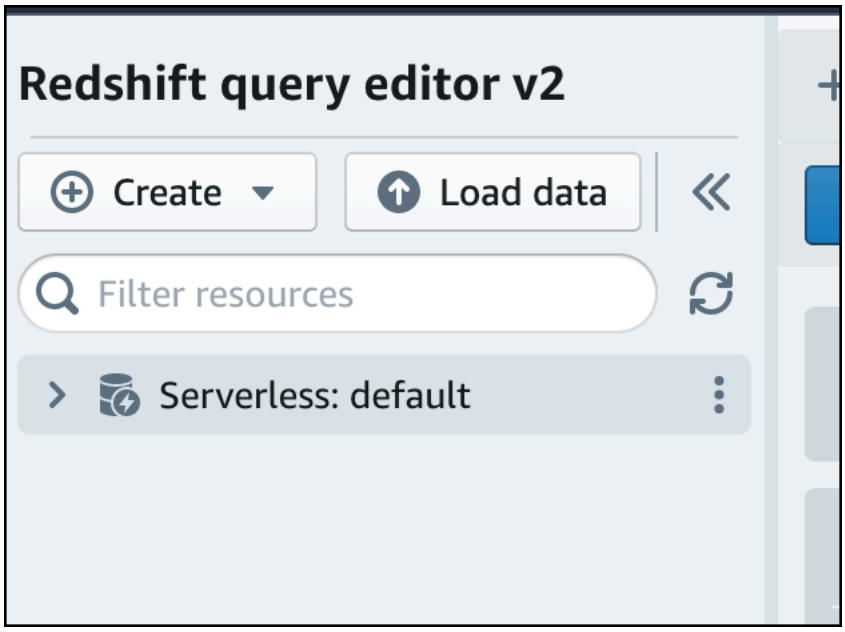
加载示例数据

现在，使用 Amazon Redshift Serverless 设置了数据仓库之后，您可以使用 Amazon Redshift 查询编辑器 v2 来加载示例数据。

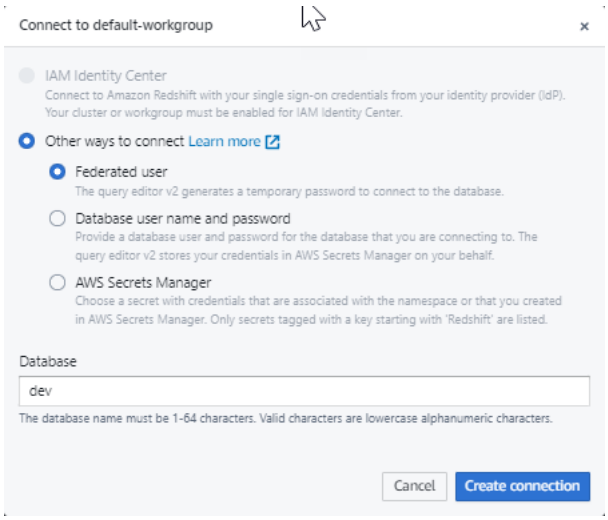
1. 要从 Amazon Redshift Serverless 控制台启动查询编辑器 v2，请选择查询数据。当您从 Amazon Redshift Serverless 控制台调用查询编辑器 v2 时，将打开一个新的浏览器选项卡，其中包含查询编辑器。查询编辑器 v2 将从客户端计算机连接到 Amazon Redshift Serverless 环境。

2. 对于本指南，您将使用您的 AWS 管理员账户和默认的 AWS KMS key。有关配置查询编辑器 v2 的信息，包括需要哪些权限，请参阅《Amazon Redshift 管理指南》中的配置您的 AWS 账户。有关将 Amazon Redshift 配置为使用客户自主管理型密钥或更改 Amazon Redshift 使用的 KMS 密钥的信息，请参阅更改命名空间的 AWS KMS 密钥。

3. 要连接到工作组，请在树视图面板中选择工作组名称。

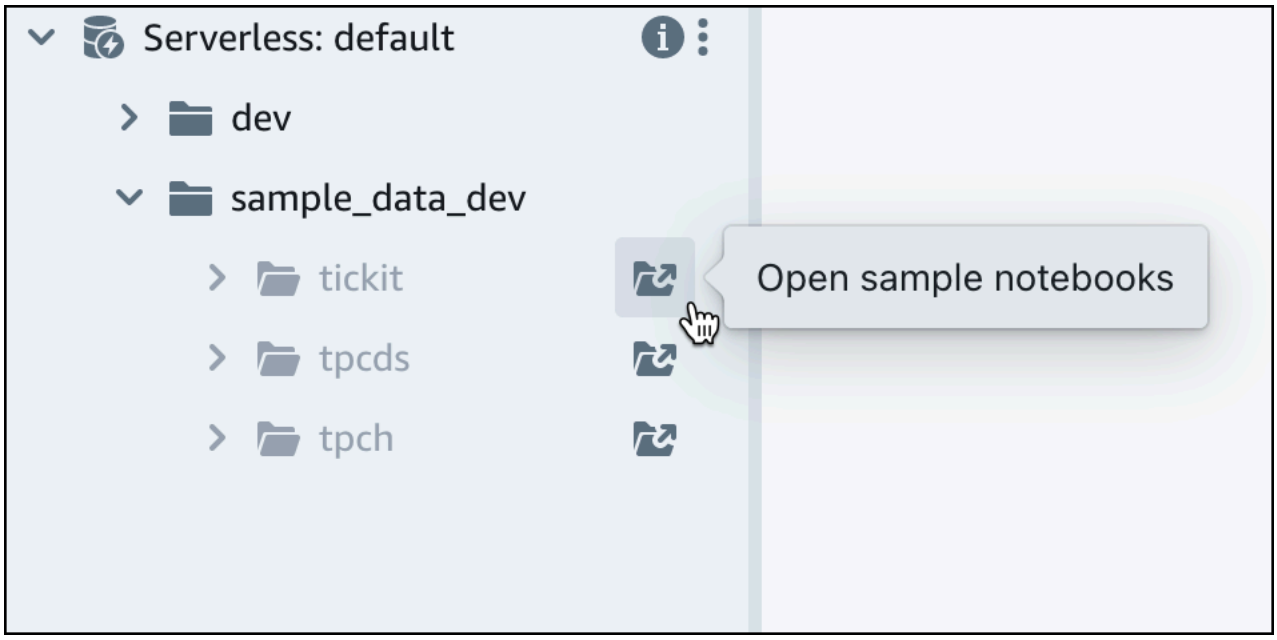


4. 在查询编辑器 v2 中首次连接到新工作组时，必须选择连接到该工作组所用的身份验证类型。在本指南中，选择联合用户，然后选择创建连接。



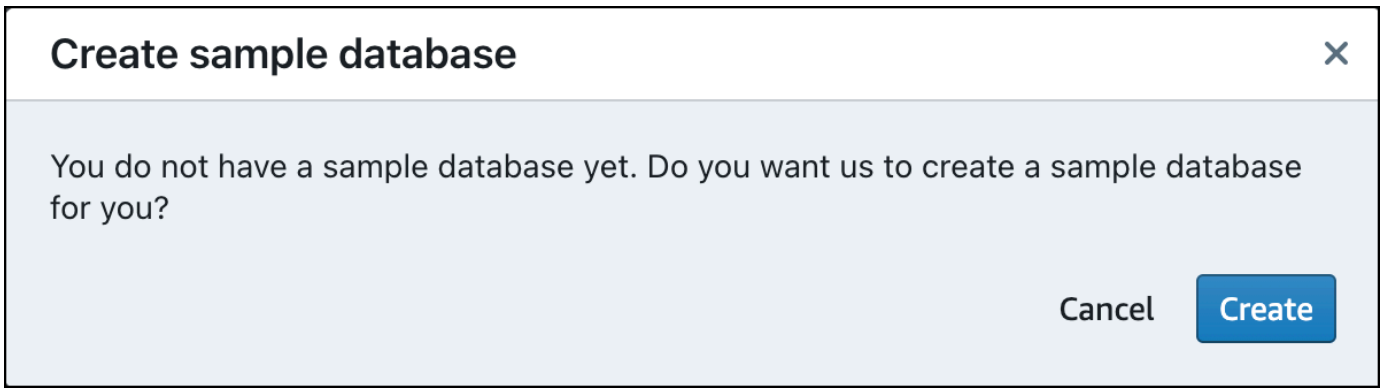
连接之后，您可以选择从 Amazon Redshift Serverless 或从 Amazon S3 存储桶加载示例数据。

5. 在 Amazon Redshift Serverless 原定设置工作组下，展开 sample_data_dev 数据库。有三个示例架构对应于三个示例数据集，您可以将这些示例数据集加载到 Amazon Redshift Serverless 数据库中。选择要加载的示例数据集，然后选择打开示例笔记本。



