AWS Well-Architected Framework

Financial Services Industry Lens

Financial Services Industry Lens: AWS Well-Architected Framework

Table of Contents

Financial Services Industry Lens - AWS Well-Architected Framework

Publication date: **May 15, 2024** (*Document revisions*)

This document describes the Financial Services Industry Lens for the AWS Well-Architected Framework. The document describes general design principles, as well as specific best practices and guidance for the six pillars of the Well-Architected Framework.

Introduction

The financial services industry includes financial services firms, independent software vendors (ISVs), market utilities, and infrastructures that supply essential services to countries around the world. The industry consists of organizations that provide the main mechanisms for:

- Paying for goods and services
- Financial markets and asset trading
- Serving as intermediates between savers and borrowers (channeling savings into investment)
- Insuring against and dispersing risk

The AWS Well-Architected Framework helps you understand the pros and cons of decisions you make while building systems on AWS. By using the Framework, you learn architectural best practices for designing and operating reliable, secure, efficient, cost-effective, and sustainable systems in the cloud. The Framework provides a way for you to consistently measure your architectures against best practices and identify areas for improvement. We believe that having well architected systems greatly increases your security, reliability, and the likelihood of business success.

In this lens, we focus the Well-Architected Framework on how to design, deploy, and architect financial services industry (FSI) workloads that promote the resiliency, security, cost savings, and operational performance in line with risk and control objectives that you define, including those that help you align with the regulatory and compliance requirements of supervisory authorities.

All customers should begin with the best practices and questions outlined in the AWS Well-Architected Framework whitepaper. This document provides additional best practices that are

focused on the technical architectures and workloads that are associated with financial services institutions.

The Financial Services Industry Lens identifies best practices for security, data privacy, and resiliency that are intended to address the requirements of financial institutions based on our experience working with financial institutions worldwide. It provides guidance on guardrails for technology teams to implement and confidently use AWS to build and deploy applications. This Lens describes the process of building transparency and auditability into your AWS environment. It also offers suggestions for controls to help you expedite adoption of new services into your environment while managing the cost of your IT services.

This document is intended for those in technology leadership roles, such as chief technology officers (CTOs), architectural leadership, developers, engineers, and operations team members, as well as individuals in the risk, compliance, and audit functions.

Lens availability

The Financial Services Industry Lens is available as an AWS-official lens in the Lens Catalog of the AWS Well-Architected Tool.

To get started, follow the steps in Adding a lens to a workload and select the **Financial Services Industry Lens**.

Design principles

The Well-Architected Framework identifies a set of four general design principles to facilitate good design in the cloud for financial services workloads.

1. **Documented operational planning**—To define your cloud-operating model, you must work with internal consumers and stakeholders to set a common goal and strategic direction. Many organizations have adopted the "Three Lines of Defense" model to improve effectiveness of risk management:
   - At the first line of defense, operational managers are responsible for initiating risk and control procedures on a day-to-day basis.
   - The second line establishes various risk management and compliance functions to help build and/or monitor the first line-of-defense controls.
   - As the third line of defense, internal auditors provide the governing body and senior management with comprehensive assurance based on the highest level of independence and objectivity within the organization.

   Establishing clear roles and responsibilities across the three lines of defense is vital to developing an effective operating model for regulated cloud adoption, see Three Lines of Defense from the Institute of Internal Auditors (IIA).
2. **Automated infrastructure and application deployment**—Automation enables you to perform and innovate quickly and scale security, compliance, and governance activities across your cloud environments. Financial services institutions that invest in automated infrastructure and application deployment are able to accelerate the rate of deployments and more simply embed security and governance best practices into their software development lifecycle.
3. **Security by design**—Financial services institutions must consider Security by Design (SbD) approach to implement architectures that are pre-tested from a security perspective. SbD helps implement the control objectives, security baselines, security configurations, and audit capabilities for applications running on AWS. Standardized, automated, prescriptive, and repeatable design templates help accelerate the deployment of common use cases as well as help align with security standards (and ease the evidence requirements for audit) across multiple workloads. For example, to protect customer data and mitigate the risk of data disclosure or alteration of sensitive information by unauthorized parties, financial institutions need to employ encryption and carefully manage access to encryption keys. SbD allows you to turn on encryption for data at rest, in transit, and if necessary, at the application level by default.

4. **Automated governance**—Humans working with runbooks and checklists often lead to delays and inaccurate results. Automated governance provides a fast, definitive governance check for applications deployment at scale. Governance at scale typically addresses the following components:
   - **Account management:** Automate account provisioning and maintain good security when hundreds of users and business units are requesting cloud-based resources.
   - **Budget and cost management:** Enforce and monitor budgets across many accounts, workloads, and users.
   - **Security and compliance automation:** Manage security, risk, and compliance at scale to verify that the organization maintains compliance, while performing against business objectives.

Scenarios

The following are common scenarios that influence the design and architecture of your financial services workloads on AWS. Each scenario includes the common drivers for the design and a reference architecture.

Topics

Financial data mesh

A commonly sought-after goal of financial services organizations is to provide access to data and to extract additional value from data that is generated or acquired across their multiple business units. For example, historical market data, alternative investment data, transaction and business process data, and third-party data sets can be combined to provide for analytics and training machine learning models.

The term *data mesh* refers to any architectural framework that enables access to a diverse set of data across the enterprise through a distributed and decentralized ownership model. A data mesh architecture effectively unites disparate data sources through centrally managed data sharing and governance guidelines. A data mesh can be used to improve data access while providing enhanced security and scalability for an enterprise. The following data mesh reference architecture is built around the following architectural principles:

- Distributed domain-driven architecture: Data management responsibility that is organized around a set of business functions or domains which are responsible for managing the lifecycle of their datasets.

- Data as a product: Each domain team manages their datasets as a product, meaning that the data is organized in a way that matches the way users consume the data. Each dataset is trustworthy, describes itself, and is fit for purpose.
- Federated data governance: Security is implemented as a shared responsibility within the organization; global standards and policies apply across domains, while each domain has its own degree of autonomy on standards and policies within the domain.
- Common access and self-serve data: Data must be quickly discoverable and consumable by subject matter experts (SMEs).

Reference architecture

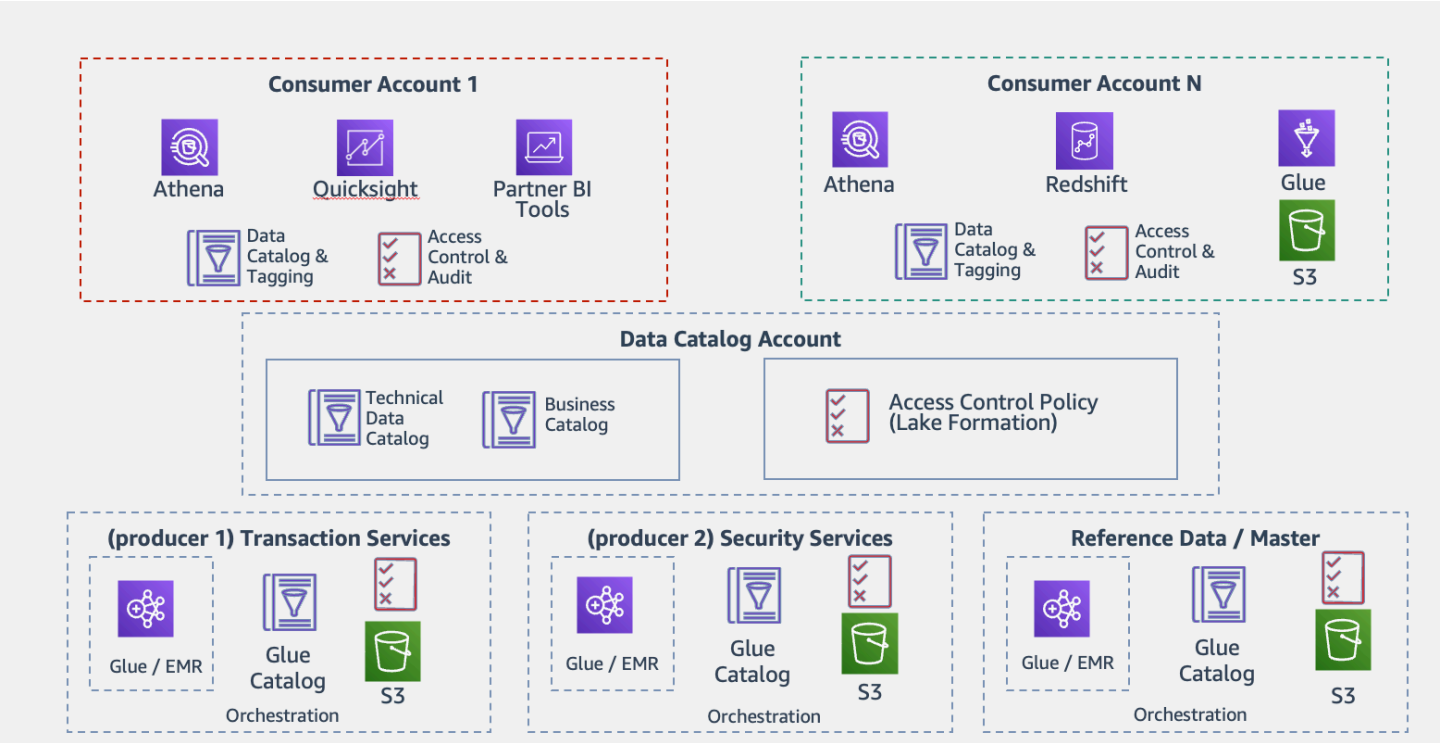

Figure 1. Financial data mesh

Architecture description

- **Producer accounts:** Business domains manage the lifecycle of their datasets in their own AWS accounts, including ETL, security, retention, and backup.
- **Catalog account:**
  - Business domains provide access to prepared datasets to a centralized catalog and access management account where Lake Formation is used to access business domain datasets.

  - The centralized catalog account manages access to business domain datasets by defining access policies to datasets from consumer accounts through Lake Formation cross-account data sharing.
- **Consumer accounts:** Data lake administrators in the consumer accounts use Lake Formation to manage granular access policies within their own account.

Artificial intelligence and machine learning

Financial institutions have used artificial intelligence and machine learning (AI/ML) technologies for years. Today, financial services organizations are harnessing the power of AI/ML to improve surveillance, reduce fraud, mitigate risk and improve compliance, enhance customer interactions, and improve operational efficiency.

Characteristics of AI/ML applications in the financial services domain

Integration of AI/ ML technologies into day-to-day operations has advanced slowly due to a lack of in-house data science and machine learning operations (MLOps) expertise and insufficient tools and services orchestrating these complex workflows. AWS provides a set of tools that make AI/ML readily accessible to any organization. Financial institutions have the following common design requirements in order to make AI/ML workloads successful in their organizations:

- **Secure ML environment**: Financial institutions have stringent security requirements for several reasons, including data protection, regulatory compliance, prevention of adversarial exploits, and to maintain trust and responsible use of AI.
- **Self-service ML capabilities:** Customers using these AWS services can enable both technical and non-technical domain experts to employ Machine Learning to foster a culture of data-driven decision-making throughout the organization.
- **Continuous integration and delivery (CI/CD):** Automate the deployment process to make it easier to roll out models into production environments and provide version control models and code artifacts.
- **Monitor ML models:** CI/CD pipelines enable continuous monitoring of deployed models, allowing teams to gather feedback, verify auditability, track performance, and make necessary adjustments.

Reference architecture

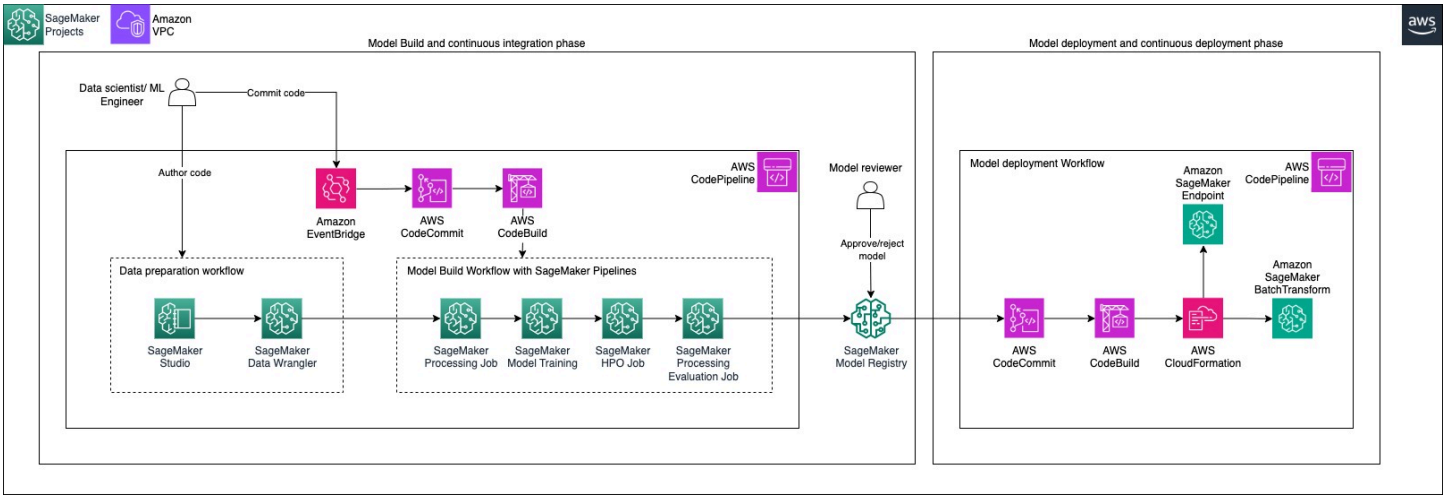

Figure 2: Reference architecture for an AI/ML pipeline

AI/ML architecture description

Business requirements phase: Define the functional requirements of the workload identifying the business problem and the desired outcomes of the AI system. Then frame the business problem by analyzing what the AI/ML application solves, what behaviors are observed, and what information should be predicted.

ML infrastructure phase: Integrate Amazon SageMaker AI with AWS networking and security services including Amazon Virtual Private Cloud (Amazon VPC), Identity and access management (IAM), AWS Key Management Service (KMS) to provide a secure environment for end-to-end machine learning workloads.

Continuous Integration Phase: SageMaker AI facilitates CI/CD by providing features like SageMaker AI Pipelines and SageMaker AI Studio. SageMaker AI Projects allows the MLOps teams to create a standardized ML experimentation environment, leverage libraries, source control repositories, and CI/CD pipelines. Data scientists can take advantage of AWS services like CodeBuild and CodeDeploy to automate the following workflows:

1. **Data preparation workflow:** Data is collected and cleaned, removing inconsistencies or errors. Then, features are selected or engineered, and the data is split into training and testing sets, to provide quality and suitability of the data for the machine learning model.

1. **Model Build Workflow:** The model build and evaluation workflow in machine learning involves two main steps. First, a model is built using a training dataset, using SageMaker AI Training, where the algorithm learns patterns and relationships from the data. Then, the model's

performance is evaluated using a separate testing dataset to assess its predictive capabilities and generalization to new and unseen data. Customers can use SageMaker AI HPO for hyper-parameter tuning in complex machine learning systems, such as deep learning neural networks, enhancing productivity by systematically exploring various combinations of hyperparameter values within specified ranges to automatically identify the best model.

Continuous delivery phase: SageMaker AI MLOps capability automates deploying and delivering machine learning models into production in a consistent manner. ML operations teams can leverage AWS continuous integration capabilities using AWS CloudFormation, CodeBuild, and CodeDeploy to automate model deployment workflows. Amazon SageMaker AI model monitoring allows customers to monitor ML applications for potential data drifts, model drifts, and bias drifts.

Model performance evaluation - evaluate the performance and accuracy of the machine learning model. Feed model drift and errors back into the model to correct it and generate more precise inferences.

Resources

Documentation

- Automate Machine Learning Workflows Tutorial

Blogs

- Building, automating, managing, and scaling ML workflows using Amazon SageMaker AI Pipelines
- Detect NLP data drift using custom Amazon SageMaker AI Model Monitor

Workshops

- Amazon SageMaker AI MLOps Workshop
- Amazon SageMaker AI MLOps: from idea to production in six steps

Cyber event recovery in financial services

Cyber threats are a growing risk to financial services organizations worldwide, and the trend is only ever increasing. These organizations are now investing increasingly on cybersecurity measures

to improve their risk posture and to implement better security practices to protect their most critical data and applications from external threats, such as ransomware and malware, and also to meet any regulatory requirements where they operate. FSI organizations are investing in building out modern cyber event recovery platforms on AWS using native AWS services as shown in the following reference architecture.

Reference architecture

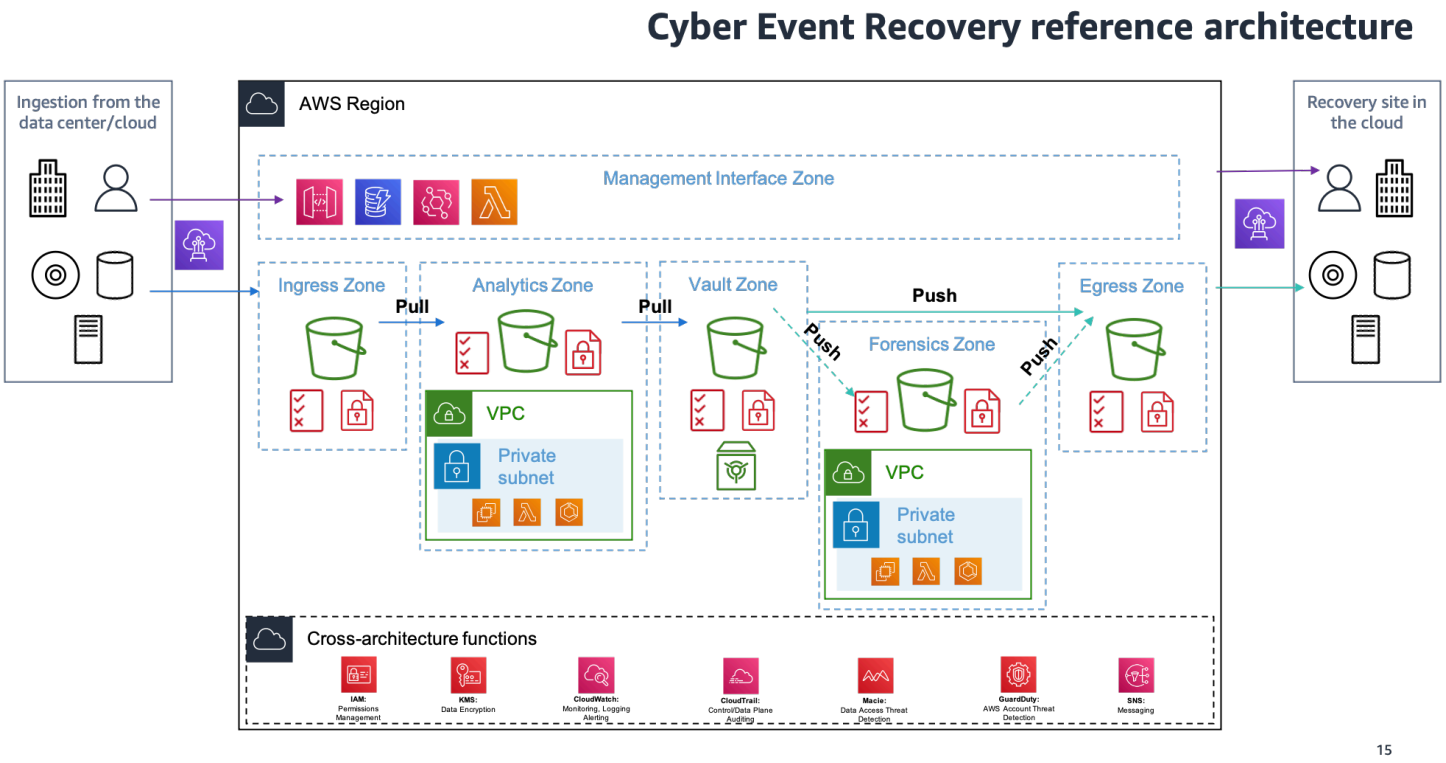

Figure 3. Cyber event recovery

Architecture description

The following components are built out as part of the cyber event data vault on AWS.

- **Ingress zone:** The raw data from the input source is first copied and stored in this zone. This zone contains different ways of sourcing data and storing it in an encrypted S3 bucket with the right security controls using IAM. It is ephemeral in nature to provide a digital air gap to the vault architecture.
- **Analytics zone:** The raw data needs to be analyzed to help prevent the transmission of corrupt data to the cyber vault. You can use services such as Amazon Macie to identify corrupt data, or write your own custom logic using AWS Lambda functions.

- **Vault zone:** Once analyzed for corruption, data is then stored in a write once, read many (WORM) compliant storage, where the data cannot be modified by anyone upon being written. This data is safe to be consumed in the event of a ransomware incident.
- **Forensics zone:** In the event of a ransomware incident, data from the vault zone can be further analyzed for anomalies before being used for recovery purposes. This is an optional step for organizations that are looking to perform more due diligence prior to the recovery process.
- **Egress zone:** The recovery process can recover the data from the vault through the egress zone. By having a separate ingress and egress zone, the vault can be secured from outside access, only providing access to the services that need it. This zone, similar to the ingress zone, is ephemeral in nature to provide a digital air gap to the vault architecture.
- **Management interface zone:** The main interface layer with the data vault, which is used to authenticate access requests, management actions, and provide the relevant status and reporting information.

For more detail, see Banking Trends 2022: Cyber vault and Ransomware.

Open banking

In open banking, banks use an API messaging framework to securely share their customer data (with consent from customers) to third-party developers and service providers, which allows for automated and secure access to the data in their core banking environment. While open banking initially started as a regulatory requirement in the United Kingdom (UK) and other regions around the world, it has now transformed into a new revenue stream for banks, as they look to monetize their data and core functionality by exposing their core environment through APIs and building new business models such as Banking as a Service (BaaS) and embedded finance on top of the APIs. Banks often choose AWS to build their open banking environment because of its inherent scalability, cost effectiveness, and the speed at which they can build. Open banking architectures supporting these use cases share the following characteristics:

- Data is shared to third parties only after consent from the customer using OAuth 2.0.
- Secure and limited third party access (with mutual Transport Layer Security (mTLS)).
- API-driven infrastructure and an elastic and scalable environment.
- Instant or near-instant access to customer account data.
- Tamper-resistant logging and audit capabilities.

Reference architecture

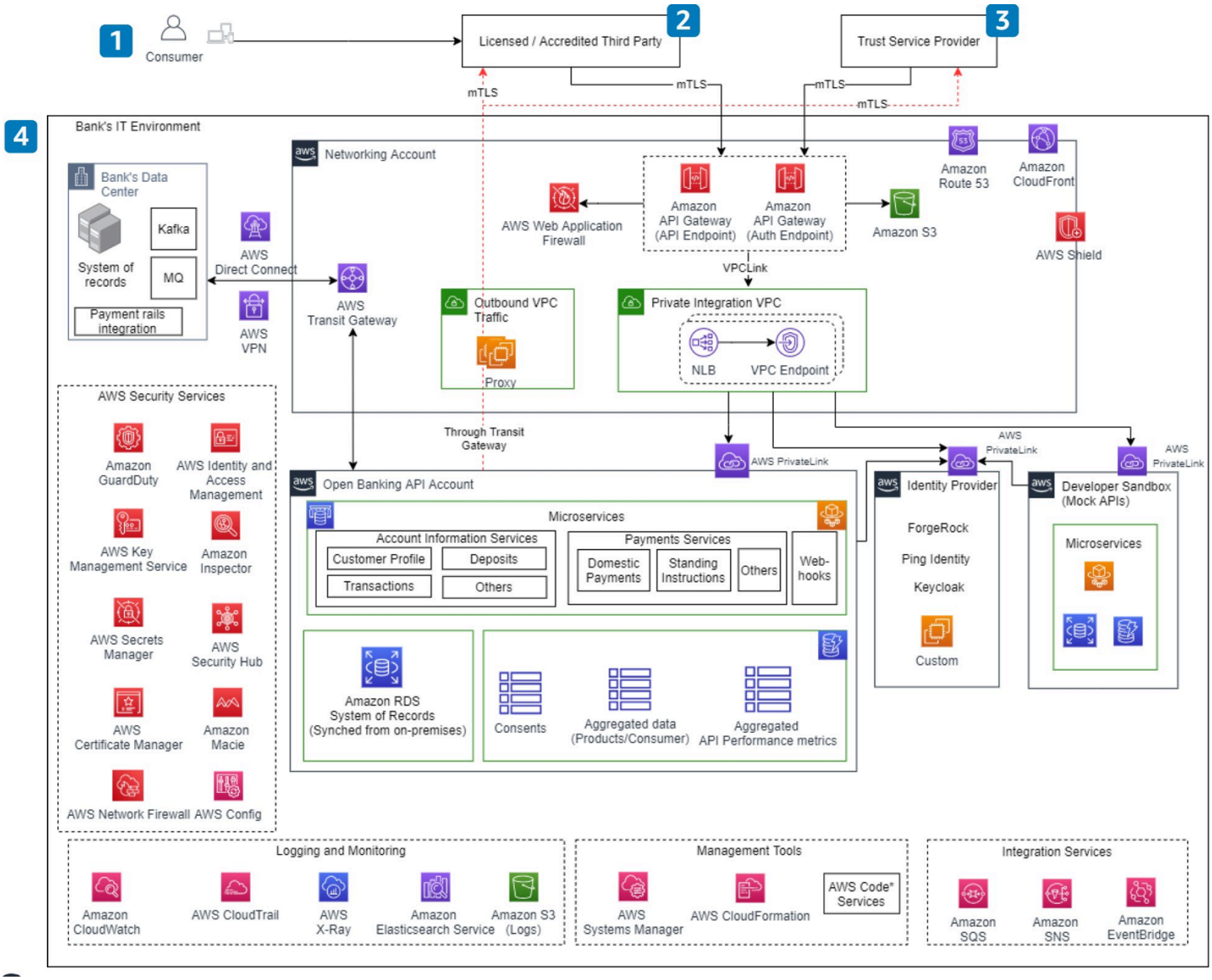

Figure 4. Reference architecture for data holder

Architecture description

1. A consumer accesses the licensed or accredited third-party application and provides consent to the third party to access consumer data or make a payment submission request.
2. Third parties in open banking can be defined as authorized institutions that provide value-added services in addition to the consumer's regular banking needs, such as accounts information (balance check, recent transactions, and statements) and payments (payment to merchants, people, and registered payees). This approach creates use cases such as spend analysis, credit decisioning, and payments for e-commerce transactions.

3. A trust service provider (TSP) is a trusted entity authorized by a supervisory government body to verify the authenticity of banks and third parties and issue digital certificates to third parties.
4. A bank's IT environment, consisting of its AWS environment and data centers, is depicted in this section. Note the breadth of AWS services that are available for banking customers.

For more information on open banking, see Open Banking on AWS.

Payments

Payment gateways facilitate financial services to make online transactions between customers and merchants. To secure these transactions, payment gateways rely on secure cloud technology. AWS provides a secure and reliable environment for payment processing and storage of payment card information, as well as a number of encryption options for sensitive data, including encryption at rest and in transit.

Encryption at rest protects stored data against unauthorized access or theft, while encryption in transit protects data as it is being transmitted between systems. AWS is certified as a PCI DSS Level Service Provider, the highest level of assessment available, which means that businesses are meeting the highest standards of security with compliance when it comes to handling credit card data.

Payment gateways can use tokenization to protect customer data by replacing the customer card data with a unique token. This token can be used for future transactions without the need to store the actual card details, making it a more secure option for customer data. Payment gateways also help merchants detect and prevent fraudulent transactions using artificial intelligence and machine learning. Payment gateway also provide analytics, flexible pricing, multi-currency support, and reconciliation reports to merchants in day-to-day business operations. Payment gateways supporting these use cases share the following characteristics:

- They provide a secure and highly available API that supports TLS 1.2 protocol for encryption.
- They have to comply with industry regulations and standards, including PCI DSS and PSD2, to protect customer data.
- They should be highly secure by following industry card standards, including features like tokenization, encryption, and fraud detection.
- They can support multiple payment methods, including debit cards, credit cards, mobile wallets, and bank transfers.

- They can help merchants with detailed analytics and reporting tools to track transactions, volumes, and key metrics.

Reference architecture

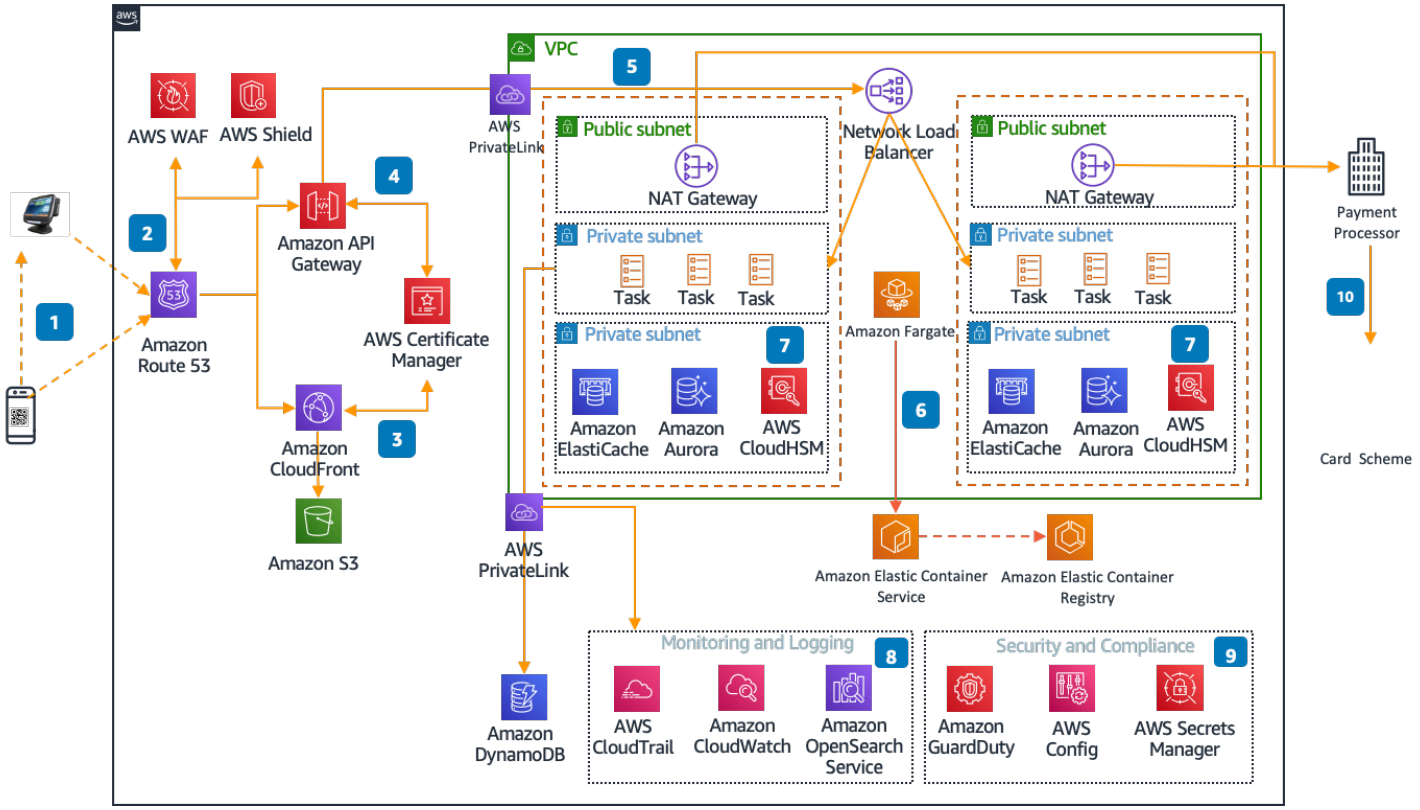

Figure 5: Architecture for QR Payments on AWS

Architecture description

1. To start, customers scan the business QR code displayed at the checkout page on a website or at the point of sale (POS) terminal.
2. Amazon Route 53 routes traffic to an Amazon API Gateway endpoint, where Amazon CloudFront distributes dynamic and static content. AWS security services, such as AWS WAF and AWS Shield, protect the web applications from common application-layer exploits and against distributed denial-of-service (DDoS) attacks.
3. CloudFront content delivery network (CDN) is used to return resources found in its cache and static resources from Amazon Simple Storage Service (Amazon S3).
4. Amazon API Gateway and Amazon CloudFront can be seamlessly integrated with AWS Certificate Manager. These services manage the complexity of creating, storing, and renewing public and private SSL/TLS X.509 certificates and keys that protect your applications.

