Design Principles and Patterns for Data Pipelines

AWS Academy Data Engineering

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Introduction

Design Principles and Patterns for Data Pipelines

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Module objectives

This module prepares you to do the following:

- Use the AWS Well-Architected Framework to inform the design of analytics workloads.
- Recount key milestones in the evolution of data stores and data architectures.
- Describe the components of modern data architectures on AWS.
- Cite AWS design considerations and key services for a streaming analytics pipeline.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Module overview

Presentation sections

- AWS Well-Architected Framework and Lenses
- The evolution of data architectures
- Modern data architecture on AWS
- Modern data architecture pipeline: Ingestion and storage
- Modern data architecture pipeline: Processing and consumption
- Streaming analytics pipeline

Activity

- Using the Well-Architected Framework

Lab

- Querying Data by Using Athena

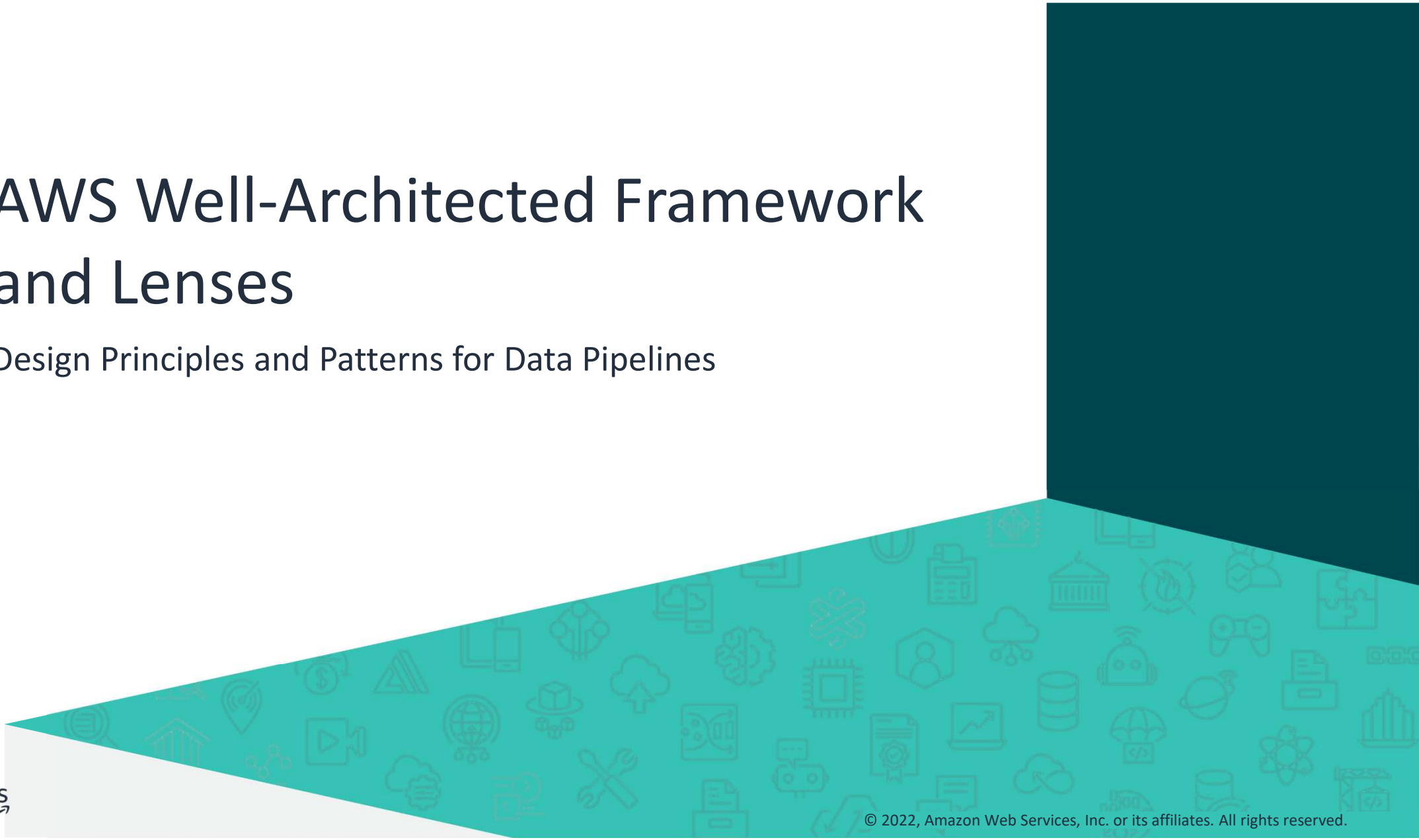
Knowledge checks

- Online knowledge check
- Sample exam question

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

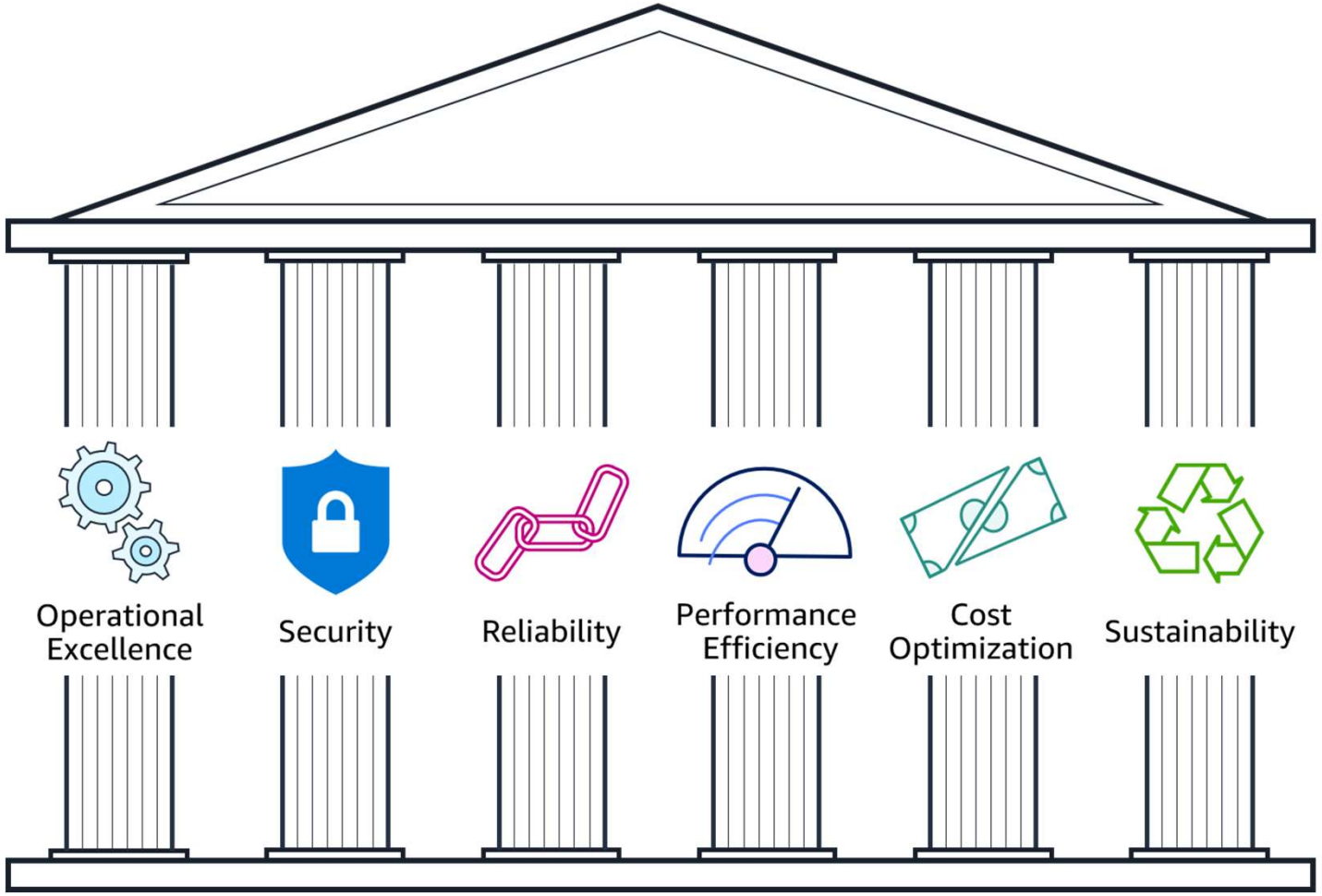
AWS Well-Architected Framework and Lenses

Design Principles and Patterns for Data Pipelines



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Well-Architected Framework pillars



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Well-Architected Framework Lenses

Well-Architected Lenses

- Extend the AWS Well-Architected Framework guidance to specific domains
- Contain insights from real-world case studies

Data Analytics Lens

- Provides key design elements of analytics workloads
- Includes reference architectures for common scenarios

ML Lens

- Addresses differences between application and machine learning (ML) workloads
- Provides a recommended ML lifecycle

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Activity: Using the Well-Architected Framework

- In this activity, you will use the Data Analytics Lens from the Well-Architected Framework to identify cloud best practices that your data engineering team should follow when building their data pipelines.
- Use the detailed instructions that are provided in your online course to complete this activity.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key takeaways: AWS Well-Architected Framework and Lenses

- The Well-Architected Framework provides best practices and design guidance across six pillars.
- The Well-Architected Framework Lenses extend guidance to focus on specific domains.
- The Data Analytics Lens provides guidance that helps with design decisions related to the elements of data (volume, velocity, variety, veracity, and value).

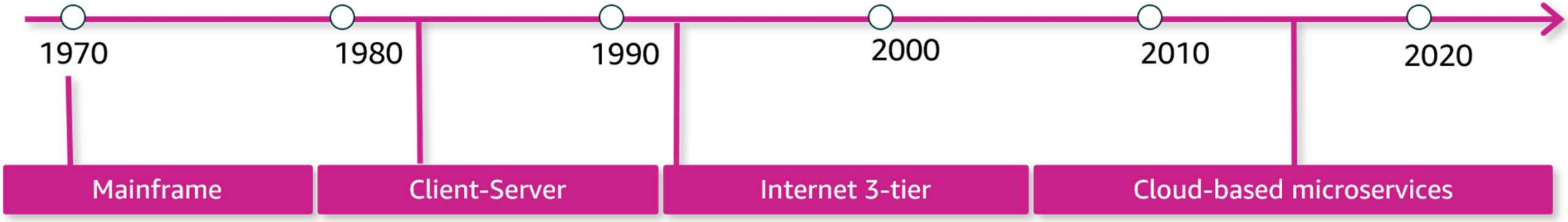
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The evolution of data architectures

Design Principles and Patterns for Data Pipelines

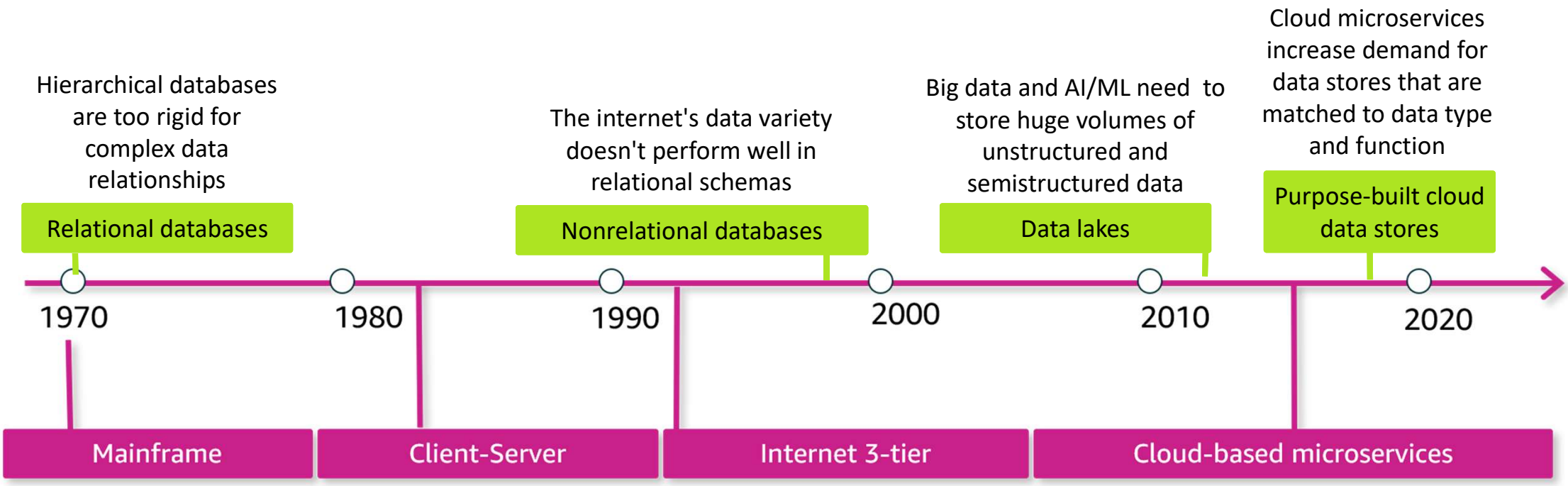
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Application architecture evolved into more distributed systems



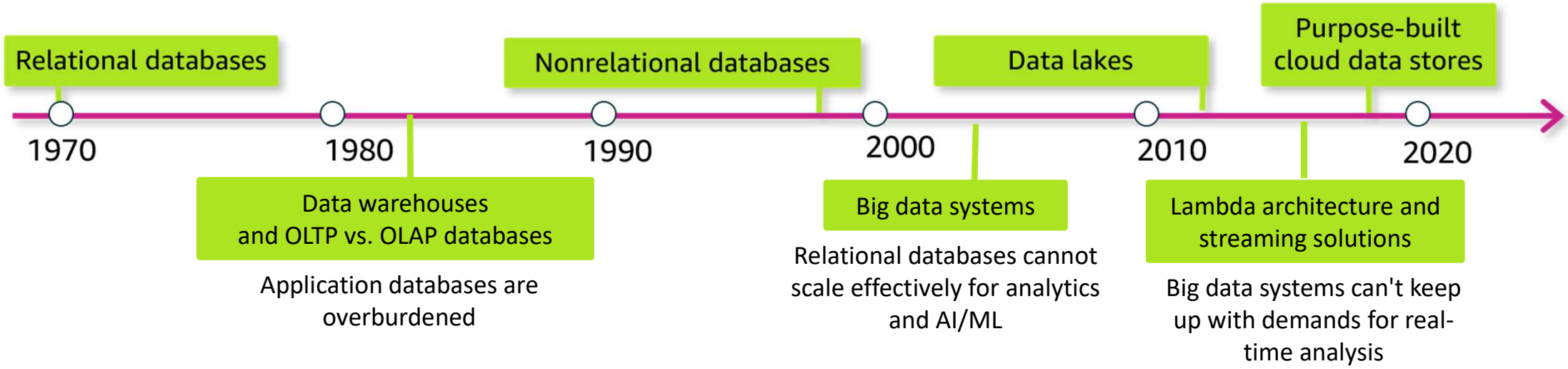
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data stores evolved to handle a greater variety of data