> **Note**
>
> SQL 笔记本是 SQL 和 Markdown 单元格的容器。您可以使用笔记本在单个文档中组织、注释及共享多个 SQL 命令。

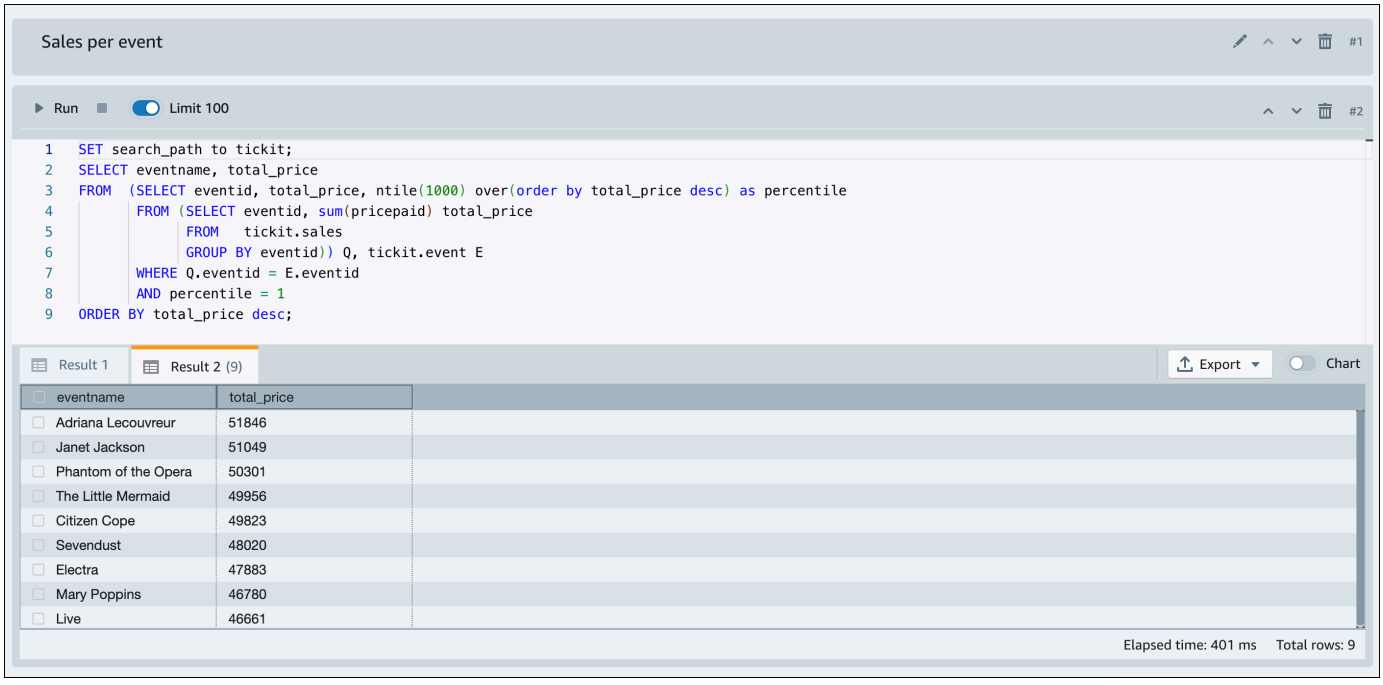
6. 首次加载数据时，查询编辑器 v2 将提示您创建示例数据库。选择创建。



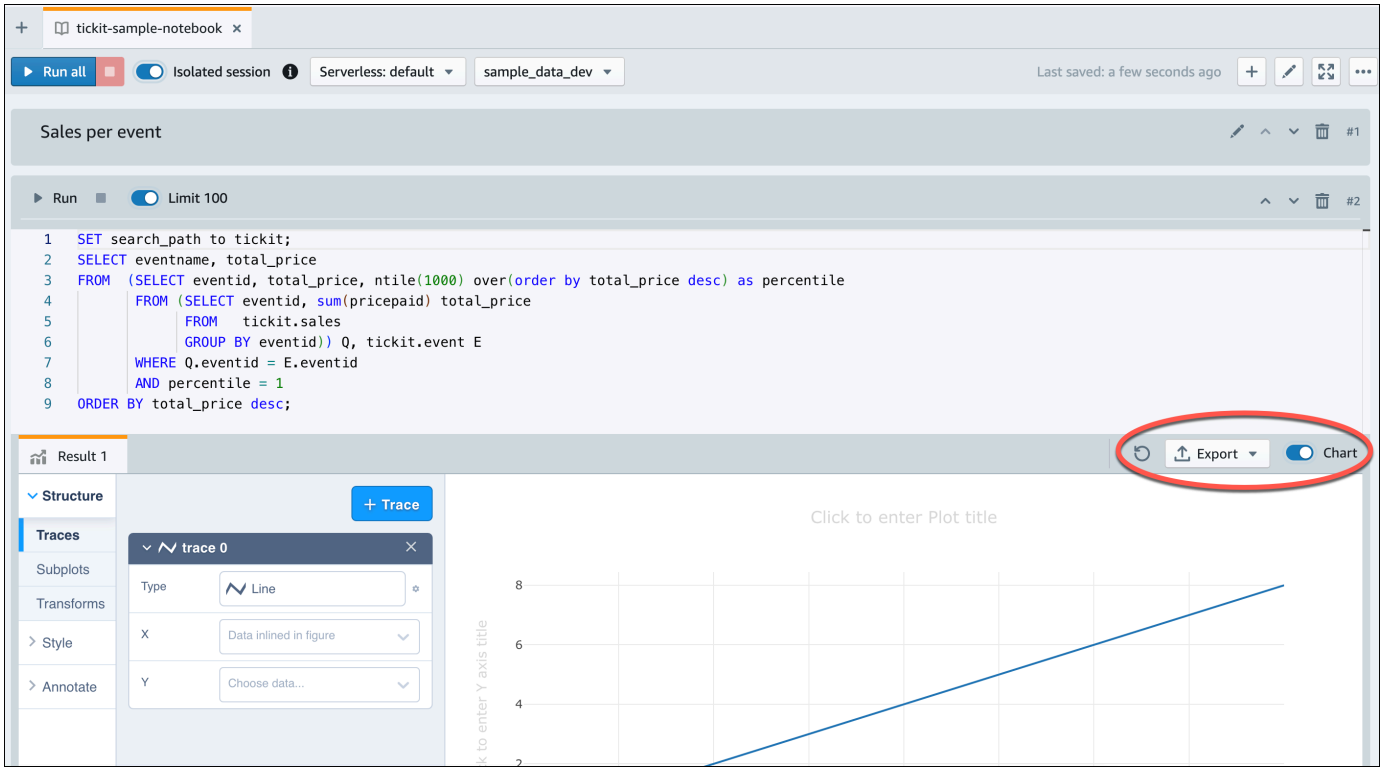
运行示例查询

设置 Amazon Redshift Serverless 之后，您可以开始在 Amazon Redshift Serverless 中使用示例数据集。Amazon Redshift Serverless 自动加载示例数据集，例如 tickit 数据集，然后您便立即可以查询数据。

- Amazon Redshift Serverless 完成示例数据的加载后，所有示例查询都会加载到编辑器中。您可以选择运行全部来运行示例笔记本中的所有查询。



您也可以将结果导出为 JSON 或 CSV 文件，或者以图表格式查看结果。



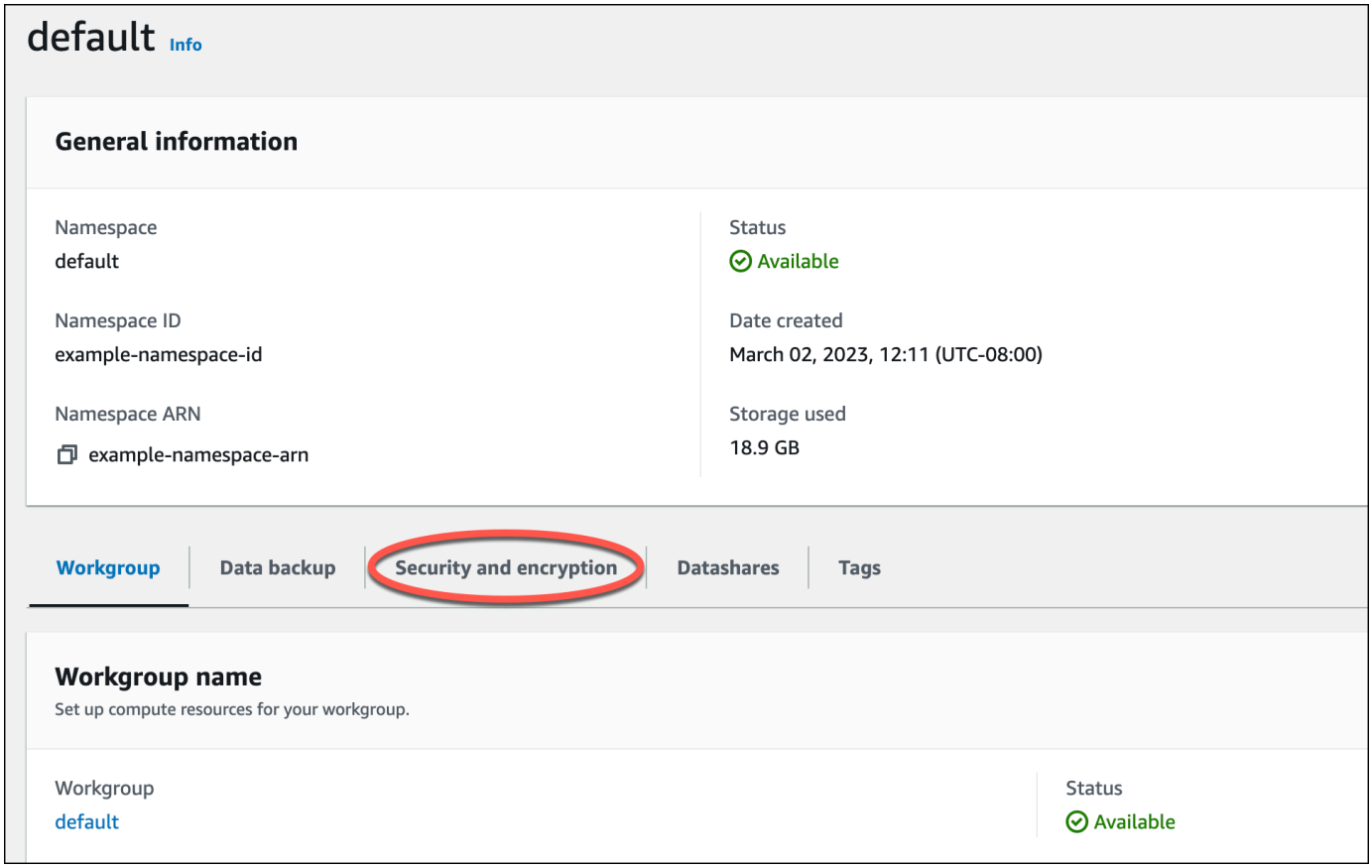
您还可以从 Amazon S3 存储桶加载数据。要了解更多信息，请参阅the section called “从 Amazon S3 加载数据”。

从 Amazon S3 加载数据

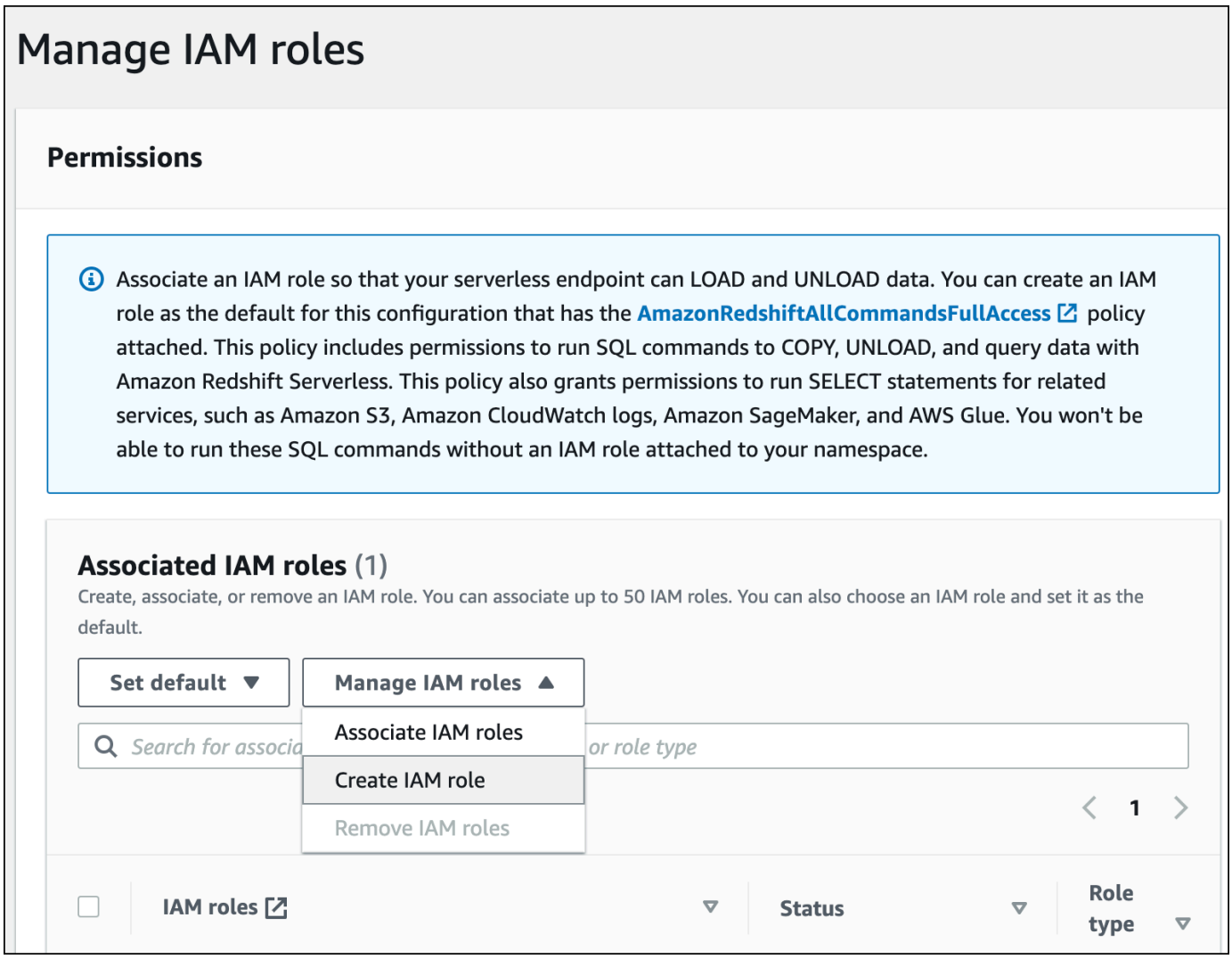
创建数据仓库后，您可以从 Amazon S3 加载数据。

此时，您拥有了一个名为 dev 的数据库。接下来，在该数据库中创建一些表，将数据上传到表，然后尝试执行查询。为方便起见，您上载的示例数据在 Amazon S3 桶中可用。