5. The request is routed through a Network Load Balancer to distribute incoming traffic across its healthy registered targets.
6. Payment request is processed at application layer using Amazon Elastic Container Service (Amazon ECS) that deploys tasks on AWS Fargate.
7. Payment transaction information is stored in Amazon Aurora or Amazon DynamoDB. Amazon ElastiCache is used as a session store to manage session information in payment processing. AWS CloudHSM is a cryptographic service for creating and maintaining hardware security modules (HSMs).
8. Service logs are collected in Amazon S3 and analyzed and monitored using Amazon OpenSearch Service.
9. At the security and compliance layer, AWS Config evaluates, assesses, and audits configurations of resources. Amazon GuardDuty monitors for malicious activity and unauthorized behavior, protecting AWS accounts and workloads. AWS Secrets Manager helps protect secrets needed to access applications, services, and IT resources.
10. Payment request outbound traffic is sent to the payment processor through a NAT Gateway that is connected to card schemes for verification.

Insurance lake

The insurance data lake provides a method for aggregating end user customer data from a large number of diverse sources, including core systems and third parties, and consolidating it within a single, secure location. The four Cs provide a best practice data lake pattern for creation of your insurance data lake:

1. **Collect:** Store all of your data in Amazon S3.
2. **Cleanse and curate:** Validate, map, transform, and log the actions performed on your data.
3. **Consume:** Derive insights from your data.
4. **Comply and secure:** Automate your audit and regulatory compliance requirements and secure your data.

Reference architecture

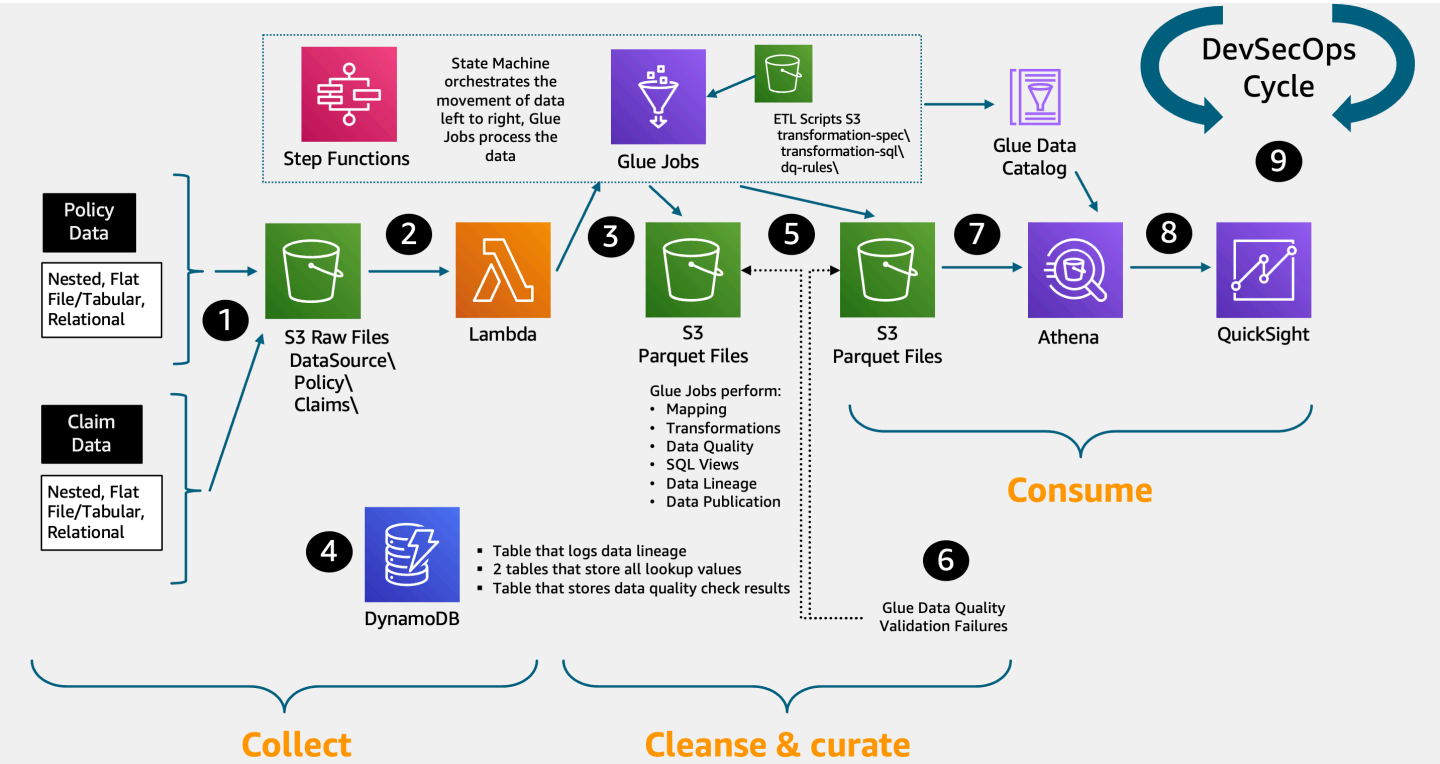

Figure 6: Insurance data lake reference architecture

Architecture description

1. Source data file is dropped into the Collect S3 bucket. Mapping file, transform file, and data quality file are present in the ETL-Scripts S3 bucket .
2. Put Event automatically initiates a Lambda function that reads metadata from the incoming source data, logs all actions, handles any errors, and starts the AWS Step Functions workflow.
3. Step Functions calls PySpark AWS Glue jobs that map the data to your pre-defined data dictionary and perform the transformations and data quality checks for both the Cleanse and Consume layers.
4. Amazon DynamoDB contains lookup values for each source data file as needed by the `lookup` and `multilookup` transforms. ETL metadata, such as job audit logs, data lineage output logs, and data quality results, are written here.
5. Cleansed and curated data is then written to compressed, partitioned Apache Parquet files in the PySpark code. The PySpark code also creates and updates AWS Glue Data Catalog databases and tables defined by your data dictionary.

6. Source data file validation failures are sent to an S3 Quarantine folder and Data Catalog table, which can populate an exception queue dashboard where a human can review and take appropriate action.
7. SQL queries can be written using the AWS Glue databases and tables.
8. QuickSight dashboards and reports can pull data from the insurance lake on a real-time or scheduled basis.
9. Full DevSecOps (everything as code and everything as automated as possible) can be managed using AWS CodePipeline and related services.

Capital markets: Market data ingestion and distribution

Capital Markets customers need access to data from a variety of sources including: market data, reference data, earnings data, alternative data, and other financial data sources. Financial data is used for making trading decisions, shaping investment strategies, providing information to regulators, and managing risk. AWS helps capital markets customers better manage and understand their data with scalable and agile cloud-based technologies. Using cloud-based solutions, customers can achieve good data governance, adhere to regulatory compliance standards, and drive profitability with financial data insights.

Reference architecture

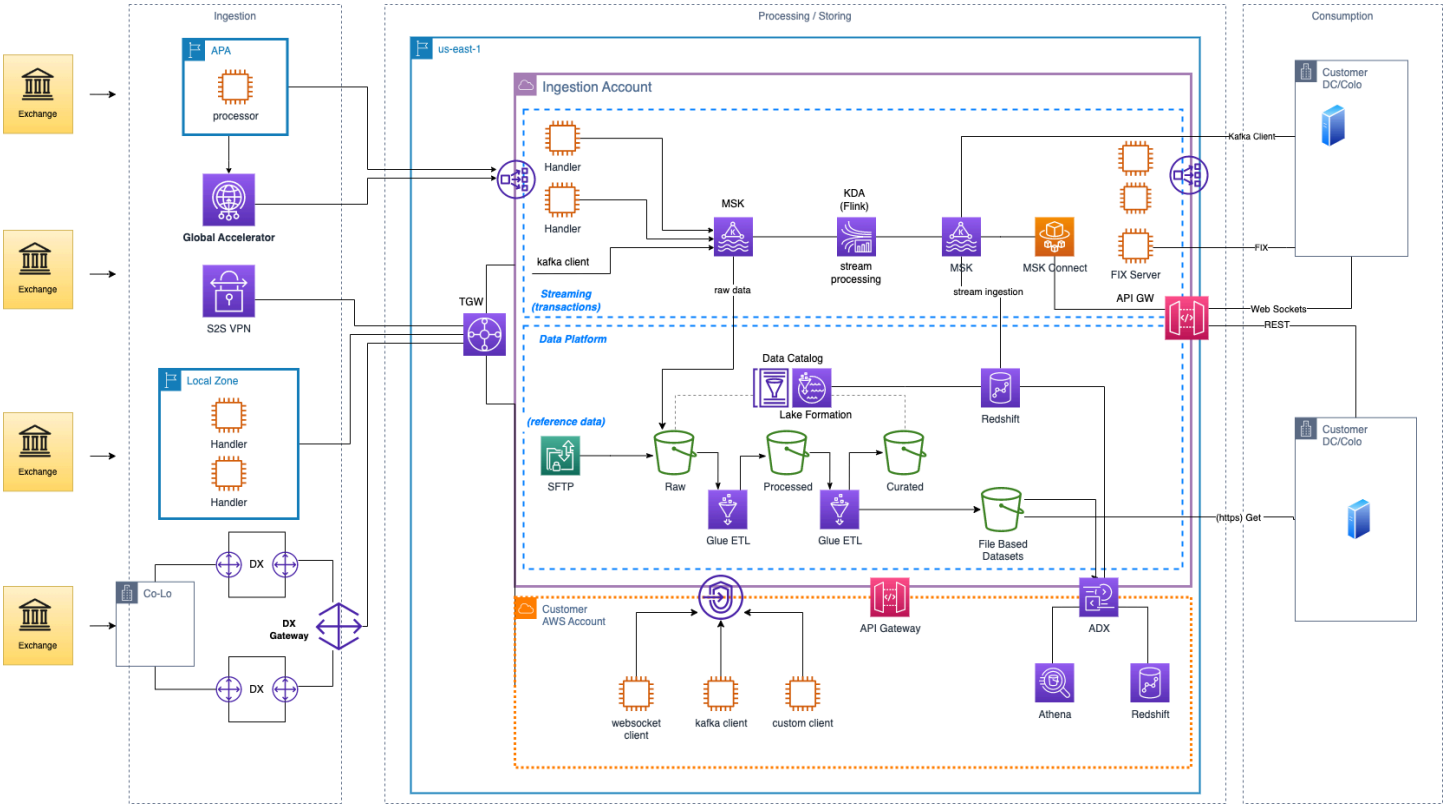

Figure 7. Market data ingest and distribution

Architecture description

The preceding architecture describes the extraction of market data from real-time and historical sources and provides a data ingestion, cataloging, and packaging workflow to provide access to market data based on customers' requests and preferences.

To begin, customers and partners can use AWS Outposts in a Co-Lo facility to connect to an exchange for real time market data and AWS Direct Connect for physical connectivity to AWS infrastructure or S2S VPNs as an alternative. They can also use AWS Local Zones for hosting workloads that require proximity to trading venues. Another example of data acquisition includes using Approved Publication Arrangement (APA) utilities for pre-trade and post-trade data.

If data is published real-time into Amazon MSK, customers have the ability to use the stream processing capabilities of Amazon Managed Service for Apache Flink (MSF) and publish the processed data to internal or external customers through MSK, WebSockets, or API Gateway depending on the use case.

Real-time data published to MSK can be ingested near-real time by Amazon Redshift which makes the data available in seconds for querying and joining with existing tables in the data warehouse. Data from MSK can also be stored in an S3 data lake.

Both batch and streaming data are added to the raw layer in the S3 data lake, where it can be cleansed, validated, and processed using AWS Glue and moved into the processed layer. Data from the processed layer can then be enriched by joining with other datasets, including reference data, and dataset ready for consumption by business users is moved to the curated layer.

Data in the data lake is governed by AWS Lake Formation, which can provide granular access controls to data, including column and row level permissions and tag-based access control. Data governed by Lake Formation can be queried by other services such as Redshift, Athena, EMR, QuickSight, etc.

Through services like AWS Data Exchange and Amazon API Gateway, the data in the data lake can be made available to other AWS accounts, whether they are part of the same organization or to external parties.

Non-AWS customers can retrieve data through API Gateway or S3 via HTTPS, depending on requirements and configuration by the customer.

For more information please refer to: Solutions for Capital Markets.

Operational excellence

The operational excellence pillar allows financial services institutions to focus on managing risks associated with operating workloads in the cloud, satisfying regulatory requirements, and becoming more agile by automating the operation and management of traditionally error-prone manual processes.

Topics

Design principles

In addition to the design principles in the AWS Well-Architected Framework whitepaper, the following design principles can help you achieve operational excellence for your financial services workloads:

- **Review applicable compliance and regulatory requirements:** Financial services institutions must be aware of all applicable regulatory and compliance obligations for their use of cloud services, and take appropriate steps to meet those obligations.
- **Evaluate legacy policies to determine relevance in the cloud:** Financial services institutions often have a robust set of operating policies that govern behaviors and decision-making for activities such as disaster recovery planning, capacity management, security and compliance guardrails, and data backup and recovery. Cloud services support new technologies, architectural patterns, and automations which are not possible or practical for on-premises environments. Policies which were originally created for on-premise environments should be revisited from a cloud perspective, rather than assumed to be necessary and relevant. Change control, for

example, should focus on changes to the architecture and configuration of the deployment pipeline, which cannot be automatically tested and reverted in the event of a failure.

- **Report service disruptions to downstream stakeholders and regulatory bodies:** Financial services institutions are required to communicate service disruptions, operational events, and failures to downstream stakeholders and regulatory bodies. They should continually monitor their workloads in the cloud and conduct root cause analysis (RCA) as an exercise in understanding the events and circumstances that led to unexpected results, as well as mitigation efforts put in place to help prevent recurrence.

Financial services workloads should be continually reviewed and prioritized with regard to their risk impact to the overall business (for example, based on their reputational, financial, or regulatory impact). Clear roles and responsibilities should be defined in the organization to understand the risks involved in the delivery of business value using cloud services.

- **Implement a risk management process:** Financial institutions have adopted a Three Lines of Defense model for risk management:
  - **First line of defense:** Operational managers perform risk and control procedures on a day-to-day basis.
  - **Second line of defense:** Various risk management and compliance functions help build and monitor the first line of defense controls.
  - **Third line of defense:** Internal auditors provide the governing body and senior management with comprehensive assurance based on the highest level of independence and objectivity within the organization.

Definitions

Operational excellence in the financial services industry is composed of the following best practice areas:

1. Organization
2. Prepare
3. Operate
4. Evolve

Organization

Best practice questions

- FSIOPS1: Have you defined risk management roles for the cloud?

FSIOPS1: Have you defined risk management roles for the cloud?

Financial institutions typically adopt a Three Lines of Defense model to improve effectiveness of risk management. The second and third lines of defense must have the appropriate skills and training necessary to understand the risks involved in the delivery of business services using cloud - services owned and managed by the first line. Establish clear roles and responsibilities both within and across the three lines of defense's functions to verify the effectiveness and auditability of the cloud operating model. Reassess these roles and responsibilities at regular intervals to keep the governance model efficient and effective.

FSIOPS01-BP01 Define roles and responsibilities across risk functions

As explained in the preceding general design principles section, financial institutions typically adopt a Three Lines of Defense model to improve effectiveness of risk management. The second and third lines of defense must have the appropriate skills and training necessary to understand the risks involved in the delivery of business services using the cloud (services owned and managed by the first line). Clear roles and responsibilities need to be established both within and across the three lines of defense's functions to verify the effectiveness and auditability of the cloud operating model. These roles and responsibilities must be reassessed at regular intervals to keep the governance model efficient and effective.

Prescriptive guidance

The roles and responsibilities of each of the three lines of defense should be clearly communicated and understood. Publishing a RACI (Responsible, Accountable, Consulted, Informed) matrix on an intranet or wiki page is a good way to reduce misunderstandings about which role owns each activity. Periodic review of these roles and responsibilities should occur more frequently immediately after they are defined or dramatically changed, and can be less frequent otherwise. The people who fill roles within the three lines of defense should be documented as well, and membership in these roles should require a standard level of training in order to consistently handle risk management.

FSIOPS01-BP02 Engage with your risk management and internal audit functions to implement a process for the approval of cloud risk controls

Significant changes in technology necessitate a refreshed assessment of new potential risks and their validations. Technology changes include migrating to the cloud, use of newer database tools, extensive mobile application usage, and AI/ML technologies. These changes may present risks to the existing control environment such that it may be unable to mitigate the original identified risks, but also may not be effective across a much broader spectrum of changes. Engagement with the risk and internal audit functions helps align with required governance obligations as cloud usage increases. This engagement needs to include documentation and demonstration by the first line, to the second and third lines, of the controls, technology, and processes that have been implemented to secure and operate the cloud environment. This process can contain a regular review cadence for new controls, so the first line can evolve their implementations as needed to quickly and safely adopt best practices for new threats.

Prescriptive guidance

All stakeholders from the three lines of defense should be invited to participate in suggesting, evaluating, and approving changes to risk controls. A periodic review of risk controls, as well as an out-of-cycle mechanism to suggest updates, should be clearly documented and understood by all stakeholders. The lifecycle of a risk control (suggestion, review, approval, training, implementation, and retirement) should also be documented and understood. Prior to implementation of a specific risk control, metrics should be identified to indicate the effectiveness of the control. These metrics should be generated and compiled automatically and should be reviewed periodically throughout the risk control's lifecycle. Thresholds that indicate effectiveness should be established, and the continued breach of those thresholds should prompt review of the risk control, with an expectation that it be updated or retired.

FSIOPS01-BP03 Implement a process for adopting appropriate risk appetites

Failures can happen at any time. The appropriate risk authority within the firm (for example, the board of directors, chief risk officers, or business risk officers) needs to evaluate the criticality of a business process (and the underlying workloads that support that process) and specify the level of availability that the firm requires for that process. This must take into consideration the potential impact that a disruption of that process has on the firm, the market, the customers, and regulatory bodies managing the financial infrastructure, as well as the cost of operating the workload in a high availability mode weighed against business agility and innovation. Working backwards from these risk appetites allows you to drive the operational priorities and the resiliency design choices

of cloud workloads supporting business services in a prioritized manner. Setting clear risk appetites allows for effective risk management and governance.

Prescriptive guidance

All workloads should be categorized based on their criticality and associated risk tolerance. In financial services organizations, this classification has often already occurred as part of disaster recovery planning, and these risk categorizations can be reused elsewhere. Once risk categories are established, requirements should be identified to be applied to workloads within each risk category. Examples of requirements might be recovery time objective (RTO) or recovery point objective (RPO) expectations, use of encryption for data in-transit and at rest, and geographies within which data must be stored. Building upon these requirements, preferred architectural patterns should be identified that help meet the needs of each risk category in an efficient and manageable way. Publishing these reference architectures is a good way to encourage their adoption, as it simplifies the use of a consistent and preferred architecture, and also provides a foundation for automation.

Prepare

Best practice questions

- FSIOPS2: Have you completed an operational risk assessment?
- FSIOPS3: Have you assessed your specific workload against regulatory needs?

FSIOPS2: Have you completed an operational risk assessment?

Financial services workloads should be continually reviewed and prioritized with regard to their risk impact to the overall business (for example, based on their reputational, financial, or regulatory impact).

FSIOPS02-BP01 Understand the Shared Responsibility Model and how it applies to services and workloads you run in the cloud

In connection with your use of the cloud, you must understand how the AWS Shared Responsibility Model affects your control environment. For example, certain controls may be the responsibility of AWS, but certain controls remain the responsibility of the financial services institution. Review the AWS Shared Responsibility Model and map AWS responsibilities and customer responsibilities according to each AWS service you use and your control environment. For those controls that

are the responsibility of AWS, you can use AWS Artifact to access audit reports and review the implementation and operating effectiveness of AWS security controls.

Prescriptive guidance

Review and understand the AWS Shared Responsibility Model, and the different demarcation points that apply to AWS infrastructure services (such as EC2), container services (such as RDS), and abstracted services (such as S3). If your organization has central functions (like a Cloud Center of Excellence or governance team), publish a shared responsibility model for your organization, which clearly defines the roles of AWS, the central team, and distributed teams.

FSIOPS02-BP02 Develop an enterprise cloud risk plan

Map the interactions between business consumers of cloud services and the internal stakeholders that shape this consumption, including risk and control considerations. Integrate across the three lines of defense functions, and provide necessary resources and training to satisfy their mandates for operating and protecting your business in the cloud while you strive to achieve your strategic goals.

This integration can be achieved by carrying out a risk-based assessment of your operating model, and is especially effective when complemented with a review of decision-making processes and authority to determine if they are cloud-appropriate. As requirements are translated into controls, pay attention to the strength of the controls to mitigate the identified risks. Another key risk factor includes the ability to control design and performance to facilitate independent assessment by internal risk management and audit functions. Focus on control design helps you incorporate key control requirements into the design from the start.

Prescriptive guidance

Evaluate existing risk models in use, and related policies, for relevance in a cloud environment. Many risk models are focused on on-premises architectures and do not account for advantages of cloud-based workloads. Reach out to your AWS account team to leverage AWS expertise in risk and compliance.

FSIOPS3: Have you assessed your specific workload against regulatory needs?

Financial services institutions must be aware of all applicable regulatory and compliance obligations for their use of cloud services, and they should take appropriate steps to meet those obligations.

FSIOPS03-BP01 Implement a process for the review of applicable compliance and regulatory requirements for your workload

Financial services institutions must be aware of all applicable regulatory and compliance obligations for their use of the cloud, and they should take appropriate steps to meet those obligations. As part of your strategy, review your migration plan and control frameworks with the relevant internal stakeholders responsible for compliance to identify any compliance requirements, including legal and regulatory requirements that apply to your use of the cloud. Note that designing a workload to meet specific technical requirements may only be one aspect of compliance, so it's important to conduct a comprehensive regulatory and compliance review. This process must include both initial design and planning, as well as pre-production readiness activities.

Prescriptive guidance

Use the AWS Compliance Center to learn about key cloud-related regulatory requirements that impact your use of the cloud, and the regulations that apply within your geography. Design a process to monitor evolving changes to compliance and regulatory obligations. Use AWS Config Conformance Packs and AWS Audit Manager to continually evaluate your compliance to applicable regulatory frameworks. If appropriate, review the AWS Sub-Processors list and sign up to be notified of changes. Use AWS Artifact to gather compliance reports that apply to your workload and geography.

Operate

Best practice questions

- FSIOPS4: How do you assess your ability to operate a workload in the cloud?
- FSIOPS5: How do you understand the health of your workload?
- FSIOPS6: How do you assess the business impact of a cloud provider service event?

FSIOPS4: How do you assess your ability to operate a workload in the cloud?

Financial services institutions often have a robust set of operating policies that govern behaviors and decision-making for activities such as disaster recovery planning, capacity management, security and compliance guardrails, and data backup and recovery. Cloud services support new

technologies, architectural patterns, and automations which are not possible or practical for on-premises environments. Policies which were originally created for on-premise environments should be revisited from a cloud perspective, rather than assumed to be necessary and relevant.

FSIOPS04-BP01 Implement a change management process for cloud resources

Cloud IT change management processes facilitate changes to IT systems in order to minimize risks to production environments while adhering to policies, audit, and risk controls. It is not uncommon, especially within financial services institutions, to see a gated change management process often requiring a review by external change advisory boards, which can take days or even weeks. As organizations take advantage of configuration management, infrastructure as code (IaC), automated testing and validation, and continuous integration and delivery, they can implement lightweight approval processes that are tightly integrated into CI/CD pipeline tools.

By automating detection and rejection of bad changes, many manual approval steps can be fully automated with a higher degree of confidence. Even in highly regulated industries where external reviews are required, such as financial services, reviews should still be integrated with the overall pipeline, even if they are manual steps initially. Regulatory requirements such as the Sarbanes-Oxley Act requires all financial reports to include an internal controls report that documents every change made to your workloads. Performing operations as code provides the capability to test, model, and simulate scenarios before rollout, which limits the potential for human error. Additionally, it satisfies regulatory requirements by providing auditors a complete record of all applied changes, including the environment in which tests and validations were run and the identity and timestamp of each change approval. This speeds up deployment cycles and innovation, while preserving security controls and guardrails.

A good change management process delivers business value while balancing risk against business value. It should do so in a way that maximizes productivity and minimizes wasted effort or cost for all participants in the process. Automation, integration, and deployment tools in the cloud allow businesses to make small, frequent changes that reduce risk and deliver business value at an increased rate. For additional guidance, see Change Management in the Cloud.

Prescriptive guidance

Financial services institutions must develop cloud capabilities in layers, producing approved, reusable artifacts at each layer, such as:

- golden Amazon Machine Images (AMIs),
- CloudFormation Templates,

- Service Catalog Products,
- container base images,
- software packages,
- and Lambda deployment packages.

Artifacts at foundational layers must go through a change control process so that they comply with enterprise guidelines, which can then be repurposed as building blocks by the rest of the organization. AWS Systems Manager Change Manager provides tracking and approval, and allows for the implementation of operational changes to application configurations and infrastructures. As the organization builds higher-level applications on a foundation of certified artifacts, you can expedite the change control process, as it only needs to focus on the higher-level artifacts, accelerating change while minimizing risk and ensuring compliance. Over time, organizations develop capabilities to administer most of the changes in automated fashion, with only a subset of changes that require manual intervention.

FSIOPS04-BP02 Implement infrastructure as code

The benefit of the cloud and infrastructure as code is the ability to build and tear down entire environments programmatically and automatically. If architected with resiliency in mind, a recovery environment can be implemented in minutes using AWS CloudFormation templates or AWS Systems Manager automation. Automation is critical for maintaining high availability and fast recovery.

Prescriptive guidance

AWS offers a wide breadth of automation tools to accomplish resiliency objectives. AWS Systems Manager helps automate complete runbooks that are used during the recovery of an application during a disaster. You can sequence a complete set of operations to automatically initiate on detection of an event. With Systems Manager automation documents, you can manage these runbooks similar to the way you manage code. You can version them and update them along with every release. This helps keep your recovery plan in sync with released code and updates to infrastructure.

FSIOPS04-BP03 Prevent configuration drift

Drift of infrastructure configuration between primary and secondary sites can lead to failure in recovery during a disaster event. Implementation of code-based management practices across your infrastructure, applications, and operational procedures provides a high degree of version control,

testing, validation, and mitigation of human error and configuration drift, which is necessary to limit the introduction of errors into your environment and to reduce the mean time to recover (MTTR).

Prescriptive guidance

Financial services institutions should monitor changes to application infrastructure by using:

- AWS CloudFormation drift detection,
- AWS CloudTrail,
- and AWS Config.

These services monitor activity within your AWS account, including actions taken through the AWS Management Console, AWS SDKs, command line tools, and other AWS services. Once detected, you can automate the reactive action by defining workflows using AWS EventBridge integration and AWS Config Rules.

FSIOPS5: How do you understand the health of your workload?

Financial services institutions are required to communicate service disruptions, operational events, and failures to downstream stakeholders and regulatory bodies. They should continually monitor their workloads in the cloud and conduct root cause analysis (RCA) as an exercise in understanding the events and circumstances that led to unexpected results, as well as mitigation efforts put in place to prevent recurrence.

FSIOPS05-BP01 Use enhanced monitoring in the cloud

High availability for financial services workloads that support critical functions requires the ability to detect failures and quickly recover from them. You can understand the operational state of your workloads by defining, collecting, and analyzing metrics in the cloud that can be incorporated into your operating model. These metrics are emitted by your code, workloads, and user activity, and need to be collected in a centralized, queryable system that can be used to visualize and examine real-world performance data. This is important for diagnosing issues that are often not clear from looking at just at application logs, Amazon CloudWatch, or system logs in isolation.

Prescriptive guidance

Review Monitoring and Observability to familiarize yourself with the capabilities of AWS services. Financial institutions require logs and metrics for two distinct use cases: operational analysis

(such as troubleshooting during an incident) and regulatory compliance. Application logs can be collected with Amazon CloudWatch Logs and stored in a centralized AWS account dedicated to logging. Access to the dedicated logging AWS account should be limited and based on least privilege, and the data can be shared in a read-only manner to other AWS accounts for analysis.

If immutable log storage is required for regulatory or corporate policy compliance, use Amazon S3 Object Lock

or Amazon S3 Glacier Vault Lock for WORM storage.

Use AWS tools such as OpenSearch or Amazon Athena, or third party tools such as Splunk, Datadog, or Sumo Logic, to provide indexing, search, analysis, and visualization capabilities.

Use CloudWatch Events for metrics and CloudWatch anomaly detection to detect changes in trends and send alerts to Operations teams.

AWS X-Ray helps you understand how your application and its underlying services perform to identify and troubleshoot performance issues and errors.

