Track 6 | Session 1

進入 AI 領域的第一步驟 - 資料平台的建置

Jayson Hsieh

Senior Solutions Architect
Amazon Web Services

aws SUMMIT

Data is a strategic asset for every organization

> “The world’s most valuable resource is no longer oil, but data.”

David Parkins, 2017, The Economist

Amazon.com lowers costs and gains faster insights with AWS data analytic offerings

amazon.com

Challenge

Amazon needed to analyze a massive amount of data to find insights, identify opportunities, and evaluate business performance.

Including catalog browsing, order placement, transaction processing, delivery scheduling, video services, and Prime registration

- 50 petabytes of data and 75,000 tables
- Processing 600,000 user analytics jobs each day
- Data is published by more than 1,800 teams
- 3,300+ data consumer teams analyze this data

The Oracle data warehouse did not scale for PB level data, was difficult to maintain, and was costly.

amazon.com

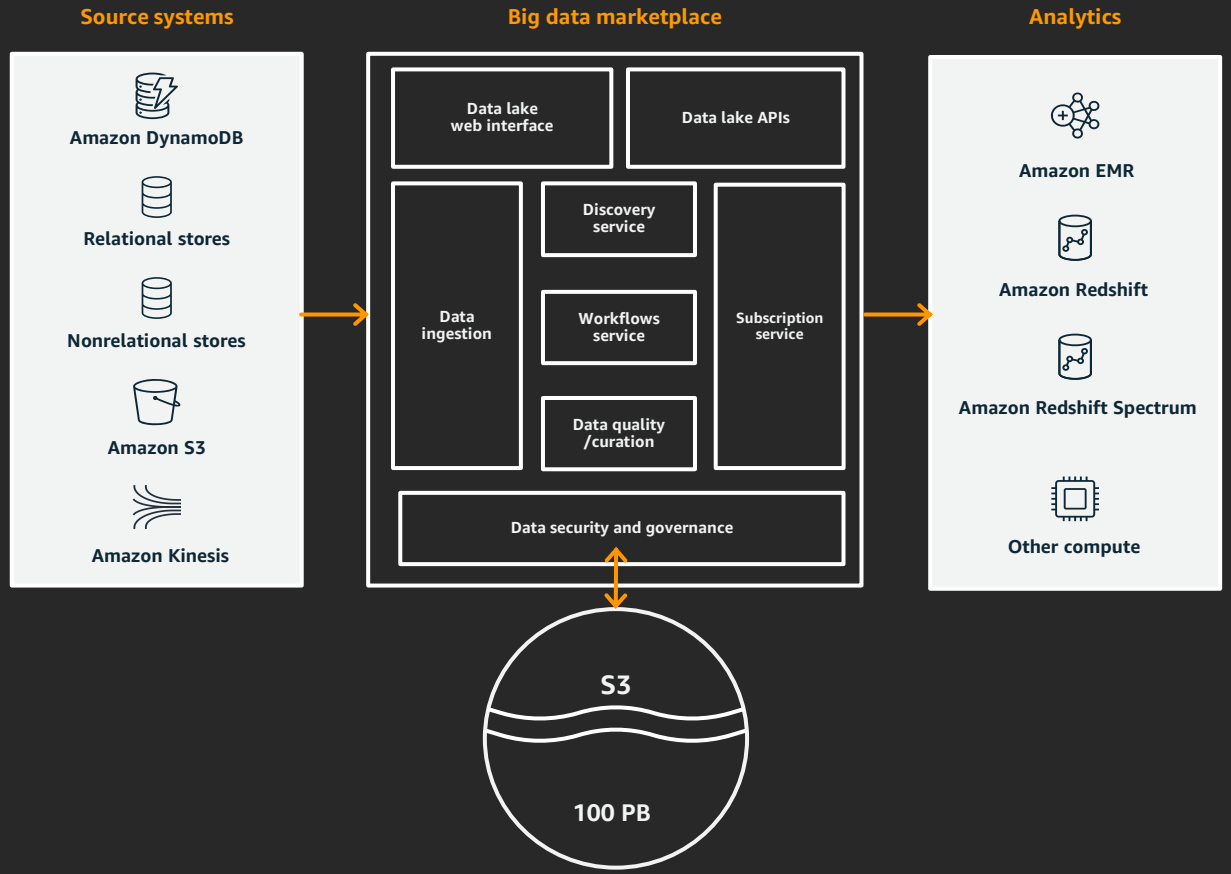
Amazon uses an AWS data lake

Solution

Amazon deployed a data lake with Amazon S3, and it now runs analytics with Amazon Redshift, Amazon Redshift Spectrum, and Amazon EMR.