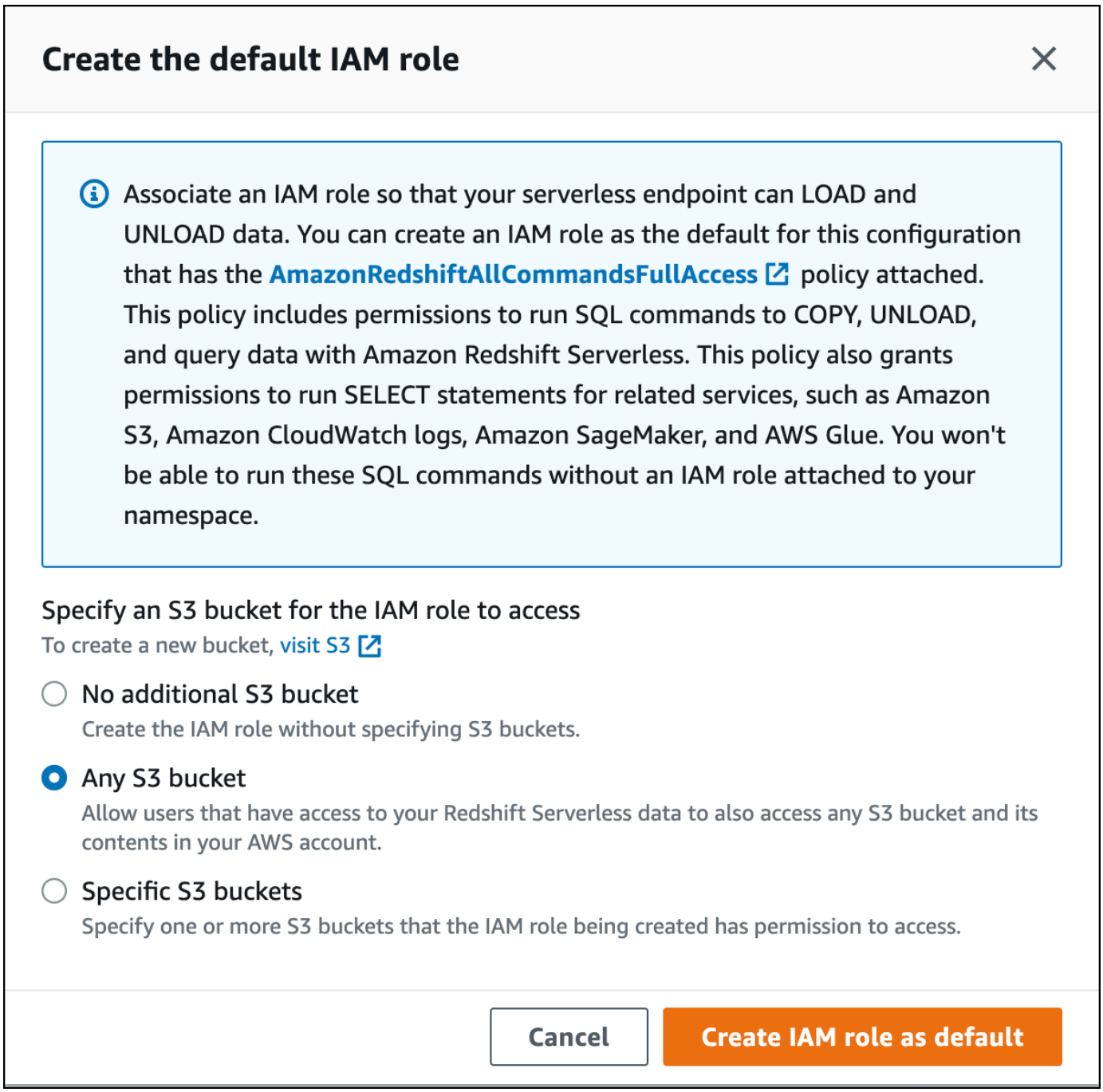
1. 在从 Amazon S3 中加载数据之前，您必须先创建具有必要权限的 IAM 角色并将其附加到无服务器命名空间。为此，请返回 Redshift Serverless 控制台并选择命名空间配置。在导航菜单中，选择您的命名空间，然后选择安全和加密。然后选择管理 IAM 角色。



2. 展开管理 IAM 角色菜单，然后选择创建 IAM 角色。



3. 选择您要授予此角色的 S3 存储桶访问权限级别，然后选择创建 IAM 角色作为默认角色。

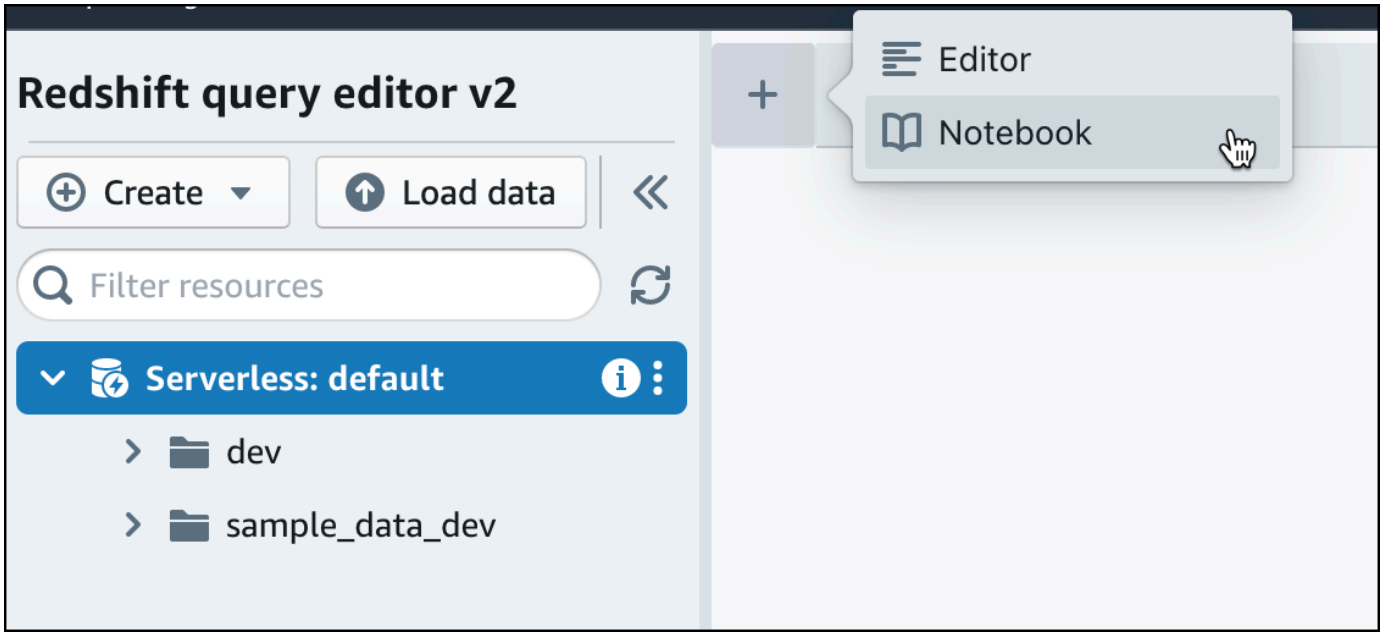


4. 选择保存更改。现在您可从 Amazon S3 加载示例数据。

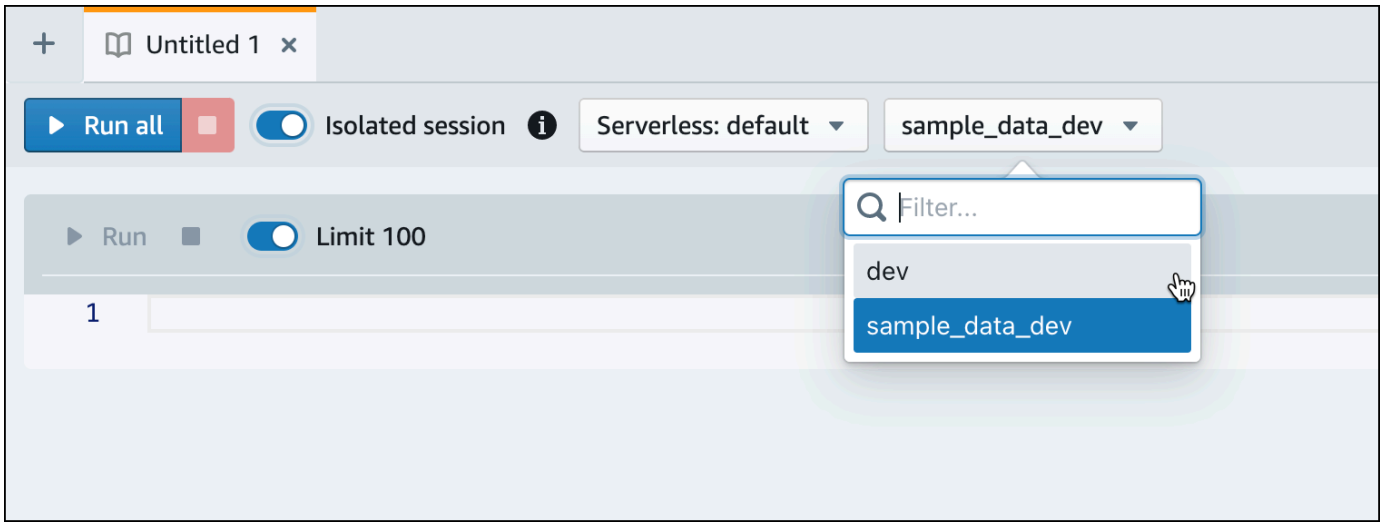
以下步骤使用公共 Amazon Redshift S3 存储桶中的数据，不过您可以使用自己的 S3 存储桶和 SQL 命令重复相同的步骤。

从 Amazon S3 加载示例数据

1. 在查询编辑器 v2 中，选择 + “添加”，然后选择笔记本以创建新的 SQL 笔记本。



2. 切换到 dev 数据库。



3. 创建表。

   如果您使用查询编辑器 v2，请复制并运行下列 create table 语句，以在 dev 数据库中创建表。有关语法的更多信息，请参阅《Amazon Redshift 数据库开发人员指南》中的 CREATE TABLE。

```
create table users(
```

```
userid integer not null distkey sortkey,
username char(8),
firstname varchar(30),
lastname varchar(30),
city varchar(30),
state char(2),
email varchar(100),
phone char(14),
likesports boolean,
liketheatre boolean,
likeconcerts boolean,
likejazz boolean,
likeclassical boolean,
likeopera boolean,
likerock boolean,
likevegas boolean,
likebroadway boolean,
likemusicals boolean);

create table event(
eventid integer not null distkey,
venueid smallint not null,
catid smallint not null,
dateid smallint not null sortkey,
eventname varchar(200),
starttime timestamp);

create table sales(
salesid integer not null,
listid integer not null distkey,
sellerid integer not null,
buyerid integer not null,
eventid integer not null,
dateid smallint not null sortkey,
qtysold smallint not null,
pricepaid decimal(8,2),
commission decimal(8,2),
saletime timestamp);
```

4. 使用查询编辑器 v2，在笔记本中创建一个新的 SQL 单元格。

5. 现在，您可以在查询编辑器 v2 中使用 COPY 命令，将大型数据集从 Amazon S3 或 Amazon DynamoDB 加载到 Amazon Redshift 中。有关 COPY 语法的更多信息，请参阅《Amazon Redshift 数据库开发人员指南中的 COPY。

   您可以使用公共 S3 存储桶中的一些示例数据运行 COPY 命令。在查询编辑器 v2 中运行以下 SQL 命令。

```
COPY users
FROM 's3://redshift-downloads/tickit/allusers_pipe.txt'
DELIMITER '|'
TIMEFORMAT 'YYYY-MM-DD HH:MI:SS'
IGNOREHEADER 1
REGION 'us-east-1'
IAM_ROLE default;

COPY event
FROM 's3://redshift-downloads/tickit/allevents_pipe.txt'
DELIMITER '|'
TIMEFORMAT 'YYYY-MM-DD HH:MI:SS'
IGNOREHEADER 1
REGION 'us-east-1'
IAM_ROLE default;

COPY sales
FROM 's3://redshift-downloads/tickit/sales_tab.txt'
DELIMITER '\t'
TIMEFORMAT 'MM/DD/YYYY HH:MI:SS'
IGNOREHEADER 1
REGION 'us-east-1'
IAM_ROLE default;
```

6. 加载数据后，在笔记本中创建另一个 SQL 单元格，然后尝试执行一些示例查询。有关使用 SELECT 命令的更多信息，请参阅《Amazon Redshift 开发人员指南》中的 SELECT。要了解示例数据的结构和架构，请使用查询编辑器 v2 进行探索。

```
-- Find top 10 buyers by quantity.
SELECT firstname, lastname, total_quantity
FROM   (SELECT buyerid, sum(qtysold) total_quantity
        FROM  sales
        GROUP BY buyerid
        ORDER BY total_quantity desc limit 10) Q, users
WHERE Q.buyerid = userid
ORDER BY Q.total_quantity desc;

-- Find events in the 99.9 percentile in terms of all time gross sales.
SELECT eventname, total_price
FROM  (SELECT eventid, total_price, ntile(1000) over(order by total_price desc) as
 percentile
       FROM (SELECT eventid, sum(pricepaid) total_price
             FROM   sales
             GROUP BY eventid)) Q, event E
       WHERE Q.eventid = E.eventid
       AND percentile = 1
ORDER BY total_price desc;
```

现在，在加载了数据并运行了一些示例查询后，您可以探索 Amazon Redshift Serverless 的其他领域。请参阅以下列表，详细了解如何使用 Amazon Redshift Serverless。

- 您可以从 Amazon S3 存储桶加载数据。有关更多信息，请参阅从 Amazon S3 加载数据。
- 您可以使用查询编辑器 v2，从小于 5MB 的本地字符分隔文件中加载数据。有关更多信息，请参阅从本地文件加载数据。
- 您可以使用具有 JDBC 和 ODBC 驱动程序的第三方 SQL 工具连接到 Amazon Redshift Serverless。有关更多信息，请参阅连接到 Amazon Redshift Serverless。
- 还可使用 Amazon Redshift 数据 API 连接到 Amazon Redshift Serverless。有关更多信息，请参阅使用 Amazon Redshift Data API。
- 您可以通过 CREATE MODEL 命令，将 Amazon Redshift Serverless 中的数据与 Redshift ML 结合使用来创建机器学习模型。请参阅教程：构建客户流失模型，了解如何构建 Redshift ML 模型。
- 您可以从 Amazon S3 数据湖中查询数据，而无需将任何数据加载到 Amazon Redshift Serverless 中。有关更多信息，请参阅查询数据湖。