You can also experience these capabilities in your own AWS account by running the One Observability Workshop, where you learn about AWS observability functionalities on Amazon CloudWatch, AWS X-Ray, Amazon Managed Service for Prometheus, Amazon Managed Grafana, and AWS Distro for OpenTelemetry. This workshop deploys a microservice-based application and guides you in discovering actionable insights through various monitoring tools. Upon conclusion, the learner is expected to have a clear understanding of logging, metrics, and traces, as well as techniques for using them across a variety of workload types.

For critical or regulated workloads workloads, Enterprise Support customers should consider subscribing to AWS Incident Detection and Response.

AWS Incident Detection and Response offers eligible AWS Enterprise Support customers proactive engagement and incident management to reduce the potential for failure and accelerate recovery of critical workloads from disruption. It achieves these objectives by fostering joint preparation with AWS to develop runbooks and response plans customized to the context of each workload onboarded to the service. Onboarded workloads are monitored by a team of Incident Management Engineers (IMEs) to detect and engage you on a call bridge within five minutes of a critical alarm.

AWS Incident Detection and Response begins with a review of your workloads for reliability and operational excellence. AWS experts work with you to define critical metrics and alarms that provide improved visibility into the application and infrastructure layers of your workloads,

which makes it easier to find and prioritize issues during an incident. AWS Incident Management Engineers continually monitor your workloads, detect critical incidents, and engage you on a call bridge with the right AWS experts to accelerate the recovery of your workloads. All incidents are managed with the highest level of severity and escalation, and AWS remains engaged until the incidents are resolved. Lessons learned from previous incidents inform improvements to response plans and workload architecture, which drives a continuous improvement cycle to improve the resiliency of your workloads.

FSIOPS05-BP02 Monitor cloud provider events

Financial institutions should use the AWS Health Dashboard, which provides information and remediation guidance when AWS is experiencing events that may impact workloads. The dashboard displays relevant and timely information to help manage events in progress, and provides proactive notifications to help plan for scheduled activities. With AWS Health Dashboard, alerts are generated by changes in the health of the AWS resources used in your applications, giving you event visibility and guidance to help quickly diagnose and resolve issues. Enterprise support and business support accounts who have access to the AWS Health API can use this API to integrate the information from AWS Health Dashboard into the centralized monitoring system and define a consistent and comprehensive alerting mechanism.

Prescriptive guidance

AWS Health provides ongoing visibility into your resource performance and the availability of your AWS services and accounts. You can use AWS Health events to learn how service and resource changes might affect your applications running on AWS. AWS Health provides relevant and timely information to help you manage events in progress. AWS Health also helps you be aware of and prepare for planned activities. The service delivers alerts and notifications initiated by changes in the health of AWS resources, which provides event visibility and guidance to help accelerate issue resolution. AWS Health provides information about service operations, such as operational issues, planned maintenance, and planned software lifecycle events.

For comprehensive visibility into AWS Health event details, such as affected resource IDs, current status (open or closed), and resource status, use AWS Health endpoints, such as the AWS Health API, the `aws.health` source in Amazon EventBridge, and the AWS Health Dashboard. These endpoints provide the most detailed and real-time information about ongoing events and changes that might affect your workloads.

AWS User Notifications notifies you through additional UX channels (email, chat, or push notifications to the AWS Management Console mobile application). AWS Health event notifications

don't contain as much detailed data as the endpoints listed previously. However, they provide a simple and effective way to notify stakeholders of issues and changes. Based on rules that you create, User Notifications creates and sends a notification when an event matches the values that you specify in a rule. You can select which UX delivery channels a notification is sent to and set up aggregation to reduce the number of notifications generated for specific events. Notifications are also visible in the AWS Management Console Notifications Center. For example, you can receive chat notifications if you have resources in your AWS account that are scheduled for updates, such as EC2 instances. For more detail, see Getting started with AWS User Notifications.

You can integrate AWS Health events with Jira and ServiceNow to receive operational and account information, prepare for scheduled changes, and manage AWS Health events using the AWS Service Management Connector. The Service Management Connector integration with AWS Health can use AWS Health events sent through EventBridge to automatically create, map, and update JIRA tickets and ServiceNow incidents.

You can use organizational view and delegated administrator access to manage AWS Health events across the organization within Jira and ServiceNow and incorporate AWS Health information directly into your team's workflow. For more detail on ServiceNow integration using the Service Management Connector, see Integrating AWS Health in ServiceNow. For more detail on Jira Management Cloud integration using the Service Management Connector, see AWS Health.

FSIOPS6: How do you assess the business impact of a cloud provider service event?

Financial institutions should assess the business impact of cloud provider service events.

FSIOPS06-BP01 Manage cloud provider service events

Financial institutions should assess the business impact of cloud provider service events. During events, timely communication regarding business disruptions should be made to affected downstream stakeholders such as customers, partners, and regulatory bodies. These service event notices should include details of which functions are impaired or unavailable due to the event, geographies and customer segments that are affected, and remediation efforts put in place to temporarily or permanently address the issue. Financial institutions should implement push notifications to alert internal teams responsible for the impacted workloads, as well as a mechanism to collect sentiment from impacted stakeholders. Throughout the duration of a cloud provider service event, financial institutions should post updates to the service event notice, and initiate a post-event operational review at the conclusion of the event (see After a service event).

Prescriptive guidance

The following describes steps you can take to respond to a service event.

Prior to a service event Identify business outcomes and KPIs that support those outcomes, like the number of payments per minute, size of a dead letter queue, or the amount of delay between putting and getting data on streams. Map metrics to workloads, and map workloads to teams who support those workloads during a service event. Provide your teams a mechanism to receive alerts and understand the response expectations. Establish baseline thresholds for normal operation and implement a system which alert if metrics fall outside of that range. Identify a primary (and secondary if necessary) communication channel that is used to provide updates to downstream stakeholders during a service event. Document and communicate expectations. Identify teams responsible for supporting key workloads, and evaluate their access to and familiarity with the Support workflow. Support Center access may be restricted by central governance policies, and access to create Support cases should be confirmed prior to a service event in order to help avoid delays in remediation.

During a service event Use push notifications to alert the teams responsible for the affected workloads and initiate a conference bridge to address the issue. Use a ticketing system or other tracking mechanism to collect stakeholder feedback, logs, and troubleshooting notes in a single location

Check the AWS Health Dashboard to confirm whether there are any AWS service events in progress that may be related to the issues you are experiencing. Create a support case in the Support Console if you suspect the service event may be related to any AWS services, or if you require assistance in troubleshooting an AWS service. Communicate the business impact and status of remediation efforts to downstream stakeholders on an established cadence using the pre-defined communication channel.

After a service event When service is restored, submit a final notification closing the event. Conduct a post-event operational review (see FSIOPS-BP14: Conduct post-event operational reviews) and provide the product of that review (an RCA or Correction of Error (COE) report) to affected downstream stakeholders and regulatory bodies. For critical workloads, Enterprise Support customers should consider subscribing to AWS Incident Detection and Response.

Evolve

Best practice questions

- FSIOPS7: Have you developed a continuous improvement model?

FSIOPS7: Have you developed a continuous improvement model?

Financial institutions should continually assess and optimize their operational processes.

FSIOPS07-BP01 Test, model, and simulate scenarios before rollout

One of the best practices to determine if you have addressed your risk with appropriate controls is to actually run scenarios against your cloud control framework and operational procedures. Once your risk and control program is established, financial institutions should continually asses and optimize their operational processes. Regular game days for workloads deployed on AWS can help build your team's muscle memory and validate that all operational procedures are effective in supporting your recovery objectives and compliance with notification requirements to regulatory bodies. We recommend designing game days to test your risk appetite and include severe, but plausible scenarios.

Prescriptive guidance

Identify financial services compliance requirements first, and then structure your game days to meet those requirements. Align the complexity of game days with the resources available within your organization. For large organizations, game days are often scoped to a specific business unit or product team. It's acceptable to presume certain inputs from other teams during your initial game days, which can make scheduling more practical. It's more important to complete simple game days regularly, and iterate on the scope and complexity over time, than to try to run complex game days from the beginning. The most critical piece of a game day is the retrospective review of lessons learned and the iterative improvement over time. Sufficient time to accomplish this should be set aside early in the planning process so that it can occur in the days immediately following the game day.

FSIOPS07-BP02 Conduct post-event operational reviews

Post-event operational reviews should be conducted after an incident. After troubleshooting and performing repair procedures, follow-up documentation and actions should be assigned. An effective post-event review results in a list of practical actions that address each of the issues that allowed the threat actor to succeed. These actions should minimize the impact of the event and teach the wider enterprise how to prevent, detect, and respond to a similar event in the future. For significant events, a Correction of Error (COE) document should be composed to capture the root cause and take preventative actions for the future. Implementation of the preventative measures should be measured in future operations meetings.

Prescriptive guidance

Post-event operational reviews are comprised of two components: identification of the problem (root cause analysis) and the identification of actions to help prevent a reoccurrence of the event (corrective actions). Identify a mechanism, such as an ITSM tool or ticketing system, to track root cause analysis efforts and associated corrective actions. Ownership for each task should be assigned to an individual, and a periodic review should be used to track status. In a large and complex environment, competing priorities and urgent activities can supersede processes such as post-event reviews that are important for long-term stability. Leaders should establish a culture which prioritizes these reviews, and should encourage teams to set aside a recurring time to spend on analysis and corrective actions.

Key AWS services

- **Management and governance**
  - AWS Config - AWS Config continually assesses, audits, and evaluates the configurations and relationships of your resources. Codify your compliance requirements as AWS Config rules and author remediation actions, automating the assessment of your resource configurations across your organization. Evaluate resource configurations for potential vulnerabilities, and review your configuration history after potential incidents to examine your security posture.
  - AWS Config Rules - AWS Config provides you with pre-built rules evaluating the configurations of your cloud resources, as well as software within managed instances, including EC2 instances and servers running on-premises, before and after provisioning. You can customize pre-built rules to evaluate your AWS resource configurations and configuration changes, or create your own custom rules on AWS Lambda that define your internal best practices and guidelines for resource configurations.
  - AWS Organizations - AWS Organizations lets you create new AWS accounts at no additional charge. With accounts in an organization, you can quickly allocate resources, group accounts, and apply governance policies to accounts or groups.
  - AWS Control Tower - Set up and govern a secure, multi-account AWS environment. AWS Control Tower simplifies AWS experiences by orchestrating multiple AWS services on your behalf while maintaining the security and compliance needs of your organization.
  - Service Catalog - Service Catalog lets you centrally manage deployed IT services, applications, resources, and metadata to achieve consistent governance of your infrastructure as code (IaC) templates. With Service Catalog, you can adhere to your compliance requirements while making sure your customers can quickly deploy the approved IT services they need.

- AWS Mainframe Modernization - AWS Mainframe Modernization is a set of managed tools providing infrastructure and software for migrating, modernizing, and running mainframe applications. Automate transforming legacy language applications into agile Java-based services with AWS Blu Age using newer web frameworks and cloud DevOps best practices. Migrate COBOL and PL/I applications with the integrated Micro Focus toolchain to preserve the programming language while modernizing infrastructure and processes for agility with cloud DevOps best practices. Create a highly available runtime environment and deploy applications in minutes with extensive automation, minimizing the administrative burden and accelerating operations.
- AWS Well-Architected Tool - The AWS Well-Architected Tool lets you review your workloads against current AWS best practices and obtain advice on how to architect your workloads for the cloud. This tool uses the AWS Well-Architected Framework.

- **Compliance**
  - AWS Compliance Center - The AWS Compliance Center is a central location to research cloud-related regulatory requirements and how they impact your industry.
  - AWS Artifact - AWS Artifact is your go-to, central resource for compliance-related information that matters to you. It provides on-demand access to security and compliance reports from AWS and ISVs who sell their products on AWS Marketplace.
  - AWS Audit Manager - Use AWS Audit Manager to map your compliance requirements to AWS usage data with prebuilt and custom frameworks and automated evidence collection.
  - AWS Backup - AWS Backup is a cost-effective, fully managed, policy-based service that simplifies data protection at scale. Examine your resources against data protection policies to maintain compliance with organizational or regulatory requirements.
- **Monitoring**
  - Amazon CloudWatch - Amazon CloudWatch collects and visualizes real-time logs, metrics, and event data in automated dashboards to streamline your infrastructure and application maintenance. Perform root cause analysis by analyzing metrics, logs, logs analytics, and user requests to speed up debugging and reduce overall mean time to resolution.
  - AWS CloudTrail - AWS CloudTrail monitors and records account activity across your AWS infrastructure, giving you control over storage, analysis, and remediation actions.
- **Deployment**
  - AWS CodeDeploy - AWS CodeDeploy is a fully managed deployment service that automates software deployments to various compute services, such as Amazon Elastic Compute Cloud (EC2), Amazon Elastic Container Service (ECS), AWS Lambda, and your on-premises servers.

Use CodeDeploy to automate software deployments, eliminating the need for error-prone manual operations.

- AWS CloudFormation - AWS CloudFormation lets you model, provision, and manage AWS and third-party resources by treating infrastructure as code.
- AWS CDK - AWS Cloud Development Kit (AWS CDK) (AWS CDK) accelerates cloud development using common programming languages to model your applications. Develop applications more efficiently using AWS CDK as the main framework to define cloud infrastructure as code.
- AWS CodeArtifact - CodeArtifact allows you to store artifacts using popular package managers and build tools like Maven, Gradle, npm, Yarn, Twine, pip, and NuGet. CodeArtifact can automatically fetch software packages on demand from public package repositories so you can access the latest versions of application dependencies.
- Amazon CodeGuru - Amazon CodeGuru is a developer tool that provides intelligent recommendations to optimize application performance, improve code quality, detect security vulnerabilities and automate code reviews.

- **Operations**
  - AWS Health Dashboard - The AWS Health Dashboard is the single place to learn about the availability and operations of AWS services. You can view the overall status of AWS services, and you can sign in to view personalized communications about your particular AWS account or organization. Your account view provides deeper visibility into resource issues, upcoming changes, and important notifications.
  - AWS User Notifications - AWS User Notifications helps users centrally set up and view notifications from AWS services, such as AWS Health events, Amazon CloudWatch alarms, or EC2 instance state change, in a consistent, human-friendly format. Users can view notifications across accounts, Regions, and services in a Console Notifications Center and configure delivery channels, like email, chat, and mobile push notifications, where they can receive these notifications.
  - AWS Incident Detection and Response - AWS Incident Detection and Response offers AWS Enterprise Support customers proactive monitoring and incident management for their selected workloads. AWS Incident Detection and Response is designed to help you reduce potential for failures on your workloads and to accelerate your recovery from critical incidents.
  - AWS Managed Services - AWS Managed Services (AMS) helps you adopt AWS at scale and operate more efficiently and securely. We leverage standard AWS services and offer guidance and performance of operational best practices with specialized automations, skills, and experience that are contextual to your environment and applications. AMS provides proactive, preventative, and detective capabilities that raise the operational bar and help reduce risk

without constraining agility, allowing you to focus on innovation. AMS extends your team with operational capabilities including monitoring, incident detection and management, security, patch, backup, and cost optimization.

- AWS Resilience Hub - AWS Resilience Hub provides a central place to define, validate, and track the resilience of your applications on AWS. AWS Resilience Hub's assessment uses best practices from the AWS Well-Architected Framework to analyze the components of an application and uncover potential resilience weaknesses as well as actionable recommendations to improve resilience.
- AWS Systems Manager - AWS Systems Manager is a secure end-to-end management solution for hybrid cloud environments. Centralize operational data in a single console and gain actionable insights across AWS services such as Amazon CloudWatch, AWS CloudTrail, and AWS Config, as well as third-party tools. Automatically resolve application issues by leveraging operational data to simply manage applications and identify issues quickly across associated AWS resource groups. Implement best practices by automating proactive processes such as patching and resource changes—as well as reactive processes—to quickly diagnose and remediate operational issues before they affect users. Remediate security events by evolving your security and compliance profiles and analyze security events after the fact to help prevent a future re-occurrence.
- Amazon EventBridge - Amazon EventBridge is a service that provides real-time access to changes in data in AWS services, your own applications, and software as a service (SaaS) applications without writing code. To get started, you can choose an event source on the EventBridge console. You can then select a target from AWS services including AWS Lambda, Amazon Simple Notification Service (SNS), and Amazon Data Firehose. EventBridge automatically delivers the events in near real-time.
- AWS Service Management Connector - AWS Service Management Connector and its integration connectors help you provision, manage, and operate AWS resources and capabilities in familiar IT service management (ITSM) tools, such as ServiceNow and Atlassian.
- Support - Support helps customers with technical issues and additional guidance to operate their infrastructures in the AWS Cloud. Customers can choose a tier that meets their specific requirements, which continues the AWS tradition of providing the building blocks of success without bundling or long term commitments.
- AWS Trusted Advisor - AWS Trusted Advisor helps you optimize costs, increase performance, improve security and resilience, and operate at scale in the cloud. Trusted Advisor continually evaluates your AWS environment using best practice checks across the categories of cost

optimization, performance, resilience, security, operational excellence, and service limits and recommends actions to remediate deviations from best practices.

- AWS Countdown - AWS Countdown is an Support offering designed for a broad range of cloud use cases, including migrations, modernizations, product launches, streaming, and go-live events. AWS Countdown helps you throughout the project lifecycle to assess operational readiness, identify and mitigate risks, and plan capacity, using proven playbooks developed by AWS experts. It empowers you with resources for operational readiness, AWS Well-Architected assessments, security reviews, and infrastructure capacity planning for your projects. AWS Countdown replaces Support's Infrastructure Event Management (IEM) service and is included with Enterprise Support.

Resources

Documents and blogs

- Build Your Own Game Day to Support Operational Resilience
- How Financial Institutions can use AWS to Address Regulatory Reporting
- How Financial Institutions can Select the Appropriate Controls to Protect Sensitive Data
- Automating and Scaling Chaos Engineering using AWS Fault Injection Service
- Goldman Sachs, an established financial services firm, transforms its operations on AWS
- Best Practices for AWS Organizations Service Control Policies in a Multi-Account Environment
- How Cover-More launched their insurance platform into a new Region and improved worldwide operations using AWS Managed Services
- How financial institutions modernize record retention on AWS
- What AWS customers need to know about DORA and the UK financial regulators' approach to outsourcing: the plan to optimize resiliency and innovation for the financial services sector
- Resilience lifecycle framework: A continuous approach to resilience improvement
- Resilience analysis framework

Whitepapers

- IIA's Three Lines of Defense model update
- Customers can achieve and test resiliency on AWS

- AWS Cloud Adoption Framework: Operations Perspective
- Designing Highly Resilient Financial Services Applications
- Running an Exchange in the Cloud
- AWS Fault Isolation Boundaries

Videos

- AWS re:Invent 2019: Leadership session: Running critical FSI applications on AWS (FSI201-L)
- Simplify the AWS Shared Responsibility Model
- AWS re:Invent 2021 - Cloud compliance, assurance, and auditing
- Simplify Operational Change Management with Change Manager
- AWS re:Invent 2021 - Intelligently automating cloud operations
- Building a Robust Monitoring Strategy - AWS Virtual Workshop
- Supports You - Getting Started with AWS Health Aware (AHA)
- AWS re:Invent 2022 - AWS Incident Detection and Response (SUP201)

Training

- AWS Well-Architected Considerations for Financial Services
- Industry Quest: Financial Services (Amazon)
- AWS Observability
- Getting Started with AWS Systems Manager
- AWS Systems Manager
- Getting Started with AWS Config
- Getting Started with AWS CloudTrail
- Introduction to Amazon CloudWatch
- Introduction to Amazon CloudWatch Logs
- AWS Cloud Essentials for Business Leaders (Financial Services)
- AWS Ramp-Up Guide: Financial Services Industry (FSI)

For Enterprise Support customers

- AWS Countdown - Plan and execute successful events with AWS Countdown, a service designed for a broad range of cloud use cases, including migrations, modernizations, product launches, streaming, and go-live events. AWS Countdown helps you throughout the project lifecycle to assess operational readiness, identify and mitigate risks, and plan capacity, using proven playbooks developed by AWS experts. AWS Countdown Premium tier provides critical support across all phases of your cloud projects from design to post-launch retrospectives. It offers designated engineers selected from a team of AWS experts who provide proactive guidance and troubleshooting. Designated engineers get involved from project inception to facilitate continuity, provide access to subject matter experts, and use support tools for faster issue resolution. They participate in critical events calls, like sales events or migration cutovers, to provide rapid issue resolution. AWS Countdown Premium helps you increase your infrastructure investment return through the acceleration of migrations and modernizations and delivery of high impact go-live events and achieve your business goals.
- Operational Excellence Deep Dive - The Operational Excellence Deep-Dive extends the coverage of the Well-Architected Operational Excellence Pillar through an expert-led engagement. The engagement is centered on a guided conversation focused on key elements of your organization, priorities, processes, tooling and culture that contribute to your operational outcomes. Insight gathered from the conversation are prioritized according to your goals and then recommendations for actions are provided that help you improve, extend and scale your operations towards delivering on your desired business outcomes.
- Incident Detection and Response - AWS Incident Detection and Response requires a paid subscription, which is added to Enterprise Support. It offers AWS Enterprise Support customers proactive monitoring and incident management for their selected workloads. AWS Incident Detection and Response is designed to help you reduce potential for failures on your workloads and to accelerate your recovery from critical incidents.
- Operations KPI workshop - The 'Operations KPI' workshop uses Amazon best practices to build a consistent approach to developing 'Key Performance Indicators' (KPIs). The centerpiece of the workshop is to create a strategy intersecting operational practices that support business needs and establish operational metrics. Customers with a business level view of operations activities based on the KPIs can determine if business needs are satisfied and identify areas needing improvement.
- Building a Monitoring Strategy workshop - The 'Building a Monitoring Strategy' workshop uses Amazon best practices to help customers build a consistent approach to the monitoring and observability of workloads. The workshop's goal is to create a strategy that aligns business and

operational metrics. The workshop helps customers identify key metrics which matter most to delivering successful business outcomes.

- Operational Readiness Review workshop - The 'Operational Readiness Review' workshop is an interactive "working backwards" session on people, process and mechanisms for customers. The workshop helps customers achieve a consistent process (including a checklist) for evaluating operational readiness of workloads prior to launch. Customers use these checklists to get visibility into risk and plan remediation's. Customers with consistent evaluation procedures gain improved confidence in meeting business outcomes.
- Incident Management workshop - The incident management workshop is a table top exercise in-which teams test their existing incident response procedures against a hypothetical incident. The engagement is an opportunity to discuss and check adoption of incident management best practices associated with people, process and tooling. Best practice matrixes and next step recommendations are created after the workshop which aim to help you respond faster, have fewer outages and increase uptime.

Security

The security pillar focuses on the ability to protect information, systems, and assets through risk assessments and mitigation strategies, while also delivering business value to the organization. In addition to the regulations that apply to any business, financial institutions are challenged with industry- specific requirements such as frequsently-changing regulations and their variation by region. Institutions that operate in more than one country or Region must also meet different requirements in different places. Financial institutions (FI) are historically a frequent target of security incidents, both because of the assets they maintain and their fundamental role in the daily functioning of a modern society. FIs have to comply with rules and laws around the protection of personal and financial information. The continuity of their operations depends directly on the security resilience they put in place.

Design principles

In addition to the design principles found in the security pillar of the AWS Well-Architected Framework, the following security design principles can help you improve the security posture of your financial services workloads:

- **Security by design**: Financial services institutions must consider a Security by Design (SbD) approach to implement architectures that are pre-tested from a security perspective. SbD helps implement the control objectives, security baselines, security configurations, and audit capabilities for applications running on AWS. Standardized, automated, prescriptive, and repeatable design templates help accelerate the deployment of common use cases as well as help align with security standards across multiple workloads. For example, to protect customer data and mitigate the risk of data disclosure or alteration of sensitive information by unauthorized parties, financial institutions need to employ encryption and carefully manage access to encryption keys. SbD allows you to turn on encryption for data at rest, in transit, and if necessary, at the application level by default.
- **Identify regulatory requirements to be implemented**: Regulators expect financial services institutions to define security objectives for workloads, and implement policies that help achieve those objectives. Regulators may also impose their own external requirements on specific workloads and expect institutions to monitor and report on their compliance with these requirements, with penalties for breaching them. Those requirements must be translated into security control objectives that are sustainable over time but flexible to adapt as regulations evolve.

- **Automated infrastructure and application deployment:** Automation helps companies to perform and innovate quickly and scale security, compliance, and governance activities across their cloud environments. Financial services institutions that invest in automated infrastructure and application deployment are able to accelerate the rate of deployments and embed security and governance best practices into their software development lifecycle.
- **Automated Governance:** Manual governance processes that rely on runbooks and checklists often lead to delays and inaccurate results. Automated governance provides a fast, definitive governance check for application deployments at scale. Governance at scale typically addresses the following components:
  - **Account management:** Automate account provisioning and maintain good security when hundreds of users and business units are requesting cloud-based resources.
  - **Budget and cost management:** Enforce and monitor budgets across many accounts, workloads, and users.
  - **Security and compliance automation:** Manage security, risk, and compliance at scale to keep the organization compliant while achieving business objectives.

Definitions

Security in the financial services industry is composed of the following best practice areas.

- Security foundations
- Identity and access management
- Detection
- Infrastructure protection
- Data protection
- Incident response
- Application security

Before you architect any workload, you need to put in place practices that influence security. You should control who can do what. In addition, you want to be able to identify security incidents, protect your systems and services, and maintain the confidentiality and integrity of data through data protection. You should have a well-defined and practiced process for responding to security incidents. These tools and techniques are important because they support objectives such as preventing financial loss or complying with regulatory obligations.

Security foundations

Questions

- FSISEC01: How does your governance enable secure cloud adoption at scale?
- FSISEC02: How do you achieve, maintain, and monitor ongoing compliance with regulatory guidelines and mandates?

FSISEC01: How does your governance enable secure cloud adoption at scale?

Cloud infrastructure provides more agility and responsiveness than traditional IT environments. This requires organizations to think differently about how they design, build, and manage applications. Cloud resources can be disposable. Because it is a pay-per-use model, it often requires a strong integration between IT governance and organizational governance. Financial services companies need to operate in a cloud environment that's agile and safe at the same time.

FSISEC01-BP01 Consider and leverage a Cloud Center of Excellence (CCoE)

When it comes to cloud adoption and governance, CCoEs (also referred as Cloud Enablement Engine (CEE)) are known drivers of change across the enterprise and the focal point for its transformation. CCoEs should have a functional model that is more aligned to provisioning and operating cloud resources, or they should act as the advisory group for cloud migrations and security baseline definitions. CCoEs help create and manage governance and security policies in collaboration with a cross-functional team and select governance tools to provide financial and risk management.

The following tenets are key guiding principles for creating a CCoE:

- The CCoE structure evolves as the organization changes.
- Treat the cloud as your product and application team leaders as the customers you are serving.
- Build company culture into everything you do.
- Organizational change management is central to business transformation. Use intentional and targeted organizational change management to change company culture and norms.
- Embrace a change-as-normal mindset. Security policies and procedures must be flexible enough to keep up with the changes in applications, IT systems, and business direction over the time and should be aligned with the financial services industry regulations and best practices.

- Operating model decisions determine how people fill roles that achieve business outcomes.

Traditionally, companies in the financial sector have distributed internal teams with distinct roles, as part of their division of duties policies. Even so, you can still get the benefits described here if the duties of a CCoE are distributed among multidisciplinary teams.

FSISEC01-BP02 Use cloud-native services for management and governance

Financial sector organizations focus on achieving security and compliance objectives in balance with faster innovation and agility. AWS Management and Governance native services takes advantage of both innovation and control as you can provision resources and applications to help meet your policies and operate your environment for business agility and governance control. These services are designed to make it easier to manage your AWS environment at scale, facilitating the secure adoption of cloud services without losing control of the environment growth.

The following articles and blogs provide advice for improving the overall security of your workloads and to hone the security posture of your internal IT resources.

The section Building a CCOE to transform the entire enterprise, from AWS documentation describes the benefits of creating a Cloud Center of Excellence (CCOE) within your organization. This allows you to adopt a number of policies that helps you evolve your security measures across several dimensions over time and scope.

The whitepaper Cloud Enablement Engine: A Practical Guide describes the step-by-step process for the initial setup activities for a CCOE, and the top ten best practices gleaned by AWS while working across a large number of customers.

By using a Service Catalog, your organization can create and manage catalogs of IT services that are approved for AWS. These IT services can include everything from virtual machine images, servers, software, databases, to complete multi-tier application architectures. For more information, see Manage pre-approved services for secure adoption at scale with Service Catalog.

AWS Control Tower can offer a straightforward way to set up and govern an AWS multi-account environment, following prescriptive best practices. AWS Control Tower orchestrates the capabilities of several other AWS services, including AWS Organizations, Service Catalog, and AWS IAM Identity Center. It allows you to build a landing zone in less than an hour.

Resources

Related documents:

- Using a Cloud Center of Excellence (CCOE) to Transform the Entire Enterprise
- 7 Pitfalls to Avoid When Building a CCOE
- AWS Control Tower and AWS Security Hub – Powerful Enterprise Twins

Related videos:

- Transform your organization's culture with a Cloud Center of Excellence
- How to Build Your Cloud Enablement Engine with the People You Already Have

FSISEC02: How do you achieve, maintain, and monitor ongoing compliance with regulatory guidelines and mandates?

Companies in the financial sector have more demanding compliance monitoring and implementation requirements than most other sectors of the economy. Traditional methods of compliance assessment do not keep pace with the dynamics of the agile cloud environment. For this reason, the best practices and tools required are specific to this type of environment. Regulations ensure that consumers' personal and financial data are protected. Compliance with these regulations helps prevent identity theft, fraud, and unauthorized disclosure of personal information. Compliance also helps maintain the integrity and stability of the financial markets by ensuring that institutions engage in responsible lending and investment practices and avoid excessive risk-taking. The following best practices help facilitate compliance in the cloud.

FSISEC02-BP01 Automate your compliance management

AWS has services to help you identify, optimize and remediate resource configurations for continuous compliance and operational efficiency. AWS services help customers achieve immutable resource configuration and offer configurable logging for the auditing of user and API activity. Using AWS Config and its proactive mode helps you save time and remove the risk of human error when you automate and scale compliance management. It helps FIs (mainly the first line of defense) effectively manage risk for their cloud resources.

FSISEC02-BP02 Use ready-to-deploy templates for standards and best practices

Ready-to-deploy templates are a quick and assertive way to measure what level of security is present in cloud environments. These templates are available both for best practices in technology such as database, serverless, and networking, and are aligned to frameworks that are widely

accepted and recognized. Among the most suitable templates are managed rules, AWS Config Conformance Packs in AWS Config, and AWS Security Hub standards. FIs can benefit from Conformance Packs that are available and ready to be used for alignment to the financial services industry's standards and regulatory requirements, such as PCI-DSS, NYDFS, and FFIEC.