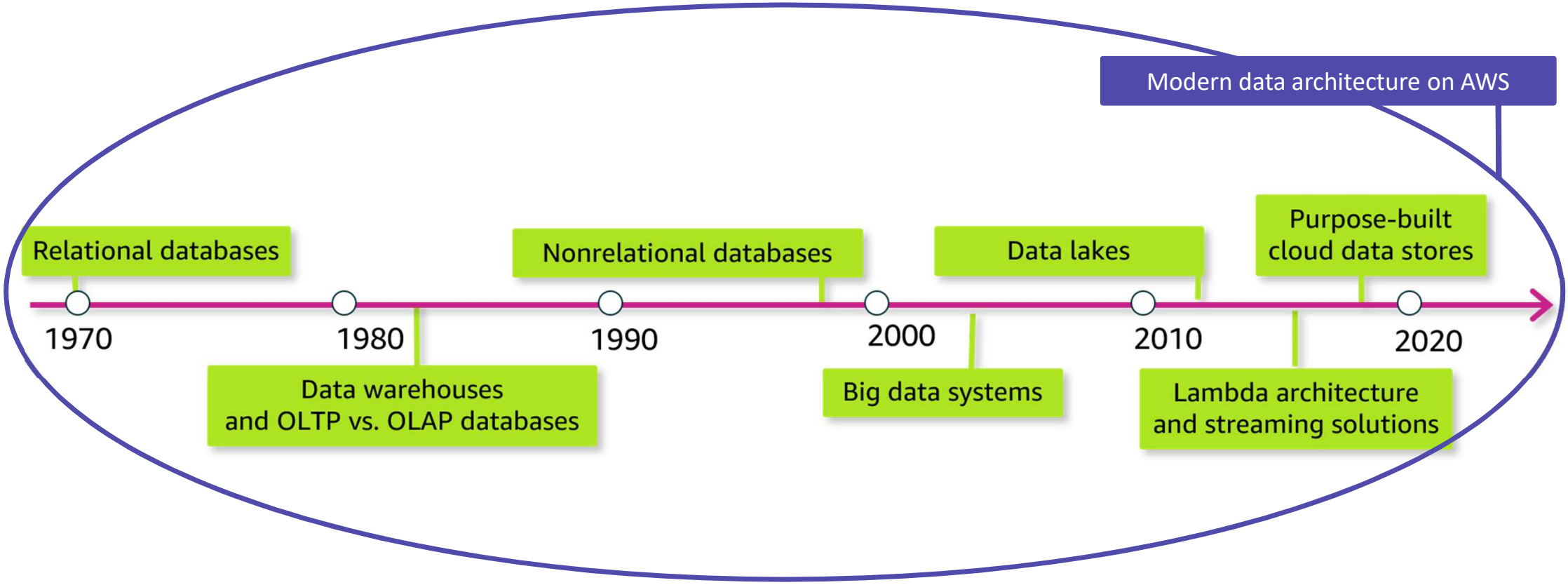
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data architectures evolved to handle volume and velocity



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Modern data architectures unify distributed solutions



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key takeaways: The evolution of data architectures

- Data stores and architectures evolved to adapt to increasing demands of data volume, variety, and velocity.
- Modern data architectures continue to use different types of data stores to suit different use cases.
- The goal of the modern architecture is to unify disparate sources to maintain a single source of truth.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Modern data architecture on AWS

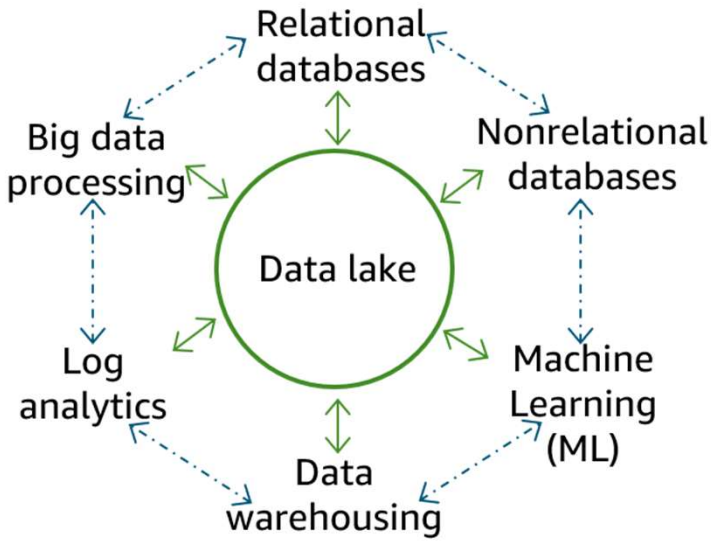
Design Principles and Patterns for Data Pipelines

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Modern data architecture

Key design considerations

- Scalable data lake
- Performant and cost-effective components
- Seamless data movement
- Unified governance

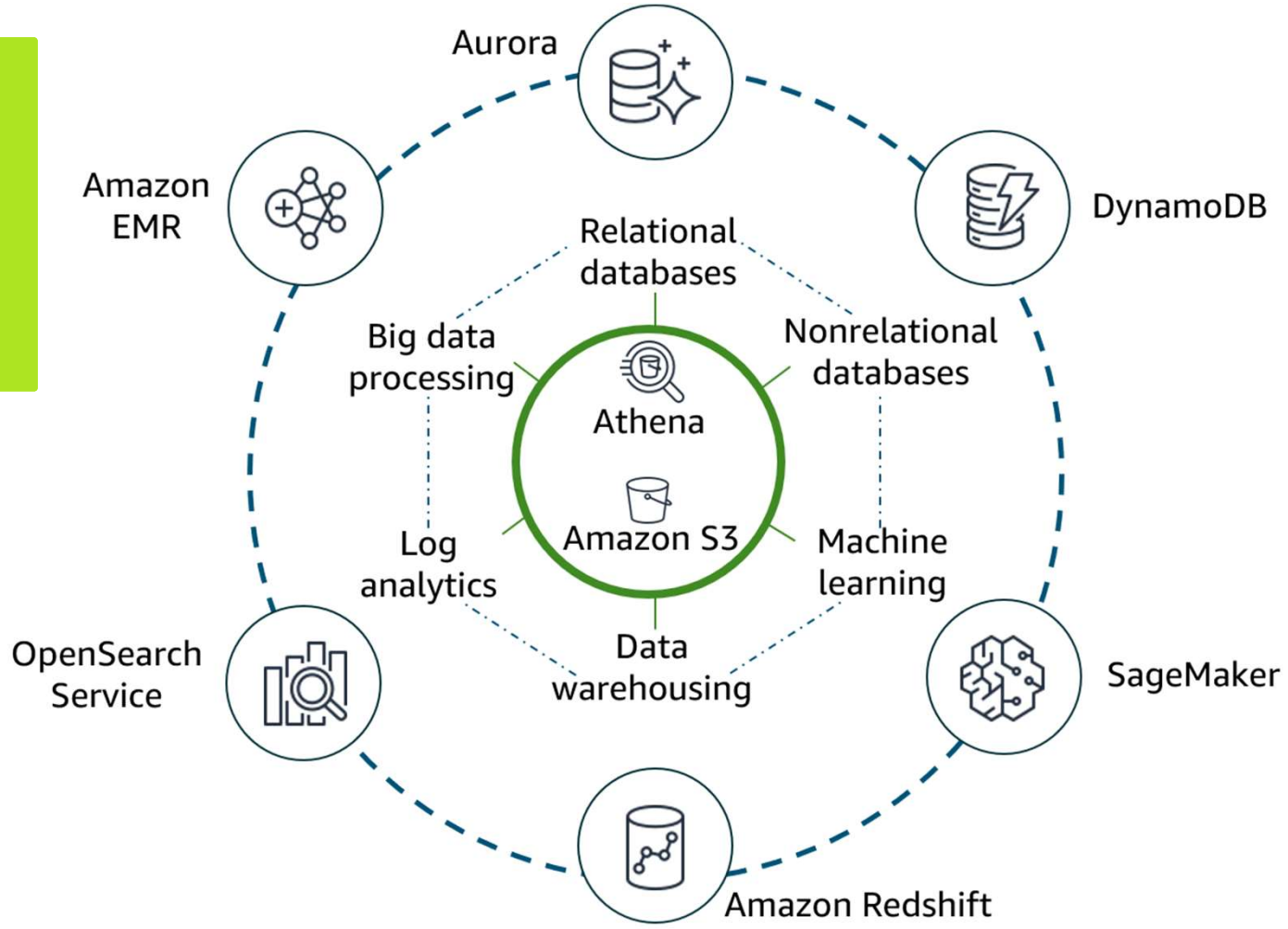


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS purpose-built data stores and analytics tools

Key design considerations

- Scalable data lake
- Performant and cost-effective components

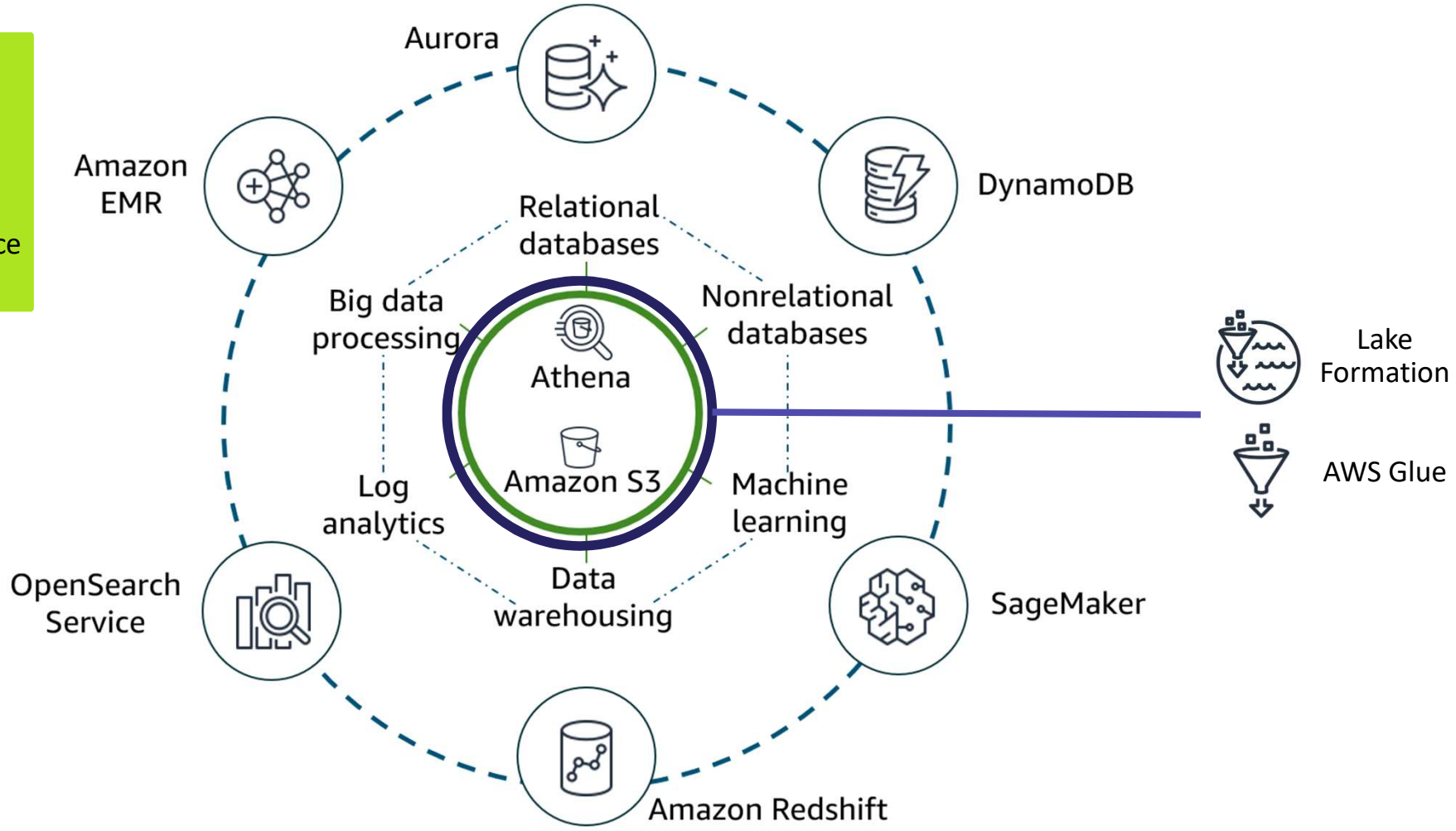


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS services to manage data movement and governance