Amazon Redshift 预置数据仓库入门

如果您是首次接触 Amazon Redshift 的用户，我们建议您先阅读以下部分以帮助您使用预置集群。Amazon Redshift 的基本流程是创建预置资源，连接到 Amazon Redshift，加载示例数据，然后对数据运行查询。在本指南中，您可以选择从 Amazon Redshift 或 Amazon S3 存储桶加载示例数据。在 Amazon Redshift 文档中，会使用示例数据来演示功能。

本教程演示如何使用 Amazon Redshift 预置集群，这些集群是您管理其系统资源的 AWS 数据仓库对象。您还可以将 Amazon Redshift 与无服务器工作组结合使用，后者是一种可根据使用情况自动扩展的数据仓库对象。要开始使用 Redshift Serverless，请参阅 Amazon Redshift Serverless 数据仓库入门。

在您创建并登录 Amazon Redshift 预置集群控制台后，您可以创建和管理 Amazon Redshift 对象，包括集群、节点和数据库。您还可以运行查询、查看查询以及执行其它 SQL 数据定义语言（DDL，Data Definition Language）和数据操纵语言（DML，Data Manipulation Language）操作。

> **Important**
>
> 您为本次练习预置的集群在真实环境中运行。只要集群运行，您的 AWS 账户就会产生费用。有关定价信息，请参阅 Amazon Redshift 定价页面。
> 为避免产生不必要的费用，请在完成练习后删除集群。本章的最后部分说明了如何执行这些操作。

登录到 AWS Management Console并打开 Amazon Redshift 控制台，网址：https://console.aws.amazon.com/redshiftv2/。

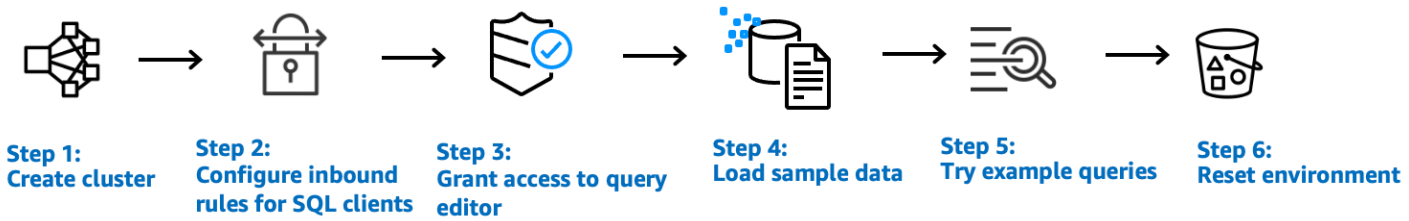
我们建议您首先进入预置集群控制面板，以便开始使用 Amazon Redshift 控制台。

根据您的配置，Amazon Redshift 预置集群控制台的导航窗格中会显示以下项目：

- Redshift Serverless – 无需设置、优化和管理 Amazon Redshift 预置集群即可访问和分析数据。
- 预置集群控制面板 – 查看 AWS 区域中集群的列表，检查集群指标和查询概览，以了解指标数据（例如 CPU 利用率）和查询信息。这些可以帮助您确定指定时间范围内的性能数据是否异常。
- 集群 – 查看此 AWS 区域中的您的集群列表，选择一个集群开始查询，或执行与集群相关的操作。您还可以从此页面创建新集群。
- 查询编辑器 – 对 Amazon Redshift 集群上托管的数据库运行查询。我们建议改为使用查询编辑器 v2。

- 查询编辑器 v2 – Amazon Redshift 查询编辑器 v2 是一个单独的基于 Web 的 SQL 客户端应用程序，可在 Amazon Redshift 数据仓库上创作和运行查询。您可以在图表中可视化结果，并通过与团队中的其他人共享查询来进行协作。
- 查询和加载 – 获取用于参考或故障排除的信息，例如最近查询的列表和每个查询的 SQL 文本。
- 数据仓库 – 创建者账户管理员可以授权使用者账户访问数据共享，也可以选择不授权任何访问权限。要使用授权的数据共享，使用者账户管理员可以将数据共享与整个 AWS 账户或账户中的特定集群命名空间相关联。管理员还可以拒绝数据共享。
- 零 ETL 集成 – 管理集成，在支持的源中，该集成使得事务数据在写入之后在 Amazon Redshift 中可用。
- IAM Identity Center 连接 – 配置 Amazon Redshift 与 IAM Identity Center 之间的连接。
- 配置 – 通过 Java 数据库连接 (JDBC) 和开放式数据库连接 (ODBC) 这两种连接将 SQL 客户端工具连接到 Amazon Redshift 集群。您还可以设置 Amazon Redshift 托管式 Virtual Private Cloud（VPC）端点。这样做会在一个基于包含集群的 Amazon VPC 服务的 VPC 与另一个运行客户端工具的 VPC 之间提供私有连接。
- AWS 合作伙伴集成 – 创建与支持的 AWS 合作伙伴的集成。
- 顾问 – 获取有关您可以对 Amazon Redshift 集群进行的更改的具体建议，以确定优化的优先级。
- AWS Marketplace – 获取有关其他工具或与 Amazon Redshift 一起使用的 AWS 服务的信息。
- 警报 – 针对集群指标创建警报，以查看指定时间段内的性能数据和跟踪指标。
- 事件 – 跟踪事件并获取有关事件发生日期、描述或事件来源等信息的报告。
- 新增功能 – 查看 Amazon Redshift 的新功能和产品更新。

在本教程中，您会执行以下步骤。



主题

- 注册 AWS
- 确定防火墙规则
- 步骤 1：创建示例 Amazon Redshift 集群
- 步骤 2：为 SQL 客户端配置入站规则

- 步骤 3：授予对 SQL 客户端的访问权限并运行查询
- 步骤 4：将数据从 Amazon S3 加载到 Amazon Redshift
- 步骤 5：使用查询编辑器尝试进行示例查询
- 步骤 6：重置环境

注册 AWS

如果您还没有 AWS 账户 账户，请先注册一个。如果您已有账户，则可以跳过此先决条件步骤，并使用您已有的账户。

1. 打开 https://portal.aws.amazon.com/billing/signup。
2. 按照屏幕上的说明操作。

   在注册时，将接到电话或收到短信，要求使用电话键盘输入一个验证码。

   当您注册 AWS 账户 时，系统将会创建一个 AWS 账户根用户。根用户有权访问该账户中的所有 AWS 服务和资源。作为最佳安全实践，请为用户分配管理访问权限，并且只使用根用户来执行需要根用户访问权限的任务。

确定防火墙规则

> **Note**
>
> 本教程假设您的集群使用默认端口 5439，并可使用 Amazon Redshift 查询编辑器 v2 运行 SQL 命令。教程中并未详细介绍您的环境中可能需要的联网配置或 SQL 客户端设置。

在某些环境中，您可以在启动 Amazon Redshift 集群时指定端口。您可以使用此端口以及集群的端点 URL 来访问集群。您还将在安全组中创建一个入站入口规则，以允许通过该端口访问您的集群。

如果您的客户端计算机位于防火墙后面，请确保您知道可用的开放端口。通过此开放端口，您可以从 SQL 客户端工具连接到集群并运行查询。如果您不知道此开放端口，则应与了解您网络防火墙规则的人员合作，以在您的防火墙中确定一个开放端口。

虽然 Amazon Redshift 默认使用端口 5439，但如果您的防火墙中未打开该端口，则无法建立连接。创建了 Amazon Redshift 集群后，则不能再更改其端口号。因此，确保指定一个在启动过程中可在您的环境中工作的开放端口。

步骤 1：创建示例 Amazon Redshift 集群

在本教程中，您将了解创建具有一个数据库的 Amazon Redshift 集群的流程。然后，您将数据集从 Amazon S3 加载到数据库的表中。您可以使用该示例集群评估 Amazon Redshift 服务。

在开始设置 Amazon Redshift 集群之前，请确保您已满足注册 AWS 和确定防火墙规则中的所有必要先决条件。

对于任何访问其它 AWS 资源上的数据的操作，您的集群需要具有权限才能代表您访问该资源和该资源上的数据。例如，使用 SQL COPY 命令从 Amazon Simple Storage Service（Amazon S3）加载数据。您通过使用 AWS Identity and Access Management（IAM）提供这些权限。您可以通过自己创建 IAM 角色并将角色附加到集群来执行此操作。有关凭证和访问权限的更多信息，请参阅《Amazon Redshift 数据库开发人员指南》中的凭证和访问权限。

要创建 Amazon Redshift 集群

1. 登录到 AWS Management Console并打开 Amazon Redshift 控制台，网址：https://console.aws.amazon.com/redshiftv2/。

   > **Important**
   >
   > 如果您使用 IAM 用户凭证，请确保您具备执行集群操作所需的权限。有关更多信息，请参阅《Amazon Redshift 管理指南》中的 Amazon Redshift 中的安全性。

2. 在 AWS 控制台上，选择要在其中创建集群的 AWS 区域。
3. 在导航菜单上，选择集群，然后选择创建集群。此时将显示创建集群页面。
4. 在集群配置部分中，指定集群标识符、节点类型和节点的值：

   - 集群标识符：在此教程中输入 **`examplecluster`**。此标识符必须是唯一的。标识符必须在 1-63 个字符之间，使用有效字符 a—z（仅小写）和 -（连字符）。
   - 选择以下方法之一以调整集群的大小：

     > **Note**
     >
     > 以下步骤假定 AWS 区域支持 RA3 节点类型。有关支持 RA3 节点类型的 AWS 区域的列表，请参阅《Amazon Redshift 管理指南》中的 RA3 节点类型概览。要详细了解每个节点类型和大小的节点规范，请参阅节点类型详细信息。

- 如果您不知道要将集群调整到多大，请选择帮我选择。执行此操作将打开大小调整计算器，该计算器将询问您有关计划存储在数据仓库中的数据的大小和查询特性的问题。

  如果您知道集群所需的大小（即节点类型和节点数），请选择 I'll choose（我会选择）。然后选择 Node Type（节点类型）和 Nodes（节点）数量来确定集群规模。

  对于本教程，请为节点类型选择 ra3.4xlarge，为节点数选择 2。

  如果可以选择可用区配置，请选择单可用区。

- 要使用 Amazon Redshift 提供的示例数据集，在示例数据中，选择加载示例数据。Amazon Redshift 会将示例数据集 Tickit 加载到默认的 `dev` 数据库和 `public` schema。

5. 在数据库配置部分中，为管理员用户名指定值。对于管理员密码，从以下选项中进行选择：

   - 生成密码 – 使用 Amazon Redshift 生成的密码。
   - 手动添加管理员密码 – 使用您自己的密码。
   - 管理 AWS Secrets Manager 中的管理员凭证 – Amazon Redshift 使用 AWS Secrets Manager 生成和管理您的管理员密码。使用 AWS Secrets Manager 生成和管理您密码的密钥会产生一定的费用。有关 AWS Secrets Manager 定价的信息，请参阅 AWS Secrets Manager 定价。

   在本教程中，使用以下值：

   - 管理员用户名：输入 **`awsuser`**。
   - 管理员用户密码：输入 **`Changeit1`** 作为密码。

6. 在本教程中，创建 IAM 角色并将其设置为集群的默认角色，如下所述。一个集群只能有一个默认的 IAM 角色集。

   a. 在集群权限下，为管理 IAM 角色选择创建 IAM 角色。

   b. 为 IAM 角色指定一个 Amazon S3 桶，以便通过以下方法访问：

      - 选择无附加 Amazon S3 桶以允许创建的 IAM 角色仅访问名为 `redshift` 的 Amazon S3 桶。
      - 选择任何 Amazon S3 桶以允许创建的 IAM 角色访问所有 Amazon S3 桶。
      - 选择特定 Amazon S3 桶以便为创建的 IAM 角色指定一个或多个 Amazon S3 桶以供访问。然后从表中选择一个或多个 Amazon S3 桶。

c. 选择创建 IAM 角色作为默认角色。Amazon Redshift 会自动创建 IAM 角色并将其设置为集群的默认角色。

由于您从控制台创建了 IAM 角色，因此它具有 `AmazonRedshiftAllCommandsFullAccess` 附加策略。这允许 Amazon Redshift 复制、加载、查询和分析 IAM 账户中 Amazon 资源的数据。

有关如何管理集群的默认 IAM 角色的信息，请参阅《Amazon Redshift 管理指南》中的创建一个 IAM 角色作为 Amazon Redshift 的默认角色。

7. （可选）在其他配置部分，关闭使用默认值以修改网络和安全、数据库配置、维护、监控和备份设置。

在有些情况下，您可以使用加载示例数据选项创建集群并希望启用增强型 Amazon VPC 路由。如果是这样，Virtual Private Cloud (VPC) 中的集群需要访问 Amazon S3 端点才能加载数据。