Prescriptive guidance

- A Conformance Pack can be deployed as is or it can be edited to include your specific resources and use cases. For more information, see Deploying a Conformance Pack Using the AWS Config Console.
- When adding a new rule, choose how it evaluates your resources, as well as how it is initiated. For more information, see Evaluation Mode and Trigger Types for AWS Config Rules.
- To determine if requirements in a standard are being met, enable the controls from AWS Security Hub standards. For more information, see Security standards and controls in AWS Security Hub.

Resources

Related documents:

- AWS Config Rules Now Support Proactive Compliance

Related videos:

- Cloud compliance, assurance, and auditing
- Setting up controls at scale in your AWS environment
- Proactive governance and compliance for AWS workloads

Identity and access management

Questions

- FSISEC03: How do you monitor the use of elevated credentials, such as administrative accounts, and guard against privilege escalation?
- FSISEC04: How do you accommodate separation of duties as part of your identity and access management design?

FSISEC03: How do you monitor the use of elevated credentials, such as administrative accounts, and guard against privilege escalation?

IAM policies are powerful and complex, so it's important to study and understand the permissions that are granted by each policy. Mitigate privilege escalation and monitor unauthorized activity in your AWS accounts.

FSISEC03-BP01 Review IAM policies and permissions

IAM policies are powerful and complex, so it's important to study and understand the permissions that are granted by each policy.

As part of the tight controls FIs implement around identity management and broader identity management policies, it is important to perform periodic reviews of your IAM roles using last accessed information to get a report about the last time that an IAM entity (user or role) attempted to access a service, and delete roles that are not in use. Before you delete a role, review its recent service-level activity by viewing service last accessed data report. Use that information to refine your policies to allow access to only the services that are in use. Repeat this process to generate a report for each type of resource in IAM.

FSISEC03-BP02 Mitigate privilege escalation

Privilege escalation refers to the ability of unauthorized users gaining access to elevated permissions, often by way of improperly written code or misconfigurations. Privilege escalation can result from misusing a number of non-administrator or non-full access permissions. To help avoid scenarios like this, pay attention to permissions that would allow the creation, change and deletion of users, roles, and policies.

As a way to help prevent privilege escalation, you should use service control policies (SCPs) to block users in your accounts, except for IAM administrators or delegated admins, from performing administrative IAM actions. Delegation is a common practice for FIs. If you want to safely delegate permissions management to trusted employees, use IAM permissions boundaries. IAM permissions boundaries allow for safe delegation of IAM permissions management while minimizing escalation of privileges. For example, developers can safely create IAM roles for Lambda functions and Amazon EC2 instances without exceeding certain permissions boundaries defined by your IAM administrators.

FSISEC03-BP03 Monitor unauthorized activity in your AWS accounts

Use the following guidelines to monitor your AWS account activity:

- Turn on AWS CloudTrail in each account, and use it in each supported Region.
- Store AWS CloudTrail log in a centralized logging account with very restricted access.
- Periodically examine CloudTrail log files. Use Amazon GuardDuty, which provides threat detection by continually analyzing AWS CloudTrail events, VPC Flow Logs and DNS logs.
- Enable Amazon GuardDuty in each account, and use it in each supported Region to automatically detect CloudTrail management events that can lead to IAM privilege escalation and other IAM finding types.
- Enable Amazon S3 bucket logging to monitor requests made to each bucket.
- If you believe there has been unauthorized use of your account, pay attention to temporary credentials that have been issued. If temporary credentials have been issued that you don't recognize, disable their permissions.
- View the last accessed information for IAM through the Management Console, CLI or AWS API.

Administrators can configure roles to require identities to pass a custom string that identifies the person or application that is performing actions in AWS when the role is assumed. This identity information is stored as the source identity in AWS CloudTrail. Administrators can review this activity in CloudTrail, and they can view the source identity information to determine who or what performed actions with assumed role sessions.

It is also a good practice to periodically review IAM policies as well as setting restrictive user access on a need to know basis. You can prevent IAM user and roles from making specified changes, through Service Control Policies (SCPs) and set Permissions boundaries for IAM entities.

Resources

Related documents:

- How to use trust policies with IAM roles
- Monitor and Notify on AWS Account Root User Activity

Related videos:

- AWS re:Inforce 2022 - Security best practices with AWS IAM

FSISEC04: How do you accommodate separation of duties as part of your identity and access management design?

FSISEC04-BP01 Implement the principle of separation of duties

Separation of duties, as it relates to security, has two primary objectives. The first objective is the prevention of conflict of interest, abuse, and errors. The second objective is the detection of control failures that include security breaches, information theft, and circumvention of security controls. While robust automation of infrastructure and application deployments helps reduce the need for human access, there can be instances where individuals need to complete key functions. For users with increased privileges, it is important to distribute system administration activities, so no one administrator can hide their activities or control an entire system. Separation of duties can help mitigate risk on critical tasks by ensuring different people are required to perform a task where the requestor and the approver can't be the same person. A common example is the use of an approver during the running of an automation on AWS Systems Manager. This principle can be used to implement numerous tasks including controlling access to your cloud resources.

FSISEC04-BP02 Use AWS Config to view historical IAM configuration and changes over time

Use AWS Config to view the IAM policy that was assigned to an IAM user, group, or role at any time in which AWS Config was recording. This information can help you determine the permissions that belonged to a user at a specific time. For example, it allows you to view whether a user had permission to modify settings on a specific date in the past.

FSISEC04-BP03 Set up alerts for IAM configuration changes and perform audits

Set up alerts to notify on IAM configuration changes including when an IAM user is created or when conflicting permissions are added to a user or role, such as being able to approve its own requests on a given workflow. This is helpful for monitoring activities by users with increased privileges. The added notification can be set up using a combination of AWS CloudTrail, Amazon CloudWatch, and Amazon SNS.

Prescriptive guidance

- To manage changes for an entire organization or for a single AWS account, you can use Change Manager, a capability of AWS Systems Manager. For more details see, Setting up Change Manager at AWS Systems Manager.

- AWS Config is a service that helps you manage compliance state changes for resources. For more details, see Viewing AWS Resource Configurations and History.
- An approval process for changes can be deployed using AWS Step Functions. To review the step-by- step tutorial, see Deploying an Example Human Approval Project.

Resources

Related documents:

- Apply the principle of separation of duties to shell access to your EC2 instances
- How to Record and Govern Your IAM Resource Configurations Using AWS Config

Related videos:

- Least Privilege & Separation of Duties for AWS ACM Private CA

Detection

Questions

- FSISEC05: How are you monitoring your ongoing cloud environment for potential threats?
- FSISEC06: How do you address emerging threats?
- FSISEC07: How are you inspecting your financial services infrastructure and network for unauthorized traffic?

FSISEC05: How are you monitoring your ongoing cloud environment for potential threats?

Financial services organizations require in-depth visibility into the security of their infrastructure and applications. Achieving this high level of visibility requires the collection of logs and audit trails and the reservation of these logs for analytics and reporting. AWS services and partners' cloud-native solutions help you implement real-time monitoring in your environment for security threats and alerting on threats once detected.

FSISEC05-BP01 Track configuration changes

As part of monitoring the environment against threats, it is critical to identify changes in the security settings that keep the environment protected. One of the benefits of the cloud is being able to maintain full visibility of what is changing in the environment. Establishing a security baseline of the deployed resources is key for a FIs first line of defense to manage the risk of its infrastructure, as well as to track changes over time.

Use AWS Config to audit and evaluate the configuration settings of your AWS resources. AWS Config continually tracks the configuration changes that occur in your resources, and by using AWS Config Managed Rules, it checks to see if these changes comply with the your defined desired state. This allows you to identify and correct configuration deviations as soon as they happen, and also helps the second and third lines of defense respond quickly.

FSISEC05-BP02 Detect unusual and unauthorized activity early

Cloud processing of large event data helps detect unauthorized activity early, which is crucial in a financial institution's incident response strategy.

Threat detection services like Amazon GuardDuty can continually monitor for unauthorized behavior to protect your AWS accounts and workloads by focusing on indication of compromise of credentials, resources, accounts or buckets. Enable Amazon GuardDuty on all of the accounts in your AWS Organization and for all of the AWS Regions, as it can detect unintended activities in unused Regions as well.

AWS Security Hub provides you with a comprehensive view of the security state in AWS and helps you check your environment against security industry standards and best practices. The activities surrounding Amazon GuardDuty and AWS Security Hub must also be tracked and analyzed using AWS CloudTrail, and they can feed a normalized central data-lake of your security-related information on Amazon Security Lake.

Detecting malware in your environment is essential. Consider enabling malware protection in Amazon GuardDuty to identify your resources that are at risk or have already been compromised by malware. Whenever Amazon GuardDuty detects suspicious behavior on an EC2 instance or a container workload, malware protection automatically initiates an agentless scan on the EBS volume attached to the resource to detect the presence of malware.

Additionally, you should also consider scanning data coming in through third party sources and often landing in your S3 buckets, as they may expose you to potentially malicious files, objects that

may be infected with malware, ransomware, or viruses. To do this, leverage AWS Partner solutions found in the AWS Marketplace.

AWS CloudTrail insights helps AWS users identify and respond to unusual activity associated with API calls by continually analyzing CloudTrail management events, and should be enabled in your trails.

You can track configuration changes at the edge with AWS Config, by recording and tracking CloudFront distribution settings changes.

Resources

Related documents:

- Cloud security software - AWS Marketplace
- GuardDuty Malware Protection FAQ

Related videos:

- The top 7 ways to operationalize AWS Security Hub

FSISEC06: How do you address emerging threats?

Security-focused enterprises are improving threat identification and remediation with DevSecOps. This approach accelerates application development and identifies threats early, and security testing is performed at each step of the software development lifecycle. Applying a DevSecOps framework is critical for an FI's software development, meeting the needs of a rapidly-changing product and a highly- regulated environment.

FSISEC06-BP01 Automate remediation of common vulnerabilities and exposures (CVEs)

Scanning servers for common vulnerabilities is a long-standing best practice. However, in the cloud, you should not only automate the evaluation of operating environments and applications, but also remediate known and emerging security vulnerabilities automatically. For example, you can use Amazon Inspector service to automatically scan servers in production, publish security findings to an Amazon Simple Notification Service (SNS) topic, run an AWS Lambda function from those notifications to examine the findings, and implement the appropriate remediation based on the type of issue.

FSISEC06-BP02 Perform static analysis on all code deploys

As part of a DevSecOps strategy, you can secure your application deployments by integrating preventive and detective security controls within the pipeline. One of the key benefits of static code analysis is that you can learn about security vulnerabilities prior to provisioning AWS resources, which can help reduce costs and risk.

FSISEC06-BP03 Conduct regular penetration testing

Simulating security incidents inside the AWS environment helps you have a better understanding of your security posture. Financial services organizations perform penetration testing of web applications most often when a new application is launched or when it's first migrated to the cloud. Some may even conduct penetration testing periodically every year. Run penetration testing regularly after every major release that involves significant re-architecture changes. Major releases might introduce vulnerabilities that didn't exist earlier.

FSISEC06-BP04 Deploy web application firewalls

AWS WAF is an application firewall service for HTTP applications that applies a set of rules to an HTTP conversation. You can buy managed rule sets from the AWS Marketplace that protect against application vulnerabilities, such as the Open Worldwide Application Security Project (OWASP Top 10), bots, or emerging CVEs. Managed rules are automatically updated by AWS Marketplace security sellers.

Prescriptive guidance

- Automation is key to maintain continuous vulnerability management and a remediation posture. For details, see Automate vulnerability management and remediation in AWS.
- Application modernization leads to containerized applications. You can deploy vulnerability management into your CI/CD pipeline and scan container images. For more details, see Use Amazon Inspector to manage your build and deploy pipelines for containerized applications.
- From a shift left approach, apply vulnerability management in your CI/CD pipeline. For more details, see Detect security vulnerabilities and automate code reviews.

Resources

Related documents:

- Penetration Testing at AWS

- Detect Python and Java code security vulnerabilities with Amazon CodeGuru Reviewer
- Amazon Inspector FAQs

Related videos:

- AWS re:Invent 2022 - Detect vulnerabilities in AWS Lambda functions using Amazon Inspector

FSISEC07: How are you inspecting your financial services infrastructure and network for unauthorized traffic?

Monitor network traffic for expected and unexpected traffic to identify irregularities and gain key insights into the security of the system. For example, a poorly-performing network can indicate that the network is under threat, and irregular attempts to contact unexpected external systems can indicate that an internal host has been compromised.

FSISEC07-BP01 Monitor instance traffic

Amazon EC2 instances automatically track aggregate network inbound and outbound traffic with Amazon CloudWatch. Use custom metrics and push log files to Amazon CloudWatch for storage, aggregation, reporting, and alert notification. Create profiles for the expected network behavior for each EC2 instance and generate alarms when deviations are detected. For example, system or web logs sent to Amazon CloudWatch Logs could generate alarms based on the number of login failures or web request latencies. Similarly, TCP connection or outstanding connection request counts could be stored in Amazon CloudWatch and used to detect security threats like SYN flood threats.

FSISEC07-BP02 Use VPC Traffic Mirroring

Use VPC Traffic Mirroring to copy network traffic from an elastic network interface of Amazon EC2 instances and forward that traffic to security and monitoring appliances for use cases such as content inspection, threat monitoring, and troubleshooting. These security and monitoring appliances can be deployed on a fleet of instances behind a Network Load Balancer (NLB) with a User Datagram Protocol (UDP) listener. Amazon VPC traffic mirroring supports traffic filtering and packet truncation, allowing you to extract traffic that you are interested in monitoring. It also addresses challenges around having to install and run packet-forwarding agents on EC2 instances. Packets are captured at the Elastic Network Interface level, which cannot be tampered with from the user space, thus offering better security posture.

FSISEC07-BP03 Use immutable infrastructure with no human access

Immutable infrastructure is a model in which no updates, security patches, or configuration changes happen in place on production systems. If changes are needed, a new version of the architecture is built and deployed. Because changes aren't allowed in immutable infrastructure, you can be confident in the deployed system. Immutable infrastructures are more consistent, reliable, and predictable, and they simplify many aspects of software development and operations by minimizing common issues related to mutability.

Adopt immutable infrastructure practices with no human access to better adhere to your audit and compliance needs. You can version control your infrastructure, and handling failure becomes a routine and continual way of doing business.

FSISEC07-BP04 Allow interactive access for emergencies only

Tightly control and monitor interactive access to EC2 instances. Interactive access should typically be provided for emergency-only, break-glass scenarios.

Test and review these pre-staged emergency user accounts, which normally are highly privileged and could be limited to read only. Limit the time duration of break-glass procedure and the password time duration. Have a ticketing system with procedures requiring that an acceptable form of authentication be provided by the requester and recorded before the accounts are made available. This helps control and reduce the account's misuse, having only pre-approved personnel who complete a certain emergency task. The break-glass accounts and distribution procedures must be documented and tested as part of implementation and carefully managed to provide timely access when needed. A special audit trail needs to be in place to monitor such emergency access for later audit and review.

Use AWS Systems Manager Session Manager to provide an interactive, one-click browser-based shell to your Amazon EC2 instances, on-premises instances, and virtual machines (VMs). Session Manager provides secure and auditable instance management without the need to open inbound ports, maintain bastion hosts, or manage SSH keys.

Prescriptive guidance

- Publish and view statistical graphs of your own metrics with Amazon CloudWatch. For more details, see Publishing custom metrics.
- You can use the CloudWatch feature of Anomaly Detection, which analyzes past metric data to create a model of expected values. The steps for that implementation is described in the following documentation: Implement CloudWatch alarms based on anomaly detection.

- Enable traffic mirroring to analyze the selected traffic from a mirror source sent to a mirror target. For more information, see Get started with Traffic Mirroring.
- To adopt a strategy of immutable servers, see the following blog post: Create immutable servers using EC2 Image Builder and AWS CodePipeline.

Resources

Related documents:

- Leveraging AWS CloudFormation to create an immutable infrastructure
- Managing temporary elevated access to your AWS environment

Related videos:

- AWS re:Invent 2022 - A deep dive on the current security threat landscape with AWS

Infrastructure protection

Questions

- FSISEC08: How do you isolate your software development lifecycle (SDLC) environments (like development, test, and production)?

FSISEC08: How do you isolate your software development lifecycle (SDLC) environments (like development, test, and production)?

We recommend that you separate production workloads from non-production workloads. Maintaining resource isolation between software development lifecycle (SDLC) environments reduces the chance of misuse and accidents in production environments. This is an important guidance for all financial institutions, including those that are subject to Payment Card Industry Data Security Standard (PCI DSS).

FSISEC08-BP01 Implement a multi-account strategy

Using multiple AWS accounts to help isolate and manage your business applications and data can help you optimize across most of the AWS Well-Architected Framework pillars, including

operational excellence, security, reliability, and cost optimization. We recommend organizing your overall AWS environment with a multi-account strategy. The extent to which you use these best practices depends on your stage of the cloud adoption journey and specific business needs.

We recommend that you isolate production workload environments and data in production accounts housed within production OUs, under your top-level workload-oriented OUs. Apart from production OUs, we recommend that you define one or more non-production OUs that contain accounts and workload environments that are used to develop and test workloads.

Having different accounts dedicated to different SDLC environments provides a natural isolation in managing privileges in IAM. AWS Organizations facilitates the management of account hierarchy. Define service control policies (SCPs) to limit the actions a user can perform inside these accounts. For example, you could minimize changes in production to CloudTrail logging, help prevent internet gateways set up in a VPC, or help prevent modifying AWS Config tracking.

To offer a straightforward way to set up and govern an AWS multi-account environment that follows prescriptive best practices, AWS has created AWS Control Tower, which extends the capabilities of AWS Organizations. To help keep your organizations and accounts from *drift*, or divergence from best practices, AWS Control Tower applies comprehensive controls (sometimes called *guardrails*). For more detail, see Limitations and quotas in AWS Control Tower.

FSISEC08-BP02 Enforce network isolation

Some financial industry regulators require the implementation of techniques such as Zero Trust or microsegmentation in their regulated entities. In addition to IAM isolation, enforce clear separation of resources between production and non-production environments. Using different accounts helps create the highest form of isolation possible on AWS. However, you may need to reach resources across accounts, especially when accessing shared services such as logging and security services.

VPC Peering connects resources in two VPCs (in the same account or between different accounts) without the need of additional gateways or VPN connections, and it makes the peered network visible to each other. This requires complete network trust between the two VPCs, and better alternatives exist depending on your use case. If the objective is to access only a few services in the other VPC, use AWS PrivateLink, which provides connectivity over an internal network without VPN and limits network exposure. Service publishers also have to specify which IAM principals can consume these endpoints and attach an IAM resources policy specifying what actions are allowed. If more extensive cross-VPC access is needed, separation and private connectivity can be also established with AWS Transit Gateways.

Resources

Related documents:

- Best Practices for Organizational Units with AWS Organizations
- Supporting Data Residency Requirements by Extending AWS Control Tower Governance to Non-supported Regions
- The AWS Security Reference Architecture
- Zero Trust architectures: An AWS perspective

Related videos:

- AWS Summit DC 2022 - Integrating AWS services and Zero Trust networks
- AWS re:Invent 2020: Zero Trust: An AWS perspective

Data protection

Questions

- FSISEC09: How are you managing your encryption keys?
- FSISEC10: How are you handling data loss prevention in the cloud environment?
- FSISEC11: How are you protecting against ransomware?

FSISEC09: How are you managing your encryption keys?

In addition to implementing the data protection recommendations applicable to any company seen in the AWS Well-Architected Framework Security Pillar, financial institutions often have additional industry-specific requirements that can influence the management of cryptographic keys.

FSISEC09-BP01 Consider compliance obligations regarding location of cryptographic keys

AWS Key Management Service (AWS KMS) uses an *envelope encryption* strategy, which consists of encrypting plaintext data with a data key, and then encrypting the data key with another key. AWS KMS keys are created in AWS KMS and never leave AWS KMS unencrypted.

AWS KMS supports three types of keys: customer-managed keys, AWS managed keys, and AWS owned keys (for more information, see the AWS KMS concepts). For many FSI customers, customer-managed keys are the preferred option, because they allow for control of the permissions to use keys from their applications or AWS services. It also provides added flexibility for key generation and storage.

Although it's less common, AWS customers who have a compliance or regulatory need to store and use their encryption keys on-premises or outside of the AWS Cloud can do so by using external key stores.

Prescriptive guidance

- Work backwards from your company's compliance objectives and security standards in order to determine the right encryption method for your use case.
  - Leverage AWS audit reports, available for download at AWS Artifact, to understand the controls implemented by AWS, and tested for operating effectiveness by third-party auditors on AWS KMS.
  - Review the list of services that you are using for your workload to understand how AWS KMS integrates with the service.
  - Review AWS Encryption SDK with AWS KMS integration if your application needs to encrypt data client-side.
- Evaluate the differences between different key types in AWS KMS.
- When using customer managed keys, consider the default key store to provide the best balance between agility, security, data sovereignty, and availability.
- Consider using custom key stores with AWS CloudHSM or the external key store to adhere to specific compliance obligations.

Resources

Related documents:

- How Financial Institutions can Select the Appropriate Controls to Protect Sensitive Data
- Announcing AWS KMS External Key Store (XKS)

Related videos:

- AWS re:Invent 2022 – Protecting secrets, keys, and data: Cryptography for the long term

- AWS re:Invent 2022 – AWS data protection: Using locks, keys, signatures, and certificates
- AWS re:Invent 2022 – Introducing AWS KMS external keys

FSISEC10: How are you handling data loss prevention in the cloud environment?

Data loss as part of a security event, accident or business process can affect both your operation and state of compliance. The following recommendations can help with the protection from theft and inadvertent or malicious loss.

FSISEC10-BP01 Prevent modifications and deletions of logs and data

Financial services agencies around the world, including the Securities and Exchange Commission (SEC) and the Financial Industry Regulatory Authority (FINRA) in the US, have created rules that require a broker-dealer to maintain and preserve electronic records exclusively in a non-rewriteable, non-erasable format, also known as a write once, read many (WORM) format.

For object data, Amazon S3 Object Lock allows you to store objects using a WORM model. You can use WORM protection for scenarios where it is imperative that data is not changed or deleted after it has been written. With S3 Object lock, you can securely deliver logs to a designated S3 bucket, and use the S3 Object Lock feature to make the logs immutable. It blocks object version deletion during a customer- defined retention period so that you can enforce retention policies. In conjunction with S3 versioning, which protects objects from being overwritten, you're able to keep objects immutable for as long as S3 Object Lock protection is applied.

For file data, use SnapLock, a feature on Amazon FSx for NetApp ONTAP that allows you to store files using a WORM model, helping prevent accidental or malicious attempts at modification and deletion for a customizable retention period. You can also back up data on FSx for ONTAP using AWS Backup and WORM-protect your backups using AWS Backup Vault Lock.

FSISEC10-BP02 Limit and monitor key deletes

Once encrypted, the data is protected by cryptographic keys that must be kept as long as the data is to be accessed. Only key administrators should perform key deletion. Review all destruction requests within the safety window, as a key cannot be destroyed immediately. Instead, it is disabled, which prevents use, and is deleted at the expiry of the window.

To help validate that the key deletion won't impact your company, set up an alarm that detects use of an AWS KMS key pending deletion.

Prescriptive guidance

- Make sure that the Amazon S3 buckets are configured to use the Object Lock feature to help prevent the objects they store from being deleted, and help meet regulatory compliance needs.
- Make sure that Amazon S3 object versioning is enabled for your Amazon S3 buckets in order to preserve and recover overwritten and deleted Amazon S3 objects as an extra layer of data protection or data retention.
- Set up AWS Config managed rule to identify Amazon S3 buckets that do not have versioning enabled, and implement automatic remediation to configure versioning on non-compliant Amazon S3 buckets.
- Implement backup and restore processes to help you restore data to a point in time before data corruption, modification or destruction. AWS provides several solutions for backups to integrate with your operational and security incident recovery procedures.
  - Use AWS Backup with AWS Organizations to centrally deploy data protection policies to configure, manage, and govern your backup activities across your AWS accounts and resources.
  - Beyond creating and storing your backups, AWS Backup Audit Manager can continuously evaluate backup activity and generate audit reports that can help you demonstrate compliance with regulatory requirements. These reports also provide you with more visibility into your backup activities, helping you monitor your operational posture and identify failures that may need further action.
- Deleting an AWS KMS key is destructive and potentially dangerous. After an AWS KMS key is deleted, you can no longer decrypt the data that was encrypted under that AWS KMS key, which means that data becomes unrecoverable.
  - Delete an AWS KMS key only when you are sure that you don't need to use it anymore.
  - If you are not sure, consider disabling the AWS KMS key instead of deleting it.
  - Control access to key deletion by creating fine-grained access control policies and allow only authorized principals with the ability to schedule key deletion.
- Create an alarm to detect and notify on AWS KMS key deletion events.
- Create an alarm to detect usage of an AWS KMS key that is scheduled for deletion.

Resources

Related documents:

- How to manage retention periods in bulk using Amazon S3 Batch Operations

Related videos:

Data protection strategies for the cloud - AWS Online Tech Talks

FSISEC11: How are you protecting against ransomware?

Ransomware refers to a business model and a wide range of associated technologies that bad actors use to extort money. The bad actors use a range of tactics to gain unauthorized access to their victims data and systems, including exploiting unpatched vulnerabilities, taking advantage of weak or stolen credentials, and using social engineering. Access to the data and systems is restricted by the bad actors, and a ransom demand is made for the safe return of these digital assets.

FSISEC11-BP01 Prevent malware infiltration by securing compute resources

To detect malware that may be the source of a ransomware incident, enable malware protection in Amazon GuardDuty. This feature automatically initiates an agentless scan on the Amazon Elastic Block Store (EBS) volumes attached to the impacted EC2 instance or container workload to detect the presence of malware.

FSISEC11-BP02 Prevent threats from accessing your data stores

Scoping access to data based on the principal of minimum privileges helps prevent as well as limit the blast radius of an exploit. An effective data classification scheme, along with enforcement and monitoring based on that scheme can help prevent an bad actor from having accessing and encrypting your data.

Network isolation and segregation is another effective protection as compromised systems cannot reach deep into your network. Leverage the best practices recommended in the Infrastructure protection section to funnel access to data stores over a private network, from a limited number of hosts.

FSISEC11-BP03 Use frequent backups to recover from a threat

Because ransomware makes itself known quickly, incorporate short-lived anti-ransomware backups into your backup cycle. AWS take snapshots of data stores, so back up often and keep these around for only a few days to limit costs.

For more information on how to protect from Ransomware at AWS, see Ransomware Risk Management on AWS Using the NIST Cyber Security Framework (CSF).

Prescriptive guidance

- Use Amazon S3 Object Lock for object storage immutability and ransomware protection within cloud storage.
- Implement backup and restore processes to help you restore data to a point in time before data corruption, modification or destruction. AWS provides several solutions for backups to integrate with your operational and security incident recovery procedures.
  - Use AWS Backup with AWS Organizations to centrally deploy data protection policies to configure, manage, and govern your backup activities across your AWS accounts and resources.
  - Enable AWS Backup Vault Lock, which enforces WORM (write-once-read-many) setting for the backups you store and create in a backup vault.
- Because many ransomware events arise from unintended disclosure of static IAM access keys, AWS recommends that you use IAM roles that provide short-term credentials, rather than using long-term IAM access keys. This includes using identity federation for your developers who are accessing AWS, using IAM roles for system-to-system access, and using IAM Roles Anywhere for hybrid access.
- Enable Amazon S3 protection in Amazon GuardDuty. With Amazon S3 protection, GuardDuty monitors object-level API operations to identify potential security risks for data in your Amazon S3 buckets.

This includes findings related to anomalous API activity and unusual behavior related to your data in Amazon S3, and can help you identify a security event early on.

- Enable Amazon GuardDuty Malware Protection across all AWS accounts in your organization, to help you detect the potential presence of malware by scanning the Amazon EBS volumes that are attached to the Amazon EC2 instances and container workloads.

Resources

Related documents:

- Protecting against ransomware
- GuardDuty findings that initiate Malware Protection scans
- Ransomware Risk Management on AWS Using the NIST Cyber Security Framework (CSF)
- Ransomware mitigation: Top 5 protections and recovery preparation actions

- Workshop: Ransomware on S3 - Simulation and Detection

Related videos:

- What is Amazon GuardDuty Malware Protection? | Amazon Web Services
- AWS re:Invent 2021 - Backup, disaster recovery, and ransomware protection with AWS

Incident response

Incident response is an integral part of a cyber security strategy, both on-premises and in the cloud. It is important to know which controls and capabilities are available, review topical examples for resolving potential concerns, and identify remediation methods that use automation to improve response speed and consistency. You should also understand and fulfill your compliance and regulatory requirements as they relate to building a security incident response program.

- As organizations grow and evolve over time, so does the threat landscape, making it important to continually review your incident response capabilities. Game days simulate a failure or event to test systems, processes, and team responses. This helps you understand where improvements can be made and can help develop organizational experience in dealing with events.
- Conduct game days regularly so that your team builds muscle memory on how to respond.

Questions

- FSISEC12: How are you meeting your obligations for incident reporting to regulators?

FSISEC12: How are you meeting your obligations for incident reporting to regulators?

Various regulations require that the banking organizations and managed service providers notify the regulators as soon as a cyber security incident has been discovered, such as the Final Issuances published by the Office of the Comptroller of the Currency (OCC), Security and Exchanges Commision (SEC) Cybersecurity Disclosure or the Network and Information Systems (NIS) regulation.

FSISEC12-BP01 Regularly review your incident response plan for regulatory compliance

Organizations that are operating in multiple Regions need to be aware the regulatory requirements of the regions they are operating in and any local data residency requirements (such as GDPR). With local data residency requirements, you cannot copy the data to a different Region for analysis purposes. In this case, you may need to consider the latency aspects if you have a global team that needs to access and analyze data from a different Region. Consider setting up a local incident response team that can act on the incident in a timely manner and report to local regulators as necessary.

As mentioned before, as part of your incident response plan, you should develop playbooks to standardize response process for cybersecurity incidents. With the ever-changing regulatory requirements of the financial industry and the dynamic nature of cloud environments, it is important to establish a process that reviews the playbooks in use to perform incident or recovery communications as required.

Prescriptive guidance

- Create your own playbooks to facilitate responses during cybersecurity incidents. Refer to building incident response playbooks for AWS for sample playbooks.
- Use AWS Compliance Center for information on regulatory responsibilities that can be related to incident responses.