Key design considerations

- Seamless data movement
- Unified governance



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

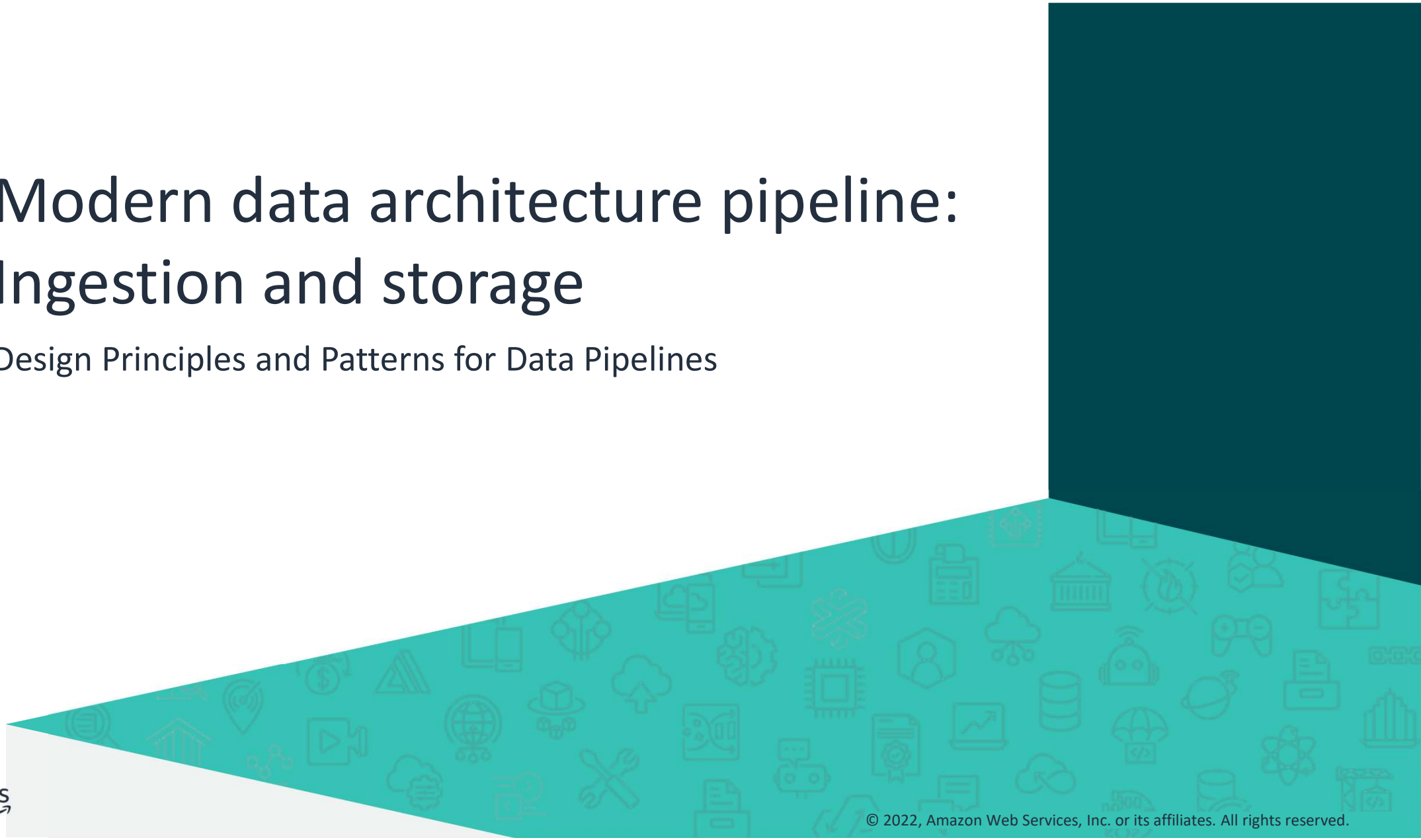
Key takeaways: Modern data architecture on AWS

- A centralized data lake provides data that can be available to all consumers.
- Purpose-built data stores and processing tools integrate with the lake to read and write data.
- The architecture supports three types of data movement: outside in, inside out, and around the perimeter.
- AWS services that are key to seamless access to a centralized lake include Amazon S3, Lake Formation, and AWS Glue.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

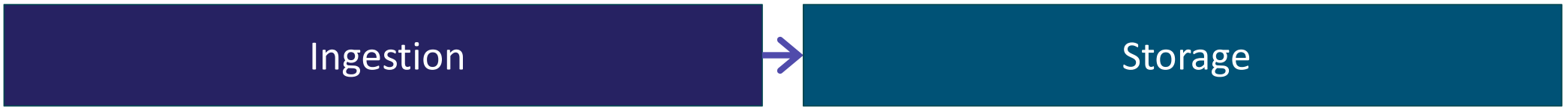
Modern data architecture pipeline: Ingestion and storage

Design Principles and Patterns for Data Pipelines



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Ingestion and storage layers in the reference architecture



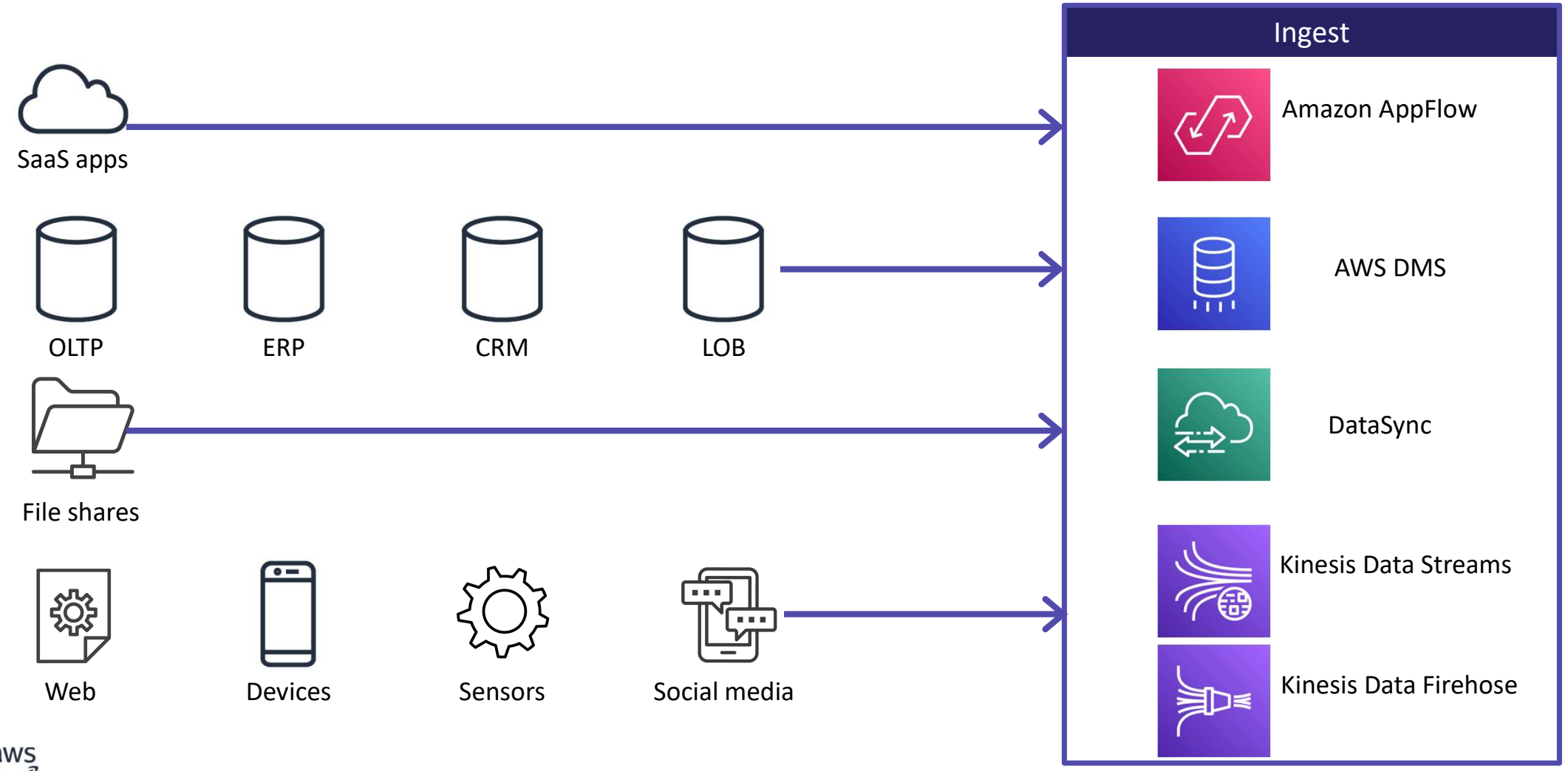
- Matches AWS services to data source characteristics
- Integrates with storage

- Provides durable, scalable storage
- Includes a metadata catalog for governance and discoverability of data

aws

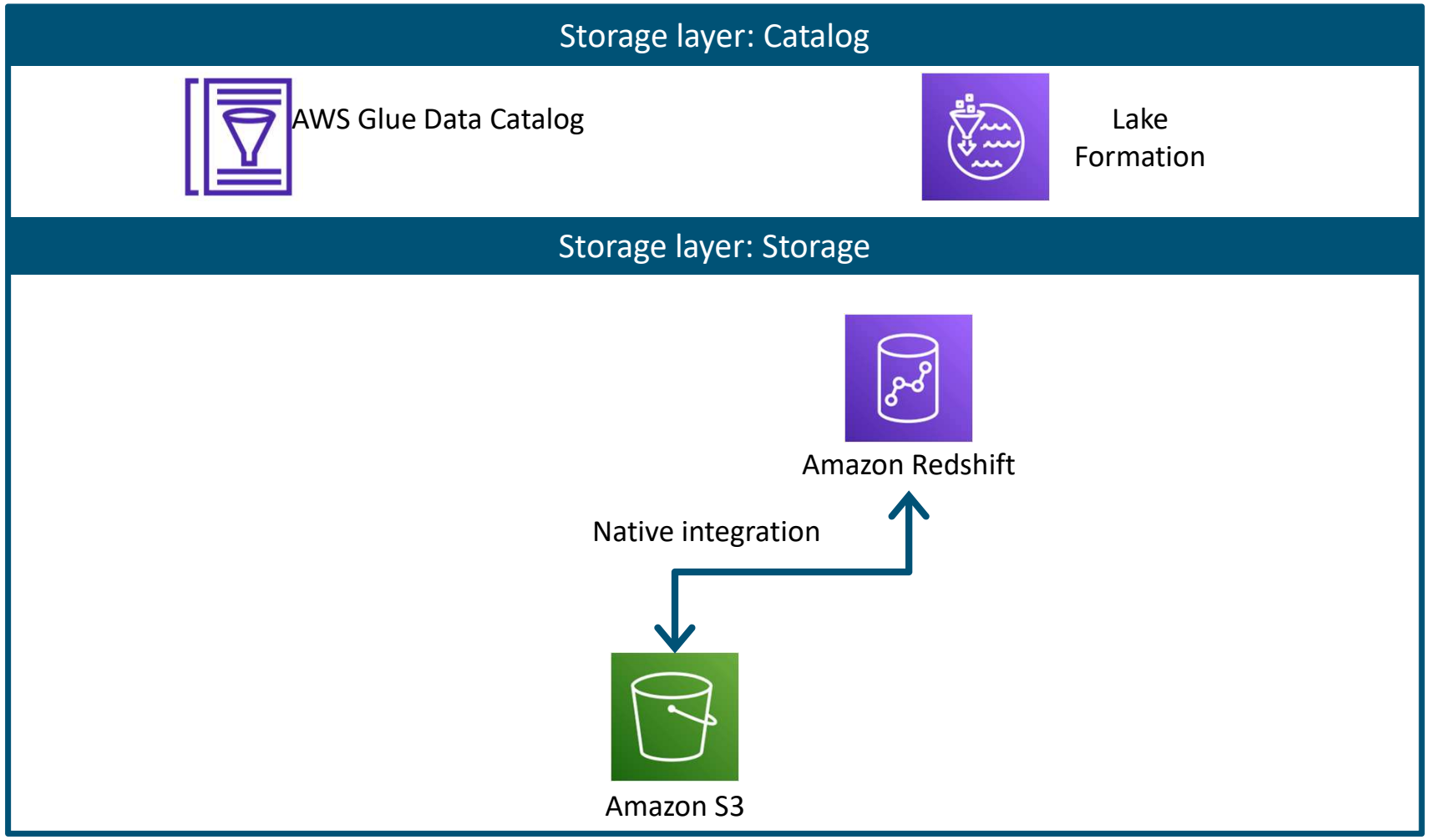
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Matching ingestion services to variety, volume, and velocity



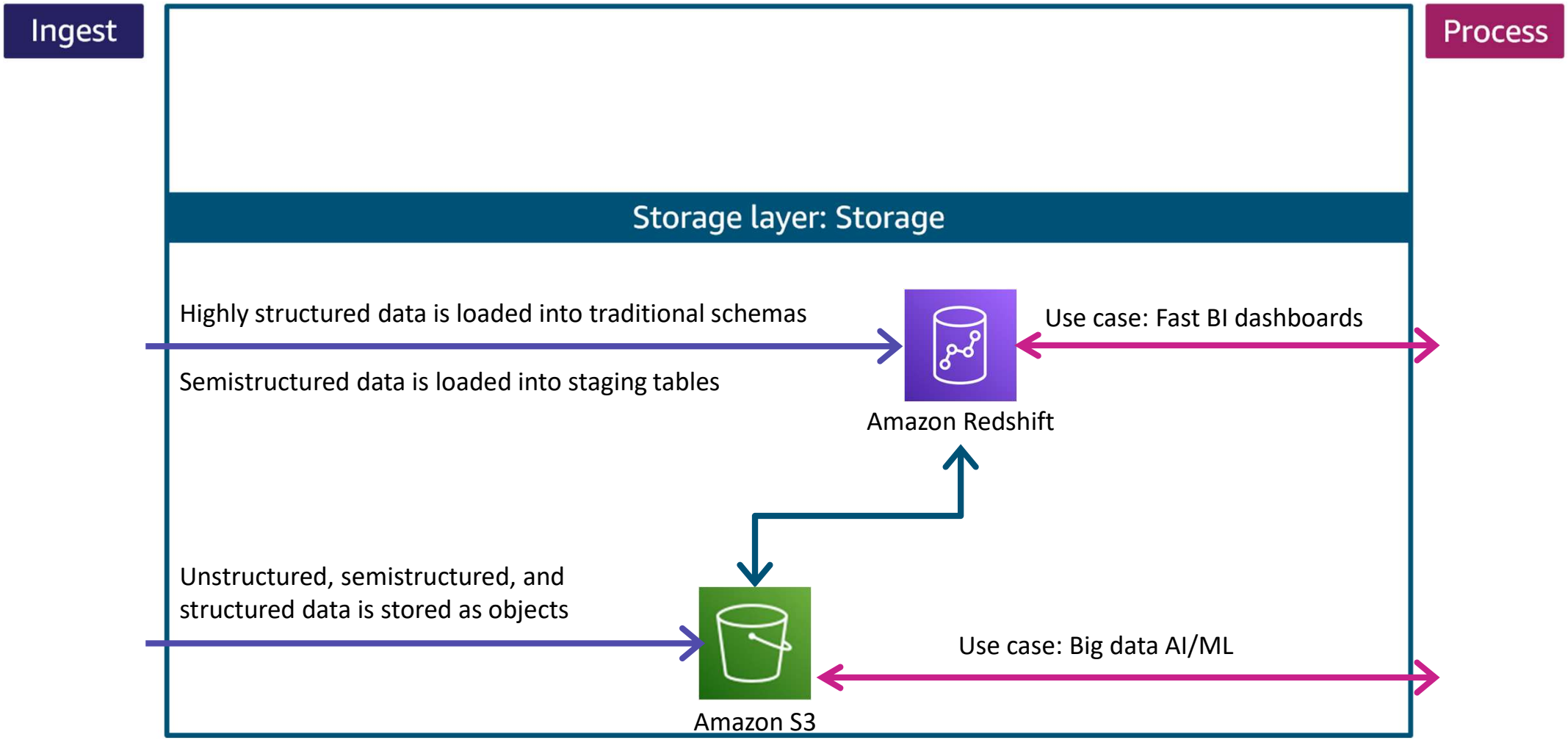
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Modern data architecture storage layer