为使集群可公开访问，您可以执行以下两项操作之一。您可以在 VPC 中配置网络地址转换 (NAT) 地址，以便集群访问互联网。或者，您可以在 VPC 中配置 Amazon S3 VPC 端点。有关增强型 Amazon VPC 路由的更多信息，请参阅《Amazon Redshift 管理指南》中的增强型 Amazon VPC 路由。

8. 选择创建集群。在集群页面上，等待您的集群创建并处于 `Available` 状态。

步骤 2：为 SQL 客户端配置入站规则

> **Note**
>
> 建议您跳过此步骤，使用 Amazon Redshift 查询编辑器 v2 访问您的集群。

在本教程后面部分中，您将从基于 Amazon VPC 服务的 Virtual Private Cloud (VPC) 内访问集群。但是，如果您从防火墙外部使用 SQL 客户端来访问集群，请确保授予入站访问权限。

检查防火墙并授予对集群的入站访问权限

1. 如果需要从防火墙外部访问集群，请检查防火墙规则。例如，您的客户端可以是 Amazon Elastic Compute Cloud (Amazon EC2) 实例或外部计算机。

有关防火墙规则的更多信息，请参阅《Amazon EC2 用户指南》中的安全组规则。

2. 要从 Amazon EC2 外部客户端进行访问，请向附加到集群的安全组添加一个允许入站流量的入口规则。您可以在 Amazon EC2 控制台中添加 Amazon EC2 安全组规则。例如，CIDR/IP 192.0.2.0/24 允许该 IP 地址范围内的客户端连接到您的集群。找到适合您环境的正确 CIDR/IP。

步骤 3：授予对 SQL 客户端的访问权限并运行查询

对于 SQL 客户端，要查询 Amazon Redshift 集群托管的数据库，您有多种选择：这些指令包括：

- 连接到您的集群，然后使用 Amazon Redshift 查询编辑器 v2 运行查询。

  如果您使用查询编辑器 v2，则无需下载和设置 SQL 客户端应用程序。您可以从 Amazon Redshift 控制台启动 Amazon Redshift 查询编辑器 v2。
- 使用 RSQL 连接到您的集群。有关更多信息，请参阅《Amazon Redshift 管理指南》中的 使用 Amazon Redshift RSQL 连接。
- 通过 SQL 客户端工具（如 SQL Workbench/J）连接到您的集群。有关更多信息，请参阅《Amazon Redshift 管理指南》中的使用 SQL Workbench/J 连接到集群。

本教程使用 Amazon Redshift 查询编辑器 v2 作为一种简单的方法，在由 Amazon Redshift 集群托管的数据库上运行查询。创建集群后，您可以立即运行查询。有关使用 Amazon Redshift 查询编辑器 v2 时注意事项的详细信息，请参阅《Amazon Redshift 管理指南》中的使用查询编辑器 v2 时的注意事项。

授予对查询编辑器 v2 的访问权限

管理员第一次为您的 AWS 账户 配置查询编辑器 v2 时，他们会选择用于加密查询编辑器 v2 资源的 AWS KMS key。Amazon Redshift 查询编辑器 v2 资源包括保存的查询、笔记本和图表。默认情况下，AWS 拥有的密钥用于加密资源。或者，管理员可在配置页面中为密钥选择 Amazon 资源名称（ARN）来使用客户托管密钥。配置账户后，AWS KMS 加密设置无法更改。有关更多信息，请参阅《Amazon Redshift 管理指南》中的配置您的 AWS 账户。

如要访问查询编辑器 v2，您需要相应权限。管理员可将 Amazon Redshift 查询编辑器 v2 的 AWS 托管式策略之一附加到 IAM 角色或用户，以授予权限。这些 AWS 托管式策略使用不同的选项编写，可控制标记资源允许共享查询的方式。您可以使用 IAM 控制台 (https://console.aws.amazon.com/iam/) 附加 IAM 策略。有关这些策略的详细信息，请参阅《Amazon Redshift 管理指南》中的访问查询编辑器 v2。

您还可以根据提供的托管式策略中允许和拒绝的权限创建您自己的策略。如果您使用 IAM 控制台策略编辑器创建自己的策略，请选择 SQL Workbench 作为您在可视化编辑器中创建策略的服务。查询编辑器 v2 使用可视化编辑器和 IAM policy simulator 中的服务名称 AWS SQL Workbench。

有关更多信息，请参阅《Amazon Redshift 管理指南》中的 使用查询编辑器 v2。

步骤 4：将数据从 Amazon S3 加载到 Amazon Redshift

创建集群后，您可以将数据从 Amazon S3 加载到数据库表。您可以通过多种方法从 Amazon S3 加载数据。

- 您可以使用 SQL 客户端运行 SQL CREATE TABLE 命令在数据库中创建表，然后使用 SQL COPY 命令从 Amazon S3 加载数据。Amazon Redshift 查询编辑器 v2 是一个 SQL 客户端。
- 您可以使用 Amazon Redshift 查询编辑器 v2 加载向导。

本教程演示如何使用 Amazon Redshift 查询编辑器 v2 运行 SQL 命令来创建表和复制数据。从 Amazon Redshift 控制台导航窗格启动查询编辑器 v2。在查询编辑器 v2 中，使用管理员用户 `awsuser`，创建与 `examplecluster` 集群和名为 `dev` 的数据库的连接。对于本教程，在创建连接时，请选择使用数据库用户名的临时凭证。有关使用 Amazon Redshift 查询编辑器 v2 的详细信息，请参阅《Amazon Redshift 管理指南》中的连接到 Amazon Redshift 数据库。

使用 SQL 命令从 Amazon S3 加载数据

在查询编辑器 v2 查询编辑器窗格上，确认您已连接到 `examplecluster` 集群和 `dev` 数据库。接下来，在数据库中创建一些表，然后将数据上传到表中。对于本教程，您加载的数据存储在一个可从多个 AWS 区域访问的 Amazon S3 存储桶中。

以下过程将会创建表并从公共 Amazon S3 存储桶加载数据。

使用 Amazon Redshift 查询编辑器 v2 复制并运行以下 create table 语句，以在 `dev` 数据库的 `public` 架构中创建表。有关语法的更多信息，请参阅《Amazon Redshift 数据库开发人员指南》中的 CREATE TABLE。

使用 SQL 客户端（例如查询编辑器 v2）创建和加载数据

1. 输入以下 SQL 命令以 CREATE `sales` 表。

```
  drop table if exists sales;
  create table sales(
salesid integer not null,
listid integer not null distkey,
sellerid integer not null,
```

```
 buyerid integer not null,
 eventid integer not null,
 dateid smallint not null sortkey,
 qtysold smallint not null,
 pricepaid decimal(8,2),
 commission decimal(8,2),
 saletime timestamp);
```

2. 输入以下 SQL 命令以 CREATE `date` 表。

```
drop table if exists date;
create table date(
 dateid smallint not null distkey sortkey,
 caldate date not null,
 day character(3) not null,
 week smallint not null,
 month character(5) not null,
 qtr character(5) not null,
 year smallint not null,
 holiday boolean default('N'));
```

3. 使用 COPY 命令从 Amazon S3 中加载 `sales` 表

> **Note**
>
> 我们建议使用 COPY 命令将大型数据集从 Amazon S3 加载到 Amazon Redshift 中。有关 COPY 语法的更多信息，请参阅《Amazon Redshift 数据库开发人员指南中的 COPY。

要加载示例数据，为您的集群提供代表您访问 Amazon S3 的身份验证。如果您在创建集群时选择了创建 IAM 角色作为默认角色，就可以通过引用所创建并设置为集群 `default` 的 IAM 角色来提供身份验证。

使用以下 SQL 命令加载 `sales` 表。您可以选择从 Amazon S3 存储桶下载并查看 sales 表的源数据。。

```
COPY sales
    FROM 's3://redshift-downloads/tickit/sales_tab.txt'
    DELIMITER '\t'
    TIMEFORMAT 'MM/DD/YYYY HH:MI:SS'
    REGION 'us-east-1'
```

```
    IAM_ROLE default;
```

4. 使用以下 SQL 命令加载 date 表。您可以选择从 Amazon S3 存储桶下载并查看 date 表的源数据。。

```
COPY date
    FROM 's3://redshift-downloads/tickit/date2008_pipe.txt'
    DELIMITER '|'
    REGION 'us-east-1'
    IAM_ROLE default;
```

使用查询编辑器 v2 从 Amazon S3 加载数据

此部分将介绍如何将您自己的数据加载到 Amazon Redshift 集群中。使用加载数据向导时，查询编辑器 v2 可简化数据的加载。在查询编辑器 v2 加载数据向导中生成和使用的 COPY 命令，支持从 Amazon S3 加载数据的 COPY 命令语法可使用的多个参数。有关 COPY 命令及其用于从 Amazon S3 中复制加载的选项的信息，请参阅《Amazon Redshift 数据库开发人员指南》中的 Amazon Simple Storage Service 中的 COPY 命令。

要将您自己的数据从 Amazon S3 加载到 Amazon Redshift，Amazon Redshift 需要一个 IAM 角色，该角色具有从指定的 Amazon S3 桶加载数据所需的权限。

要将您自己的数据从 Amazon S3 加载到 Amazon Redshift，可以使用查询编辑器 v2 加载数据向导。有关如何使用加载数据向导的信息，请参阅《Amazon Redshift 管理指南》中的从 Amazon S3 加载数据。

在集群中创建 TICKIT 数据

TICKIT 是一个示例数据库，您可以选择将其加载到 Amazon Redshift 集群中，以学习如何在 Amazon Redshift 中查询数据。您可以通过以下方式创建一组完整的 TICKIT 表，并将数据加载到集群中：

- 在 Amazon Redshift 控制台中创建集群时，您可以选择同时加载示例 TICKIT 数据。在 Amazon Redshift 控制台上，依次选择集群和创建集群。在示例数据部分，选择加载示例数据，在集群创建过程中，Amazon Redshift 自动将其示例数据集加载到您的 Amazon Redshift 集群 dev 数据库。
- 要连接到现有集群，请执行以下操作：
  - 在 Amazon Redshift 控制台中，从导航栏中选择集群。
  - 从集群窗格中选择您的集群。
  - 选择查询数据，在查询编辑器 v2 中查询。

  - 在资源列表中展开 examplecluster。如果这是您首次连接到集群，则会出现连接到 examplecluster。选择数据库用户名和密码。将数据库保留为 **dev**。指定 **awsuser** 作为用户名，指定 **Changeit1** 作为密码。
  - 选择创建连接。
- 使用 Amazon Redshift 查询编辑器 v2，您可以将 TICKIT 数据加载到名为 sample_data_dev 的示例数据库。在资源列表中选择 sample_data_dev 数据库。在 tickit 节点旁边，选择打开示例笔记本图标。确认您要创建示例数据库。
- Amazon Redshift 查询编辑器 v2 创建示例数据库以及名为 tickit-sample-notebook 的示例笔记本。您可以选择全部运行来运行此笔记本，以查询示例数据库中的数据。

要查看有关 TICKIT 数据的详细信息，请参阅《Amazon Redshift 数据库开发人员指南》中的示例数据库。

步骤 5：使用查询编辑器尝试进行示例查询

要设置和使用 Amazon Redshift 查询编辑器 v2 来查询数据库，请参阅《Amazon Redshift 管理指南》中的使用查询编辑器 v2。

现在，尝试一些示例查询，如下所示。要在查询编辑器 v2 中创建新查询，请选择查询窗格右上角的 + 图标，然后选择 SQL。将出现一个新的查询页面，您可以在其中复制和粘贴以下 SQL 查询。

> **Note**
>
> 请务必先运行笔记本中的第一个查询，即使用以下 SQL 命令将 `search_path` 服务器配置值设置为 `tickit` 架构：
>
> ```
> set search_path to tickit;
> ```

有关使用 SELECT 命令的更多信息，请参阅《Amazon Redshift 数据库开发人员指南》中的 SELECT。

```
-- Get definition for the sales table.
SELECT *
FROM pg_table_def
WHERE tablename = 'sales';
```

```
-- Find total sales on a given calendar date.
SELECT sum(qtysold)
FROM   sales, date
WHERE  sales.dateid = date.dateid
AND    caldate = '2008-01-05';
```

```
-- Find top 10 buyers by quantity.
SELECT firstname, lastname, total_quantity
FROM   (SELECT buyerid, sum(qtysold) total_quantity
        FROM  sales
        GROUP BY buyerid
        ORDER BY total_quantity desc limit 10) Q, users
WHERE Q.buyerid = userid
ORDER BY Q.total_quantity desc;
```

```
-- Find events in the 99.9 percentile in terms of all time gross sales.
SELECT eventname, total_price
FROM  (SELECT eventid, total_price, ntile(1000) over(order by total_price desc) as
 percentile
       FROM (SELECT eventid, sum(pricepaid) total_price
             FROM   sales
             GROUP BY eventid)) Q, event E
       WHERE Q.eventid = E.eventid
       AND percentile = 1
ORDER BY total_price desc;
```