Resources

Related documents:

- General Data Protection Regulation (GDPR) Center

Related videos:

- Introduction to AWS Compliance Center

Key AWS services

- **Security foundations**

  - AWS Control Tower: Set up and govern a secure, multi-account AWS environment
  - AWS Organizations: Centrally manage your environment as you scale your AWS resources
  - AWS Config: Assess, audit, and evaluate configurations of your resources
- **Identity and access management**
  - AWS Identity and Access Management (IAM): Control users' access to and usage of AWS. Create and manage users and groups and grant or deny access. Enforce strong authorization and authentication.
  - AWS Identity Center: Centrally manage workforce access to multiple AWS accounts and applications
  - Amazon Cognito: Implement secure, frictionless customer identity and access management that scales
  - AWS Secrets Manager: Quickly rotate, manage and retrieve database credentials, API keys, and other secrets through their lifecycle
- **Detection**
  - AWS Security Hub: Automate AWS security checks and centralize security alerts
  - Amazon GuardDuty: Protect your AWS accounts with intelligent threat detection
  - Amazon CodeGuru: Automate code reviews and optimize application performance with ML-powered recommendations
  - Amazon Inspector: Automated and continual vulnerability management at scale
  - Amazon CloudWatch: Observe and monitor resources and applications on AWS, on premises, and on other clouds
- **Infrastructure protection**
  - Amazon VPC: Define and launch AWS resources in a logically isolated virtual network
  - AWS Network Firewall: Deploy network firewall security across your VPCs
  - AWS Verified Access: Provide secure access to corporate applications without a VPN
  - AWS WAF: Protect your web applications from common exploits
- **Data protection**
  - AWS CloudTrail: Track user activity and API usage
  - AWS Key Management Service (KMS): Create and control keys used to encrypt or digitally sign your data
  - Amazon Macie: Discover and protect your sensitive data at scale
  - AWS Backup: Centrally manage and automate data protection

- Amazon S3 Glacier: Long-term, secure, durable storage classes for data archiving at the lowest cost and milliseconds access

- **Incident response**
  - AWS Lambda: Run code without thinking about servers or clusters
  - AWS Audit Manager: Continuously audit your AWS usage to simplify how you assess risk and compliance with regulations and industry standards.
  - AWS GameDay: Fun, gamified, hands-on learning
  - AWS Compliance Center: Research cloud-related regulatory requirements

Reliability

The reliability pillar provides guidance to help customers apply best practices in the design, delivery, and maintenance of AWS environments. The reliability pillar provides best practices on how a system can recover from infrastructure or service disruptions, dynamically acquire computing resources to scale demand, and mitigate disruptions caused by events such as misconfigurations or transient network issues.

The technology systems of financial institutions are complex and highly interconnected to each other, and to non-financial entities. The proper functioning of many industries depends on certain types of workloads, for example, payment processing, trading and settlement, market data, custody and entitlement management, and financial messaging. Regulators continue to focus on the resilience of financial institutions through bodies such as the Basel Committee on Banking Supervision, Board of Governors of the Federal Reserve System, RegSCI, Bank of England and other regulatory bodies, issuing policies and guidance that the financial services institutions need to adhere to.

In this section, we provide in-depth best practices that financial institutions can use with AWS services to construct highly available, resilient, and scalable solutions at lower costs compared to traditional on-premises IT. To discuss these best practices, we use the concept of service availability interchangeably with the Recovery Time Objective (RTO) and Recovery Point Objective (RPO). An introduction to the concept of service availability and its relation to the recovery objectives can be found in the Well-Architected Reliability Pillar.

Design principles

Financial institutions can leverage AWS services to provide the levels of resilience and availability that their workloads need based on their criticality. The AWS Global infrastructure is built around Regions, Availability Zones (AZs), Local Zones, and edge locations. Our AWS services are of global, Regional, or zonal nature. For example, Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Elastic Block Store (Amazon EBS) are zonal services. A zonal service is one that provides the ability to specify which Availability Zone the resources are deployed into.

These services operate independently in each Availability Zone within a Region, and more importantly, fail independently in each Availability Zone as well. This means that components of a service in one Availability Zone don't have dependencies on components in other Availability Zones. We can do this because a zonal service has zonal data planes. Services like Amazon

Simple Storage Service (Amazon S3), Amazon Simple Queue Service (Amazon SQS) and Amazon DynamoDB are Regional services.

Regional services are services that AWS has built on top of multiple Availability Zones so that customers don't have to figure out how to make the best use of zonal services. We logically group together the service deployed across multiple Availability Zones to present a single Regional endpoint to customers. In addition to Regional and zonal AWS services, there is a small set of AWS services like IAM and Amazon Route 53, that do not have control planes and data planes that exist independently in each Region. Because their resources are not Region-specific, they are commonly referred to as *global*. Global AWS services still follow the conventional AWS design pattern of separating the control plane and data plane in order to achieve static stability. The significant difference for most global services is that their control plane is hosted in a *single* AWS Region, while their data plane is globally distributed. Therefore, when building critical workloads on AWS it is important to understand the services fault isolation boundary and how the boundary defines the resilience of your workload.

The global infrastructure outlined gives AWS the ability to provide fault isolation to its customers. The disruption of a zonal resource has no impact on resources in other Availability Zones. The disruption of a Regional service has no impact on services in other AWS Regions. For global services, mitigation techniques such as splitting the control plane and data plane mean that the services core functionality continues to operate when the control plane is disrupted, as they can operate independently of one another.

Definitions

1. Foundations: The scope of foundational requirements extends beyond a single workload or project. Before architecting any system, foundational requirements that influence reliability should be in place.
2. Workload architecture: A reliable workload starts with upfront design decisions for both software and infrastructure. Your architecture choices impact your workload behavior across all six Well-Architected pillars.
3. Change management: Changes to your workload or its environment must be anticipated and accommodated to achieve reliable operation of the workload. Changes include those imposed on your workload such as spikes in demand, as well as those from within, such as feature deployments and security patches.
4. Failure management: Failures are a given, and everything eventually fails over time. This is a given, whether you are using the highest-quality hardware or lowest cost components.

"Everything fails all the time. We needed to build systems that embrace failure as a natural occurrence." — Werner Vogels

5. Reliability: Reliability is the ability of a workload to perform its intended function correctly and consistently when it's expected to. This includes the ability to operate and test the workload through its total lifecycle.
6. Resilience: Resilience is the ability of a workload to recover from infrastructure or service disruptions, dynamically acquire computing resources to meet demand, and mitigate disruptions, such as misconfigurations or transient network issues.
7. Embedded Metric Format: EMF is a part of Amazon CloudWatch that helps you ingest complex, high-cardinality application data as logs and generate actionable metrics from them. By using this format to send logs from resources such as Lambda functions and containers, you can create custom metrics without having to instrument or maintain separate code, while gaining powerful analytical capabilities on your log data.

Note: Definitions 1–4 are the domain definitions for the Well-Architected Reliability Pillar.

Design for resilience

AWS offers capabilities that can be leveraged to provide different levels of resilience in the cloud based on your business requirements. When building a workload in the AWS Cloud, AWS is responsible for the resilience of the cloud. This means, we are responsible for the resilience of the services and infrastructure offered in the AWS Cloud. This infrastructure is composed of the hardware, software, networking, and facilities that run AWS Cloud services.

The implementation, configuration, and operation of your applications on AWS is your responsibility. The AWS Cloud services that you choose to consume, how you configure them, how you manage change and failure, and how you plan for disaster recovery are some of your key responsibilities that contribute to the resilience of your system. As a user of AWS, you are responsible for how you configure the services and resources you build into your systems. For example you can make the decision to deploy an Amazon RDS database with a synchronous replica, or as a standalone instance. You are also responsible for establishing monitoring for your system so you can understand when it is not meeting your customers' expectations or delivering business value.

This responsibility determines the amount of configuration work, testing mechanisms, recovery mechanisms, operational tooling, and observability logic that you can design into your workload to make it resilient.

Financial institutions should consider the following when building resilient workloads in the cloud:

- Software development lifecycle
- Resilience requirement planning
- Resilience architecture
- Observability
- Data backup and retention

Software development lifecycle

Best practice questions

- FSIREL01: Have you planned for events that impact your software development infrastructure and challenge your recovery and resolution plans?
- FSIREL02: Are you practicing continuous resilience to ensure that your services meet regulatory availability and recovery requirements?

FSIREL01: Have you planned for events that impact your software development infrastructure and challenge your recovery and resolution plans?

Financial services institutions are increasingly relying on continuous integration (CI) and deployment (CD) pipelines to accelerate development and deployment. Often the only way to change production systems is through the pipeline to ensure that quality controls, security guard rails, and standards are maintained as part of the change management process.

FSIREL01-BP01 Treat your CI/CD tools as critical workload components for recovery

If key elements of an SDLC environment, such as the CI/CD pipeline, are impacted, you might not be able to commit new code, change configurations, pull containers, or upload application artifacts, which can result in an outage of your workload. Understand the entire dependencies of your SDLC and plan for disruption of the critical components that the SDLC relies on. Consider replicating your SDLC environment and supporting services in another Region, which allows you to continually replicate source code, application, and container repositories.

FSIREL02: Are you practicing continuous resilience to ensure that your services meet regulatory availability and recovery requirements?

Your workload, and the environment in which it operates, is constantly changing. To keep pace, resiliency practices should not be considered a one-time effort. Make resilience a regular part of your feature delivery and operational cadence throughout a workload's lifetime.

FSIREL02-BP01 Practice regular resilience testing

Resilience is not a one-time effort. Resilience should be part of your day-to-day operations and practiced continuously. Perform chaos engineering experiments and scenario testing like AZ Availabilty Power Interruption or Cross-Region connectivity faults regularly to increase your team's understanding of how your workload behaves in adverse conditions such as excessive load, slow or failed network links, or a combination of adverse conditions. Continuous testing for resilience helps you to anticipate, observe, and respond to faults, as well as find blind spots that you didn’t know existed. By practicing continuous resilience testing and chaos engineering, your teams can improve observability and gain confidence in their ability to quickly detect and recover from incidents as recovery procedures are practiced and improved.

FSIREL02-BP02 Implement an operational readiness review process

To capture learnings from previous incidents and minimize reoccurrence across teams, implement an operational readiness review process within your organization. As part of your incident analysis process, identify key questions that, if asked prior to the incident, may have prevented the incident from occurring. Maintain a list of these key questions so that, as new features are released, your developers can refer back to the list and make sure that they don't repeat the same mistakes that have disrupted other workloads.

Resilience requirement planning

Best practice questions

- FSIREL03: How are your business and regulatory requirements driving the resilience of your workload?

FSIREL03: How are your business and regulatory requirements driving the resilience of your workload?

FSIREL03-BP01 Use business criticality to drive recovery objectives

Financial institutions scrutinize their most critical functions where a disruption to the function could cause harm to consumers, policy holders, participants, or industry integrity. This harm could mean that customers are unable to quickly recover (for example, when a firm is unable to put a client back into the correct financial position after a disruption or if they exceed the allowed disruption time). Resilience requirements should guide the development and operation of workloads that deliver or support these functions. Resilience requirements should be written to verify that the workload implementing the requirements is able to meet impact tolerances. In capturing resilience requirements, financial institutions must also consider any regulatory requirements concerning resilience.

The resilience of a workload should be defined by the business sponsoring the workload and is usually presented as RTO and RPOs plus a service-level objective (SLO). The criticality of a workload should therefore drive the investment for automated recovery of the workload. Example SLOs and mappings to resilience tiers are shown in Table 1 and 2.

Table 1 – Example resilience tiering for SLO

Availability SLO	Resilience tier	Acceptable downtime per year
99.99%	Platinum - Tier 1	52.60 minutes
99.90%	Gold - Tier 2	8.77 hours
98%	Silver - Tier 3	7.31 days

Table 2 – Example resilience tiering for RTO and RPO

Tier	Max RTO	Max RPO	Criteria	Cost
Platinum - Tier 1	15 minutes	30 seconds	Mission-critical workloads	$$$

Tier	Max RTO	Max RPO	Criteria	Cost
Gold - Tier 2	15 minutes – 8 hours	2 hours	Important, but not mission-c ritical workloads	$$
Silver - Tier 3	6 hours – a few days	24 hours	Noncritical workloads	$

FSIREL03-BP02 Apply fine grained workload resilience requirements

It's common to initially think of a workload's availability as a single target for the workload as a whole. However, upon closer inspection, we frequently find that certain functions of a workload have different availability requirements. For example, some systems might prioritize the ability to receive and store new data ahead of retrieving existing data. Other systems prioritize real-time operations over operations that change a system's configuration or environment. The Well-Architected reliability pillar outlines a few of the ways that you can decompose a single workload into constituent parts-per-function and evaluate the availability requirements for each. The benefit of decomposing is to focus efforts on availability according to the specific needs of and the value delivered by the individual function, rather than engineering the whole system to the strictest requirement.

Developing a system to the highest levels of availability can be expensive. Being able to address the resilience of individual workload functions can allow you to justify the investment based on the value of the function. With the functions measured by their criticality, you can also make informed trade-offs such as degrading the performance of less critical functions to maintain performance of the workload's most critical functions.

FSIREL03-BP03 Use past examples of market volatility in determining peak loads

In financial services workloads, even ones that do not directly provide services for traders such as settlement and clearing, market volatility creates peak demand requirements with a long-tail. The peak volume of an extreme event is much higher than one would expect to model a normal distribution, and thus typical p95 and p99 metrics are insufficient for estimating peak load. Determine if the workloads have dependencies on market volatility, and adjust load testing scenarios based on historical peaks, allowing you to determine how the workload performs in unexpected situations. It is common that financial services workloads are subject to dramatic increases in demand. The scaling response to the increase in demand must keep up with the

change in demand. For example, automatic scaling can take several minutes for a workload to be ready to receive traffic, and may exceed the ability to respond to customer requests in the expected timeframe, resulting in missed SLAs. For mission critical workloads, consider concepts like static stability and graceful degradation so that the workload continues to perform within acceptable limits, even under extreme load.

FSIREL03-BP04 Model failures to identify resilience requirements

Resilience requirements, like other system requirements, can be tested and should be documented in response to a business need. A resilience requirement must be met by the workload in order to achieve the RTO, RPO, and availability objective of the business function the workload supports. The resilience requirement does this by defining a control, which must be designed and implemented to mitigate the impact of a failure somewhere within the workload, with the workload's dependencies, or in the workload's environment. Use modeling techniques (for example, failure modes and effects analysis (FMEA)), combined with Operational Readiness Reviews (ORR), to anticipate the scenarios that could disrupt the workload's ability to meet its objectives. Create resilience requirements to mitigate any harm anticipated by the failure modeling analysis.

Resilience architecture

Best practice questions

- FSIREL04: Does the resilience and the architecture of your workload reflect the business requirements and resilience tier?
- FSIREL05: Is the resilience of the architecture addressing challenges for distributed workloads across AWS and an external entity?

FSIREL04: Does the resilience and the architecture of your workload reflect the business requirements and resilience tier?

Understanding how AWS services can impact your workload's availability is an important step in determining the resilience of your architecture.

FSIREL04-BP01 Use best practices to implement highly resilient critical workloads

Financial services institutions must be compliant with regulatory frameworks that define policies towards the resilience and operational excellence of their mission critical or core workloads.

Workloads designated by regulators and financial institutions as critical are therefore subject to greater scrutiny from regulators because financial services institutions must demonstrate that they can recover operations within reasonable recovery times and with little or no data loss.

To achieve these targets, you must mitigate scenarios that may disrupt your system by anticipating the scenarios, being able to monitor for their occurrence, and having pre-arranged responses in place. Adopting processes like ORRs, predictive monitoring with leading indicators, and consistent deployments are just some of the best practices that can be used to mitigate common scenarios. Additional workload design patterns for resilient systems can be found in the The Amazon Builders' Library.

FSIREL04-BP02 Provide external dependency accessibility from failover environments

FSI workloads often rely on many external service integrations with partner firms or online services from other departments in the same firm. While your workload may be able to resume service in a different failover environment, confirm that the system is able to operate with its dependencies from the failover environment. Make your dependencies accessible from the failover environment, and verify that the workload is able to function despite any changes in network attributes, such as latency.

Tightly coupled dependencies may need to be failed over in advance of your workload's failover. This slows down the recovery of your workload as it waits for its dependencies to become available. Coordinate your disaster recovery failover to expedite this process and bring down the recovery time to within acceptable ranges.

FSIREL04-BP03 Decouple your dependencies

Design your workload so that it is able to function despite impairment to dependencies, like external service integrations with partner firms, as well as services from other departments in the same firm. Decouple your workload from its dependencies so that it has static stability and continues functioning, or at least fails gracefully, even when its dependencies are impaired. Workload code should be reviewed and tested with the consideration that any API call to an external dependency may time out with no response, or return an unexpected error. Use chaos engineering to perform experiments where the workload's functionality is observed during simulation dependency disruption.

FSIREL05: Is the resilience of the architecture addressing challenges for distributed workloads across AWS and an external entity?

FSIREL05-BP01 Evaluate the resilience of cross-cloud application architectures

Understand the characteristics of your application components and how each component that is consumed across clouds may impact your system as a whole. Use failure mode and effects analysis (FMEA) to consider the severity and plausibility of possible failure modes, including application-level failures and service provider failures based on the provider's service event history. Consider if the added complexity of deployment across different types of environments adds to or reduces overall resilience.

FSIREL05-BP02 Address hybrid resiliency

Use AWS Direct Connect to provide a consistent network experience rather than internet-based connections. Achieve highly resilient network connections between Amazon Virtual Private Cloud (Amazon VPC) and your on-premises infrastructure by using multiple redundant Direct Connect connections. Use AWS Direct Connect Resiliency Toolkit to help you choose the right resiliency model. The AWS Direct Connect Failover Testing feature allows you to test the resiliency of your AWS Direct Connect connection by disabling the Border Gateway Protocol session between your on-premises networks and AWS.

Observability

Best practice questions

- FSIREL06: To mitigate operational risks, can your workload owners detect, locate, and recover from gray failures?
- FSIREL07: How do you monitor your resilience objectives to achieve your strategic objectives and business plan?
- FSIREL08: How do you monitor your resources to understand your workloads health?

FSIREL06: To mitigate operational risks, can your workload owners detect, locate, and recover from gray failures?

Failures, such as loss of network connectivity, is often considered in a binary nature where the connectivity is functioning normally or not functioning at all. However there are non-binary

failures called *gray failures*, which are defined by the characteristic of differential observability, meaning that different entities observe the failure differently. Gray failures can be subtle and difficult to detect. An example of a gray failure with network connectivity is a 40% packet loss of all TCP packets over a network link. Another example is intermittent failure on one or more servers behind a load balancer where some requests fail, but not enough to initiate the load balancer's health check. Overall service health metrics may be based on aggregate metrics, such as average response time from the load balancer, which may obscure localized failures.

FSIREL06-BP01 Monitor indicators aside from system metrics that can signal client impairment

Capture data that measures the experience of your workload's clients to understand when anomalies are affecting the customer experience with a workload function. Such measures are often collected as percentiles to prevent outliers when trying to understand the impact over time and how it's spread across your workload's clients. Examples of such metrics may be the 99th percentile of latency from the load balancer, a deviation in the number of requests being received over time, or the number of unsuccessful responses returned to the client. Highly visible workload owners should also have a means to monitor sudden increases in inbound customer support requests, and complaints on social media channels. Have a way for users to send feedback directly from within the service, or adjacent channels that can be monitored by service owners in near real-time.

FSIREL06-BP02 Have a way to find outliers hiding in aggregate metrics

Wherever system dashboards and monitors are reporting on aggregate results across a fleet of resources, be sure that system operators can also break out metrics and find outliers. Use tools like Amazon CloudWatch Contributor Insights and CloudWatch RUM to be able to ask questions like: "Who are the top 10 clients with high error rates?" And: "Do those top 10 clients share a common root cause?"

FSIREL06-BP03 Use anomaly detection to detect unusual changes in user engagement metrics

FSI workload owners should monitor for anomalies in metric data such as the number of user requests that receive a timely and successful response, and user session dropout rates (the number of users that began a multi-step process, such as a payment flow, but didn't finish). With Amazon CloudWatch you can enable anomaly detection on various metrics, which continually analyzes the

metrics, determines normal baselines, and surfaces anomalies that can in turn be used to initiate a CloudWatch alarm.

FSIREL06-BP04 Have a way to manually route away during failure

There may be a need to fail away from a primary system to its secondary, either because a system that depends on your workload needs to failover, or due to an unexpected, undetected impact to your primary system. In such cases you may need to manually override the status of health checks and route traffic away from the sources of a gray failure. You can use services such as Route 53 Application Recovery Controller and its feature zonal shift with routing controls. Also consider having a way to manually control and override the responses from each health check target, providing you with full control when a workload is considered unhealthy and initiated to route around the faulty resources.

FSIREL06-BP05 Establish baselines for expected network traffic

To understand conditions of high or unexpected network traffic, you must establish a steady state of metrics for the expected data flows between your workload and its users as well as between the components within your workload. This baseline should initiate an operational response when a workload is suddenly seeing abnormal traffic throughput that exceeds the expected steady state ranges. Understanding the steady state is key in creating the knowledge of normal communication patterns between and within the workload components. Knowing which network communications patterns are outside of normal ranges helps operations teams troubleshoot and isolate impacted components.

FSIREL07: How do you monitor your resilience objectives to achieve your strategic objectives and business plan?

FSIREL07-BP01 Monitor and validate your RPO

RPO is the maximum amount of data loss allowed as the result of a system failure expressed in units of time. Online Transaction Processing (OLTP) systems within financial services institutions typically leverage continuous data replication to a failover environment, where the RPO is a function of the latency of the data replication. AWS database services such as Amazon RDS and Amazon DynamoDB offer continuous data replication and also provide replication latency metrics that can be continuously monitored. RPO can be further verified by continuously adding synthetic records into the transaction stream and validating that each synthetic record was received, processed, and replicated within the RPO target limit. Furthermore CloudWatch alarms should be configured to alert whenever replication delays are routinely exceeding the system RPO limits.

FSIREL07-BP02 Monitor and validate your RTO

RTO is often defined as the maximum amount of time allowed for a system to resume its normal operations after a failure. RTO is measured and validated by testing system recovery processes and directly measuring the time it takes to recover. To be able to provide audit evidence for proof of DR and recovery exercises, you have to understand your workload's dependency chains to prove that if any of its dependencies fail, your service can stay within the boundary of the defined RTO.

FSIREL08: How do you monitor your resources to understand your workloads health?

High availability for applications requires the ability to detect failures and recover quickly. Workloads must be configured to emit the relevant telemetry to detect failures, so that operational processes can capture and react to these events.

FSIREL08-BP01 Use a single pane of glass for monitoring

Amazon CloudWatch provides robust monitoring, allowing you to organize the data to escalate detected issues as quickly as possible. Without adequate processes in place, you may miss leading indicators of problems. A single pane of glass and standardizing cloud monitoring standards across your organization can help avoid information silos and simplify the analysis of monitoring data. Combining monitoring of AWS system metrics and workload logs enables analysts to cross-reference signals and log information across dependent systems. Frequently, issues surface in invoking systems, and IT professionals spend time parsing logs on the invoking systems instead of on the dependent systems where the error originated. Consider embedding metrics in logs with Embedded Metric Format (EMF), which allows you to quickly dive from the single pane of glass to the most granular entity of your workload. More information on building efficient dashboards for operational visibility can be found in the The Amazon Builders' Library.

FSIREL08-BP02 Alert on the absence of an event

The absence of monitoring data can indicate an underlying issue. Implement controls that alert on missed reporting intervals. Treat missing data as a security breach, and raise alarms appropriately.

FSIREL08-BP03 Identify metrics and validate alerts through load testing

Workloads must be load-tested regularly to validate scaling and resilience. Identify key metrics (for both components that auto scale with demand and for static resources such as relational databases) that correlate with capacity constraints and customer outages during these load tests.

As part of your load-testing, validate these metrics and associated alerts, ensuring that alerts are issued as expected. Perform load tests in lower environments to identify indicators for alerting and automated remediation. Validation of your indicators and alerts through load testing minimize your Mean Time to Detection (MTTD), giving your recovery mechanisms more time to respond and increasing the workload's availability.

FSIREL08-BP04 Use distributed tracing tools for service-oriented architectures

As systems become more distributed with the implementation of microservices architectures, the challenge of identifying performance bottlenecks increase. Use workload performance monitoring tools such as AWS X-Ray to trace and provide telemetry across multiple systems and on a transaction-by-transaction basis. Adopt tools like AWS X-Ray and Open Telemetry as integrated tools that provide tracing and data as transactions span across multiple services.

Backup and retention

Best practice questions

- FSIREL09: How are you backing up data in the cloud?
- FSIREL10: How are backups retained?

FSIREL09: How are you backing up data in the cloud?

Not all backups are created equal, and not all have equal value. Ensure that the data you're backing up, and the way in which it is stored, is commensurate with the value of the data backup.

FSIREL09-BP01 Implement a backup strategy

A comprehensive backup strategy is an essential part of an organization's data protection plan to withstand, recover from, and reduce any impact that might be sustained due to a security event. You should create an extensive backup strategy that defines which data must be backed up, how often data must be backed up, and monitoring of backup and recovery tasks. It is equally important to highlight which data should not be backed up; your backup strategy should balance the cost of implementing a backup strategy and the cost of backup retention with the value of the backups. If data is non-essential or could be reconstructed from other sources, make it clear to teams that not everything has to be backed up.

FSIREL09-BP02 Maintain backups in a secondary Region

When you develop a comprehensive strategy for backing up and restoring data, consider backing up your data into another AWS Region allowing you to recover quickly in the case of a disaster recovery scenario. For those applications with criticality, requiring them to operate in multiple Regions makes sure that you replicate your backups from the primary to the secondary Region. Copying backups between Regions can be done using custom tooling or the original features of various AWS services such as Amazon RDS. Alternatively, management of backups between Regions, including the management of encryption keys for cross-Region replication, can be automated and performed using AWS Backup.

FSIREL10: How are backups retained?

FSIREL10-BP01 Understand requirements for data backup and retention

An important task of determining the resilience requirements of a workload is to identify data backup and retention needs. Financial institutions may have standards for backup and retention of data in their systems, which may be informed by regulatory requirements. Financial services customers must understand the requirements that apply to the workloads that are running in their environments.

FSIREL10-BP02 Back up logs as part of the backup strategy

In addition to the backup of workload data and databases, the system logs may also fall under regulatory requirements. Include the AWS CloudTrail, CloudWatch Logs, workload, and system logs in the log backup plan. In AWS, customers use Amazon S3, Amazon S3 Glacier, Amazon EBS snapshots, and Amazon RDS snapshots for backups of AWS services, and AWS Storage Gateway for on-premises backup to AWS. The AWS Backup service centralizes the management of the backups across the AWS environment by creating tag-based policies to manage the backups.

FSIREL10-BP03 Incorporate anti-ransomware backups into your backup strategy

In addition to the normal backup cycle, short-lived anti-ransomware backups need to be inserted into the backup cycle. Define a frequency and retention time on how long these ransomware backups should be held that aligns with your corporate security strategy. While a Regional copy of the data is sufficient for most cases, you can consider replicating backups with AWS Backup into another Region and AWS account. For a more detailed discussion around preventing ransomware, see Protecting resources.

FSIREL10-BP04 Create lifecycle policies for backups

Based on regulatory requirements, create lifecycle policies to retain and purge data in AWS. You can use a lifecycle policy in Amazon S3 to allow for the automation of migration of data to the most appropriate storage tier. AWS Backup allows for the management of retention of data across the environment through tag-based policies. AWS Backup also provides you with a Vault Lock mechanism to help prevent changes to backup lifecycles, as well as help prevent manual deletion of backups, helping you to align with your compliance requirements.

FSIREL10-BP05 Use Glacier Vault Lock and S3 Object Lock for WORM storage

Financial institutions often need to retain records for many years in write-once indelible storage. FSIs can use Glacier Vault Lock and S3 Object Lock mode to store data using a write-once-read-many (WORM) model. Amazon S3 Object Lock has been assessed by Cohasset Associates for use in environments that are subject to SEC 17a-4, CFTC, and FINRA regulations. The Amazon S3 Object Lock mode applied to an object stops users from modifying that object. To track which objects have S3 Object Lock, you can refer to an Amazon S3 inventory report that includes the status of objects. Amazon S3 Object Lock helps you adhere to regulatory requirements that require WORM storage, or add another layer of protection against object changes and deletion. For more information about how Amazon S3 Object Lock relates to these regulations, see the Cohasset Associates Compliance Assessment for Amazon S3 whitepaper. AWS also has partners that specialize in legal hold search and archive solutions that are compatible with AWS, and often built on top of AWS WORM features. Refer to the AWS Partners website for information.

Key AWS services

- **Resilient architecture**
  - **Amazon S3:** Leverage Amazon S3 object storage and replication to provide durability and resilience of your data on AWS. It is available Regionally (resilient against events that impact an entire Availability Zone) and also supports cross-Regional replication for geographic isolation.
  - **Amazon EC2 Auto Scaling:** Maintain workload availability and automatically add or remove Amazon EC2 instances according to conditions you define. You can also use the dynamic and predictive scaling features of Amazon EC2 Auto Scaling to respond to changing demand as well as schedule the right number of Amazon EC2 instances based on predicted demand to scale faster.

- **Amazon Route 53:** Use the 100% availability of Route 53's data plane to direct traffic based on latency, proximity, and workload health checks to enable a variety of low-latency, fault-tolerant architectures.
- **AWS Direct Connect:** Connect your data centers to AWS over dedicated, private, and consistent connections using Direct Connect.
- **Amazon Virtual Private Cloud (VPC)**: Provision a logically isolated section of AWS where you can launch AWS resources.
- **Amazon CloudFront:** You can cache your content in CloudFront's edge locations worldwide and reduce the workload on your origin by only fetching content from your origin when needed. You can use CloudFront's native origin failover capability to automatically serve your content from a backup origin when your primary origin is unavailable.
- **Amazon RDS Multi-AZ or Amazon Aurora:** Use Amazon RDS or Aurora Multi-AZ deployments to provide enhanced availability for production database workloads. Amazon RDS synchronously replicates data from a primary instance to a secondary in a different AZ which runs on a fault-isolated and independent infrastructure. In case of infrastructure failure, Amazon RDS automatically fails over to the standby so that you can resume database operations. These database services can also be configured to asynchronously replicate your data to additional AWS Regions to support multi-Region architectures.
- **Amazon DynamoDB:** Amazon DynamoDB is a fully managed NoSQL database service that provides fast and predictable performance with seamless scalability. DynamoDB automatically spreads the data and traffic for your tables over a sufficient number of servers to handle your throughput and storage requirements and your data that is stored is automatically replicated across multiple Availability Zones in an AWS Region. DynamoDB also supports Global Tables to give you the ability to store your data across multiple AWS Regions.
- **AWS Shield and AWS Shield Advanced:** AWS Shield is a managed service that provides protection against distributed denial of service (DDoS) exploits for workloads running on AWS. AWS Shield Advanced provides additional protections against more sophisticated and larger exploits for your workloads running on Amazon EC2, Elastic Load Balancing (ELB), Amazon CloudFront, AWS Global Accelerator, and Route 53.
- **AWS Lambda:** AWS Lambda lets you run code without provisioning or managing servers. AWS Lambda is designed to use replication and redundancy to provide high availability for both the service itself and for the Lambda functions it operates. There are no maintenance windows or scheduled downtimes for either.