Benefits

Amazon doubled the data stored from 50 PB to 100 PB, lowered costs, and was able to gain insights faster.



Customers want more value from their data

Growing exponentially

From new sources

Increasingly diverse

Used by many people

Analyzed by many applications

Common analytics use cases – which do you need?

Data warehouse modernization

Big data and data lakes

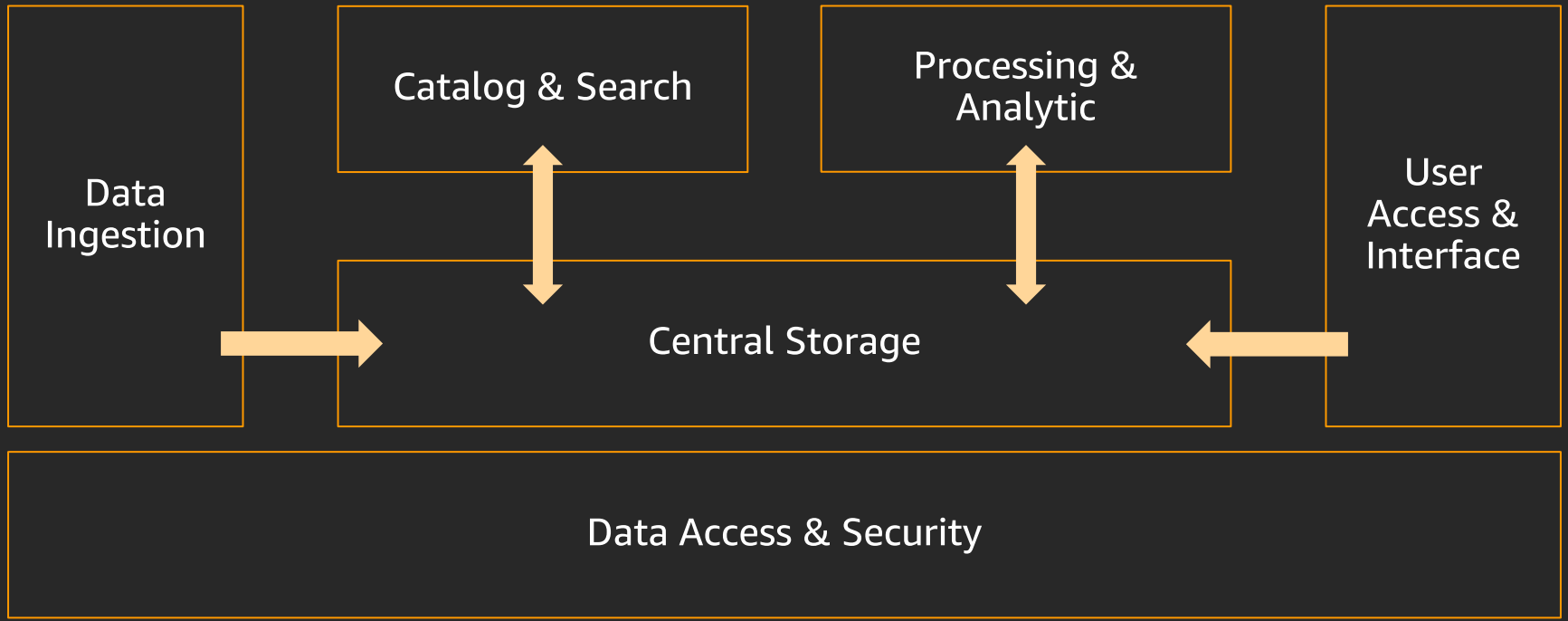
Real-time streaming and analytics