步骤 6：重置环境

在前面的步骤中，您已成功创建了 Amazon Redshift 集群，将数据加载到表中，并使用 Amazon Redshift 查询编辑器 v2 等 SQL 客户端查询了数据。

完成本教程后，我们建议您请通过删除您的示例集群来将您的环境还原至先前的状态。在将 Amazon Redshift 服务删除之前，您需要继续为其付费。

但是，如果您打算执行其他 Amazon Redshift 指南中的任务或者运行命令以在数据仓库中定义和使用数据库中介绍的任务，则需要将示例集群保持在运行状态。

删除集群

1. 登录到 AWS Management Console并打开 Amazon Redshift 控制台，网址：https://console.aws.amazon.com/redshiftv2/。
2. 在导航菜单上，选择集群以显示集群的列表。
3. 选择 `examplecluster` 集群。对于操作，选择删除。此时将显示删除 examplecluster？页面。
4. 确认要删除的集群，取消选中创建最终快照设置，然后输入 **`delete`** 以确认删除。选择删除集群。

在集群列表页面上，集群状态会在集群被删除时进行更新。

完成本教程后，您可以在用以了解有关 Amazon Redshift 的其他资源中详细了解 Amazon Redshift 以及后续步骤。

运行命令以在数据仓库中定义和使用数据库

Redshift Serverless 数据仓库和 Amazon Redshift 预置数据仓库都包含数据库。启动数据仓库后，您可以使用 SQL 命令管理大多数的数据库操作。除了少数例外，所有 Amazon Redshift 数据库的 SQL 功能和语法都是一样的。有关 Amazon Redshift 中可用的 SQL 命令的详细信息，请参阅《Amazon Redshift 数据库开发人员指南》中的 SQL 命令。

当您创建数据仓库时，大多数情况下，Amazon Redshift 还会创建默认的 dev 数据库。连接到 dev 数据库后，您可以创建另一个数据库。

以下各节将为您演练使用 Amazon Redshift 数据库时的常见数据库任务。这些任务以创建数据库开始，如果您继续到最后一项任务，则可以通过删除数据库来删除所创建的全部资源。

本部分中的示例假定以下内容：

- 您已创建了 Amazon Redshift 数据仓库。
- 您已从 SQL 客户端工具（例如 Amazon Redshift 查询编辑器 v2）建立了到数据仓库的连接。有关查询编辑器 v2 的更多信息，请参阅《Amazon Redshift 管理指南》中的使用 Amazon Redshift 查询编辑器 v2 查询数据库。

主题

连接到 Amazon Redshift 数据仓库

要连接到 Amazon Redshift 集群，请从 Amazon Redshift 控制台集群页面展开连接到 Amazon Redshift 集群，然后执行以下操作之一：

- 选择查询数据可使用查询编辑器 v2 对 Amazon Redshift 集群托管的数据库运行查询。创建集群后，可以使用查询编辑器 v2 立即运行查询。

  有关更多信息，请参阅《Amazon Redshift 管理指南》中的使用 Amazon Redshift 查询编辑器 v2 查询数据库。
- 在使用客户端工具中，选择您的集群，然后通过复制 JDBC 或 ODBC 驱动程序 URL，使用 JDBC 或 ODBC 驱动程序从您的客户端工具连接到 Amazon Redshift。从您的客户端计算机或实例上使用此 URL。对应用程序进行编码以使用 JDBC 或 ODBC 数据访问 API 操作，或使用支持 JDBC 或 ODBC 的 SQL 客户端工具。

  有关如何查找集群连接字符串的更多信息，请参阅查找集群连接字符串。
- 如果 SQL 客户端工具需要驱动程序，您可以选择 JDBC 或 ODBC 驱动程序，下载特定于操作系统的驱动程序，以从客户端工具连接到 Amazon Redshift。

  有关如何为 SQL 客户端安装适当的驱动程序的更多信息，请参阅配置 JDBC 驱动程序版本 2.0 连接。

  有关如何配置 ODBC 连接的更多信息，请参阅配置 ODBC 连接。

要连接到 Redshift Serverless 数据仓库，请从 Amazon Redshift 控制台的 Serverless 控制面板页面执行以下操作之一：

- 使用 Amazon Redshift 查询编辑器 v2 对 Redshift Serverless 数据仓库托管的数据库运行查询。创建数据仓库后，您可以使用查询编辑器 v2 立即运行查询。

  有关更多信息，请参阅使用 Amazon Redshift 查询编辑器 v2 查询数据库。
- 通过复制 JDBC 或 ODBC 驱动程序 URL，使用 JDBC 或 ODBC 驱动程序从您的客户端工具连接到 Amazon Redshift。

  要处理您的数据仓库中的数据，您需要使用 JDBC 或 ODBC 驱动程序，从您的客户端计算机或实例进行连接。对应用程序进行编码以使用 JDBC 或 ODBC 数据访问 API 操作，或使用支持 JDBC 或 ODBC 的 SQL 客户端工具。

  有关如何查找连接字符串的更多信息，请参阅《Amazon Redshift 管理指南》中的连接到 Redshift Serverless。

创建 数据库

验证数据仓库已启动并运行以后，您就可以创建数据库了。您将在此数据库中实际创建表、加载数据和运行查询。一个数据仓库可托管多个数据库。例如，您可以在同一个数据仓库中创建一个名为 SALESDB 的数据库用于销售数据，并创建一个名为 ORDERSDB 的数据库用于订单数据。

要创建名为 **SALESDB** 的数据库，请在 SQL 客户端工具中运行以下命令。

```
CREATE DATABASE salesdb;
```

> **Note**
>
> 运行该命令后，请确保在 SQL 客户端工具中，刷新数据仓库的对象列表，以查看新的 salesdb。

对于本练习，我们将接受默认设置。有关更多命令选项的信息，请参阅《Amazon Redshift 数据库开发人员指南》中的 CREATE DATABASE。要删除数据库及其内容，请参阅《Amazon Redshift 数据库开发人员指南》中的 DROP DATABASE。

创建 SALESDB 数据库后，可以从 SQL 客户端连接到新数据库。请使用与当前连接所用的相同连接参数，但将数据库名称改为 SALESDB。

创建用户

默认情况下，只有您在启动数据仓库时创建的管理员用户才有权访问数据仓库中的默认数据库。要向其他用户授予访问权限，请创建一个或多个账户。数据库用户账户在数据仓库中的所有数据库中全局适用，不属于单个数据库。

使用 CREATE USER 命令可创建新的用户。在创建新用户时，您指定新用户的用户名称和密码。我们建议您为用户指定密码。密码长度必须为 8–64 个字符，并且必须包含至少一个大写字母、一个小写字母和一个数字。

例如，要创建名为 **GUEST** 并且密码为 **ABCd4321** 的用户，请运行以下命令。

```
CREATE USER GUEST PASSWORD 'ABCd4321';
```

要以 GUEST 用户的身份连接到 SALESDB 数据库，请在创建用户时使用相同的密码，例如 ABCd4321。

有关其他命令选项的信息，请参阅《Amazon Redshift 数据库开发人员指南》中的 CREATE USER。

创建架构

创建新数据库后，您可以在当前数据库中创建新 schema。schema 是包含命名数据库对象（例如表、视图和用户定义的函数 (UDF)）的命名空间。一个数据库可以包含一个或多个 schema，每个 schema 只属于一个数据库。两个 schema 可以具有共享相同名称的不同对象。

您可以在同一数据库中创建多个 schema，以便按照所需的方式组织数据，或对数据进行功能分组。例如，您可以创建一个 schema 来存储所有暂存数据，并创建另一个 schema 来存储所有报告表。您还可以创建不同的 schema 来存储与同一数据库中的不同业务组相关的数据。每个 schema 都可以存储不同的数据库对象，例如表、视图和用户定义的函数 (UDF)。此外，您还可以使用 AUTHORIZATION 子句创建 schema。此子句授予指定的用户所有权，或者设置指定的 schema 可以使用的最大磁盘空间量的配额。

Amazon Redshift 会为每个新的数据库自动创建一个名为 `public` 的架构。如果在创建数据库对象时未指定 schema 名称，则这些对象会进入 `public` schema。

要访问 schema 中的对象，请使用 `schema_name.table_name` 表示法限定对象名称。schema 的限定名称由以点分隔的 schema 名称和表名称组成。例如，您可能具有一个包含 `price` 表的 `sales` 模式，以及一个同样包含 `price` 表的 `inventory` schema。当您引用 `price` 表时，您必须将其限定为 `sales.price` 或者 `inventory.price`。

以下示例为用户 GUEST 创建了一个名为 **SALES** 的 schema。

```
CREATE SCHEMA SALES AUTHORIZATION GUEST;
```

有关更多命令选项的信息，请参阅《Amazon Redshift 数据库开发人员指南》中的 CREATE SCHEMA。

要查看数据库中的 schema 列表，请运行以下命令。

```
select * from pg_namespace;
```

该输出值应该类似于以下内容。

```
  nspname             | nspowner |         nspacl
----------------------+----------+--------------------------
  sales               |  100     |
```

```
  pg_toast            |   1      |
  pg_internal         |   1      |
  catalog_history     |   1      |
  pg_temp_1           |   1      |
  pg_catalog          |   1      | {rdsdb=UC/rdsdb,=U/rdsdb}
  public              |   1      | {rdsdb=UC/rdsdb,=U/rdsdb}
  information_schema  |   1      | {rdsdb=UC/rdsdb,=U/rdsdb}
```

有关如何查询目录表的更多信息，请参阅《Amazon Redshift 数据库开发人员指南》中的查询目录表。

使用 GRANT 语句为用户授予 schema 的权限。

以下示例授予 GUEST 用户使用 SELECT 语句，在 SALES 架构中的所有表或视图中选择数据的权限。

```
GRANT SELECT ON ALL TABLES IN SCHEMA SALES TO GUEST;
```

以下示例一次性授予 GUEST 用户所有可用的权限。

```
GRANT ALL ON SCHEMA SALES TO GUEST;
```

创建表

创建新数据库后，创建表以存放您的数据。在创建表时指定列信息。

例如，运行以下命令创建一个名为 **DEMO** 的表。

```
CREATE TABLE Demo (
  PersonID int,
  City varchar (255)
);
```

默认情况下，新的数据库对象（例如表）是在数据仓库的创建期间，在名为 public 的默认架构中创建的。您可以使用另一个 schema 来创建数据库对象。有关 schema 的更多信息，请参阅《Amazon Redshift 数据库开发人员指南》中的管理数据库安全。

您还可以使用 schema_name.object_name 表示法来创建表，以在 SALES schema 中创建表。

```
CREATE TABLE SALES.DEMO (
  PersonID int,
  City varchar (255)
```

```
);
```

要查看和检查架构及其表，您可以使用 Amazon Redshift 查询编辑器 v2。或者，您可以使用系统视图查看 schema 中的表列表。有关更多信息，请参阅 查询系统表和视图。

Amazon Redshift 使用 encoding、distkey 和 sortkey 列进行并行处理。有关如何设计包含这些元素的表的更多信息，请参阅设计表的 Amazon Redshift 最佳实践。

在表中插入数据行

创建表后，向该表中插入数据行。

> **Note**
>
> INSERT 命令将行插入到表中。要进行标准的批量加载，请使用 COPY 命令。有关更多信息，请参阅使用 COPY 命令加载数据。

例如，要将值插入 DEMO 表中，运行以下命令。

```
INSERT INTO DEMO VALUES (781, 'San Jose'), (990, 'Palo Alto');
```

要对特定架构中的表插入数据，请运行以下命令。

```
INSERT INTO SALES.DEMO VALUES (781, 'San Jose'), (990, 'Palo Alto');
```

从表中选择数据

创建表并填充数据以后，可使用 SELECT 语句显示表中包含的数据。SELECT * 语句会返回表中所有数据的所有列名和行值。使用 SELECT 是验证最近添加的数据是否正确插入表中的绝佳方法。

要查看您在 **DEMO** 表中输入的数据，请运行以下命令。

```
SELECT * from DEMO;
```

结果应该类似以下内容：

```
 personid |   city
```

```
----------+-----------
      781 | San Jose
      990 | Palo Alto
(2 rows)
```

有关使用 SELECT 语句查询表的更多信息，请参阅 SELECT。

加载数据

本指南中的多个示例使用 TICKIT 示例数据集。您可以下载文件 tickitdb.zip，其中包含各个示例数据文件。然后，您可以将示例数据上传到您自己的 Amazon S3 存储桶中。

要加载数据库的示例数据，请首先创建表。然后使用 COPY 命令加载包含存储在 Amazon S3 桶中的示例数据的表。有关创建表和加载示例数据的步骤，请参阅步骤 4：将数据从 Amazon S3 加载到 Amazon Redshift。