- **Monitoring**

- **CloudWatch:** Amazon CloudWatch is the principal monitoring service for AWS Cloud resources and the workloads that you run on AWS.
- **Amazon VPC Flow Logs: Amazon** VPC Flow Logs is a feature that enables you to capture information about the IP traffic going to and from network interfaces in your VPC. Amazon VPC Flow Logs can be monitored through CloudWatch.

- **Backup and retention**
  - **Amazon S3 Glacier**: Amazon S3 Glacier, is an extremely low-cost storage service optimized for infrequently used data, or cold data.
  - **Amazon EBS snapshots,** and **Amazon RDS snapshots**: Snapshots for both Amazon RDS and Amazon EBS allow point-in-time recovery of the data stored in them. They can be configured to run automatically or at a scheduled time.
  - **AWS Backup:** AWS Backup is a centralized backup service that simplifies and provides a cost-effective way for you to back up your workload data across AWS services in the AWS Cloud and on-premises. Storage volumes, databases, and file systems are backed up to a central place where you can configure and audit the AWS resources you are backing up, automate backup scheduling, set retention policies, and monitor recent backup and restore activity.

Resources

Refer to the following resources to learn more about our best practices related to reliability.

Documents and blogs

- Banking Trends 2022: Cyber Vault and Ransomware
- Implement an SQL Server HA/DR Solution on AWS Outposts
- Disaster Recovery Compliance in the Cloud, part 1: Common Misconceptions
- Disaster Recovery Compliance in the Cloud, part 2: A Structured Approach
- Chaos Engineering in the Cloud
- Amazon Builders Library
- Understand Resiliency Patterns and Trade-offs to Architect Efficiently in the Cloud
- Rapidly recover from an application failure in a single az

Whitepapers

- AWS Fault Isolation Boundaries
- Availability and Beyond

Partner solutions

- Cohasset Associates Compliance Assessment for Amazon S3 whitepaper.

Videos

- Building Confidence through chaos engineering on AWS

Performance efficiency

The performance efficiency pillar focuses on the efficient use of resources to meet requirements, and to maintain and improve that efficiency as demands change and technologies evolve.

Key topics include selecting the right infrastructure based on workload requirements, monitoring performance, and making informed decisions to maintain efficiency.

Performance optimization should be a continuous, data-driven process of confirming business requirements, monitoring and measuring workload performance, identifying under-performing components and adjusting the infrastructure to meet evolving requirements. By reviewing your choices on a cyclical basis, you can take advantage of the continually evolving AWS Cloud.

Design principles

In addition to the design principles in the AWS Well-Architected Framework whitepaper, the following design principles can help you achieve performance efficiency for your financial services workloads.

Consider both internal and external requirements

Regulators expect financial services institutions to define operational performance objectives for workloads, and implement policies that achieve those objectives. Regulators may also impose their own Key Performance Indicator (KPI) requirements on systemically-important workloads, such as Open Banking interfaces, or trading transaction reporting and expect institutions to monitor and report on their compliance with these requirements, with penalties for breaches. The objectives must define both qualitative and quantitative measures of operational performance and thereby explicitly state the performance standards that the workload intends to meet.

Architect for performance-driven workloads

Some financial services workloads, for example high-frequency trading systems and risk calculation engines, are particularly performance sensitive, with factors such as speed of completion and latency of response directly impacting the profitability of the system. Systems with considerations like these need to prioritize performance over other factors such as cost-efficiency or reliability, considering the trade-offs required to achieve their performance goals while also preserving non-functional requirements such as transactional consistency and recoverability. See the the section called “Trade-offs” section of this pillar for more detail.

Use managed services Leverage AWS cloud services to allow teams to use a wide range of technologies, to experiment with options and achieve their performance goals, while maintaining overall control. You can reduce the time it takes to configure, and invest in operations and on-going management, reducing operational overhead and using the right tool for the job.

Definitions

Focus on the following areas to achieve performance efficiency in the cloud:

- Selection
- Review
- Monitoring
- Trade-offs

Gather data on all aspects of the architecture for a data-driven approach to building a high-performance architecture, from the high-level design to the selection and configuration of infrastructure services and components from compute to storage and networking.

Reviewing these architectural choices on a regular basis helps you take advantage of continually evolving AWS cloud capabilities and match your workload requirements with available services and features.

Monitoring the performance of your workload continuously makes you aware of deviances from expected performance, and able to take timely action. It is also important to plan for the future performance of the system by performing load tests of projected future loads, and running game days of exceptional circumstances in order to understand the behavior and limits of the system. Performance can degrade unexpectedly as workloads grow.

However, be aware of some constraints AWS places on testing of this type, as running load tests on Amazon Web Services can initiate security mechanisms. For more information, see the Amazon Elastic Compute Cloud testing policy. In particular, Penetration testing can be run only on permitted AWS services and Distributed Denial of Service (DDoS) testing must be performed by a pre-approved AWS Partner.

Finally, make trade-offs in your architecture to improve performance, such as using compression to reduce the size of data stored and transiting your network, caching frequently used data in dedicated services or relaxing consistency requirements, prioritizing your most important requirements.

Selection

Best practice questions

- FSIPERF01: How do you select the best performing architecture?
- FSIPERF02: How do you select your compute architecture?
- FSIPERF03: How do you select your storage architecture?
- FSIPERF04: How do you select your network architecture?

FSIPERF01: How do you select the best performing architecture?

Performance objectives for workloads can vary depending on the criticality of the workload. While more stringent performance requirements are expected for critical systems such as core banking, payments processing, trade performance, and market data feeds, all cloud workloads benefit from defining performance requirements.

FSIPERF01-BP01 Use internal and external risk to determine performance requirements

External regulatory, as well as internal risk requirements, are often a good place to start for performance requirements. For some systems, regulators release sector-wide guidance including potential stress tests. For others, regulators require that financial institutions have the capability to deliver on the operational resilience and the performance targets they have set for themselves.

FSIPERF01-BP02 Factor in rate of increase in load and scale-out intervals

Identify the upper bounds of the peak load against a system, as well as the amount of time needed to reach peak load. Load tests often overlook the rate of increase in traffic and create tests that scale up too quickly or too slowly. If the load test ramps up too quickly, the system may not be able to add capacity rapidly enough to meet the demand, which degrades performance and introduces errors. Load tests need to be run periodically and with every major release of the system.

FSIPERF01-BP03 Benchmark your solution

Benchmark your existing solution and its components in order to understand their performance characteristics and capacity to exceed their current profiles. AWS services like AWS Lambda and CloudWatch can be useful tools for building, running and monitoring a load testing environment

due to their low overhead for setup and extensive scaling capabilities. For more information, see AWS Prescriptive Guidance for load testing and Distributed Performance Testing.

FSIPERF02: How do you select your compute architecture?

FSIPERF02-BP01 Select your compute architecture based on workload requirements

The optimal compute solution for a particular architecture depends on the workload deployment method, degree of automation, usage patterns, and configuration. Third-party solutions can bring their own requirements for infrastructure, which must also be considered. Different compute solutions may be chosen for each step of a process. Selecting the wrong compute solutions for an architecture can lead to lower performance efficiency.

Some financial services computing workloads, like risk modeling, are typically loosely coupled and can benefit from event-driven architectures leveraging the scaling capacity of AWS serverless compute options like AWS Lambda and AWS Fargate, combined with messaging services including Amazon SQS and Amazon EventBridge to decouple components. These serverless solutions minimize the overhead of capacity management, automatically scaling in or out to meet demands.

Containerized infrastructure can enable financial services institutions to achieve their goals for speed and scalability by providing a standardized environment to leverage across multiple solutions, and supporting the development of microservice-based architectures. Where scale is the primary factor, AWS serverless container compute engine, AWS Fargate, can be used with both Amazon Elastic Container Service (Amazon ECS) and Amazon Elastic Kubernetes Service (Amazon EKS), removing the overhead of managing and provisioning compute resources.

For solutions with more specific performance requirements, or needing to run on virtual machines as their compute solution, AWS offers a wide range of Amazon Elastic Compute Cloud (Amazon EC2) instance types, which you can use to select the configuration that is best suited to your needs at any given time. This allows you to both take advantage of the latest CPU technologies as they are released without consideration for prior investment, and choose instance types with features that best suit your workload's requirements, for example instance variants optimized for network, storage, or compute performance.

The Financial Services Grid Computing on AWS whitepaper explores this topic in more detail for specific workloads.

FSIPERF03: How do you select your storage architecture?

AWS offers a wide range of storage options and as with compute, the best performance is obtained when targeting the specific storage needs of an application.

FSIPERF03-BP01 Select your storage architecture based on workload requirements

When you select a storage solution, verify that it aligns with your access patterns to achieve the desired performance. It is simple to experiment with different storage types and configurations without having to make commitments.

Financial services grid compute workloads can take advantage of Amazon FSx for Lustre, which provides a fully managed file system that's optimized for the performance and costs of workloads requiring file system access across thousands of Amazon EC2 instances, optionally backed by an S3 bucket, which makes it simple for clients to persist input and results of the calculations.

Consider whether your solutions can make use of caching services to improve performance, by storing frequently used data in memory, or bringing data closer to consumers. Many AWS services offer features for caching or dedicated services including Amazon ElastiCache, and Amazon File Cache.

Financial services solutions have historically made use of databases as a key component, often to verify transactional integrity, and here AWS also offers a wide range of database options. Select database options that align with your performance requirements, using different database technologies for different purposes, such as Amazon Timestream time-series database for storing ticking market data, rather than a one-size-fits-all use of traditional relational databases.

FSIPERF03-BP02 Consider changing needs over the entire lifecycle of your data

Financial services workloads often have requirements to keep data available for many years to help meet regulatory requirements, leading to significant amounts of data being retained. Amazon S3 and Amazon S3 Glacier storage classes provide the optimal solution for many data retention requirements with their almost unlimited capacity and predictable performance. Consider the use of the services' own lifecycle policies (supported by Amazon Elastic File System and Amazon S3 among others) to help meet your requirements. These services offer integrated lifecycle-based policies for moving data between tiers of storage based on access patterns and user-defined requirements. If the features of a single service do not meet your requirements, combine multiple storage services to satisfy requirements, rather than selecting a single storage service to help

meet your requirements, for example persisting Amazon FSx for Lustre file systems to Amazon S3 for long-term, low-cost retention. Note that costs for the service remain low, provided that the services are restricted to a single AWS Region.

FSIPERF04: How do you select your network architecture?

Use performance requirements to drive the selection of network components and architecture.

FSIPERF04-BP01 Use AWS services to optimize your network routes

Proximity to data sources, both internal and external, and the distance between components can be a key factor for financial services workloads, like high-frequency automated trading systems, so make use of AWS services to sit your solution as close as possible to dependencies. Where this location is outside of an AWS Region, make use of AWS edge location solutions such as AWS Outposts and AWS Local Zones to deploy workloads in the most suitable location, making the trade-off that not all AWS services may be compatible with these. For example Low Latency Trading has strict latency service level agreements (SLAs), where a millisecond can make the difference between completing a transaction or missing an opportunity, and due to these low latency requirements, brokers' low latency trading systems must be in close proximity to the exchanges.

Use AWS Direct Connect to provide the shortest and most reliable path to AWS resources for components hosted outside of AWS. Use Amazon CloudFront to cache static content closer to use cases, and AWS Global Accelerator to route connections to the closest possible source, leveraging the AWS backbone network and bringing your solutions closer to markets, users, and data. When using multiple AWS Regions, use Route 53 latency-based routing to serve requests from the AWS Region with the lowest latency.

FSIPERF04-BP02 Use Amazon EC2 instances and features to optimize your networking

Consider network performance when selecting Amazon EC2 instances, with specific network optimized variants indicated by the n-suffix, and bare metal instances offering direct access to the underlying host, further optimizing the networking stack.

Within an Amazon VPC, when inter-process communication latency, throughput, and consistency is a consideration, use Amazon EC2 Placement Groups to have greater control over the location of your virtual instances and optimize network communication, resulting in improved network performance reduction in latency and increased packet processing rates. The use of cluster

placement groups is covered in greater detail in the Crypto market-making latency and Amazon EC2 shared placement groups blog post on optimizing market-making systems.

Monitoring

Best practice questions

- FSIPERF05: How do you evaluate compliance with performance requirements?

FSIPERF05: How do you evaluate compliance with performance requirements?

Here are several methods for doing so:

- Monitoring of your workload at multiple levels helps verify that your resources are performing as expected and you are aware of deviations.
- Consider all dimensions of the solution for monitoring, for example client-side and server-side metrics, application metrics and infrastructure metrics, technical and functional metrics.
- Monitor for failure rates and alert when they are above expected values.
- Identify KPIs and create threshold alerts for them and determine what actions to take (like autoscaling) when thresholds are breached - this allows you to observe the overall health of your system and identify non-binary, or grey, failure states.
- Provide visibility of data loss in your metrics, for example by monitoring for lost messages.
- Where possible capture inter-solution and inter-process communication streams to aid with the reproduction of issues.

FSIPERF05-BP01 Use Application Performance Monitoring (APM) tools

Use APM tools to provide your organization the capability to verify that application performance meets its defined requirements. AWS offers features and services to monitor and subsequently right-size the cloud services that you need to meet performance requirements.

For example, you can monitor and set alarms on latency and error rates for each user request using Amazon CloudWatch metrics and alarms, or on your downstream dependencies, or on the success and failure of key operations. Amazon CloudWatch Synthetics can be used to create *canaries*, configurable scripts that run on a schedule, or to monitor your endpoints, and APIs.

The required level of monitoring generates huge amounts of data, which can be challenging for operation teams to store, analyze, and visualize, so make use of services including Amazon Managed Service for Prometheus to monitor and alert on containers, Amazon Managed Grafana to visualize metrics and logs, and the wide range of features found in Amazon CloudWatch, to provide the appropriate tools for monitoring your systems without the overhead of managing additional infrastructure. Teams need training to update their skills and processes and take full advantage of this new fidelity of insight.

FSIPERF05-BP02 Integrate performance testing into the release cycle

Rather than considering performance testing to be a separate part of the workload release cycle, integrate performance testing into your release process and CI/CD tooling. This allows you to record and evaluate performance metrics across releases, being aware of and taking action when metrics change as early as possible.

FSIPERF05-BP03 Verify consistency and failure recovery during load tests

You must verify data consistency and recovery during periods of high load. Ensuring that your workload's RTO and RPO is still valid under the highest load can uncover gaps in your architecture and operational resilience.

FSIPERF05-BP04 Understand performance of the system under peak load and in failure scenarios

Include testing of common failure scenarios in your performance testing suites to understand your workload behaviour in these situations and determine areas for improvement.

Extend the range of performance testing scenarios to cover testing at loads beyond current peak loads, and testing the scaling processes themselves of the application to understand how the environment behaves under increasing load.

Under common or anticipated failure scenarios, workloads should exhibit predictable failure patterns with performance degrading gracefully using techniques such as fail-open behavior, and the transformation of hard dependencies into soft dependencies.

FSIPERF05-BP05 Include dependencies in your load tests

Financial institutions need to map resources they need to continuously deliver their important business services. These resources are your people, processes, technology, facilities, and

information, including third-party service providers. This mapping allows the identification of operational dependencies, vulnerabilities, and threats. Incorporating the dependencies of your workload (such as on financial messaging providers) as part of your performance tests enables you to demonstrate the overall resiliency of your workload.

Trade-offs

Best practice questions

- FSIPERF06: How do you make trade-offs in your architecture?

FSIPERF06: How do you make trade-offs in your architecture?

Financial services workloads often have to make trade-offs in their architecture to meet their most important goals and KPIs, where performance of the system is deemed more important than other factors, or vice-versa.

FSIPERF06-BP01 Understand your priorities and architect to meet them

For example, a low-latency trading system needs to preserve the performance of the system above all other factors, and be prepared to compromise on the cost of infrastructure to meet their goals. In this situation it is still important not to compromise on availability, and this may require significant investment in parallel, independent, deployments for example an independent deployment of the application stack in multiple AWS Availability Zones or Regions rather than a failover architecture.

Within the workload it may be necessary to trade-off between persistent capacity and elasticity to make sure that the application always has the ability to handle peak workloads without needing timed or reactive scaling up. Consider how much of your peak workload you need to be able to service at any time.

When choosing services consider performance determinism. AWS serverless services like AWS Lambda and AWS Fargate can bring significant performance benefits due to their ability to scale elastically on demand, without intervention, but this is often coupled with less fine control over the underlying environment, for example CPU clock speed, and this can introduce an element of variability into workload performance. Where the workload performance must be as consistent as possible, consider using Amazon EC2, where you get the widest choice, and greatest level of control, over the production environment. For example, using Amazon EC2 directly enables the use

of ENA Express, to increase network throughput and reduce latency, but brings restrictions on the Amazon EC2 instances that support this feature.

Consider trade-offs in your application architecture. For example, to preserve network latency you may choose to use certain services and configurations that are more complex to implement and maintain, but offer better performance, such as using VPC Peering instead of AWS Transit Gateway to minimize the number of network hops for your most critical traffic. For optimal connectivity to on-premises workloads consider the best position for your AWS Direct Connect Gateway to bring it closest to the most sensitive workloads.

Key AWS services

Compute

- Amazon Elastic Compute Cloud (Amazon EC2) offers the broadest and deepest compute environment, with over 500 instance types and choices employing the latest processor, storage, networking, operating system, and purchase model to help you best match the needs of your workload.
- Amazon EC2 Spot Instances let you take advantage of unused Amazon EC2 capacity in the AWS cloud at a discount compared to On-Demand Instances prices. You can use Spot Instances for various stateless, fault-tolerant, or flexible applications such as big data, containerized workloads, CI/CD, web servers, high-performance computing (HPC), and test and development workloads.
- Amazon EC2 Auto Scaling lets you automatically add or remove Amazon EC2 instances using scaling policies that you define to service established or real-time demand patterns. The fleet management features of Amazon EC2 Auto Scaling help maintain the health and availability of your fleet.
- AWS Compute Optimizer can help you to avoid overprovisioning or under provisioning three types of AWS resources—Amazon EC2 instances, Amazon Elastic Block Store (Amazon EBS) volumes, and AWS Lambda functions—based on your utilization data.

Storage

- Amazon FSx for Lustre provides fully managed shared storage with the scalability and performance of the popular Lustre file system.

Networking

- AWS Global Accelerator is a networking service that helps you improve the availability, performance, and security of your public applications.
- Amazon CloudFront is a content delivery network (CDN) service built for high performance, security, and developer convenience.
- AWS Direct Connect provides the shortest path to your AWS resources. While in transit, your network traffic remains on the AWS global network and never touches the internet, reducing the chance of hitting bottlenecks or unexpected increases in latency.
- AWS Outposts is a family of fully managed solutions delivering AWS infrastructure and services to on-premises or edge locations for a truly consistent hybrid experience. Outposts supports workloads and devices requiring low latency access to on-premises systems, local data processing, data residency, and application migration with local system interdependencies.
- AWS Local Zones are a type of infrastructure deployment that places compute, storage, database, and other select AWS services close to large population and industry centers.

Monitoring

- Amazon CloudWatch collects and visualizes real-time logs, metrics, and event data in automated dashboards to streamline your infrastructure and application maintenance.
- Amazon Managed Service for Prometheus is a Prometheus-compatible service that monitors and provides alerts on containerized applications and infrastructure at scale, integrated with Amazon Elastic Kubernetes Service (Amazon EKS), Amazon Elastic Container Service (Amazon ECS), and AWS Distro for OpenTelemetry.
- Amazon Managed Grafana is a fully managed service for Grafana, a popular open-source analytics environment that lets you query, visualize, understand, and receive alerts about your metrics no matter where they are stored.
- AWS Distro for OpenTelemetry can collect metadata from your AWS resources and managed services to correlate application performance data with underlying infrastructure data, reducing the mean-time-to-problem resolution.
- AWS X-Ray provides a complete view of requests as they travel through your application and filters visual data across payloads, functions, traces, services, APIs, and more with no-code and low-code motions.
- Amazon DevOps Guru uses machine learning (ML) to detect abnormal operating patterns so you can identify operational issues before they impact your customers.

Operations

- AWS Trusted Advisor provides recommendations that help you follow AWS best practices, identifying ways to optimize your AWS infrastructure, improve security and performance, reduce costs, and monitor service quotas.

Resources

Refer to the following resources to learn more about our best practices related to performance efficiency of financial services industry solutions.

Documentation and blogs

- Rethinking the low latency trade value proposition using AWS Local Zones
- How to improve FRTB's Internal Model Approach implementation using Apache Spark and Amazon EMR
- How cloud increases flexibility of trading risk infrastructure for FRTB compliance
- Crypto market-making latency and Amazon EC2 shared placement groups
- CloudFront FSI Service Spotlight
- Automating and Scaling Chaos Engineering using AWS Fault Injection Service

Whitepapers

- Financial Services Grid Computing on AWS

Partner solutions

- AWS STAC-M3 benchmark results: Low-latency tick analytics made easy
- Scaling and Managing TIBCO DataSynapse GridServer on AWS

Reference architectures

- High Performance Computing on AWS
- Running SAS Grid on AWS

- General HPC Architecture on AWS

Videos

- NYSE: Protecting markets through real-time data processing
- Nasdaq: Moving mission-critical, low-latency workloads to AWS
- HSBC Uses Serverless to Process Millions of Transactions in Real Time
- FINRA Collects, Analyzes Billions of Brokerage Transaction Records Daily Using AWS
- How FINRA operates PB-scale analytics on data lakes with Amazon Athena
- How Morgan Stanley leveraged Amazon EC2 Spot to Scale on Demand
- Risk calculations using HPC and Spot Instances with Morgan Stanley
- DBS Bank: Scalable Serverless Compute Grid on AWS
- Temenos: Building Serverless Banking Software at Scale
- How AWS Helped a Financial Services Company Adopt a Serverless Architecture to Effectively Scale
- How a Financial Services Company Addressed a 4X Increase in Call Volume with Cloud

Cost optimization

The cost optimization pillar provides guidance to help customers apply best practices in the design, delivery, and maintenance of their AWS environments, with the most effective use of services and resources at a minimal cost. Cost optimization includes the continual process of refinement and improvement of a system over its entire lifecycle.

Many financial services institutions have low latency requirements for trading, high performance compute requirements, and constantly changing security and regulatory requirements. These, coupled with extrinsic economic drivers, create a demand for robust and dynamic cost optimization processes across all of the organization's workloads. Many financial services companies run at enterprise scale, and thus run hundreds of diverse workloads with large IT and Security staff.

Compared to small and mid-sized businesses (SMBs), that typically run an on-premises model where budgets are relatively fixed and organizations strive to utilize allocated budgets, larger enterprise workloads running on AWS allow customers the freedom to continuously evolve and optimize their resources, usage, and processes to stay efficient. That said, even SMBs can maximize their ROI by migrating their workloads to the AWS cloud and implementing cost optimization best practices.

From the initial design of your first proof-of-concept to the ongoing operation of extensive production workloads, adopting the cost optimization practices outlined in this whitepaper allows you to build and operate cost-aware systems that achieve your business outcomes while minimizing costs, thus allowing your business to maximize its return on your IT investments.

For most financial services customers, there is a need to understand cost both at an application-level and at a workload-level to improve and optimize your costs associated with the workload.

Design considerations

Similar to the other pillars of the Well-Architected Framework, there are several trade-offs to consider for cost optimization. For example, whether you plan to optimize for your speed-to-market, versus optimizing for cost. In some cases, it's best to optimize for speed — going to public quickly, shipping new features, or meeting a deadline — rather than investing in upfront cost optimization. Because you can optimize over time, you can gradually migrate your workload to achieve higher cost optimization once your initial products are launched.

Due to the agility of many enterprise customers' operations, investing in reserved instances (RIs) and AWS Compute Savings Plans (SPs) and usage of spot instances can provide your team with direct ways to save on costs, even if you do not utilize them 100%. Nonetheless, capacity planning and usage forecasting is important for managing your commitment plans.

Design decisions are sometimes directed by haste rather than data, and the temptation always exists to overcompensate for worst case scenarios, rather than spend your time benchmarking for a most cost-optimal deployment. Overcompensation can lead to over-provisioned and under-optimized deployments. However, you may find this to be a reasonable approach if you begin your cloud migration by lifting and shifting resources from your on-premises environment to the cloud. Once you have stabilized your cloud-based workloads post-migration, then you can develop a program to optimize them over time afterwards.

Investing the right amount of effort in a cost optimization strategy up front allows you to realize the economic benefits of the cloud more readily, by ensuring a consistent adherence to best practices and avoiding unnecessary over provisioning. The following sections provide techniques and best practices for the initial and ongoing implementation of Cloud Financial Management and cost optimization for your workloads

Adopting the practices in this pillar helps you build architectures that can:

- Move your usage and costs in line with demand.
- Use appropriate services and resource types to minimize costs.
- Analyze, attribute, and forecast costs.
- Reduce costs over time.

Cost optimization is a continual process of refinement and improvement of a system over its entire lifecycle. A cost-optimized system aims to fully utilize all resources, achieve an outcome at the lowest possible price point, and meet your workload's functional and business requirements.

Design principles

In addition to the design principles described in the cost optimization pillar of the AWS Well-Architected Framework whitepaper, the Financial Services Industry Lens identifies the following design principles to facilitate good design in the cloud for your financial services workloads.

These design principles can help you to build and operate cost-aware workloads that achieve business outcomes while minimizing costs and allowing your organization to maximize your return on investment:

- **Monitor cost and resource utilization:** Financial services workload usage can be cyclical and can have usage spikes during specific days like month-end or quarter-end, or it can be intra-day during specific hours. AWS provides customers with a number of usage monitoring services that can scale your operations up and down as demand conditions require. Monitor cost at an application-level, and a workload-level on a regular basis, and optimize usage of resources and cost.
- **Define recovery objectives per workload:** FSI customers have workloads with varying levels of recovery objectives (RTO and RPO). A cost-conscious design approach considers the recovery objectives per workload before suggesting an appropriate DR strategy (Backup and restore, pilot light, warm standby or active/active).
- **Operational efficiencies:** FSI customers may request customization of their workloads to achieve certain business outcomes. Using AWS analytics tools, you can track, attribute, and charge back your IT costs to each responsible FSI business unit and organizational group. This encourages accountability among the teams and leads to better departmental management of usage costs.
- **Data transfer cost**: In many scenarios, the customer workloads can be running in multiple Regions. Monitor data transfer and storage egress costs for the workload.
- **Adopt cloud financial management:** Due to the size of many financial services enterprise customers, the benefits of aligning your IT organization with a Cloud Financial Management approach helps you save on the costs of both infrastructure and operations. To enable this capability, invest in knowledge building programs, resources, and processes to help become a more cost-efficient organization.

Definitions

There are five focus areas for cost optimization in the cloud as described in the cost optimization pillar of AWS Well Architected Framework. These focus areas are applicable to all types of workloads including financial services. In this section, we have listed financial services-specific best practices.

- Practice Cloud Financial Management
- Expenditure and usage awareness
- Cost-effective resources

- Manage demand and supplying resources
- Optimize over time

Practice Cloud Financial Management (CFM)

Cloud Financial Management (CFM) allows finance, product, technology, and business organizations to manage, optimize, and plan costs as they grow their usage and scale on AWS. The primary goal of CFM is to allow customers to achieve their business outcomes in the most cost-efficient manner and accelerate economic and business value creation while finding the right balance between agility and control. AWS CFM offers a set of capabilities to manage, optimize, and plan for cloud costs while maintaining business agility. CFM is paramount not only to effectively manage costs, but also to verify that investments are driving expected business outcomes. The four pillars of the Cloud Financial Management Framework in the AWS Cloud are *see*, *save*, *plan*, and *run*. Each of these pillar areas has a set of activities and capabilities.

Following the best practices in CFM is essential for managing costs in your financial services workloads.

These Cloud Financial Management best practices help you establish cost transparency to control your resources and plan your spend to optimize your return on investments.

Best practice questions

- FSICOST01: Is your cloud team educated on relevant technical and commercial optimization mechanisms?
- FSICOST02: Do you apply the Pareto-principle (80/20 rule) to manage, optimize, and plan your cloud usage and spend?
- FSICOST03: Do you use automation to drive scale for Cloud Financial Management practices?

FSICOST01: Is your cloud team educated on relevant technical and commercial optimization mechanisms?

Due to the size of many financial services enterprise customers, the benefits of aligning your IT organization with a Cloud Financial Management approach helps you save on the costs of both infrastructure and operations. To enable this capability, invest in knowledge building programs, resources, and processes to help become a more cost-efficient organization.

FSICOST01-BP01 Evangelize cloud education among all (including non-technical) staff and stakeholders

A company-sponsored cloud training program exists, and is required for all cloud stakeholders regardless of their seniority or affiliated organization.

- Digital training: On-demand courses so your team can learn about the latest services when and where it's convenient.
- Classroom training: In-person and virtual training from instructors who teach your team in a hands-on learning environment. For new employees, it should be part of their onboarding training, and should be mandatory training on a yearly basis for all existing employees and contractors.
- AWS Certification: Validate technical skills and cloud expertise to grow your career and business.

FSICOST02: Do you apply the Pareto-principle (80/20 rule) to manage, optimize, and plan your cloud usage and spend?

Investing the right amount of effort in a cost optimization strategy up front allows you to realize the economic benefits of the cloud more readily, by ensuring a consistent adherence to best practices and avoiding unnecessary over provisioning. CFM is paramount not only to effectively manage costs, but also to verify that investments are driving expected business outcomes.

FSICOST02-BP01 Apply the Pareto-principle 80/20 rule for your CFM efforts

No matter your organization size, pay specific attention to your capacity investment while developing CFM-related concepts. Here are some examples of CFM activities to apply the 80/20 rule to create an optimal input/output solution.

- Cost allocation: Start with default allocation opportunities (per AWS account, AWS-generated `createdBy` tag), then follow up by tagging all AWS services that support tagging, check overall percentage of cost allocation. In case you reach 80% cost allocation, check if equal allocation of the unallocated 20% of costs is acceptable for your organization (for example, splitting AWS service cost equally between business units or teams). Before spending time and budget on a third-party solution (for example, telemetry) ensure that shared resources you aim to allocate are substantial (for example, over 20% of monthly bill).