Operational and search analytics

Self-service business analytics

Acquisition of third-party data for analysis

Data Architectures



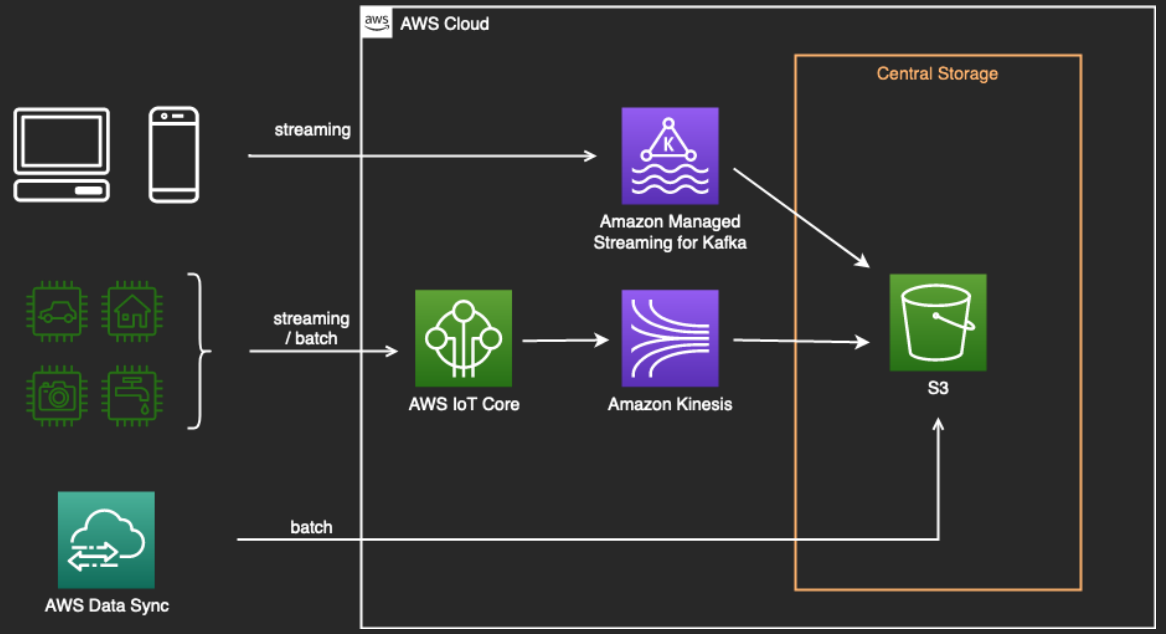
General Data Platform Design Principles

aws SUMMIT

Automate data ingestion

Ingestion of data should be automated using triggers, schedules, and change detection

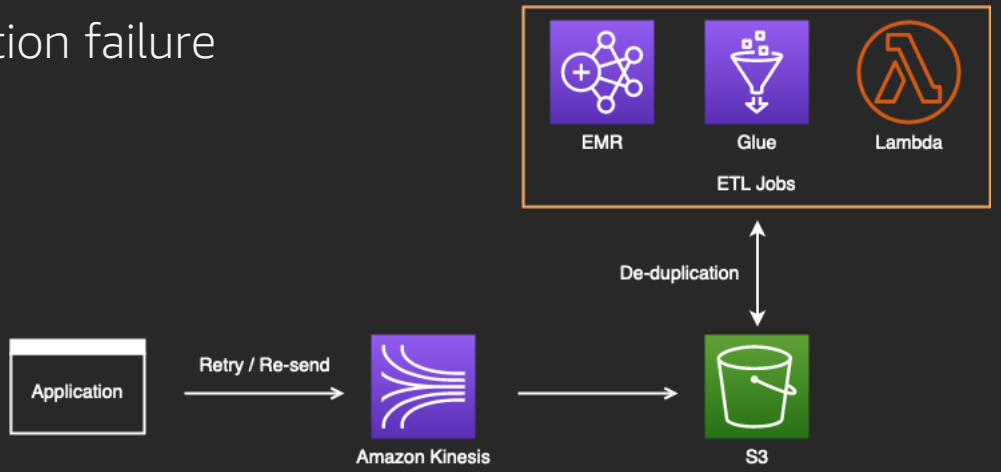
- Eliminates error-prone manual processes
- Allows data to be processed as it arrives