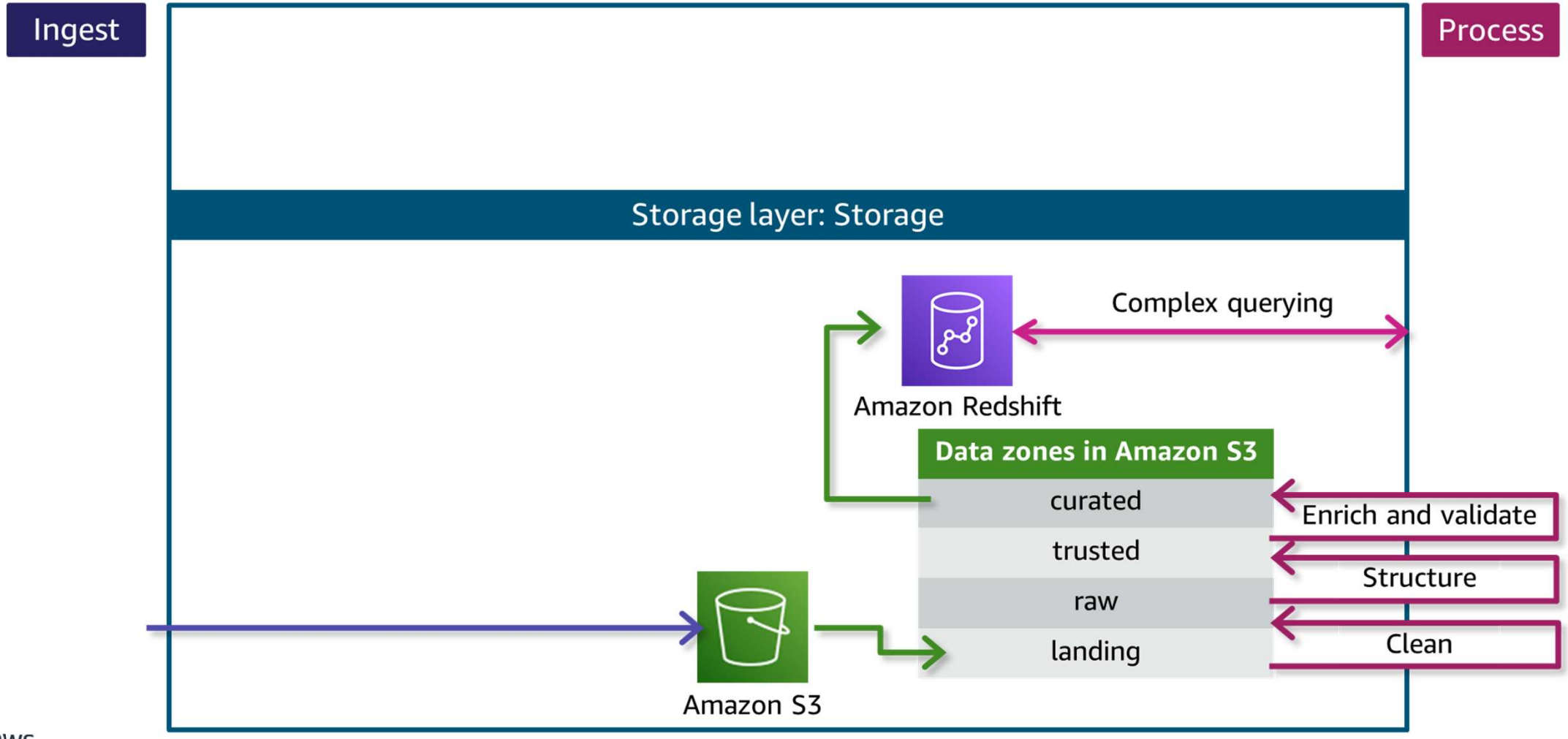
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Storage for variety, volume, and velocity



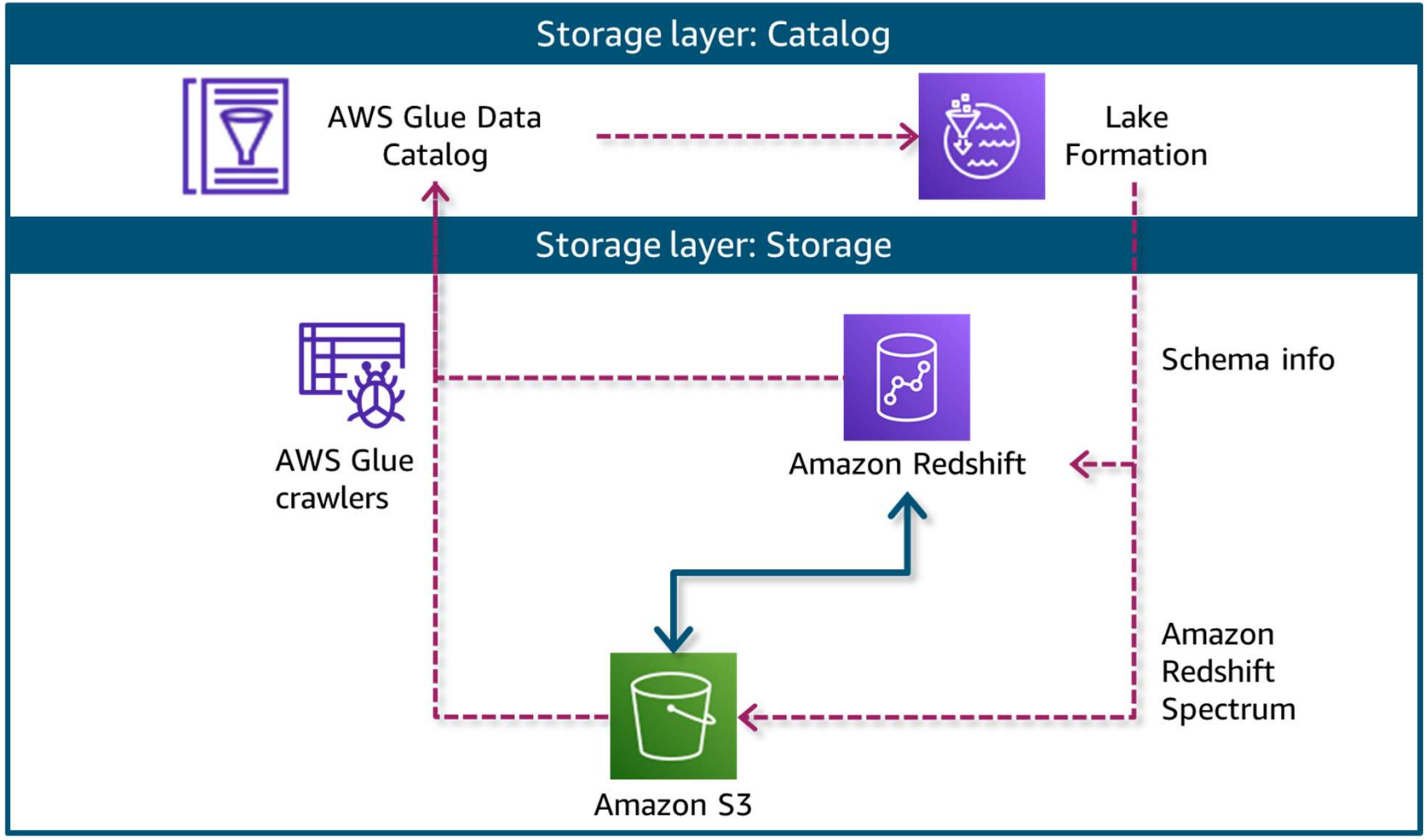
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Storage zones for data in different states



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Catalog layer for governance and discoverability



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

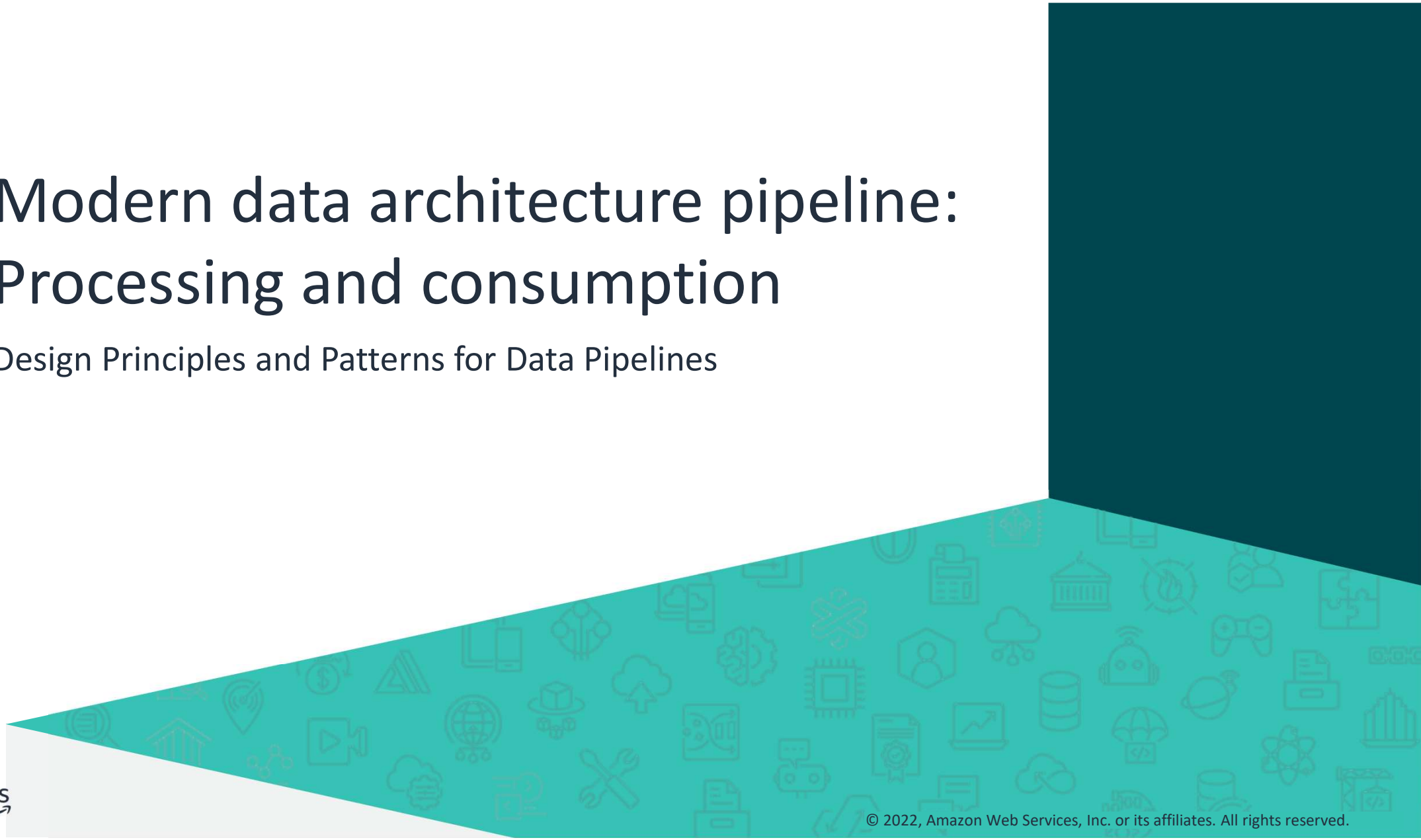
Key takeaways: Modern data architecture pipeline: Ingestion and storage

- The AWS modern data architecture uses purpose-built tools to ingest data based on characteristics of the data.
- The storage layer uses Amazon Redshift as its data warehouse and Amazon S3 for its data lake.
- The Amazon S3 data lake uses prefixes or individual buckets as zones to organize data in different states, from landing to curated.
- AWS Glue and Lake Formation are used in a catalog layer to store metadata.
- With the catalog, Amazon Redshift Spectrum can query data in Amazon S3 directly.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

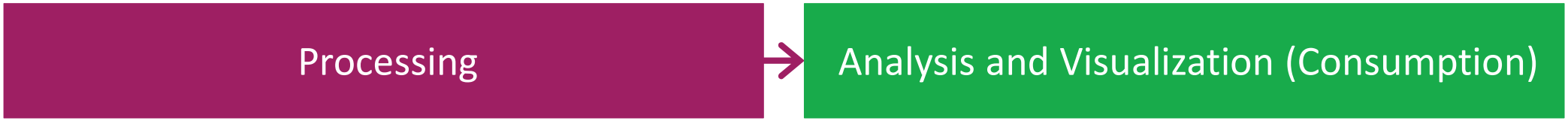
Modern data architecture pipeline: Processing and consumption