- Cost optimization: Incorporate implementation of low-hanging cost optimization recommendations (from Cost Explorer or AWS Trusted Advisor) into daily activities of your teams. Centralized teams evaluate and book SP and RI quarterly, decentralized teams perform instance rightsizing and modernization weekly. CFM practitioners report it is more efficient to spend 30 minutes per week rather than one day per month. While implementing cost optimization that require technical changes, pay attention to long term benefits – one-time adjustments provide reoccurring savings. Evaluate time and capacity invested into technical adjustments versus cost saving for at least the next 24 months. These types of calculations help prioritize activities with the highest impact.

FSICOST03: Do you use automation to drive scale for Cloud Financial Management practices?

Automation can drastically reduce the cost of the CFM. You can provision resources using auto scaling or using managed services, set budgets to meet, and alerts to inform users on cost utilization.

FSICOST03-BP01 Use automation to drive scale for Cloud Financial Management practices

Automation can drastically reduce the cost of the CFM. You can provision resources using auto scaling or using managed services, set budgets to meet, and alerts to inform users on cost utilization.

Automating operations reduces the frequency of manual tasks, improves efficiency, and benefits enterprises by delivering a consistent and reliable experience when deploying, administering, or operating workloads. You can free up human resources from manual operational tasks and use them for higher value tasks and innovations, thereby improving business outcomes. Enterprises require a proven, tested way to manage their workloads in the cloud. That solution must be secure, fast, and cost effective, with minimum risk and maximum reliability.

Expenditure and usage awareness

Understanding your organization's costs and drivers is critical for managing your cost and usage effectively, and identifying cost-reduction opportunities. Organizations typically operate multiple workloads run by multiple teams. These teams can be in different organization units, each with

its own revenue stream. The capability to attribute resource costs to the workloads, individual organization, or product owners drives efficient usage behavior and helps reduce waste. Accurate cost and usage monitoring allows you to understand how profitable organization units and products are, and allows you to make more informed decisions about where to allocate resources within your organization. Awareness of usage at all levels in the organization is key to driving change, as change in usage drives changes in cost. Consider taking a multi-faceted approach to becoming aware of your usage and expenditures. Your team must gather data, analyze, and then report.

Best practice questions

- FSICOST04: How do you promote cost-awareness within your organization?
- FSICOST05: How do you track anomalies in your ongoing costs for AWS services?
- FSICOST06: How do you track your workload usage cycles?

FSICOST04: How do you promote cost-awareness within your organization?

Awareness of usage at all levels in the organization is key to driving change, as change in usage drives changes in cost. Consider taking a multi-faceted approach to becoming aware of your usage and expenditures. Your team must gather data, analyze, and then report.

FSICOST04-BP01 Promote a culture of transparency on costs

To promote transparency and accountability of costs, it is important to have standard mechanisms that show or charge back the costs to business units or applications. Companies use tags to allocate cost to teams, business units, or organizations within an enterprise and to observe trends. Enforce a tagging taxonomy with tag policies within pipelines that deploy infrastructure as code (IAC) and govern using SCPs at the organization-level and configuration across all AWS accounts. For more information on tags, see: Using AWS cost allocation tags.

In the large organization, some teams are very advanced in cost optimization and they are aware of cost impacts while other teams are not that mature. Hence, team cooperation, sharing importance of Cloud Finance Management, Cloud Center of Excellence is extremely important to promote a culture of cost optimization. For more information on tags, see this article: Using AWS cost allocation tags.

FSICOST05: How do you track anomalies in your ongoing costs for AWS services?

Understanding your organization's costs and drivers is critical for managing your cost and usage effectively, and identifying cost-reduction opportunities. Accurate cost and usage monitoring allows you to make more informed decisions about where to allocate resources within your organization.

FSICOST05-BP01 Be aware of anomalies and periodically review your architecture

Anomalies can drive up cost. Set up AWS Cost Anomaly Detection to detect and alert on anomalous spend patterns in your deployed AWS services. Cost Anomaly Detection automatically determines thresholds each day by adjusting for organic growth and seasonal trends (like usage increases from Sunday to Monday or increased spend at the beginning of the month) through machine learning models. Financial systems usually integrate with several other third-party systems, and Cost Anomaly Detection can integrate with these systems as well.

FSICOST06: How do you track your workload usage cycles?

Financial services workload usage can be cyclical and can have usage spikes during specific days like month-end or quarter-end, or it can be intra-day during specific hours. AWS provides customers with a number of usage monitoring services that can scale your operations up and down as demand conditions require. Monitor cost at an application-level, and a workload-level on a regular basis, and optimize usage of resources and cost.

FSICOST06-BP01 Monitor your workload usage cycle around times of higher and lower utilization (quarter-end, year-end, weekends, and holidays) to identify ways to reduce your costs

You may have workload usage cycles for week-end or month-end, and quarter-end have more usage of resources. In some cases, there could be higher usage due to events like the start of trading hours, holidays shopping, and so on. Monitoring usage and corresponding events are helpful to optimize cost and architecture. You can choose to shutdown unused instances, for example Amazon EC2 servers for development, or QA on Friday, and bring them back up on Monday.

Cost-effective resources

Using the appropriate services, resources, and configurations for your workloads is key to cost savings. Consider the following when creating cost-effective resources:

You may employ internal guardrails (built using AWS Organization Service Control Policies) to allow a limited set of services to be provisioned to contain costs. If a workload requires services outside of the allow list, they need to centralized, created, and shared with an individual account, or created by an administrator.

Best practice questions

- FSICOST07: Are you using all the available AWS credit and investment programs?
- FSICOST08: Are you monitoring usage of Savings Plans regularly?
- FSICOST09: Are you using the cost advantages of tiered storage?
- FSICOST10: Do you use lower cost Regions to run less data-intensive or time-sensitive workloads?
- FSICOST11: Do you use cost tradeoffs of various AWS pricing models in your workload design?
- FSICOST12: Are you saving costs by adopting a set of modern microservice architectures?
- FSICOST13: Do you use cloud services to accommodate consulting or testing of projects?
- FSICOST14: How do you measure the cost of licensing third-party applications and software?

FSICOST07: Are you using all the available AWS credit and investment programs?

Multiple credit options are available, such as migrations, digital innovation, cloud economics, and prototyping to activate credits.

FSICOST07-BP01 Use AWS credit programs such as Migration Acceleration Plan, Digital Innovation, and Activate to save costs and drive cloud adoption

Multiple credit options are available, such as: Migrations, Digital Innovation, Cloud Economics, Prototyping to Activate credits. Different departments work in silos, and often the credits earned by the workload in one department need to be publicized for consumption across the other units. Ensure that the workloads are leveraging these credits across the organization. Purchasing third-party products or even data from AWS Marketplace. Talk to your account team to get relevant

information on a regular basis for available credit programs. For example, **Activate** provides up to $100K in credits for startups.

FSICOST08: Are you monitoring usage of Savings Plans regularly?

Capacity planning and usage forecasting is important for managing your commitment plans. Gain better control of the flexibility of Savings Plan usage and manage costs with regular monitoring on a regular cadence over quarterly basis, or reviews at regular time intervals.

FSICOST08-BP01 Sign up for a compute savings plan for discounts on compute versus on-demand pricing

Financial systems usually have a predicted usage pattern. Sign up for a compute savings plan, as they offer discounts on compute of up to 72% compared to on-demand pricing. The most flexible type of Savings Plan applies across the core compute services (Amazon EC2, AWS Fargate, and AWS Lambda) and across Amazon EC2 instance size, operating system, tenancy, Availability Zone, and Region. This flexibility accommodates continuously evolving workloads and avoids unused commitment. Instead of a single monolithic savings plan, opt for smaller concurrent active Savings Plans, which are additive to reduce commitment risk, increase discount coverage, and relieve the burden of long-range usage predictions. Gain better control of the flexibility of Savings Plan usage and manage costs with regular monitoring on a regular cadence over quarterly basis, or reviews at regular time intervals.

Understand how Savings Plans can also be shared across all accounts within an AWS Organization or consolidated billing family.

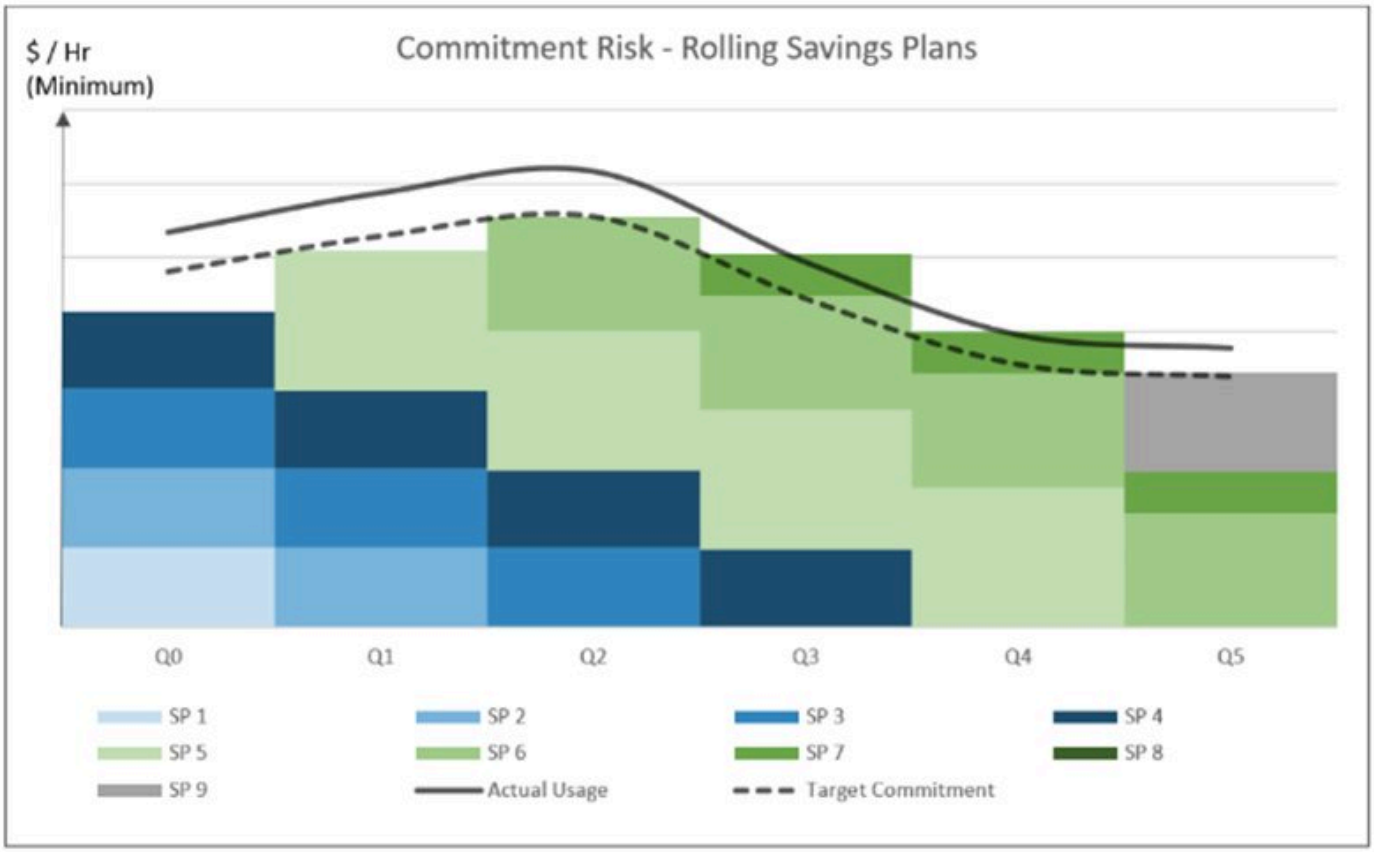

Figure 2: When Savings Plan 3 expires at the start of Q3, it is replaced with a much smaller Savings Plan 7, and when Savings Plan 4 expires at the start of Q4, no Savings Plan is purchased to replace it. As a result, over-commitment is reduced.

FSICOST09: Are you using the cost advantages of tiered storage?

FSI companies usually have long retention policies for their regulatory and audit requirements. They usually span multiple years and might even be able to take up to a day or two to be able to retrieve old data. Understand and use the cost advantages of tiered storage.

FSICOST09-BP01 Define data retention policies to select the right storage type for your data lifecycle

FSI companies usually have long retention policies for their regulatory and audit requirements. They usually span multiple years and might even be able to take up to a day or two to be able to retrieve old data. Defining data retention policies and corresponding architecture to transfer data from main storage to archival storage is important. This can be achieved by transferring data from RDS database to S3, or creating snapshot and storing it for better cost efficiencies.

FSICOST10: Do you use lower cost Regions to run less data-intensive or time-sensitive workloads?

FSI companies usually have to plan their Disaster Recovery (DR) and also run a cadence of dry runs for regulatory purposes, and typically opt to setup their DR site in an alternate AWS Region. Depending on the SLA for latency, data sovereignty and compliance needs, you could run DR in a less costly Region.

FSICOST10-BP01 Use less costly Regions for disaster recovery and test platforms

FSI companies usually have to plan their Disaster Recovery (DR) and also run a cadence of dry runs for regulatory purposes, and typically opt to setup their DR site in an alternate AWS Region. Depending on the SLA for latency, data sovereignty, and compliance needs, you could run DR in a less costly Region. Consider cheaper Regions for non-production environments.

FSICOST11: Do you use cost tradeoffs of various AWS pricing models in your workload design?

Cloud cost is an important part of the design and architecture process and is used in making trade-offs between quality, performance, security and other non-functional requirements. Cloud cost is considered when selecting AWS services (using building block services such as Amazon EC2 versus using managed services such as Amazon ECS).

FSICOST11-BP01 Identify pricing models and savings plans for your selected AWS services when designing your architecture

Cloud cost is an important part of the design and architecture process and is used in making trade-offs between quality, performance, security and other non-functional requirements. Cloud cost is considered when selecting AWS services (using building block services such as Amazon Elastic Compute Cloud versus using managed services such as Amazon Elastic Container Service).

Cost factors that go into the selection of cloud resources based on the level of cost optimization provided by pricing models or AWS services include: Savings Plans, Reserved Instances, Amazon EC2 Spot Instances, or Amazon S3 Intelligent-Tiering. Cost trade-offs also include resource-level decisions based on performance (for example, selecting an XL instead of a 2XL resource size).

Product designs take the pricing structure of AWS services into account (for example, Elastic Load Balancing charges for elasticity and inter-Availability Zone data transfer charges). Design activities also include cost estimation for the services being built using the AWS Pricing Calculator, AWS

Price List API, or third-party pricing tools, or they might involve building and deploying proof of concepts to measure actual costs.

The cost of the new workload is measured on an ongoing basis during the workload's entire lifecycle, and unexpected cost variances are used to influence future product changes in the workload. Here are several examples:

- **Managed services:** AWS managed services helps reduce operational overhead to maintain servers, apply patches, and add high availability, security etc. Plan to use as many managed services as possible to reduce operational cost.
- **Serverless architecture:** FSI companies often have the need to set up automation for processing events and workflows for technology operations. If you use EC2 instances or databases, you are likely not using 100% of the compute capacity at all times. Many customers only use 10–20% of the available capacity in their EC2 fleet at any point in time. This average is also affected by High Availability and Disaster Recovery requirements, which typically result in idle servers waiting for traffic from failovers. In serverless models such as AWS Lambda or DynamoDB, you pay per-request and by duration of time. Additionally, serverless architectures can lower the overall Total Cost of Ownership (TCO) since many of the networking, security, and DevOps management tasks are included in the cost of the service.
- **Caching data:** Most of the fintech customers use the API heavily. So to optimize on time and money, implement caching mechanisms like caching at the edge or caching data in in-memory cache and so on. This depends on the type of the APIs and how APIs are designed. In the case of static data, you can cache at the edge for long-term, and for dynamic content you can cache in in-memory stores or for a short duration.
- **Right storage selection:** Select the right storage mechanism to optimize cost across metrics, such as storage, IOPS, and data throughput. You can use a combination of the Amazon S3 family of products or AWS database products such as: Amazon Redshift, Amazon RDS, Amazon FSx, Amazon EBS , or Amazon EFS. For more information about these services, see: Amazon Storage overview and AWS Database.
- **Choosing the right instances and usage of Spot Instances:** Choose the right instances, and choose Spot Instances if possible to optimize the cost. You can mix and match with Spot Instances and on-demand capacity. You can use a base amount of capacity with On-Demand Instances, and use Spot Instances for spikes in demand.
- **CPU architecture:** If your application is not dependent on a specific CPU architecture like ARM versus x86, you might consider Graviton-based instances. Many AWS services, including Amazon EC2, Amazon Aurora, Amazon ElastiCache, Amazon EMR, AWS Lambda, and AWS Fargate,

support AWS Graviton-based instances with significant price performance benefits. For more information, see Getting started with Graviton.

FSICOST12: Are you saving costs by adopting a set of modern microservice architectures?

Financial institutions are moving from monolithic legacy systems such as mainframes into modern microservices architectures, giving them the flexibility of provisioning multiple environments to develop features rapidly, instead of waiting for the single monolith environment to be available, giving them greater agility and faster time-to-market.

FSICOST12-BP01 Migrate your mainframe and on-premises infrastructure to adopt a cloud-based microservices approach

Financial institutions are moving from monolithic legacy systems such as mainframes into modern microservices architectures, giving them the flexibility of provisioning multiple environments to develop features rapidly, instead of waiting for the single monolith environment to be available, giving them greater agility and faster time-to-market. Quantifying this gain is important for stakeholder buy-in.

FSICOST13: Do you use cloud services to accommodate consulting or testing of projects?

Some financial services institutions hire contractors during specific months, or for a project. Procuring a new machine, and ensuring that it is meeting the compliance standards of a financial services institution can be resource intensive. Using a service like Amazon WorkSpaces for end-user computing can help with cost-efficient utilization of resources.

FSICOST13-BP01 Set up pay-as-you-go services when team expands for certain duration

Some financial services institutions hire contractors during specific months, or for a project. These contractors can work on a project for a short duration, like 6 months to a year. Procuring a new machine, and ensuring that it is meeting the compliance standards of a financial services institution can be resource intensive. Using a service like Amazon WorkSpaces for end-user computing can help with cost-efficient utilization of resources. You can create workspaces per your internal standards, and provision it for a new resource.

FSICOST14: How do you measure the cost of licensing third-party applications and software?

If you are using third-party software, understand the specific licensing terms of each third-party vendor.

FSICOST14-BP01 Consider the cost of licensing third-party applications and software

If you are using third-party software, understand the specific licensing terms of each third-party vendor. AWS offers both Dedicated Hosts that have pre-installed virtualization software (Hypervisor) whereas bare metal servers do not have pre-installed virtualization software. Choosing the right instance type specific to the licensing terms may reduce your third-party licensing costs.

Generally, third-party software applications and associated support can provide your workload with a lower overall cost of ownership than in-house created applications. Because software vendors have a much broader perspective of customer requirements, their software can more economically support a wider range of use cases than an in-house developed solution. A software support agreement reduces your technical debt if and when new workload features are needed.

Optimize over time

You can optimize cost over time by reviewing new services and implementing them in your workload. As AWS releases new services and features, it is a best practice to review your existing architectural decisions to ensure that they remain cost effective. As your requirements change, be aggressive in decommissioning resources, components, and workloads that you no longer require. Consider the following best practices to help you optimize over time. While optimizing your workloads over time and improving your CFM culture in your organization, evaluate the cost of effort for operations in the cloud, review your time-consuming cloud operations, and automate them to reduce human efforts and cost by adopting related AWS services, third-party products, or custom tools (like AWS CLI or AWS SDKs).

Best practice questions

- FSICOST15: Have you reviewed your ongoing cost structure tradeoffs for your current AWS services lately?

- FSICOST16: Are you continuously assessing the ongoing costs and usage of your cloud implementations?
- FSICOST17: Are you continually reviewing your workload to provide the most cost-effective resources?
- FSICOST18: Do you have specific workload modernization or refactoring goals in your cloud strategy?
- FSICOST19: Do you use the cloud to drive innovation and operational excellence of your business model to impact both the top and bottom line?

FSICOST15: Have you reviewed your ongoing cost structure tradeoffs for your current AWS services lately?

You can optimize cost over time by reviewing new services and implementing them in your workload. As AWS releases new services and features, it is a best practice to review your existing architectural decisions to ensure that they remain cost effective.

FSICOST15-BP01 Monitor and optimize your ongoing costs, ROIs, and tradeoffs against alternative AWS services on a periodic basis to maintain your lowest cost of ownership

Financial services institutions add new human resources periodically, like contractors, vendors, or FTEs, so it is necessary to maintain a cost-aware culture. There are also enhancements from AWS on cost-related services. You should conduct periodic workshops, sessions on effective ways to measure, monitor and optimize cost to spread awareness of cost optimization to existing resources, as well as new resources on the team. The frequency of such workshops should be at least once every six months. Every six months, or during the session, you should recognize cost optimization wins and recognize individual people driving or contributing to the cost optimization. This drives cost-optimization culture in a team.

FSICOST16: Are you continuously assessing the ongoing costs and usage of your cloud implementations?

There is a process to examine existing cloud spend, and identify cost optimization opportunities using manual analysis, or the use of tools (AWS Billing and Cost Management and AWS Cost Management tools, AWS Partner tools, open-source tools, or DIY tools). As your requirements

change, be aggressive in decommissioning resources, components, and workloads that you no longer require.

FSICOST16-BP01 Use AWS cost management tools to perform retrospective, audit-based cost optimization on existing cloud workloads

There is a process to examine existing cloud spend, and identify cost optimization opportunities using manual analysis, or the use of tools (AWS Billing and Cost Management and Cost Management tools, AWS Partner tools, open-source tools, or DIY tools). Cost optimization opportunities are identified, prioritized, and implemented in a continuous, programmatic manner, ensuring all cloud workloads run as lean as possible while meeting all functional and non-functional requirements.

FSICOST17: Are you continually reviewing your workload to provide the most cost-effective resources?

There are multiple factors that affect the architecture, for example, new enhancements related to business requirements, re-architecting your workload to improve efficiency, new services released by AWS, price changes by AWS, or your team creating an MVP product with services without considering costs. It is necessary to continually review the architecture and resources used by your workload.

FSICOST17-BP01 Assess workload architecture to identify the most cost-effective resources

There are multiple factors that affect the architecture, for example, new enhancements related to business requirements, re-architecting your workload to improve efficiency, new services released by AWS, price changes by AWS, or your team creating an MVP product with services without considering costs. It is necessary to assess the architecture and resources used by workload, for example, usage of serverless technologies, managed services to reduce the operational overhead, or AWS Graviton-based instances that meet your needs. Alternatively, you can refactor your monolithic application to run as microservices. Most of the FSI systems are API-driven, so splitting them across a number of diverse services helps procurement, and the right-sizing of related resources.

FSICOST18: Do you have specific workload modernization or refactoring goals in your cloud strategy?

In traditional financial institutions, databases and core banking solutions are key cost drivers. Improve your total cost of ownership (TCO) by refactoring your lift and shift strategies to continue your modernization activities where you can improve performance while reducing your costs.

FSICOST18-BP01 Define ambitious modernization strategy to become truly AWS optimized

In traditional financial institutions databases and core banking solutions are key cost drivers. Improve your Total Cost of Ownership (TCO) by refactoring your lift and shift strategies to continue your modernization activities where you can improve performance while reducing your costs. The Operational Excellence pillar helps you define which workloads are suitable for refactoring. In the case of core banking systems provided by a vendor, start a dialog with your vendor to build a roadmap for workload modernization to make them cost-efficient. Also concentrate on modernization of workloads that interact with databases and core banking systems (for example, customer-facing web-pages, and apps). Leverage the AWS service WorkSpaces for remote diagnostics.

FSICOST19: Do you use the cloud to drive innovation and operational excellence of your business model to impact both the top and bottom line?

Today, technology and digital solutions are an integral part of FSI operations, however IT cost is not the biggest block within all expenditures in the profit and loss of FSI customers (personnel and marketing have greater impacts on cost). Using AWS Cloud solutions and services to change the way you operate impacts your profitability in the short and long term.

FSICOST19-BP01: Use AWS cloud services to change the way you reduce cost and improve agility in your infrastructure

Today, technology and digital solutions are an integral part of FSI operations, however IT cost is not the biggest block within all expenditures in the profit and loss of FSI customers (personnel and marketing have greater impacts on cost). Using AWS cloud solutions and services to change the way you operate impacts your profitability in the short and long term. Think big and explore regularly with your AWS Account Management team to test and launch new use cases and

solutions. For example, you may boost your IT teams' productivity by exploring Amazon Q Developer. With Intelligent Document Processing, you can automatically process financial or insurance documents using AI and free up capacity on your service teams.

Key AWS services

The following is a list of AWS services that are relevant for financial services customers.

- AWS Cost Explorer
- AWS Budgets
- AWS Cost and Usage Report
- AWS Billing and Cost Management Conductor
- AWS Cost Anomaly Detection
- AWS Cost Categories
- AWS Application Cost Profiler
- AWS Purchase Order Management
- AWS Billing and Cost Management Console
- Reserved Instance Reporting
- AWS Customer Carbon Footprint Tool
- CUDOS dashboards
- Saving Plans - Compute, Amazon EC2, and Amazon SageMaker AI
- Reserved Instances & Nodes - Amazon EC2, Amazon RDS, Amazon Redshift, Amazon ElastiCache, Amazon OpenSearch Service
- AWS Billing and Cost Management and AWS Cost Allocation Tags
- Cost Optimization Hub

Resources

Refer to the following resources to learn more about our best practices related to cost optimization for financial services customers. For more information, refer to the FSI Blog and Amazon Connect Resources.

Documents and blogs

- Cloud Financial Management Presentations relevant to financial services workloads
- Benefits and Customer Use Cases for Cost Optimization
- How Medibank achieved cost visibility and control on AWS
- Verisk Cost-Management Case Study
- Wealthfront Cost Reduction Cast Study
- AWS Cloud Financial Management Guide (PDF)

Whitepapers

- Best Practices for Tagging AWS Resources
- The Hartford: Total Cloud Migration
- Understand Your Amazon EKS Spend and Enable FinOps for Kubernetes with Anodot

Partner solutions

There are many AWS Partner solutions helping our customers to provide cost insights. These partners use AWS services and are well-integrated with AWS. Financial services customers can use these third-party products in addition to AWS services to optimize cost. These products are listed in the AWS Marketplace, which can help with seamless billing and discounts, if applicable.

Videos

- AWS CFM Talks
- AWS Financial Services Costs Management

Training materials

- Skill Builder - Cloud for Finance Professionals
- AWS Well-Architected Cost Optimization Workshop

Sustainability

The sustainability pillar provides you with the discipline to address the long-term environmental, economic, and societal impact of your business activities. You can find extensive prescriptive guidance on your implementation's sustainability in the Well-Architected Sustainability Pillar whitepaper.

Financial institutions must focus on sustainability within their cloud operating model to reduce their impact on the environment and to encourage sustainable practices. Focusing on these areas helps financial institutions adapt their workloads to financial services industry sustainability best practices, to adopt new environmentally friendly technology trends, and to plan for the business impacts of potential future regulatory requirements.

Sustainability topics

The sustainability pillar includes the following topics on AWS cloud-based architectures. Keep these topics in mind when developing your workload and also when assessing the sustainability performance of your workloads.

- **Cloud sustainability:** Compare cloud capabilities against on-premises or hosted sustainability performance.
- **Shared responsibility model:** You *and* AWS are responsible for your cloud sustainability performance. AWS is responsible for providing the most sustainable infrastructure possible while you are responsible for judiciously developing workloads that take advantage of the most sustainable options provided by AWS.
  - **Sustainability of the cloud**: AWS' responsibility to you
  - **Sustainability in the cloud**: Your responsibility
- **Design principles for sustainability in the cloud:** This lens pillar and the sustainability whitepaper offer an extensive set of design principles and best practices for achieving the best possible outcomes.
- **Improvement processes:** After performing a Well-Architected Framework review, a number of improvement recommendations are provided by the AWS Well-Architected Tool. You can use this process to implement sustainability improvements to your workloads.
- **Best practices for sustainability:** The sustainability pillar and this lens pillar provide recommendations for sustainable practices across six areas of your critical workload infrastructure.

Design principles

The following section defines a set of design principles for financial services sustainability best practices. Typically, some areas of financial services workload performance must be very low latency (for example, trading and investing where microseconds can cost $1MMs) while other areas have no concerns about speed. In many areas of financial services, data retention for at least seven years is critical. However, maintaining low latency storage strategies for six-year-old insurance data is very wasteful.

- **Implement low-latency workloads only for time-critical performance.** Trading generally requires high-performance compute, networks, and storage, while banking and insurance typically do not.
- **Use tiered storage for data requiring long-term archive storage.** Financial records typically must be stored for at least seven years. Amazon S3 storage classes can better align resource usage to retrieval needs. Amazon S3 Standard is not recommended for long-term storage without any requirement for timely data retrieval. Amazon S3 Glacier storage classes offer long-term, secure, durable storage classes for data archiving.
- **Region selection is a complex factor for implementing financial workloads.** While selection of low-carbon Regions is generally recommended for processing of financial data, sometimes data residency requirements stipulate the use of higher carbon storage. Also, some financial data's low latency requirements drive the choice of Regions with a higher carbon footprint due to greater network latency requirements. The selection of the best Region might be driven by taking into account a variety of reasons.
- **Back up data only when it's difficult to recreate.** Too often data is backed up in multiple locations and in the wrong tiers of storage. Use smart backup and storage disciplines to reduce your workload's overall carbon footprint.

Definitions

- Region selection
- Alignment to demand
- Software and architecture
- Data
- Hardware and services
- Process and culture

Region selection

The choice of Region for your workload significantly affects its KPIs, including performance, cost, and carbon footprint. To effectively improve these KPIs, you should choose Regions for your workloads based on both business requirements and sustainability goals.

Best practice questions

- FSISUS01: How do you select the most sustainable Regions in your area?
- FSISUS02: How do you address data sovereignty regulations for location of sustainable Region?
- FSISUS03: How do you select a Region to optimize financial services workloads for sustainability?

FSISUS01: How do you select the most sustainable Regions in your area?

The choice of Region for your workload significantly affects its KPIs, including performance, cost, and carbon footprint. To effectively improve these KPIs, you should choose Regions for your workloads based on both business requirements and sustainability goals.

FSISUS01-BP01 Select a Region with lower environmental impact that meets your financial services industry business and compliance considerations

Prescriptive guidance

The following guidance is provided to aid your selection of most sustainable Regions in your area:

- Shortlist potential Regions based on the following topics:
  - Data security and privacy issues
  - Regulatory compliance requirements
  - The operational efficiency of your workloads
  - Local data sovereignty concerns (see FSISUS02)
  - A number of services and features that optimize sustainability
- Select Regions by market-based or location-based methods in line with your financial services industry's internal relevant sustainability guidelines that are used to track and to compare your organization's year-to-year emissions.