Design ingestion for failures and duplicates

Ingestion triggered from requests and events must be idempotent

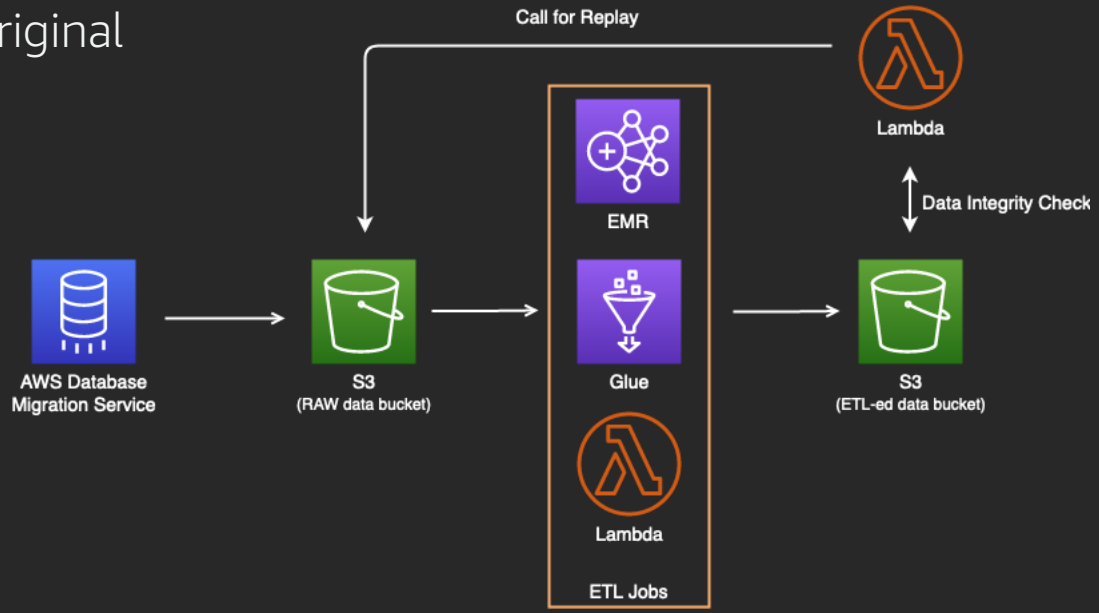
- Appropriate retries
- Deal with message duplication failure



Preserve original source data

Having raw data in its pristine form allows you to repeat the ETL process in case of failures

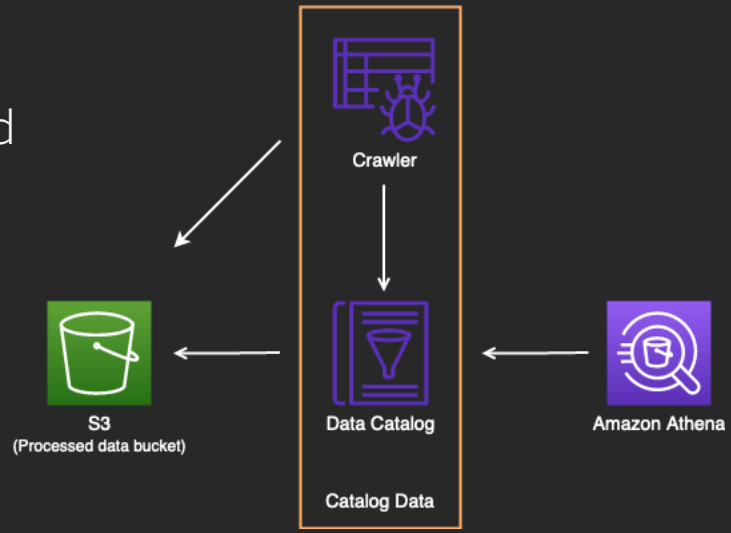
- No transformation of the original data files
- Allow replay data pipeline