Design Principles and Patterns for Data Pipelines



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Processing and consumption layers in the reference architecture



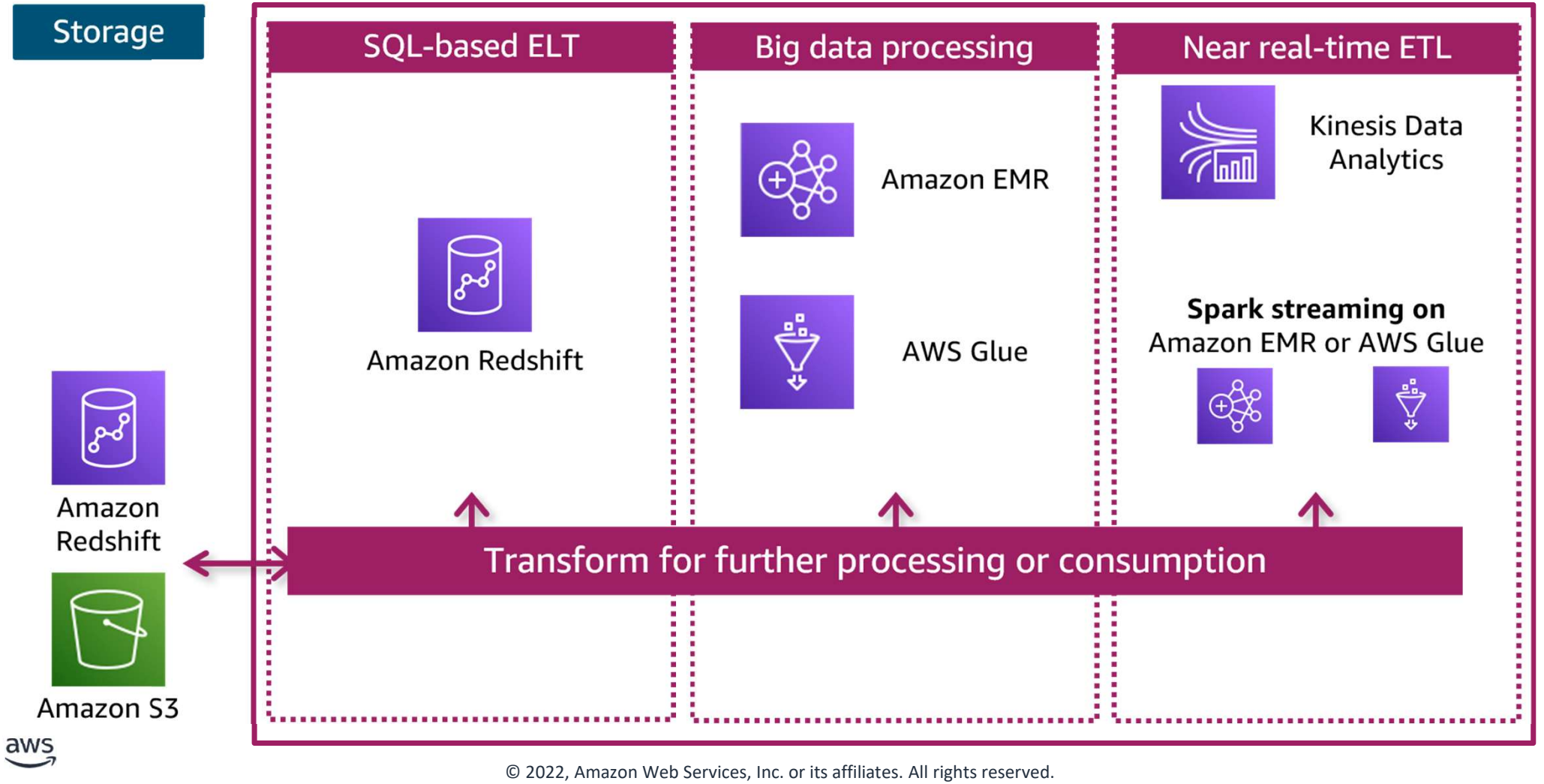
- Transforms data into a consumable state
- Uses purpose-built components

- Democratizes consumption across the organization
- Provides unified access to stored data and metadata

aws

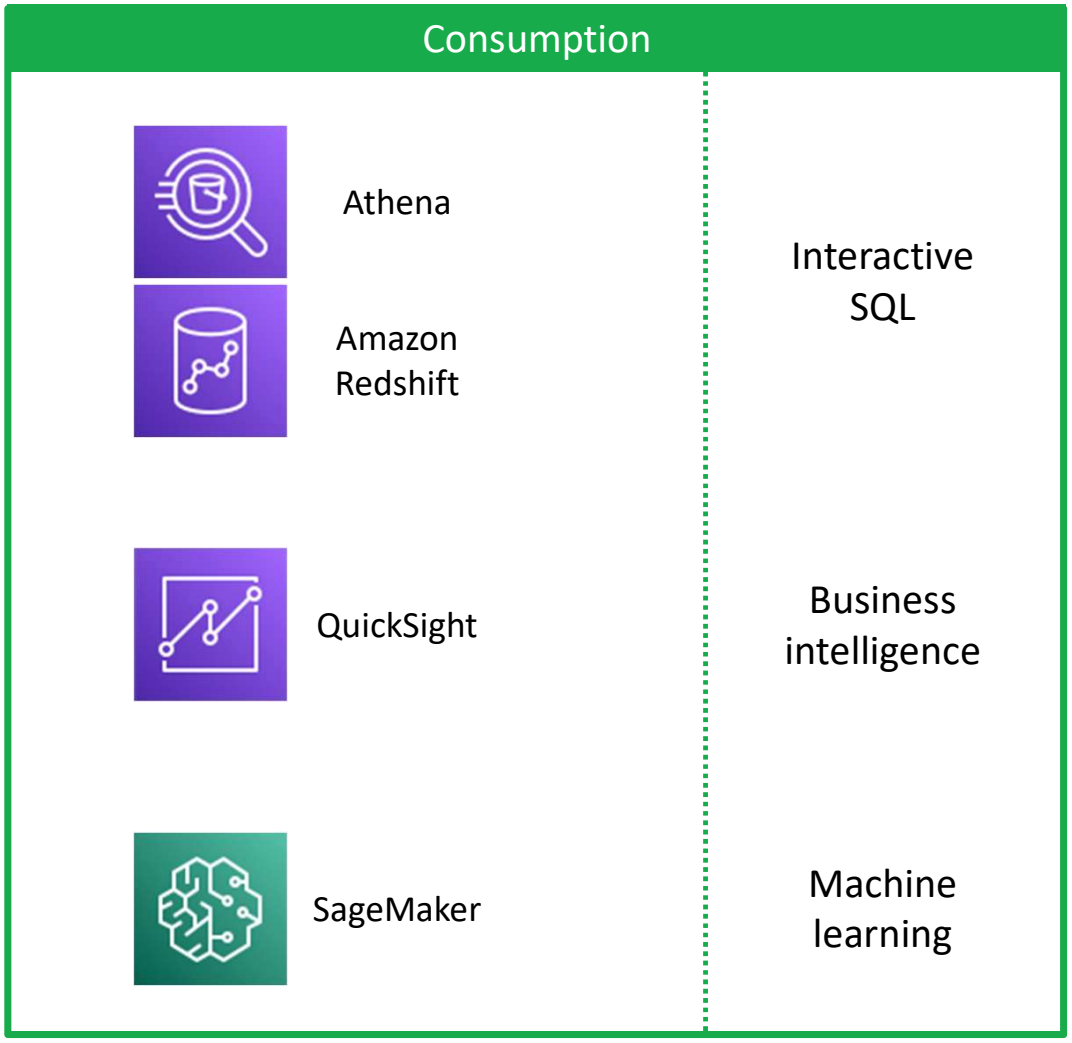
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Modern architecture pipeline: Processing



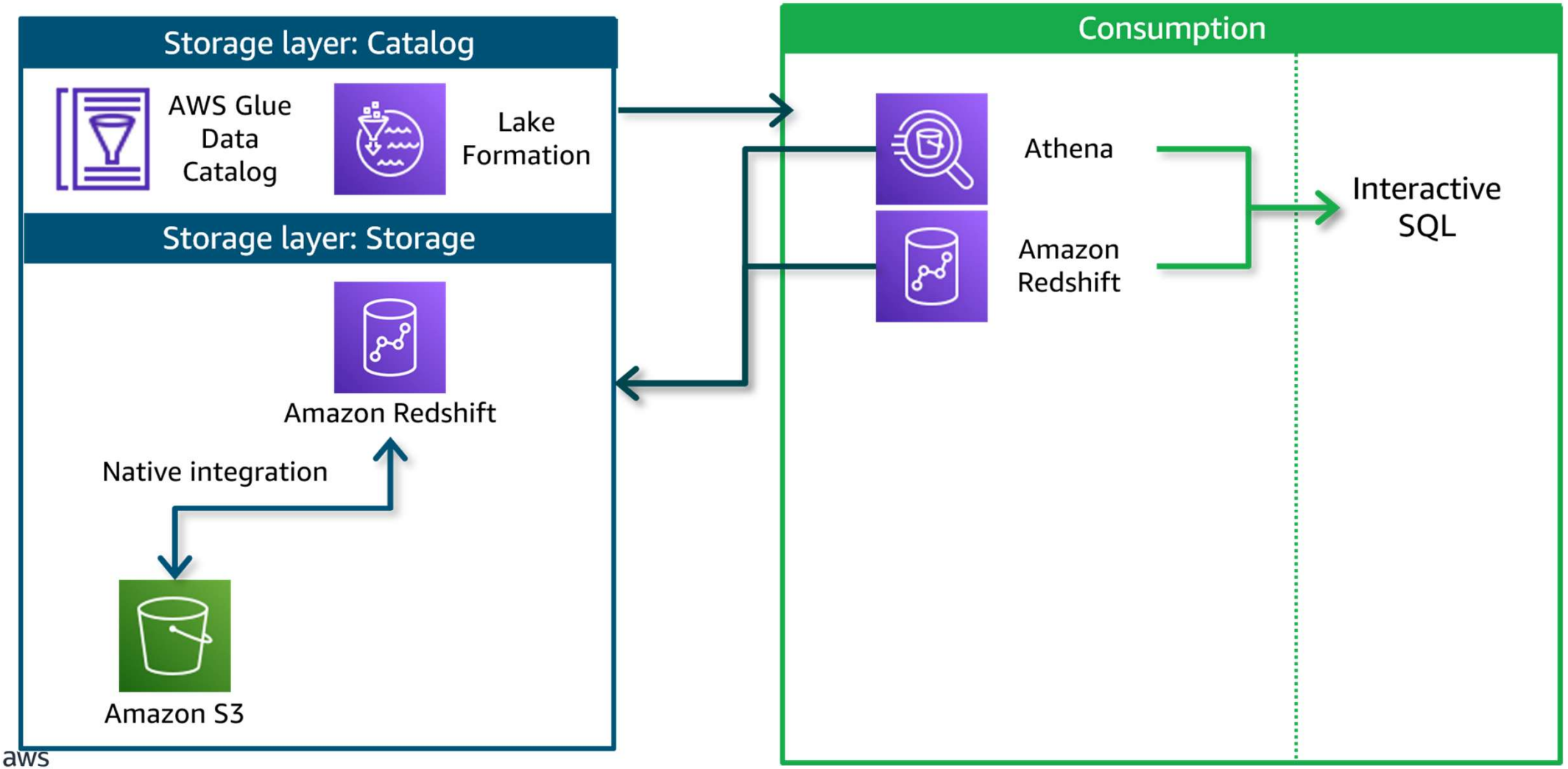
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Modern data architecture: Consumption layer



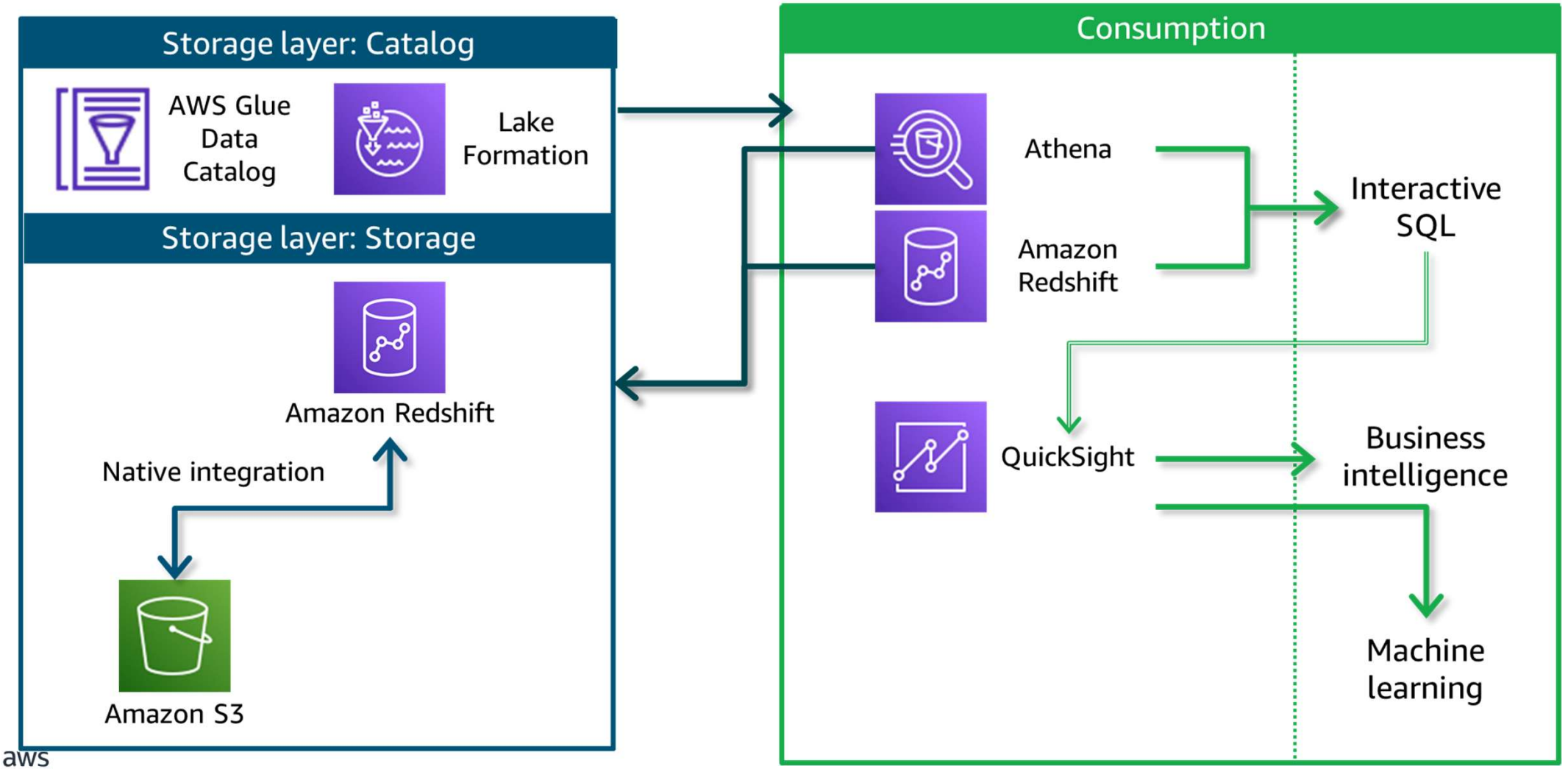
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Consuming data by using interactive SQL