- Wherever possible, choose a Region that provides better than 95% renewable energy, using the market-based method and low grid carbon intensity, as well as using a typical location-based method.

FSISUS02: How do you address data sovereignty regulations for location of sustainable Region?

While selection of low-carbon Regions is generally recommended for processing of financial data, sometimes data residency requirements stipulate the use of higher carbon storage.

FSISUS02-BP01 Run workloads and store restricted data in required country and unrestricted in sustainable Region selected by following SUS01 guidance

Prescriptive guidance

The following guidance provides insights into data sovereignty regulations.

- Review data sovereignty regulations and identify workloads and data that can be run in sustainable Regions. You may need to separate your data and processing to take advantage of data and processes using lower carbon resources where data residency is not required, while accessing higher carbon resources when data residency is a requirement.
- Choose a sustainable Region following the guidance provided in FSISUS01.
- Run your workloads and store data whenever you are not restricted to specific locations using more sustainable Regions.

FSISUS03: How do you select a Region to optimize financial services workloads for sustainability?

Financial institutions must focus on sustainability within their cloud operating model to reduce their impact on the environment and to encourage sustainable practices. Focusing on these areas helps financial institutions adapt their workloads to financial services industry sustainability best practices, to adopt new environmentally friendly technology trends, and to plan for the business impacts of potential future regulatory requirements. The selection of the best Region might be driven by taking into account a variety of reasons.

FSISUS03-BP01 Select Regions that offer services required by financial services organizations and hardware that maximizes the reduction of your carbon footprint

Prescriptive guidance

Recommended guidance for customer architecture includes:

- Develop a list of all services required by financial services workloads.
- Select a Region using guidance from FSISUS01-BP01.
- Develop a cross-reference of sustainable Regions chosen according to the services that are offered within each Region as well as the variety and types of sustainable hardware offered in the Region.

Alignment to demand

Best practice questions

- FSISUS04: How do you prioritize business critical functions over non-critical functions?
- FSISUS05: How do you define, review, and optimize network access patterns for sustainability?

FSISUS04: How do you prioritize business critical functions over non-critical functions?

Determine what is defined as a business-critical process and workload, and protect and prioritize it. Model and prioritize individual functions and workloads by recording relevant metadata, such as interdependencies, SLAs for particular flows, and nuances of user access.

FSISUS04-BP01 Actively manage each business function and the allocation and configuration of resources

Prescriptive guidance

- Use Amazon ECS Spot compute for non-critical workloads such as end-of-month reconciliations.
- Use Amazon EC2 Dedicated Hosts queues for priority jobs such as order initiation.
- Use Amazon ECR Lifecycle Policies for ephemeral ETL data such as ingestion ledgers.

- Develop architecture strategies that use built-in queueing and buffering to offload non-critical tasks.

FSISUS04-BP02 FSI workloads serve the highest common denominator of application demands

Systems in financial services are built to serve the highest level of performance for retention, availability, and integrity. This leads to workloads that often exceed performance expectations or might not be respectful of ancillary or critical jobs and workflows. Breaking down a system into its component parts allows for a more fine-grained view of resource consumption and the trade-offs possible to balance SLAs against your sustainability goals.

Prescriptive guidance

Provide prioritization advice to customers on the following topics:

- **Prioritize at the organizational level:** Determine what is defined as a business-critical process and workload, and protect and prioritize it.
- **Prioritize at the SCP/OU level:** Restrict AWS usage-based metrics on your Organizational Units' (OU) profiles and requirements. Batch-running processes that have extended SLAs can have dedicated accounts and permissions to restrict and reduce their carbon impact; for example, select serverless preferences, choose specific instance types, or operate during specific processing hours. Development and test instances should have enforced central guardrails to limit Amazon EBS attachments, or automatically pause and resume resources as needed.
- **Prioritize at the account level:** Model and prioritize individual functions and workloads by recording relevant metadata, such as interdependencies, SLAs for particular flows, and nuances of user access. For example, investigations and warm access commonly take longer at a bank than its typical 35-day retention period.
- **Prioritize at the resource or tag level:** Use tags to group and aggregate the management and reporting of resources. You may only have one critical flow but you likely monitor dozens of processes and receive millions of Event Notifications. Create a prioritization schema to determine which process matter most to your workload operations.
- **Prioritize at the job or object level:** Not all jobs are born equal. Use mechanisms such as graceful termination of non-critical jobs and active workload management to help you prioritize at the job and object levels.

FSISUS05: How do you define, review, and optimize network access patterns for sustainability?

Assess and optimize network access patterns for sustainability. Pay attention to redundant layers and redirects or patterns generating excessive and unnecessary data movement.

FSISUS05-BP01 Analyze network access patterns to identify the places that your customers are connecting from

Prescriptive guidance

Remove redundant layers and redirects, use pagination and local caching mechanisms to reduce data movement, and consider separating workloads that serve different users.

FSISUS05-BP02 Avoid common architectural misconfigurations

Problem statement

In financial services organizations, it's common to hairpin large amounts of traffic through on-premises networks, have largely redundant layers of control using trusted private networks, and sometimes include untrusted public traffic.

A simple example of this is using AWS Direct Connect where performance is often degraded as FSI organizations insist that all inbound and outbound traffic originates from their network.

Another common mistake is to serve both OLAP and OLTP workloads from the same database or cluster, which normally span two or more completely different geographic locations. Both of these patterns generate excessive and unnecessary data movement.

Prescriptive guidance

Identify poor architectural choices and risky configurations as good candidates for remediation. Assess your workflows from the perspective of varying demand over time, so select scalable AWS services over fixed ones. Do not underestimate your network requirements, especially for peak loads. Provide sufficient failover resources to support your operations in case of partial outages.

Software and architecture

Best practice questions

FSISUS06: How do you monitor and minimize resource usage for financial services workloads?

Monitor and analyze your financial services' usage patterns to minimize resource usage. Identify services that are not required to be operational at all times or that can be scaled up and down based on user access patterns.

FSISUS06-BP01 Actively monitor your FSI resource usage

- Monitor and analyze your financial services' usage patterns to minimize resource usage.
- Identify services that are not required to be operational at all times, or that can be scaled up and down based on user access patterns.
- For example, many consumer-based services can be scaled down or turned off during off-peak hours.

Prescriptive guidance

- Remove underutilized software modules and combine these functions into other software services.
- Minimize the average resource demand required per unit-of-work using automatic scaling services, serverless transaction processing, or shutting down your resources when usage patterns permit.
- Use queue-driven architectures, pipeline management, and On-Demand Instance workers to maximize your utilization for batch processing.

FSISUS07: How do you optimize batch processing components for sustainability?

Because batch processing is often found within many workloads across financial systems, verify that the minimum number of resources are consumed by batching transactions together while meeting your customer SLA and system requirements.

FSISUS07-BP01 Optimize your batch processing systems

Because batch processing is often found within many workloads across financial systems, verify that the minimum number of resources are consumed by batching transactions together while meeting your customer SLA and system requirements.

Prescriptive guidance

- Queue up several requests together that don't require immediate processing.
- Increase serialization to flatten utilization across your pipeline.
- Modify the capacity of individual components to prevent idling resources waiting for input.
- Create buffers and establish rate limiting to smooth the consumption of external services.
- Use the most efficient available hardware and services to optimize your software.
- If possible, schedule jobs during times of day where carbon intensity for power is lowest.

FSISUS08: How do you optimize your resource usage?

Review and optimize your resource usage by implementing either a pub/sub or pull mechanism instead of relying on a polling approach.

FSISUS08-BP01 Use event-driven architecture

Implement either a pub/sub or pull mechanism instead of using a polling approach.

Prescriptive guidance

- Implement event-driven architecture where possible to avoid idling of resources running and waiting for state changes.
- If event-driven architecture is not possible, modify the capacity of individual components to prevent idling downstream resources waiting for input.

- Avoid polling APIs or queues, instead have components and services subscribe to events or be notified of changes to reduce the idling of resources.

FSISUS09: How do you optimize areas of your code that use the most resources?

Analyze and optimize your code's efficiency to improve resource utilization.

FSISUS09-BP01 Monitor and optimize areas of code that are the most compute resource-intensive

Prescriptive guidance

- Use CodeGuru and Amazon Q Developer to optimize your code's efficiency.
- If possible, choose the most efficient OS and programming languages to run your code.
- Remove unnecessary code such as modules that perform sorting or formatting.

FSISUS10: Have you selected the storage class with the lowest carbon footprint?

Data is at the heart of strategic innovations for the financial services industry. This can have many use cases ranging from providing hyperpersonalised experiences for customers, training machine learning models to better understand risk and fraud detection. Each use case requires different levels of data availability, processing, and storage and therefore varies in storage technologies from transactional databases, to data lakes and data warehouses. These come with various considerations from a sustainability perspective.

FSISUS10-BP01 Balance your data performance requirements against its carbon footprint

Prescriptive guidance

To balance data performance requirements against its carbon footprint:

- Define proxy metrics to monitor the business outcome of the data-involved service in relation to their environmental impact. An example proxy metric could be efficiency of the AI/ML service to help detect fraud faster (with the associated cost saving) and the carbon footprint of training

and storing the data. These proxy metrics then become the vehicle to balance your performance requirements against its carbon footprint. Proxy metrics can be collected by importing AWS Cost and Usage Report as well as Amazon CloudWatch metrics into Amazon S3 and monitored using Amazon Athena and Amazon QuickSight.

- Use the right storage class for Amazon S3 Storage Classes based on the data performance requirements. The storage class impacts the environmental impact of the dataset through its access patterns and its architecture. For example, in Amazon S3 One Zone-IA, energy and server capacity are reduced because data is stored only within one Availability Zone. Amazon S3 Storage Classes can be configured at the object level and a single bucket can contain objects stored across all of the storage classes.
  - Learn more about Amazon S3 Storage Classes and their use cases.
  - You can also use Amazon S3 lifecycle policies to transition objects automatically between storage classes without application changes. In general, you have to make a trade-off between resource efficiency, access latency, and reliability when considering these storage mechanisms.
- For storage systems that are a fixed size, such as Amazon EBS or Amazon FSx, monitor the available storage space and automate storage allocation on reaching a threshold. You can use Amazon CloudWatch to collect and analyze different metrics for Amazon EBS and Amazon FSx.
- Avoiding the backup of unnecessary data can help lower cost and reduce the storage resources used by the workload. Only back up data that has business value or is needed to satisfy compliance requirements. Use AWS Backup for Amazon EFS or Amazon S3 Glacier Storage options for backup of infrequently accessed data.

Data types may include the following:

- Real-time analytics for financial services, including banking, payments, insurance, and markets.
- Unstructured data such as biometrics, facial images, and documents.
- Structured data like fund movements or, transaction attempts.

FSISUS10-BP02 Separate data into hot, warm, cold storage

Prescriptive guidance

- Implement a data classification policy to understand its criticality to business outcomes and choose the right energy-efficient storage tier. Determine criticality, confidentiality, integrity, and availability of data based on risk to the organization.

- Evaluate your data characteristics and access pattern to collect the key characteristics of your storage needs. Key characteristics to consider include:
  - **Data type:** Structured, semistructured, unstructured
  - **Data growth:** Bounded, unbounded
  - **Data durability:** Persistent, ephemeral, transient
  - **Access patterns:** Reads or writes, frequency, spiky, or consistent
- Use these requirements to group data into one of the data classification tiers that you adopt. For more detail on data classification categories, see the Data Classification whitepaper.
- AWS Glue Data Catalog lets you store, annotate, and share metadata in the AWS cloud while providing comprehensive audit and governance capabilities, in order to periodically audit your environment for untagged and unclassified data and tag the data appropriately.

FSISUS11: Do you store processed data or raw data?

FSISUS11-BP01 Use processed data to reduce your storage footprint

Often raw data from your data sources may include a large number of observations from streaming data sources that continually produce data, or include large amounts of redundant data from a variety of sources. You can reduce your storage requirements by first processing the raw data, then storing only the results. Unless you have a raw data retention compliance policy or requirement, you can purge the raw data automatically shortly after processing to reduce your data storage requirements.

Hardware and services

Best practice questions

- FSISUS12: What is your process for benchmarking instances for existing workloads?
- FSISUS13: Can you complete workloads over more time while not violating your maximum SLA?
- FSISUS14: Do you have multi-architecture images for grid computing systems?
- FSISUS15: What is your testing process for workloads that require floating point precision?

FSISUS12: What is your process for benchmarking instances for existing workloads?

Maximizing your instance utilization is an effective and quantifiable practice that helps you meet your sustainability goals. But reaching an ideal utilization state is a process — it's uncommon for customers to achieve optimal instance utilization on their first attempt. Define a process to monitor resource utilization over time so you can benchmark performance and make the necessary adjustments to your workloads.

FSISUS12-BP01 Set appropriate instance usage goals that reflect your sustainability requirements

Prescriptive guidance

- Instance utilization goals differ for every company, but you can use common metrics that are broadly applied regardless of company size, age, industry, or domain like carbon emissions and energy consumption.
- You can use these metrics to set goals like an ideal utilization percentage, or a maximum idle instance threshold.
- It's important to set measurable instance utilization goals that apply within the context of your business to see and iterate over time.
- Setting appropriate goals provides guidance and justification for every decision that your organization makes as it collectively works toward a sustainable usage state.

FSISUS12-BP02 Track your overall process in achieving your goals

Prescriptive guidance

- It's harder to achieve goals if you are not aware of your progress and if you don't know where you are, you're unable to pivot to make the right changes in reaching your goal.
- Do this by setting a regular cadence with the appropriate stakeholders to identify the current state and creating action plans to iterate, if necessary.
- AWS provides tools to help your track your overall progress such as the AWS Customer Carbon Footprint Tool to report on emissions from your AWS usage, and specifically Amazon EC2, which follows Greenhouse Gas (GHG) Protocol standards.

- You can analyze the changes in your emissions over time and forecast how your emissions change across your sustainability journey.

FSISUS12-BP03 Monitor your individual instance performance metrics

Prescriptive guidance

- Establish a process to monitor individual instances to help you to use two major optimization approaches:
  - Using only what you need
  - Right-sizing what you do need
- Amazon CloudWatch provides a unified view of metrics that you can use to benchmark instance performance. Use both the default and custom metrics to gather the data you need to make informed decisions.
- For example, you can use the `IsIdle` default metric for Amazon EMR to identify clusters for termination. This process helps your organization adopt more optimal instances types since newer generation instances typically have better energy-to-performance ratios.
- Run performance tests specific to the processor to better understand your workloads' needs to help lower your workload's instance count by evaluating whether workloads are properly fitted to an instance family by performance metrics other than CPU, and reduce unnecessary instances.
- Establish a process to also track supply to demand with Amazon EC2 Auto Scaling. This helps keep your scaling policies dynamic and relevant to changes to your workload.
- **Implementation guidance:** Hpc 7g instance may be the obvious contender for a grid computing workload, but network constraints could cause the need for more instances. Consider switching to C7gn. Do not go after cores, as memory bandwidth, faster I/O, and higher clock speeds may be more beneficial for highly intensive financial simulations. For example, on AWS Graviton, since each vCPU is its own physical core, verify that workloads are running instances beyond 60% CPU to breakage to best assess threshold and limit over provisioning instances.
- **Service recommendations:** Use the following services to achieve these goals:
  - AWS Compute Optimizer
  - Amazon CloudWatch metrics
  - AWS Graviton Performance Runbook

FSISUS13: Can you complete workloads over more time while not violating your maximum SLA?

How do you avoid load spikes to reduce the provisioned capacity required for your workload?

Flattening the workload demand curve can help you to reduce the provisioned capacity for a workload and reduce its environmental impact. In other words, if you can afford to spread out the load over a longer period of time, rather than having a higher peak in a shorter span of time, then you lower the overall resource demand for the workload. By doing so, you lower the overall amount of provisioned capacity, and thus lower overall energy consumption to meet the workload's demand.

FSISUS13-BP01 Do not complete a customer transaction in the shortest time when not required by end users

Prescriptive guidance

If your workload does not have time-sensitive requirements, consider running them during times when public demand is lower. This distributes energy consumption to flatten the resource demand curve. Evaluate your workload requirements to assess if you are able to make this adjustment.

FSISUS13-BP02 Introduce jitter to your scheduled tasks

Prescriptive guidance

- Assess if your scheduled tasks can be distributed to run at random times during an hour or throughout the day. This minimizes the highs of peak demand load and spreads it across the day instead. Avoid using the same start minute of scheduled tasks. Doing so creates high demand for resources at a specific time, which introduces stress on energy consumption. Staggering job start times avoids load spikes and creates time-flexible workloads.
- **Implementation guidance:** Evaluate whether highly intensive computational workloads such as financial simulation can be spread over time and run fewer instances to maximize renewable energy availability. If a grid computing workload is using a third-party scheduler, prioritize workloads that need to provide calculations for regulators and trading desks that need information prior to markets opening, so workloads that are not urgent can be pushed off and worked on at a consistent rate to maximize renewable energy availability. Additionally, verify that a proper fault tolerance framework is implemented, as restarting a launch can increase launch time and energy consumption.
- **Service recommendations:** Use Amazon Simple Queue Service (Amazon SQS) achieve your goal.

FSISUS14: Do you have multi-architecture images for grid computing systems?

Multi-architecture image support for a particular workload makes it easier for you to build different images and thus different architectures and operating systems from the same source and refer to them all by the same abstract manifest. The manifest specifies the layers of system content that make up the image as well as its runtime characteristics and configuration. Having a multi-architecture image increases the flexibility of the workload thus increases the opportunity to use hardware that may be more sustainable.

FSISUS14-BP01 Use instances with higher energy efficiency

Prescriptive guidance

- AWS Graviton-based instances use up to 60% less energy than comparable EC2 instances.

FSISUS14-BP02 Design applications that can use different Amazon EC2 instance types

Prescriptive guidance

- This is what we would call a flexible workload. In contrast, inflexible workloads rely only on a few instance types. These instances types may be less energy efficient than others.
- Flexible workloads are ideal for Spot Instances. Running workloads on Spot Instances is generally considered more energy efficient than On-Demand Instances because Spot is overhead required for the Amazon EC2 On-Demand service to run.
- Use Amazon EC2's spare capacity with Spot Instances to extract the same value, which increases the total value generated from the Amazon EC2 environment as a whole.

FSISUS14-BP03 Adopt a serverless, event-driven architecture

Prescriptive guidance

- Consider using a serverless, event-driven architecture to maximize overall resource utilization. Serverless architecture removes the requirement to run and maintain physical servers since AWS handles this on your behalf.

- The cost of serverless architectures generally correlates with the level of usage, thus increases your workload's cost efficiency.
- **Implementation guidance:** Maximize energy efficiency as well as availability by building multi-architecture workloads that can run on a variety of Spot Instances. It is important to account for error precision when expanding compiler options on varying processors.
- **Service recommendations:** Use the following services to achieve your goal:
  - Amazon Simple Queue Service and Amazon EC2 Spot Instances
  - AWS CodeBuild
- Determine which of your workloads is suitable for use of floating-point accuracy, performance, and efficiency. Consider testing with a cluster of instances to see how well it performs at scale.
- **Implementation guidance:** For intensive financial simulations and calculations, test the number of bits that are required to achieve your floating point precision and consider reducing number of bits by selecting different floating-point formats, including bfloat16, that's supported by AWS Graviton.
- **Service recommendations:** Use the following services to achieve your goal:
  - AWS Batch
  - AWS Parallel Cluster

FSISUS15: What is your testing process for workloads that require floating point precision?

FSISUS15-BP01 Minimize the bit count while maintaining precision

Prescriptive guidance

Floating point precision is a way to represent real numbers in a finite binary format. It stores a number in a fixed-width field with the intent to reduce the memory bandwidth and storage requirements compared to double-precision arithmetic results. Although double-precision can sometimes lead to more accurate results, single-precision calculations can be faster and thus reduce overall energy consumption for particular workloads. Determine which of your workloads is suitable for use of floating-point accuracy, performance, and efficiency. Consider testing with a cluster of instances to see how well it performs at scale.

Implementation guidance:

- For intensive financial simulations and calculations, test the number of bits that are required to achieve your floating point precision and consider reducing number of bits by selecting different floating-point formats, including bfloat16, that's supported by AWS Graviton.
- Using floating point Quantization, you can represent numbers using lower bit-count integers or floating point numbers without incurring a significant loss in accuracy. Specifically, you can reduce resource usage by replacing the parameters in your workload with (1) half-precision (16 bit), (2) bfloat16 (16 bit, but the same dynamic range as 32 bit), or 8-bit integers instead of the usual single-precision floating-point (32 bit) values.
- **Service recommendations:** Use the following services to achieve your goal.
  - AWS Batch
  - AWS Parallel Cluster
  - Graviton3

Process and culture

Best practice questions

- FSISUS16: Do you achieve a judicious use of development resources?
- FSISUS17: How do you minimize your test, staging, sandbox instances?
- FSISUS18: How do you define the minimum requirement in response time for customers in order to maximize your green SLA?

FSISUS16: Do you achieve a judicious use of development resources?

FSISUS16-BP01 Verify that all development resources are dedicated to an active project or team

Often, project test environments and resources are set up in anticipation of an upcoming project. If that project is cancelled or never commences, some development resources could be orphaned from their original projects. To mitigate this, establish a regular review of all test resources to reduce these missing projects.

FSISUS17: How do you minimize your test, staging, sandbox instances?

FSISUS17-BP01 Use infrastructure as code (IaC) code base to snapshot your environment allowing you to decommission test infrastructure

Prescriptive guidance

Reducing the number, frequency, and use of test and staging environments can reduce your environmental impact. If you use Infrastructure as Code (IaC) — with AWS Event Engine or Workshop Studio — to snapshot your environments, you can break down the infrastructure once your testing is complete. This allows you to reduce the unneeded resources. If the test environment is required later, you can use IaC to restore it when needed.

Instead of creating separate instances to test several environments, use snapshots to test only the required workload using the same instance. You can queue your testing based on development priorities to reduce the use of test and staging instances.

FSISUS18: How do you define the minimum requirement in response time for customers in order to maximize your green SLA?

FSISUS18-BP01 Use a green SLA

Prescriptive guidance

The Institute of Electronics and Electrical Engineers standards body has created a set of recommendations known as the *green SLA* that offsets the responsiveness of system to meet customer requirements against the need to reduce environmental impacts. For more information, see Providing green SLAs in High Performance Computing clouds.

Key AWS services

AWS services promoting sustainability practices include:

- Amazon S3, S3 Glacier, Deep Archive, Amazon S3 Intelligent-Tiering, One Zone – Tiered storage classes
- High performance computing *(link is missing)*
- Infrastructure as Code (IaC)

- AWS Event Engine *(link needs to be fixed)*
- AWS Batch
- AWS Parallel Cluster
- Amazon Simple Queue Service and Amazon EC2 Spot Instances
- AWS CodeBuild
- Amazon Simple Queue Service (Amazon SQS)
- AWS Compute Optimizer
- Amazon Cloudwatch metrics
- AWS Graviton Performance Runbook
- Amazon EC2 Auto Scaling
- Amazon CloudWatch
- AWS Customer Carbon Footprint Tool
- Amazon CodeGuru
- Amazon Q Developer
- Amazon ECS Spot
- Amazon EC2 Dedicated Hosts
- Amazon ECR Lifecycle Policies
- *The following links are missing from the Word document and need to be provided or the links removed.*
- AWS Data Exchange and Amazon Open Data Initiative (ESG & Alternative Data Sourcing)
- SageMaker AI Family (risk assessment, predictive analytics, and forecasting)
- Graviton - EC2 (60% less energy for the same performance)
- IoT (tracking environmental change, for example, office based heat, light)
- Amazon Monitron (predictive maintenance, for example, UPS and generators)
- Customer Carbon Footprint Tool (carbon measurement)
- AWS Well-Architected Tool (sustainability assessment)
- Cost Explorer (detailed usage and efficiency reporting)
- AWS Trusted Advisor (efficiency recommendations)
- Reserved Instances and Saving Plans (more with less, spread the load)

Resources

Documentation and blogs

- What to Consider when Selecting a Region for your Workloads
- How to select a Region for your workload based on sustainability goals
- Renewable energy projects on Amazon Around the Globe
- Renewable Energy Methodology
- Building Sustainable, Efficient, and Cost-Optimized Applications on AWS
- Reducing Your Organization's Carbon Footprint with Amazon CodeGuru Profiler
- Increasing sustainability for your Microsoft workloads on AWS
- Seven-step roadmap for CEOs and CFOs who are embarking on sustainability reporting journeys

Whitepapers

- What is CodeGuru Profiler?
- Providing green SLAs in High Performance Computing clouds
- AWS Graviton Performance Testing

Conclusion

The goal of the Financial Services Industry Lens for the Well-Architected Framework is to provide architectural best practices for designing and operating reliable, secure, efficient, and cost-effective regulated financial services workloads on AWS. In operational excellence, we outline best practices around how people, process, and operating models need to be aligned so that workloads running on AWS can support critical financial services business services. Architectures for financial services workloads need to incorporate security and evidence-based compliance design patterns. Financial services customers also need to continuously monitor, measure, and test failure and recovery in the cloud to achieve their business resiliency and performance objectives. These objectives can be met with significant cost savings by right-sizing and establishing governance models around consumption and monitoring of AWS resources.

This framework can improve security, resiliency, and operational efficiency for financial services customers migrating and building apps on AWS, and can also assist in meeting regulatory and compliance obligations.

Contributors

Contributors to this version of the Well-Architected FSI Lens document include:

- Amanda Anderson, Financial Services Specialist Central US, Amazon Web Services
- Jason Barto**,** Principal Solutions Architect, Amazon Web Services
- Bikash Behera, Enterprise Transformation Arch , AWS BDSI FSI, Amazon Web Services
- Sundeep Bhasin, Principal Compliance Specialist, Amazon Web Services
- Julio Carvalho, Principal Security Solutions Architect, Amazon Web Services
- Ruy Cavalcanti, Senior Security Solutions Architect, Amazon Web Services
- Peter Chung, Senior Solutions Architect, Amazon Web Services
- James Craig, Sr Partner Solutions Architect, EMEA FSI, Amazon Web Services
- Pradeep Dhananjaya, Senior Solutions Architect , AWS BDSI FSI, Amazon Web Services
- Gregg Sorrels, Senior Technical Account Manager - GFS, Amazon Web Services
- Guillermo Tantachuco, Principal Solutions Architect, Amazon Web Services
- Michael Dobson, Senior Technical Acct Mgr (MNG) , Amazon Web Services
- Laurent Domb, Chief Technologist, WWPS Federal Financials, Amazon Web Services
- Praveen Edem, Senior Solutions Architect, Amazon Web Services
- Vikram Elango, Senior AI/ML Specialist Solutions Architect, Amazon Web Services
- Romy van Es, Partner Solutions Architect, Amazon Web Services
- Cory Visi, Financial Services Industry Solutions Architect, Amazon Web Services
- Kurt Gray, Senior Manager, Solutions Architecture, Amazon Web Services
- Aravind Gopaluni, Senior Security Solutions Architect, Amazon Web Services
- Sven Hansen, AWS BDSI EA Comm Field Solutions Architect, Amazon Web Services
- Hahnara Hyun, Senior Specialist Solutions Architect, EC2 Graviton, Amazon Web Services
- Max Ivashchenko, Senior Solutions Architect, Amazon Web Services
- Anu Jayanthi, Startup Solutions Architect, Amazon Web Services
- Kenneth Jackson, Sr Mgr, Solution Architecture, Amazon Web Services
- Sudhir Kalidini – Principal Solutions Architect, Amazon Web Services
- Ligia Lopes, Sr. Manager, Public Policy, Amazon Web Services
- John Lucking, Tech Lead Insurance, BDSI FSI Business Development, Amazon Web Services

- Sumit Malik, Enterprise Support Manager (M-MNG), Amazon Web Services
- Colin Marden, Principal Solutions Architect, Amazon Web Services
- Alket Memushaj, Principal Solutions Architect, Capital Markets , AWS BDSI FSI, Amazon Web Services
- Fernando Nunes, Senior TAM (MNG), Amazon Web Services
- Mike Perna, Capital Markets Principal Solutions Architect, Amazon Web Services
- Viktoriia Potishuk, Senior Business Development Manager, Amazon Web Services
- Chintan Sanghavi, Senior Partner SA, Startup, Amazon Web Services
- Padmapriya Seshadri, Senior Solutions Architect, Amazon Web Services
- Anil Sharma, Senior Partner Solutions Architect, Atos, Migration, WW, Amazon Web Services
- T. Luke Young, Climate Change Business Development Manager, Amazon Web Services
- Darius Januskis, Senior Solutions Architect - Financial Services, Amazon Web Services
- Bruce Ross, Senior Lens Leader, Well-Architected Framework, Amazon Web Services

Contributors to the previous version of this document included:

- Arjun Chakraborty, Principal Solution Architect, AWS Financial Services
- Ilya Epshteyn, Principal Solutions Architect, AWS Financial Services
- Misha Goussev, Principal Solutions Architect, AWS Financial Services
- Som Chatterjee, Senior Technical Program Manager, AWS Commerce Platform
- James Craig, Senior Partner Solution Architect, AWS Financial Services
- Anjana Kandalam, Manager, Solutions Architecture, AWS
- Roberto Silva, Senior Solutions Architect, AWS
- Chris Redmond, Senior Consultant, Governance, Risk and Compliance, AWS Professional Services
- Pawan Agnihotri, Senior Manager, Solutions Architecture, AWS Global Financials
- Rahul Prabhakar, Global FSI Lead, AWS Security Assurance
- Jaswinder Hayre, Senior Manager, Solutions Architecture – Security, AWS
- Jennifer Code, Principal Technical Program Manager, AWS Financial Services
- Igor Kleyman, FSI Industry Specialist, AWS Security Assurance
- John McDonald, Head of Governance, Risk & Compliance – Americas, AWS Financial Services
- John Kain, Head of Banking and Capital Markets Business Development

Document revisions

To be notified about updates to this whitepaper, subscribe to the RSS feed.

Change	Description	Date
Major update	Added the sustainability pillar and numerous updates and changes throughout.	May 15, 2024
Minor update	Improved formatting of best practices.	March 3, 2022
Minor update	Updated links.	March 10, 2021
Minor update	The Reliability Pillar content adjusted for readability and clarity.	February 22, 2021
Minor updates	Updated question numbering in FSISEC and FSIREL. Minor text updates to improve accuracy.	June 3, 2020
Initial publication	Financial Services Industry Lens first published.	May 19, 2020

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

AWS Glossary

For the latest AWS terminology, see the AWS glossary in the *AWS Glossary Reference*.