Describe data with metadata

It's essential that any dataset that makes its way into a data store environment is discoverable and classified

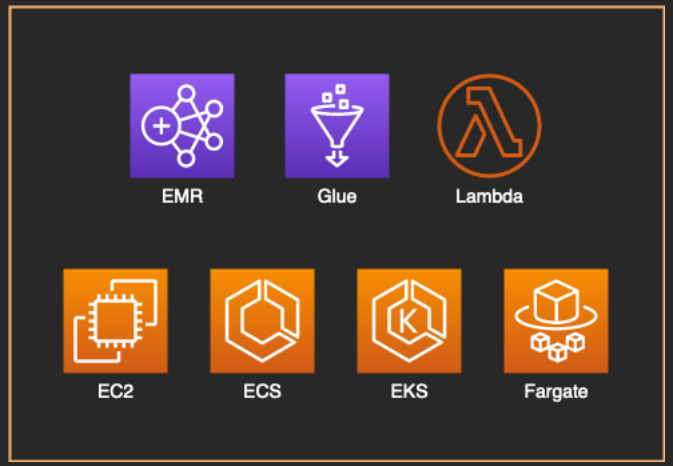
- Capture metadata for application to leverage the ingested datasets
- Ensure that this activity is well-documented and automated



Use the right ETL tool for the job

Select an ETL tool that closely meets your requirements for streamlining the workflow between the source and the destination

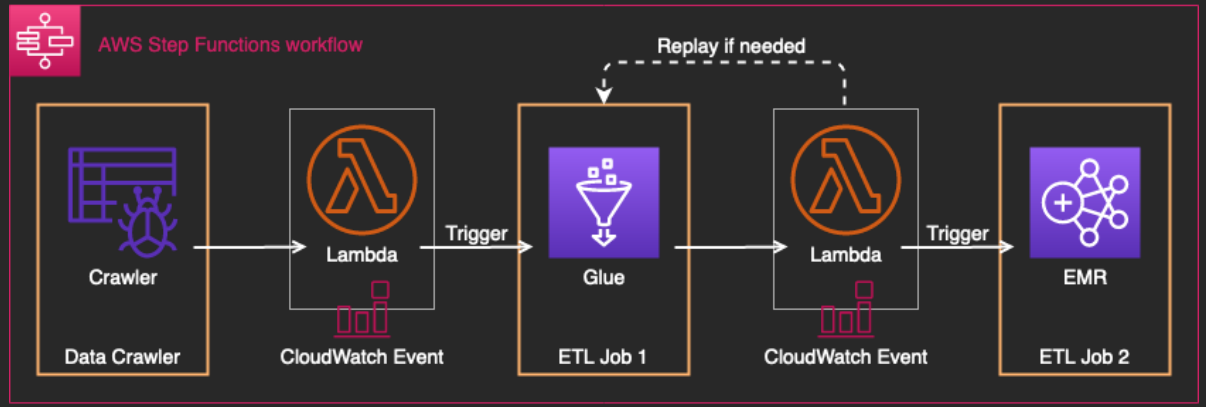
- Several options
  - Custom built to solve specific problems
  - Assembled from open source projects
  - Commercially licensed ETL platforms
- Support for complex workflows, APIs and specific languages
- Connectors to varied data stores
- Performance, budget, and enterprise scale.

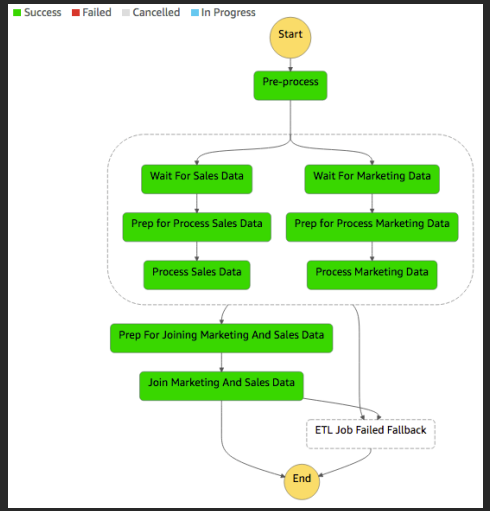


Automate ETL workflows

Chaining ETL jobs ensures the seamless execution of your ETL workflow

- Output from one process or job typically serves as an input to another
- Ensure you have the visibility of tracking and debugging any failure