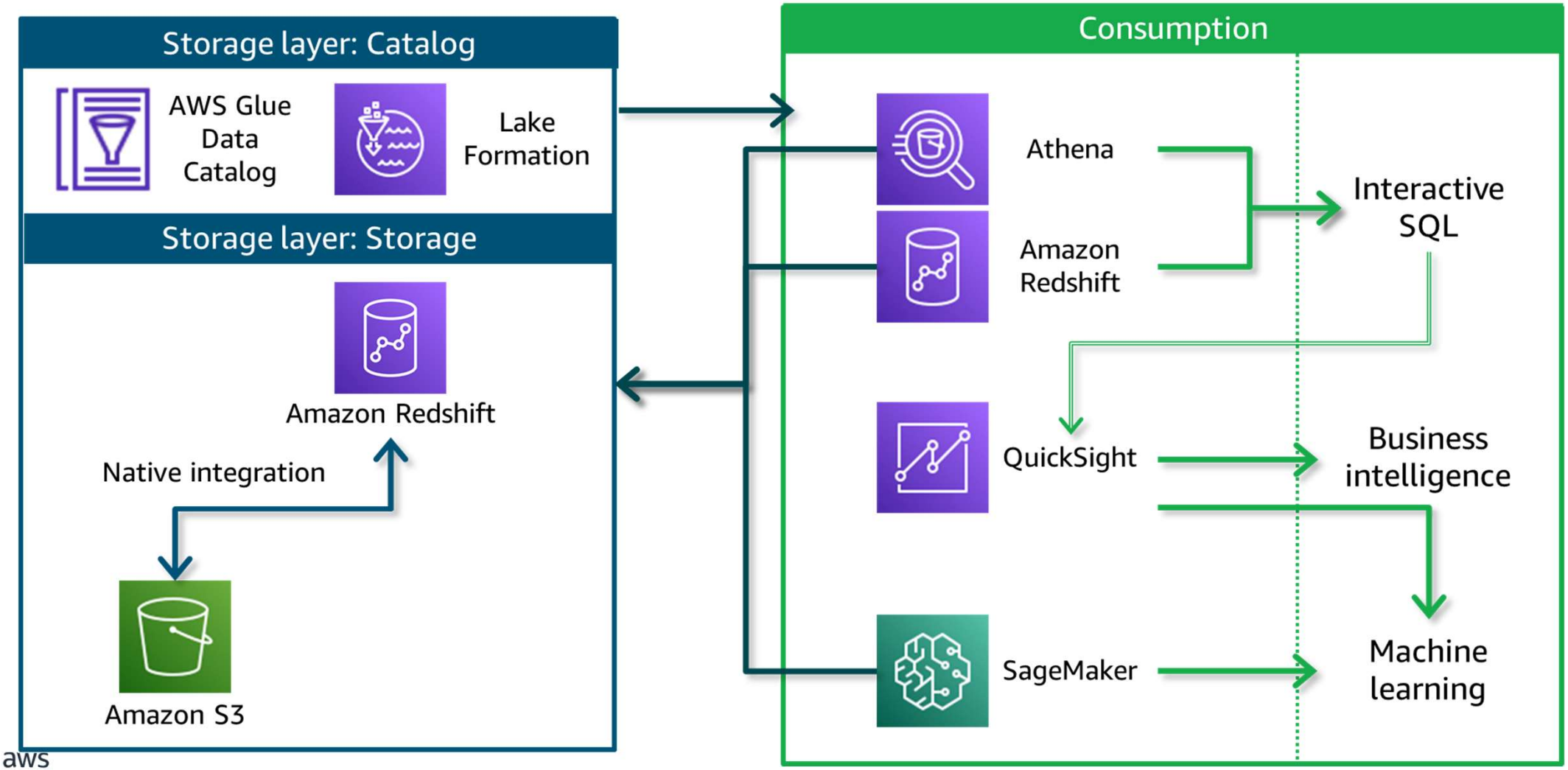
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Consuming data for business intelligence



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Consuming data for ML



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

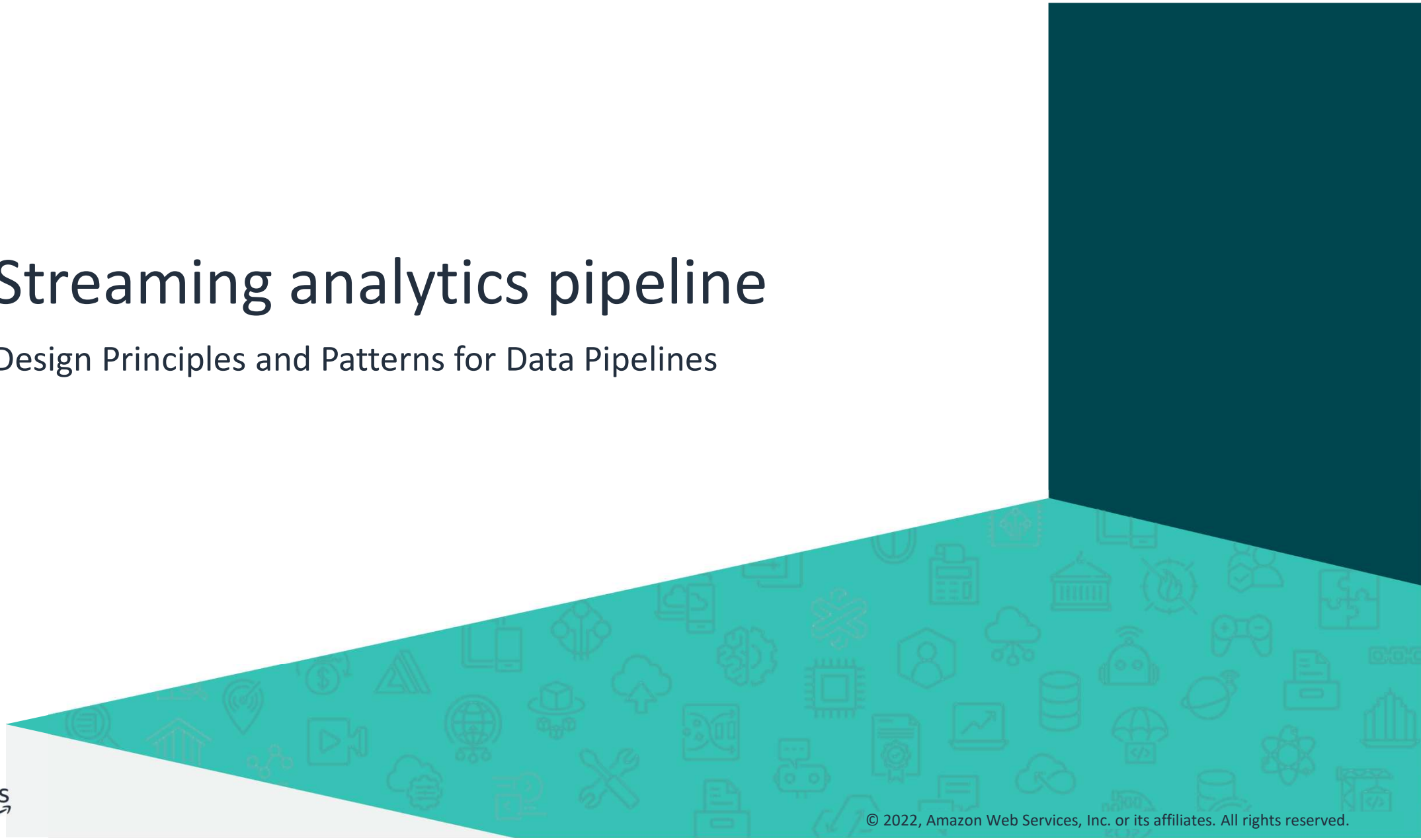
Key takeaways: Modern data architecture pipeline: Processing and consumption

- Components in the processing layer are responsible to transform data into a consumable state.
- The processing layer supports three types of processing: SQL-based ELT, big data processing, and near real-time ETL.
- The consumption layer provides unified interfaces to access all the data and metadata in the storage layer.
- The consumption layer supports three analysis methods: interactive SQL queries, BI dashboards, and ML.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

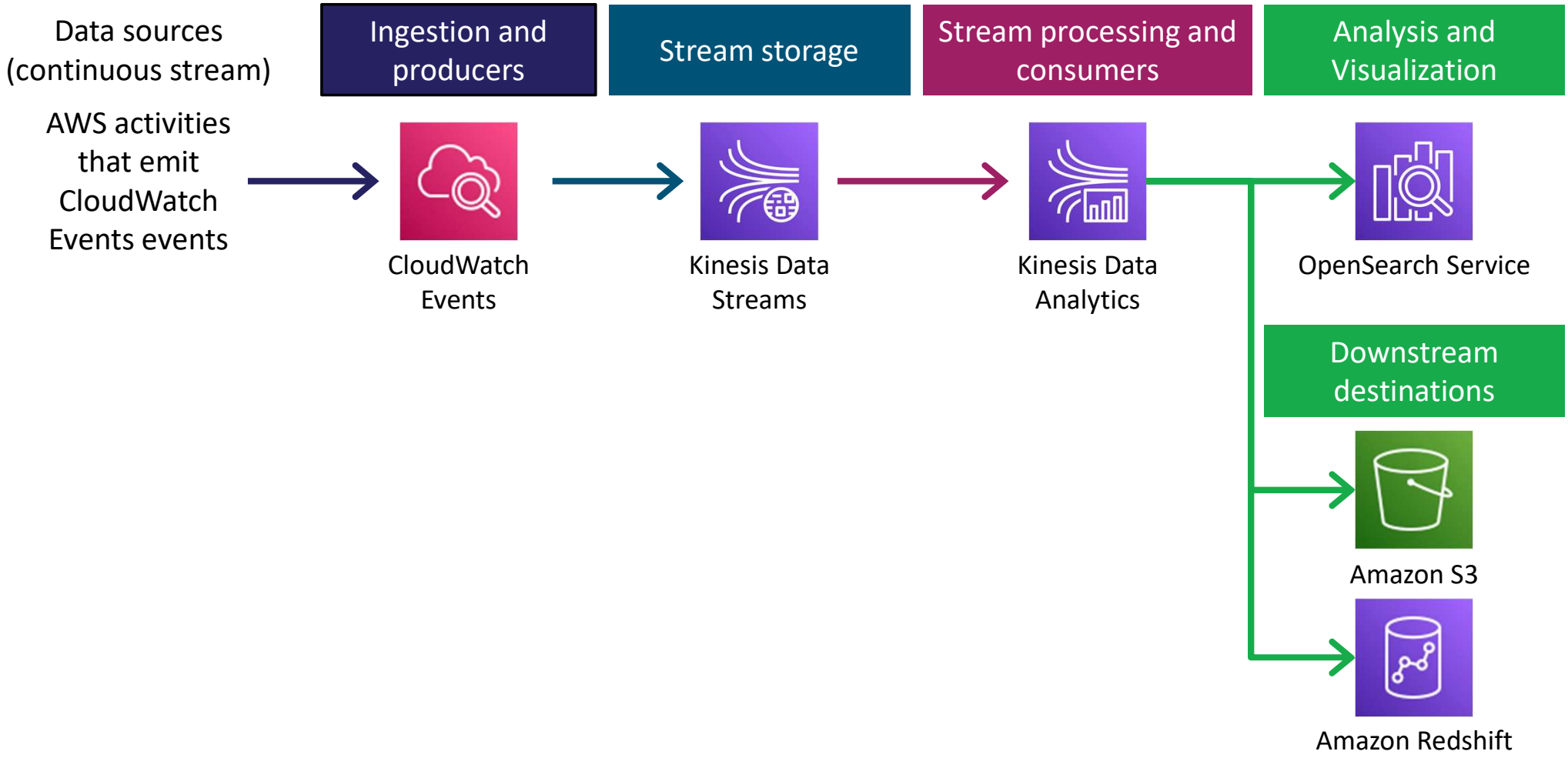
Streaming analytics pipeline

Design Principles and Patterns for Data Pipelines



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Example architecture: Stream processing pipeline



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key takeaways: Streaming analytics pipeline

- Streaming analytics includes producers and consumers.
- A stream provides temporary storage to process incoming data in real time.
- The results of streaming analytics might also be saved to downstream destinations.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Lab: Querying Data by Using Athena

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Lab introduction: Querying Data by Using Athena

- In this lab, you will learn how to use Athena and AWS Glue to query data that is stored in Amazon S3.
- You will optimize Athena queries and create views to simplify use by multiple team members.
- You will use AWS CloudFormation to integrate Athena into your infrastructure and AWS Identity and Access Management (IAM) to provide the appropriate level of access to your users.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Debrief: Querying Data by Using Athena

- How could you use Athena and AWS Glue in other businesses with analytics workloads?
- What security considerations do you need to address before you implement solutions like those in the lab with a team?
- How could you address concepts such as least privilege and data governance with these types of solutions?
- What is one way that a data analyst could run a named query that you created and provided to the team?