查询系统表和视图

除了您创建的表以外，您的数据仓库还包含多个系统表和视图。这些表和视图包含有关您的安装以及系统上运行的各种查询和进程的信息。您可以查询这些系统表和视图来收集有关数据库的信息。有关更多信息，请参阅《Amazon Redshift 数据库开发人员指南》中的系统表和视图参考。每个表或视图的说明指出了表或视图是对所有用户可见还是只对超级用户可见。以超级用户身份登录以查询只对超级用户可见的表。

查看表名称列表

要查看 schema 中所有表的列表，您可以查询 PG_TABLE_DEF 系统目录表。您可以首先检查 `search_path` 的设置。

```
SHOW search_path;
```

结果应如下所示：

```
  search_path
---------------
 $user, public
```

以下示例将 SALES schema 添加到搜索路径并显示 SALES schema 中的所有标。

```
set search_path to '$user', 'public', 'sales';

SHOW search_path;

      search_path
-----------------------
 "$user", public, sales


select * from pg_table_def where schemaname = 'sales';

 schemaname | tablename |  column  |          type          | encoding | distkey |
 sortkey | notnull
------------+-----------+----------+------------------------+----------+---------
+---------+---------
 sales      | demo      | personid | integer                | az64     | f       |
  0 | f
 sales      | demo      | city     | character varying(255) | lzo      | f       |
  0 | f
```

以下示例显示当前数据库上所有 schema 中的所有称为 DEMO 的表的列表。

```
set search_path to '$user', 'public', 'sales';
select * from pg_table_def where tablename = 'demo';

 schemaname | tablename |  column  |          type          | encoding | distkey |
 sortkey | notnull
------------+-----------+----------+------------------------+----------+---------
+---------+---------
 public     | demo      | personid | integer                | az64     | f       |
  0 | f
 public     | demo      | city     | character varying(255) | lzo      | f       |
  0 | f
 sales      | demo      | personid | integer                | az64     | f       |
  0 | f
 sales      | demo      | city     | character varying(255) | lzo      | f       |
  0 | f
```

有关更多信息，请参阅 PG_TABLE_DEF。

您还可以使用 Amazon Redshift 查询编辑器 v2 查看指定架构中的所有表，方法是首先选择要连接到的数据库。

查看用户

您可以查询 PG_USER 目录来查看所有用户的列表，还可以查看用户 ID (USESYSID) 和用户权限。

```
SELECT * FROM pg_user;

  usename   | usesysid | usecreatedb | usesuper | usecatupd |  passwd  | valuntil |
 useconfig
------------+----------+-------------+----------+-----------+----------+----------
+-----------
 rdsdb      |        1 | true        | true     | true      | ******** | infinity |
 awsuser    |      100 | true        | true     | false     | ******** |          |
 guest      |      104 | true        | false    | false     | ******** |          |
```

Amazon Redshift 在内部使用用户名称 rdsdb 执行日常管理和维护任务。您可以向 SELECT 语句添加 where usesysid > 1 来筛选查询，使其只显示用户定义的用户名称。

```
SELECT * FROM pg_user WHERE usesysid > 1;

  usename   | usesysid | usecreatedb | usesuper | usecatupd |  passwd  | valuntil |
 useconfig
------------+----------+-------------+----------+-----------+----------+----------
+-----------
 awsuser    |      100 | true        | true     | false     | ******** |          |
 guest      |      104 | true        | false    | false     | ******** |          |
```

查看最近的查询

在上一示例中，adminuser 的用户 ID（user_id）为 100。要列出 adminuser 最近运行的四次查询，您可以查询 SYS_QUERY_HISTORY 视图。

您可以使用此视图查找最近运行的查询的查询 ID（query_id）或进程 ID（session_id）。您还可以使用此视图检查完成查询花了多长时间。SYS_QUERY_HISTORY 包含查询字符串（query_text）的前 4000 个字符，以便帮助您查找特定查询。在 SELECT 语句中使用 LIMIT 字句来限制结果数量。

```
SELECT query_id, session_id, elapsed_time, query_text
FROM sys_query_history
WHERE user_id = 100
ORDER BY start_time desc
LIMIT 4;
```

结果看起来如下所示。

```
 query_id |  session_id  |  elapsed_time |   query_text
----------+--------------+---------------
+-----------------------------------------------------------------
 892      |    21046     |       55868   | SELECT query, pid, elapsed, substring
 from ...
 620      |    17635     |     1296265   | SELECT query, pid, elapsed, substring
 from ...
 610      |    17607     |       82555   | SELECT * from DEMO;
 596      |    16762     |      226372   | INSERT INTO DEMO VALUES (100);
```

确定运行的查询的会话 ID

要检索关于该查询的系统表信息，您可能需要指定与查询关联的会话 ID（进程 ID）。或者，您可能需要查找仍在运行的查询的会话 ID。例如，如果您需要取消在预置集群上运行时间过长的查询，就需要会话 ID。您可以通过查询 STV_RECENTS 系统表获取正在运行的查询的会话 ID 列表，以及相应的查询字符串。如果查询返回多个会话，您可以通过查看查询文本确定所需会话 ID。

要确定正在运行的查询的会话 ID，请运行以下 SELECT 语句。

```
SELECT session_id, user_id, start_time, query_text
FROM sys_query_history
WHERE status='running';
```

取消查询

如果您的查询运行时间过长或消耗过多资源，请取消该查询。例如，创建一个门票卖家列表，其中包含卖家名称和售出门票数。下面的查询从 SALES 表和 USERS 表中选择数据，并通过在 WHERE 子句中匹配 SELLERID 和 USERID 联接这两个表。

```
SELECT sellerid, firstname, lastname, sum(qtysold)
FROM sales, users
WHERE sales.sellerid = users.userid
GROUP BY sellerid, firstname, lastname
ORDER BY 4 desc;
```

结果看起来如下所示。

```
 sellerid | firstname | lastname | sum
```

```
----------+-----------+----------+------
  48950   |   Nayda   |   Hood   | 184
  19123   |   Scott   | Simmons  | 164
  20029   |    Drew   | Mcguire  | 164
  36791   |  Emerson  | Delacruz | 160
  13567   |   Imani   |   Adams  | 156
  9697    |  Dorian   |    Ray   | 156
  41579   | Harrison  | Durham   | 156
  15591   |  Phyllis  |  Clay    | 152
  3008    |  Lucas    | Stanley  | 148
  44956   |  Rachel   |Villarreal| 148
```

> **Note**
>
> 这是一个复杂的查询。对于本教程，您无需关注此查询的构造方式。

上一查询运行几秒钟并返回 2102 行。

假设您忘记放入 WHERE 子句。

```
SELECT sellerid, firstname, lastname, sum(qtysold)
FROM sales, users
GROUP BY sellerid, firstname, lastname
ORDER BY 4 desc;
```

结果集行数将为 SALES 表中的所有行数乘以 USERS 表中的所有行数 ((49989*3766)。这称为笛卡尔联接，我们不建议使用。结果超过 1.88 亿行，并且运行时间很长。

要取消正在运行的查询，请使用 CANCEL 命令并提供查询的会话 ID。使用 Amazon Redshift 查询编辑器 v2，在查询运行时，您可以通过选择取消按钮来取消查询。

要查找会话 ID，请启动新会话并查询 STV_RECENTS 表，如上一步所示。以下示例显示了如何使结果更易读。为此，请使用 TRIM 函数去除尾随空格并只显示查询字符串的前 20 个字符。

要确定正在运行的查询的会话 ID，请运行以下 SELECT 语句。

```
SELECT user_id, session_id, start_time, query_text
FROM sys_query_history
WHERE status='running';
```

结果看起来如下所示。

```
 user_id |   session_id  |   start_time                |   query_text
---------+---------------+-----------------------------
+-------------------------------------------------------------
 100     |    1073791534 | 2024-03-19 22:26:21.205739 | SELECT user_id, session_id,
 start_time, query_text FROM  ...
```

要取消会话 ID 为 1073791534 的查询，请运行以下命令。

```
CANCEL 1073791534;
```

> **Note**
>
> CANCEL 命令不会停止事务。要中止或回滚事务，请使用 ABORT 或 ROLLBACK 命令。要取消与某一事务关联的查询，应先取消查询，然后再中止事务。

如果取消的查询与某一事务关联，请使用 ABORT 或 ROLLBACK 命令取消事务并放弃对数据进行的所有更改：

```
ABORT;
```

除非您以超级用户身份登录，否则只能取消您自己的查询。超级用户可以取消所有查询。

如果您的查询工具不支持并发运行查询，则另外启动一个会话来取消查询。

有关取消查询的更多信息，请参阅《Amazon Redshift 数据库开发人员指南》中的 CANCEL。

使用超级用户队列取消查询

如果您的当前会话有太多查询同时运行，则您可能要等另一个查询完成之后才能运行 CANCEL 命令。在这种情况下，使用不同的工作负载管理查询队列运行 CANCEL 命令。

通过使用工作负载管理，您可以运行不同查询队列中的查询，这样就无需等待另一查询完成。工作负载管理程序会创建一个单独的队列，称为超级用户队列，可以用来进行故障排除。要使用超级用户队列，以超级用户身份登录，并使用 SET 命令将查询组设置为“superuser”。运行您的命令后，可使用 RESET 命令重置查询组。

要使用超级用户队列取消查询，请运行以下命令。

```
SET query_group TO 'superuser';
CANCEL 1073791534;
RESET query_group;
```

查询数据不在 Amazon Redshift 中

在下文中，您可以了解如何开始查询远程源上的数据，这包括 Amazon S3 数据、远程数据库管理器、远程 Amazon Redshift 数据库，以及使用 Amazon Redshift 训练机器学习（ML）模型。

主题

查询数据湖

您可以使用 Amazon Redshift Spectrum 在 Amazon S3 文件中查询数据，而不必将数据加载到 Amazon Redshift 表中。Amazon Redshift 提供了 SQL 功能，专为对存储在 Amazon Redshift 集群和 Amazon S3 数据湖中的超大型数据集进行快速在线分析处理（OLAP）而设计。您可以查询多种格式的数据，包括 Parquet、ORC、RCFile、TextFile、SequenceFile、RegexSerde、OpenCSV 和 AVRO。您可以创建外部架构和表以定义 Amazon S3 中文件的结构。然后，您可以使用外部数据目录，如 AWS Glue 或您自己的 Apache Hive 元存储。对数据目录类型进行的更改将立即对您的任何 Amazon Redshift 集群可用。

在您的数据注册到 AWS Glue Data Catalog 并启用 AWS Lake Formation 后，您可以使用 Redshift Spectrum 查询它。

Redshift Spectrum 驻留在独立于您的集群的专用 Amazon Redshift 服务器上。Redshift Spectrum 将很多计算密集型任务（如谓词筛选和聚合）推送到 Redshift Spectrum 层。Redshift Spectrum 还可以通过智能方式扩展，以利用大规模并行处理。

您可在一个或多个列上对外部表进行分区，以通过消除分区来优化查询性能。您可以使用 Amazon Redshift 表查询和联接外部表。您可以从多个 Amazon Redshift 集群中访问外部表并在同一 AWS 区域的任何集群中查询 Amazon S3 数据。更新 Amazon S3 数据文件后，立即可从您的任何 Amazon Redshift 集群查询到该数据。

有关 Redshift Spectrum 的更多信息，包括如何使用 Redshift Spectrum 和数据湖，请参阅 Amazon Redshift 数据库开发人员指南中的开始使用 Amazon Redshift Spectrum。

查询远程数据库管理器上的数据

您可以使用联合查询，将 Amazon RDS 数据库和 Amazon Aurora 数据库中的数据，与 Amazon Redshift 数据库中的数据相联接。您可以使用 Amazon Redshift 直接查询操作数据（无需移动数据）、应用转换以及将数据插入 Redshift 表中。联合查询的某些计算将分配到远程数据源。

要运行联合查询，Amazon Redshift 首先建立到远程数据源的连接。然后，Amazon Redshift 会检索有关远程数据源中表的元数据，发出查询，然后检索结果行。然后，Amazon Redshift 将结果行分配给 Amazon Redshift 计算节点，以进行进一步处理。

有关为联合查询设置环境的信息，请参阅 Amazon Redshift 数据库开发人员指南中的下列主题之一：

- 开始使用对 PostgreSQL 的联合查询
- 开始使用对 MySQL 的联合查询

访问其他 Amazon Redshift 数据库中的数据

使用 Amazon Redshift 数据共享，您可以在 Amazon Redshift 集群或 AWS 账户之间安全、轻松地共享实时数据，以用于读取目的。您可以即时、精细、高性能地访问 Amazon Redshift 集群中的数据，而无需手动复制或移动数据。您的用户可以在 Amazon Redshift 集群中查看更新的最新、最一致的信息。您可以在不同级别共享数据，例如数据库、架构、表、视图（包括常规视图、后期绑定视图和具体化视图）和 SQL 用户定义函数（UDF）。