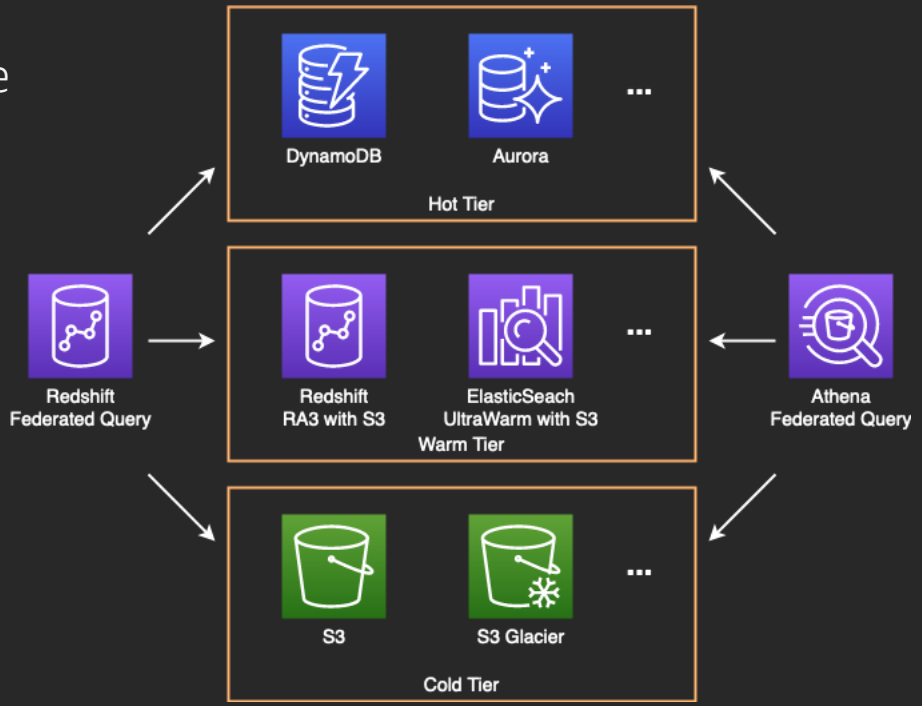




Tier storage appropriately

Store data in the optimal tier to ensure that you leverage the best features of the storage services for your analytics applications

- Two basic parameters for choosing the right data storage
  - Data format
  - Access Frequency
- Distributing your datasets into different services
  - Metadata tier & Payload tier
  - Hot, warm and cold tiers



Secure, protect, and manage your entire analytics pipeline

Both the data assets and the infrastructure for storing, processing, and analyzing data must be secured

- Implementing fine-grained controls that allow authorized users to manage particular assets
- Access roles might change at various stages of an analytics pipeline
- Ensuring that unauthorized users are blocked from taking any actions that would compromise data confidentiality and security