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Module wrap-up

Design Principles and Patterns for Data Pipelines

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Module summary

This module prepared you to do the following:

- Use the AWS Well-Architected Framework to inform the design of analytics workloads.
- Recount key milestones in the evolution of data stores and data architectures.
- Describe the components of modern data architectures on AWS.
- Cite AWS design considerations and key services for a streaming analytics pipeline.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Module knowledge check

- The knowledge check is delivered online within your course.
- The knowledge check includes 10 questions based on material presented on the slides and in the slide notes.
- You can retake the knowledge check as many times as you like.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Sample exam question

A data engineer has implemented the AWS modern data architecture as described in the Well-Architected Framework. An analyst wants to combine and explore customer sales data from the data warehouse with customer support ticket data for the last 6 months. The support ticket data is available as a JSON extract from their SaaS support system (Zendesk).

How could the engineer meet this need and simplify the effort for the engineer and the analyst? (Select TWO.)

Identify the key words and phrases before continuing.

The following are the key words and phrases:

- **AWS modern data architecture**
- **Combine and explore** customer sales in the **data warehouse** with **SaaS data** available as a **JSON extract**
- **Data for the last 6 months**
- **Simplify** the effort

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Sample exam question: Response choices

A data engineer has implemented the **AWS modern data architecture** as described in the Well-Architected Framework. An analyst wants to **combine and explore customer sales data** from the **data warehouse** with customer support ticket data for the **last 6 months**. The support ticket data is available as a **JSON extract** from their **SaaS support system** (Zendesk).

How could the engineer meet this need and **simplify** the effort for the engineer and the analyst? (Select TWO.)

Choice	Response
A	Use Amazon AppFlow to ingest the SaaS data into Amazon S3.
B	Use Kinesis Data Firehose to stream the SaaS data into Amazon S3.
C	Use OpenSearch Service to index the customer ticket data so that the analyst can search the data.
D	Create an ETL process that transforms the data that is ingested into Amazon S3 to a curated format. Then, load the data into the data warehouse.
E	Use Amazon Redshift Spectrum to write a query that includes data from the data lake and the data warehouse.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Sample exam question: Answer

The correct answers are A and E.

Choice	Response
A	Use Amazon AppFlow to ingest the SaaS data into Amazon S3.
B	
C	
D	
E	Use Amazon Redshift Spectrum to write a query that includes data from the data lake and the data warehouse.

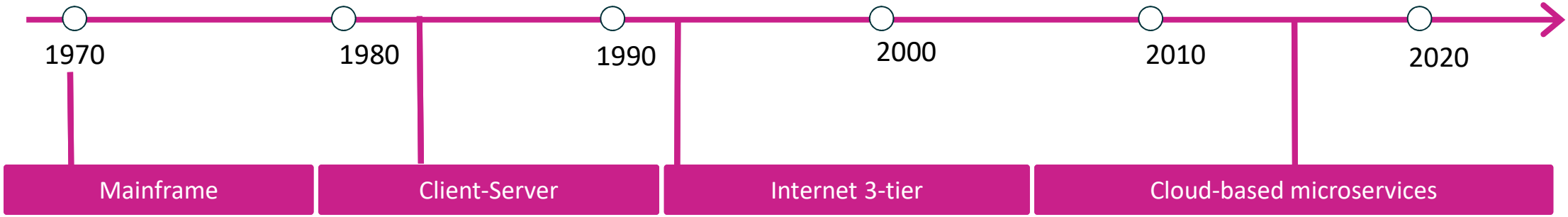
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you

Corrections, feedback, or other questions?
Contact us at https://support.aws.amazon.com/#/contacts/aws-academy.
All trademarks are the property of their owners.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Application architecture evolved into more distributed systems

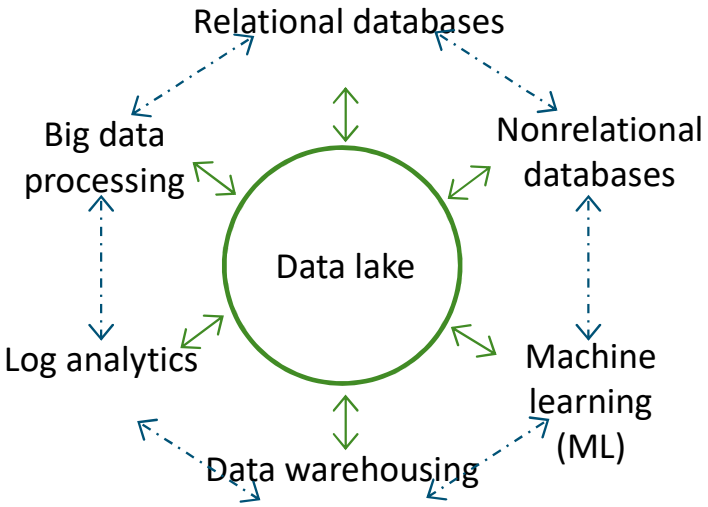


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Modern data architecture

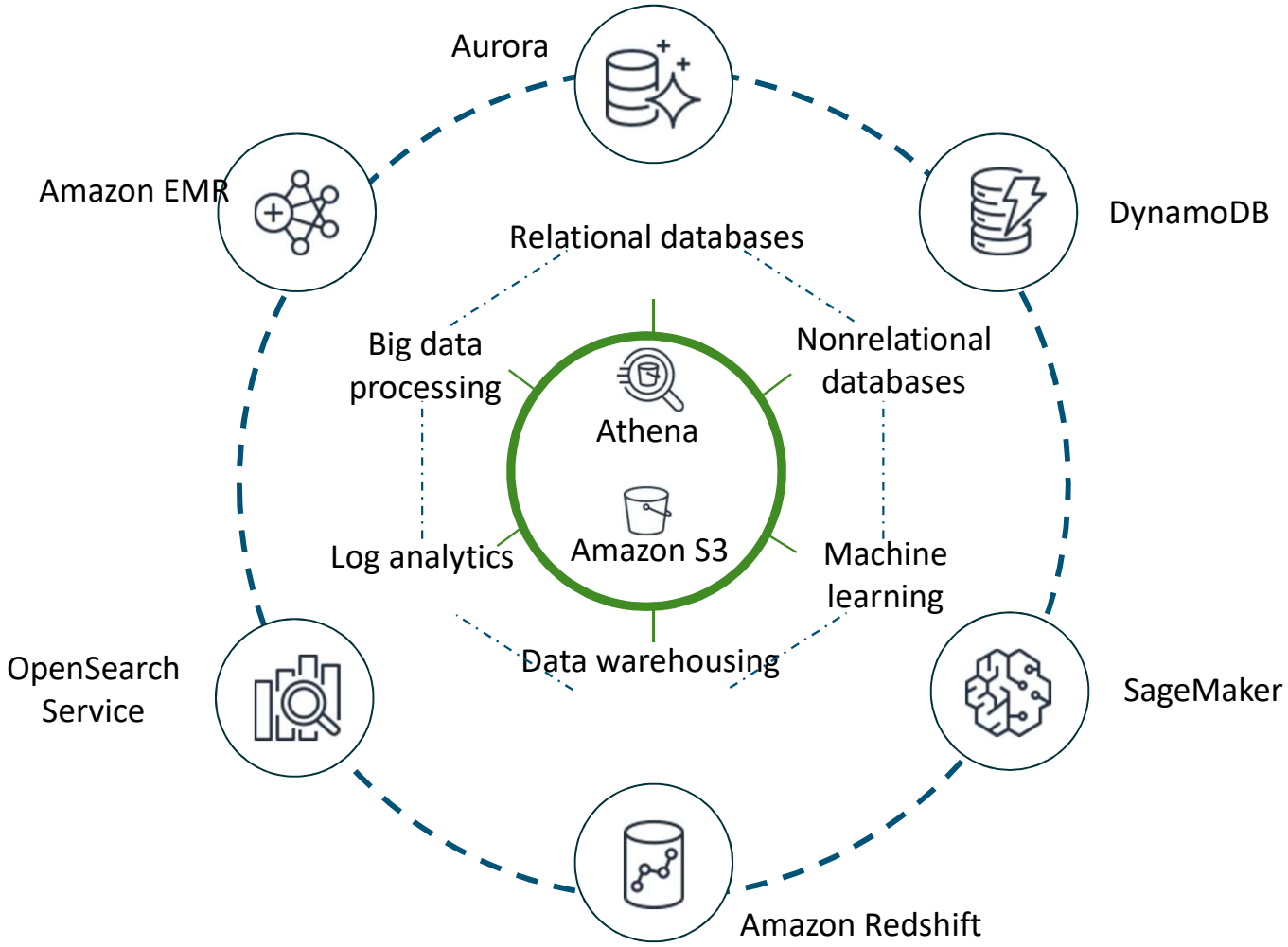
Key design considerations

- Scalable data lake
- Performant and cost-effective components
- Seamless data movement
- Unified governance



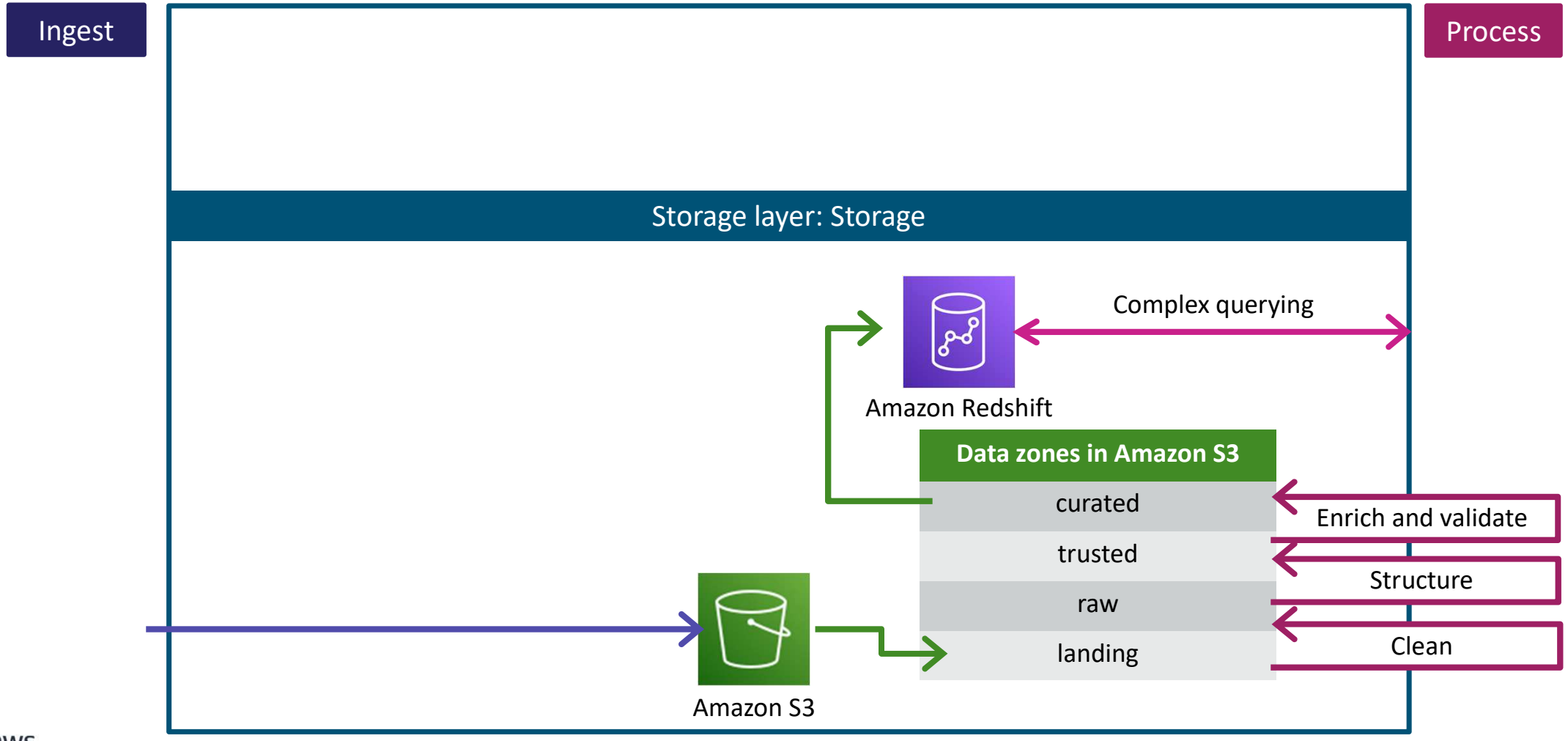
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Modern data architecture