Amazon Redshift 数据共享对于以下使用案例尤其有用：

- 集中业务关键型工作负载 – 使用与多个业务情报 (BI) 或分析集群共享数据的中央提取、转换和加载 (ETL) 集群。此方法提供读取工作负载隔离和单个工作负载的退款。
- 在环境之间共享数据 – 在开发环境、测试环境和生产环境之间共享数据。您可以在不同的粒度级别下共享数据，以提高团队敏捷性。

有关数据共享的更多信息，请参阅《Amazon Redshift 数据库开发人员指南》中的管理数据共享任务。

使用 Amazon Redshift 数据训练机器学习模型

使用 Amazon Redshift 机器学习 (Amazon Redshift ML)，您可以通过向 Amazon Redshift 提供数据来训练模型。然后，Amazon Redshift ML 将创建模型来捕获输入数据中的模式。接着，您可以使用这些模型为新输入数据生成预测，而不会产生额外的成本。通过使用 Amazon Redshift ML，您可以使用

SQL 语句训练机器学习模型，并在 SQL 查询中调用它们以进行预测。您可以通过迭代更改参数和改进训练数据来继续提高预测的准确性。

Amazon Redshift ML 使 SQL 用户能够更轻松地使用熟悉的 SQL 命令创建、训练和部署机器学习模型。借助 Amazon Redshift ML，您可以使用 Amazon Redshift 集群中的数据，通过 Amazon SageMaker AI Autopilot 训练模型并自动获得最佳模型。然后，您可以对模型进行本地化，并从 Amazon Redshift 数据库中进行预测。

有关 Amazon Redshift ML 的更多信息，请参阅《Amazon Redshift 数据库开发人员指南》中的开始使用 Amazon Redshift ML。

了解 Amazon Redshift 概念

Amazon Redshift Serverless 让您可以访问和分析数据，而无需对预置数据仓库执行任何配置操作。系统将自动预置资源，数据仓库的容量会智能扩展，即使面对要求最为苛刻且不可预测的工作负载也能提供高速性能。数据仓库空闲时不会产生费用，您只需为实际使用的资源付费。您可以在 Amazon Redshift 查询编辑器 v2 或您最喜欢的商业智能（BI，Business Intelligence）工具中，直接加载数据并开始查询。在易于使用且无需承担管理任务的环境中，享受最佳性价比，使用熟悉的 SQL 功能。

如果您是首次接触 Amazon Redshift 的用户，我们建议您先阅读以下部分：

- Amazon Redshift Serverless 功能概览 – 本主题概要介绍了 Amazon Redshift Serverless 及其关键功能。
- 服务亮点和定价 – 在该产品详细信息页面上，您可以找到有关 Amazon Redshift Serverless 的亮点和定价的详细信息。
- Amazon Redshift Serverless 数据仓库入门 – 在本主题中，您将详细了解如何创建 Amazon Redshift Serverless 数据仓库，以及如何使用查询编辑器 v2 开始查询数据。

如果您希望手动管理 Amazon Redshift 资源，则可以创建预置集群来满足自己的数据查询需求。有关更多信息，请参阅 Amazon Redshift 集群。

如果您的组织符合条件，而在您创建集群的 AWS 区域中，Amazon Redshift Serverless 不可用，则您也许可以通过 Amazon Redshift 免费试用计划创建集群。请选择生产或免费试用来回答问题您打算将此集群用于什么？ 选择免费试用时，您将创建具有 dc2.large 节点类型的配置。有关选择免费试用的更多信息，请参阅 Amazon Redshift 免费试用。有关提供了 Amazon Redshift Serverless 的 AWS 区域的列表，请参阅《Amazon Web Services 一般参考》中针对 Redshift Serverless API 列出的 Amazon Redshift 端点。

以下是 Amazon Redshift Serverless 的一些关键概念：

- 命名空间 – 数据库对象和用户的集合。命名空间将您在 Amazon Redshift Serverless 中使用的所有资源组合在一起，例如架构、表、用户、数据共享和快照。
- 工作组 – 计算资源的集合。工作组存放 Amazon Redshift Serverless 运行计算任务所用的计算资源。这些资源的示例包括 Redshift 处理单元（RPU，Redshift Processing Unit）、安全组、使用限制。您可以使用 Amazon Redshift Serverless 控制台、AWS Command Line Interface 或 Amazon Redshift Serverless API 来配置工作组的网络和安全设置。

有关配置命名空间和工作组资源的更多信息，请参阅使用命名空间和使用工作组。

以下是 Amazon Redshift 预置集群的一些关键概念：

- 集群 – 集群是 Amazon Redshift 数据仓库的核心基础设施组件。

  集群包含一个或多个计算节点。这些计算节点运行编译后的代码。

  如果集群预置有两个或更多计算节点，则一个额外的领导节点将协调这些计算节点。领导节点处理与应用程序的外部通信，例如商业智能工具和查询编辑器。您的客户端应用程序仅直接与领导节点交互。计算节点对于外部应用程序是透明的。

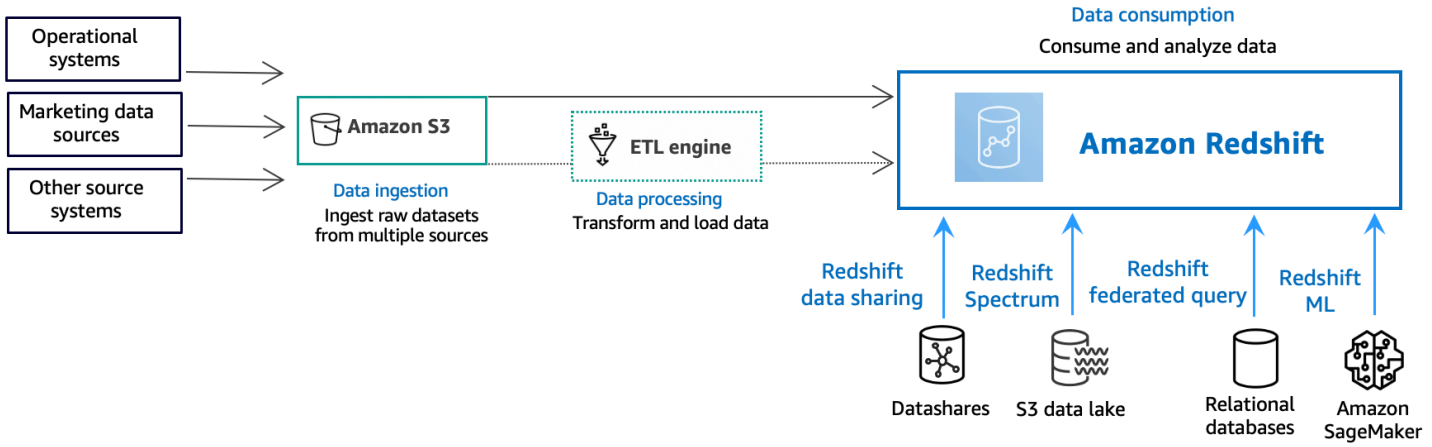
- 数据库 – 一个集群包含一个或多个数据库。

  用户数据存储在计算节点上的一个或多个数据库中。您的 SQL 客户端与领导节点进行通信，进而通过计算节点协调查询运行。有关计算节点和领导节点的详细信息，请参阅数据仓库系统架构。在数据库中，用户数据被组织成一个或多个架构。

  Amazon Redshift 是一个关系数据库管理系统（RDBMS），可与其它 RDBMS 应用程序兼容。虽然它提供了与典型 RDBMS 相同的功能，包括联机事务处理（OLTP）功能，例如，插入和删除数据。Amazon Redshift 还针对数据集的高性能批量分析和报告进行了优化。

接下来，您可以在 Amazon Redshift 中找到对典型数据处理流程的描述以及流程不同部分的描述。有关 Amazon Redshift 系统架构的更多信息，请参阅数据仓库系统架构。

以下示意图说明 Amazon Redshift 中典型的数据处理流程。



Amazon Redshift 数据仓库是一个企业级的关系数据库查询和管理系统。Amazon Redshift 支持与多种类型的应用程序（包括业务情报 (BI)、报告、数据和分析工具）建立客户端连接。在运行分析查询时，您将在多阶段操作中检索、比较和计算大量数据以产生最终结果。

在数据摄取层，不同类型的数据源会持续将结构化、半结构化或非结构化数据上载到数据存储层。该数据存储区用作暂存区，用于存储处于不同消费准备状态的数据。Amazon Simple Storage Service（Amazon S3）存储桶就是这种存储的示例。

在可选数据处理层，源数据使用提取、转换、加载（ETL）或提取、加载、转换（ELT）管道进行预处理、验证和转换。然后，使用 ETL 操作对这些原始数据集进行优化。ETL 引擎的一个示例是 AWS Glue。

在数据使用层，数据将加载到 Amazon Redshift 集群中，您可以在其中运行分析工作负载。

有关分析工作负载的示例，请参阅查询外部数据来源。

用以了解有关 Amazon Redshift 的其他资源

如需了解有关 Amazon Redshift Serverless 的更多信息，我们建议您使用以下 Amazon Redshift 资源，继续详细了解本指南中介绍的概念：

- 功能视频：这些视频可帮助您了解 Amazon Redshift 功能。
  - 要全面地了解 Amazon Redshift Serverless，请观看以下视频。Amazon Redshift Serverless 90 秒简介。
  - 要了解如何设置无服务器数据仓库和开始查询数据，请观看以下视频。Amazon Redshift Serverless 入门。
- Amazon Redshift 管理指南：本指南基于此《Amazon Redshift 入门指南》构建。它针对创建、管理以及监控 Amazon Redshift Serverless 和 Amazon Redshift 预置集群的概念和任务，提供了全面详实的信息。
- Amazon Redshift 数据库开发人员指南：本指南也基于此《Amazon Redshift 入门指南》构建。它为数据库开发人员提供全面详实的信息，帮助他们了解如何设计、构建、查询和维护构成数据仓库的数据库。
  - SQL 参考：本主题介绍 Amazon Redshift 的 SQL 命令和函数引用。
  - 系统表和视图参考：本主题介绍 Amazon Redshift 的系统表和视图。
- Amazon Redshift 教程：本主题展示了有关 Amazon Redshift 功能的教程。
  - 从 Amazon S3 加载数据：本教程介绍了如何从 Amazon S3 存储桶中的数据文件将数据加载到 Amazon Redshift 数据库表中。
  - 数据共享入门：此部分介绍如何共享和访问其他 Amazon Redshift 集群中的数据。
  - 将空间 SQL 函数与 Amazon Redshift 一起使用：本教程演示了如何在 Amazon Redshift 中使用某些空间 SQL 函数。
  - 使用 Amazon Redshift Spectrum 查询嵌套数据：本教程介绍了使用 Redshift Spectrum 通过外部表查询 Parquet、ORC、JSON 和 Ion 文件格式的嵌套数据。
  - 配置手动工作负载管理 (WLM) 队列：本教程介绍如何在 Amazon Redshift 中配置手动工作负载管理 (WLM)。
  - Amazon Redshift ML 入门：此部分向用户介绍如何使用熟悉的 SQL 命令创建、训练和部署机器学习模型。
- 新增功能：此网页列出了 Amazon Redshift 的新功能和产品更新。

文档历史记录

> **Note**
>
> 有关 Amazon Redshift 中的新特征的描述，请参阅新增特征。

下表介绍对《Amazon Redshift 入门指南》的重要文档更改。

更改	描述	发行日期
文档更新	更新了指南，纳入了有关开始使用常见数据库任务、查询数据湖、查询远程源上的数据、共享数据以及使用 Amazon Redshift 数据训练机器学习模型的新部分。	2021 年 6 月 30 日
新功能	更新了指南，对新的示例加载过程进行了描述。	2021 年 6 月 4 日
文档更新	更新了指南，删除了原始 Amazon Redshift 控制台并改进步骤流程。	2020 年 8 月 14 日
新控制台	更新了指南以描述新的 Amazon Redshift 控制台。	2019 年 11 月 11 日
新功能	更新了指南，描述了快速启动集群过程。	2018 年 8 月 10 日
新功能	更新了指南，介绍如何从 Amazon Redshift 控制面板启动集群。	2015 年 7 月 28 日
新功能	更新了指南，介绍如何使用新的节点类型名称。	2015 年 6 月 9 日
文档更新	更新了配置 VPC 安全组的屏幕截图和程序。	2015 年 4 月 30 日
文档更新	更新了屏幕截图和程序，以便与当前控制台匹配。	2014 年 11 月 12 日
文档更新	将从 Amazon S3 加载数据的信息移到单独一部分，将后续步骤部分移到最后一个步骤，让内容更醒目。	2014 年 5 月 13 日
文档更新	删除了“欢迎”页面，将这部分内容并入“入门”主页面。	2014 年 14 月 3 日

更改	描述	发行日期
文档更新	根据客户反馈和服务更新全新发布的 Amazon Redshift 入门指南。	2014 年 14 月 3 日
新指南	这是 Amazon Redshift 入门指南 的第一个版本。	2013 年 2 月 14 日