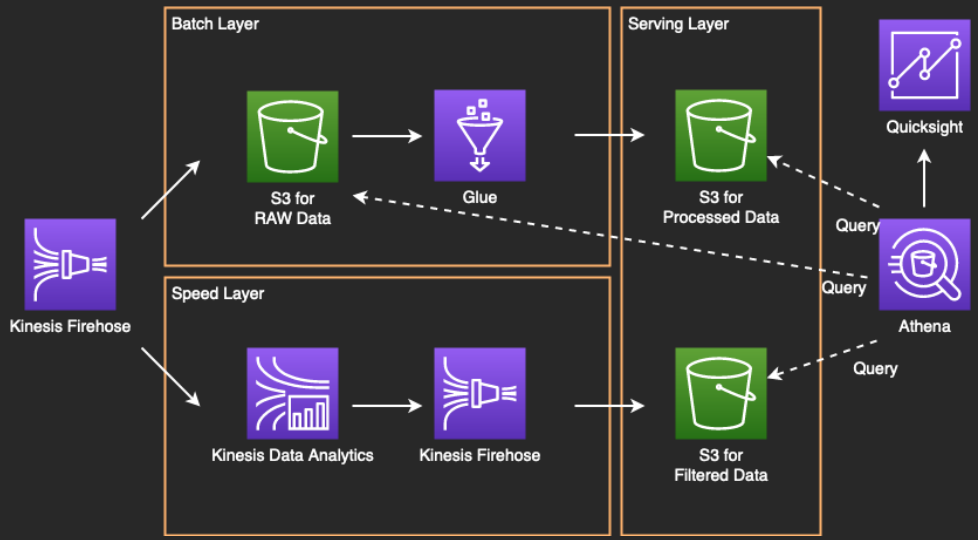
AWS Lake Formation

AWS Identity and Access Management

Design for scalable and reliable analytics pipelines

Make analytics execution compute environments reliable and scalable

- Keep up the pace of data volume and velocity
- Provide high data reliability and optimized query performance to support different analytics applications
  - batch and streaming ingest
  - fast ad hoc queries to data science



AWS supports your needs

aws SUMMIT

Why choose AWS for data lakes and analytics?

1

Easiest to build data lakes and analytics

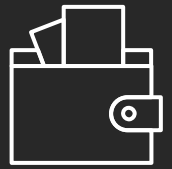
2

Most secure infrastructure for analytics

3

Most comprehensive and open

4



Most scalable and cost-effective

1. Easiest to build data lakes and analytics

The fastest way to go from zero to insights, covering all data for all users

- A single storage layer (Amazon S3) for all analytics and ML
- A service to build secure data lakes in days
- Deep integration across analytics and infrastructure (including federated queries)

2. Most secure infrastructure for analytics

Services for security and governance

Customers need to have multiple levels of security, identity and access management, encryption, and compliance to secure their data lakes

Security

- Amazon GuardDuty
- AWS Shield
- AWS WAF
- Amazon Macie
- Amazon VPC

Identity

- IAM
- AWS SSO
- Amazon Cloud Directory
- AWS Directory Service
- AWS Organizations

Encryption

- AWS Certificate Manager
- AWS Key Management Service
- Encryption at rest
- Encryption in transit
- Bring your own keys, HSM support

Compliance

- AWS Artifact
- Amazon Inspector
- AWS CloudHSM
- Amazon Cognito
- AWS CloudTrail

3. Most comprehensive and open

Data, visualization, engagement & machine learning

NEW AWS Data Exchange | Amazon QuickSight | Amazon Pinpoint | Amazon SageMaker | Amazon Comprehend | Amazon Lex | Amazon Polly | Amazon Rekognition | Amazon Translate

+ Many more

Analytics

Amazon Redshift | Amazon EMR (Spark & Hadoop) | AWS Glue (Spark & Python) | Amazon Athena | Amazon Elasticsearch Service | Amazon Kinesis Data Analytics

Data lake infrastructure & management

Amazon S3/ Amazon S3 Glacier | AWS Lake Formation | AWS Glue

Data movement

AWS Database Migration Service | AWS Snowball | AWS Snowmobile | Amazon Kinesis Data Firehose | Amazon Kinesis Data Streams | Amazon Managed Streaming for Apache Kafka

4. Most scalable, cost-effective, high-performance infrastructure for analytics

On-Demand, Reserved, and Spot Instances to reduce costs

100 Gbps-bandwidth network interfaces for performance

Industry-leading choice of 200+ instance types to meet workload needs

Five highly available storage tiers and intelligent tiering

Learn analytics with AWS Training and Certification

Resources created by the experts at AWS to help you build and validate data analytics skills

New free digital course: **Data Analytics Fundamentals**

Classroom offerings, including **Big Data on AWS**, feature AWS expert instructors and hands-on labs

Validate expertise with the **AWS Certified Big Data—Specialty** exam or the new **AWS Certified Data Analytics—Specialty** beta exam

Visit aws.amazon.com/training/paths-specialty/

Thank you!

Jayson Hsieh
hsiej@amazon.com

aws SUMMIT