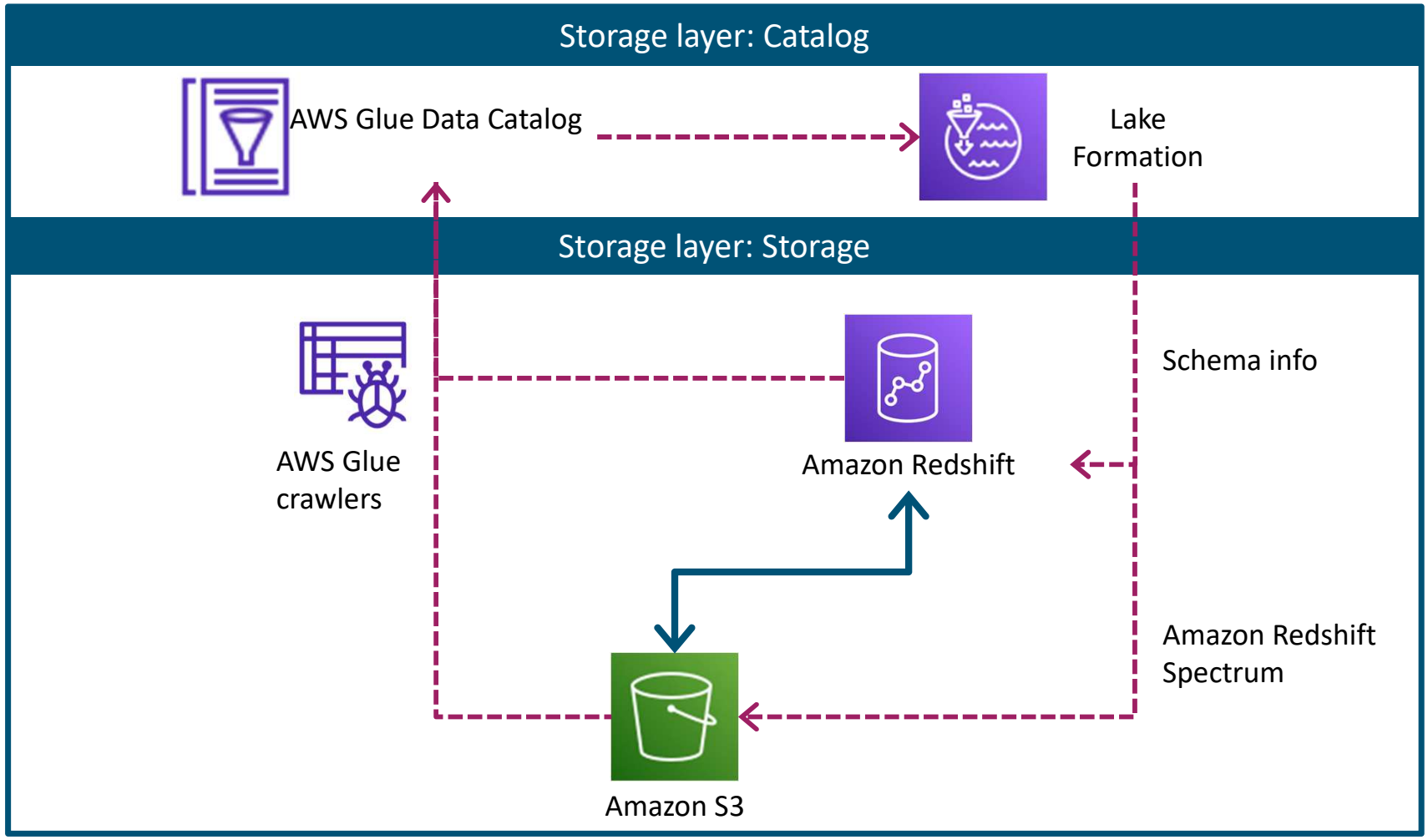
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Storage zones for data in different states



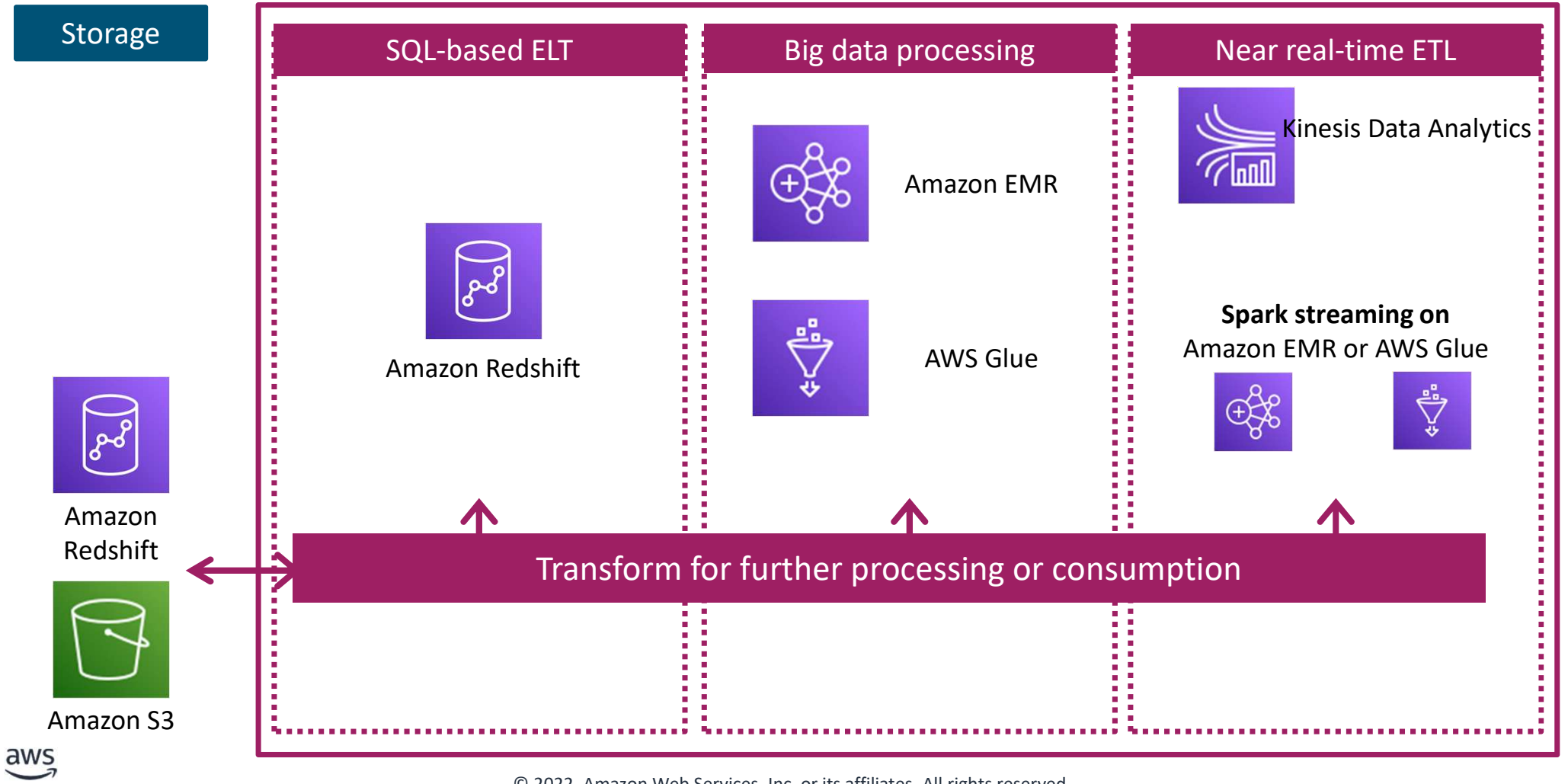
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Catalog layer for governance and discoverability



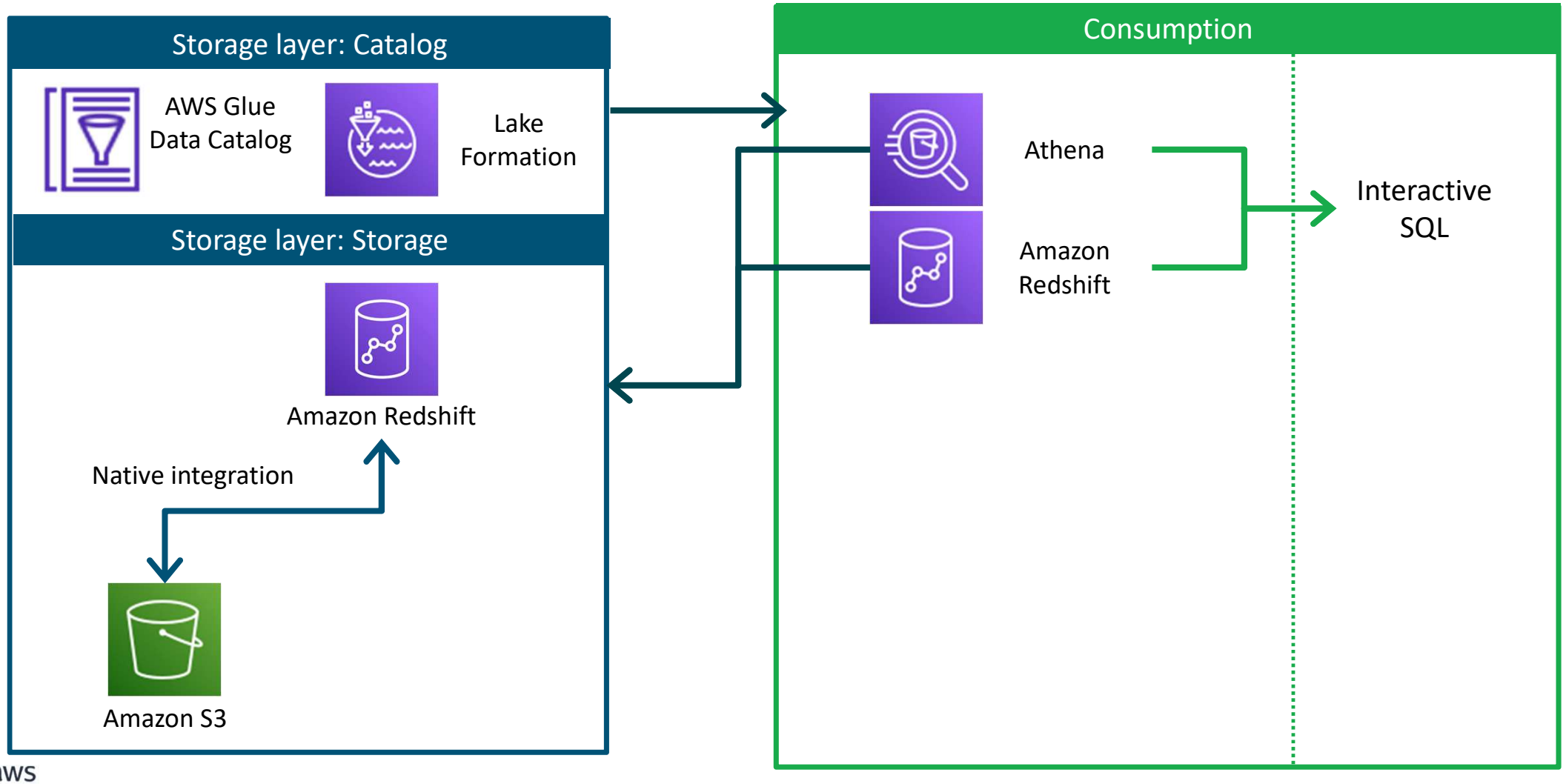
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Modern architecture pipeline: Processing



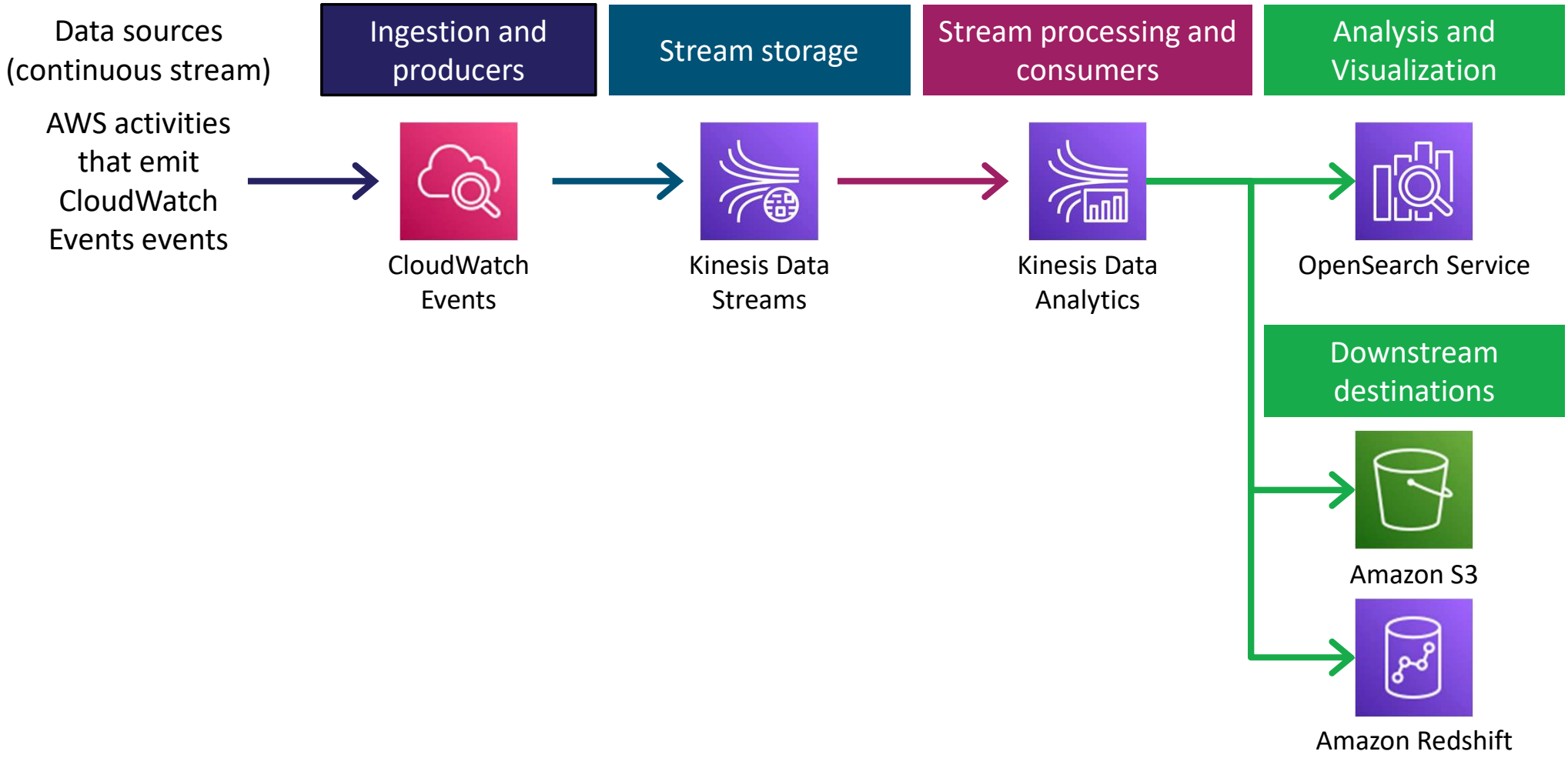
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Consuming data



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Example architecture: Stream processing pipeline



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.