

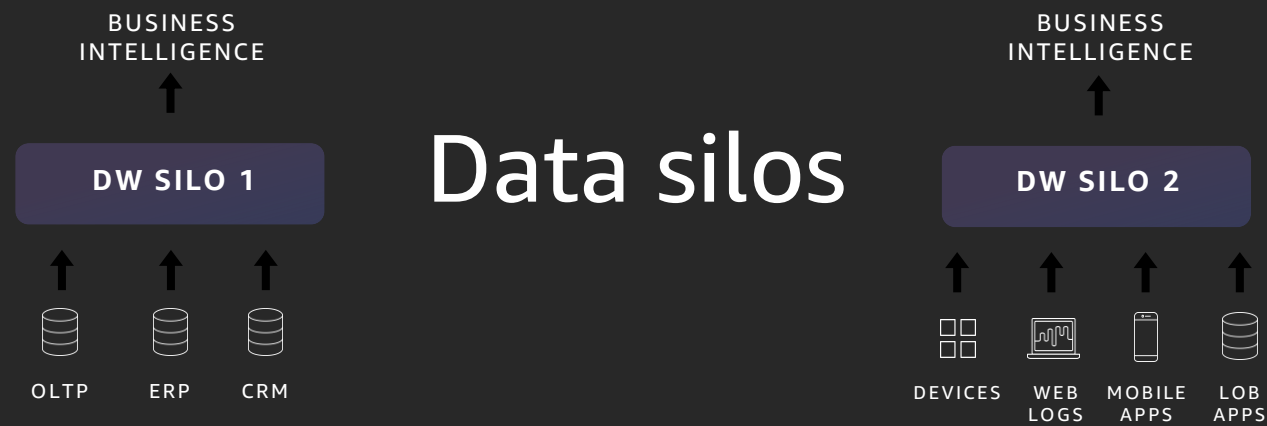
Track 6 | Session 2

搭建現代化的資料數據湖

Young Yang
ML Specialist SA
Amazon Web Services

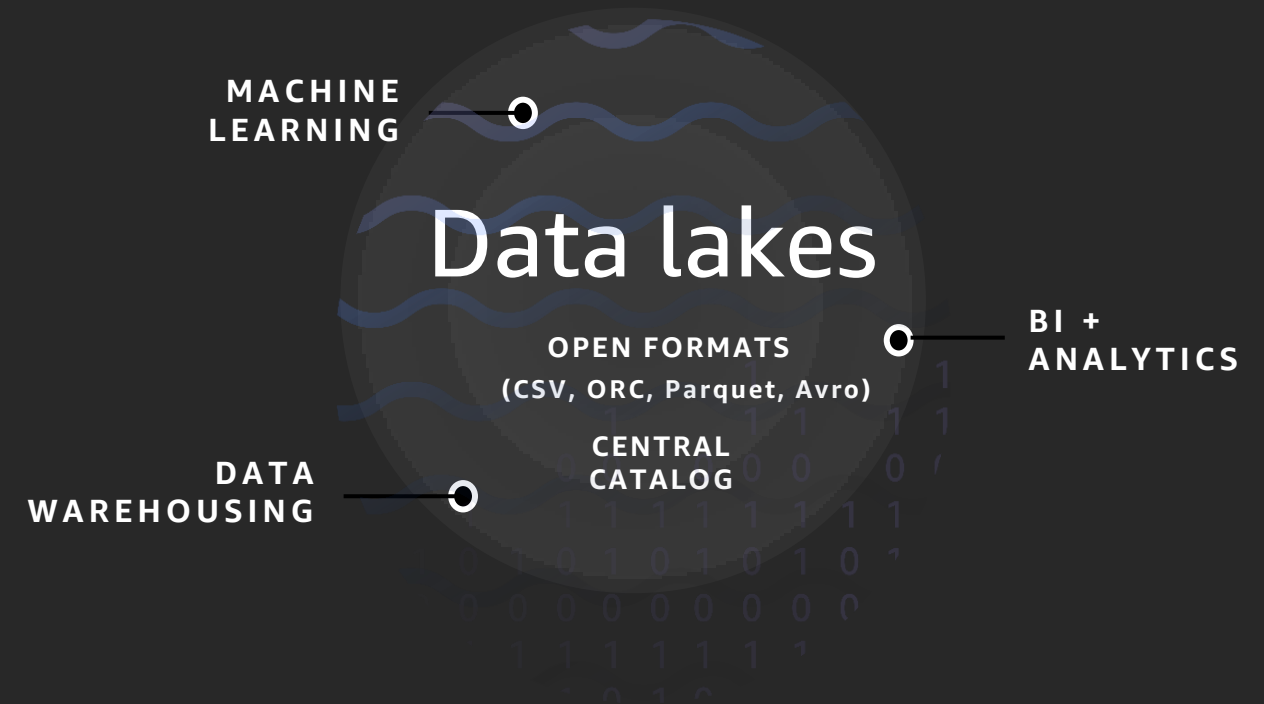
AWS offers a modern data platform

Old guard data patterns

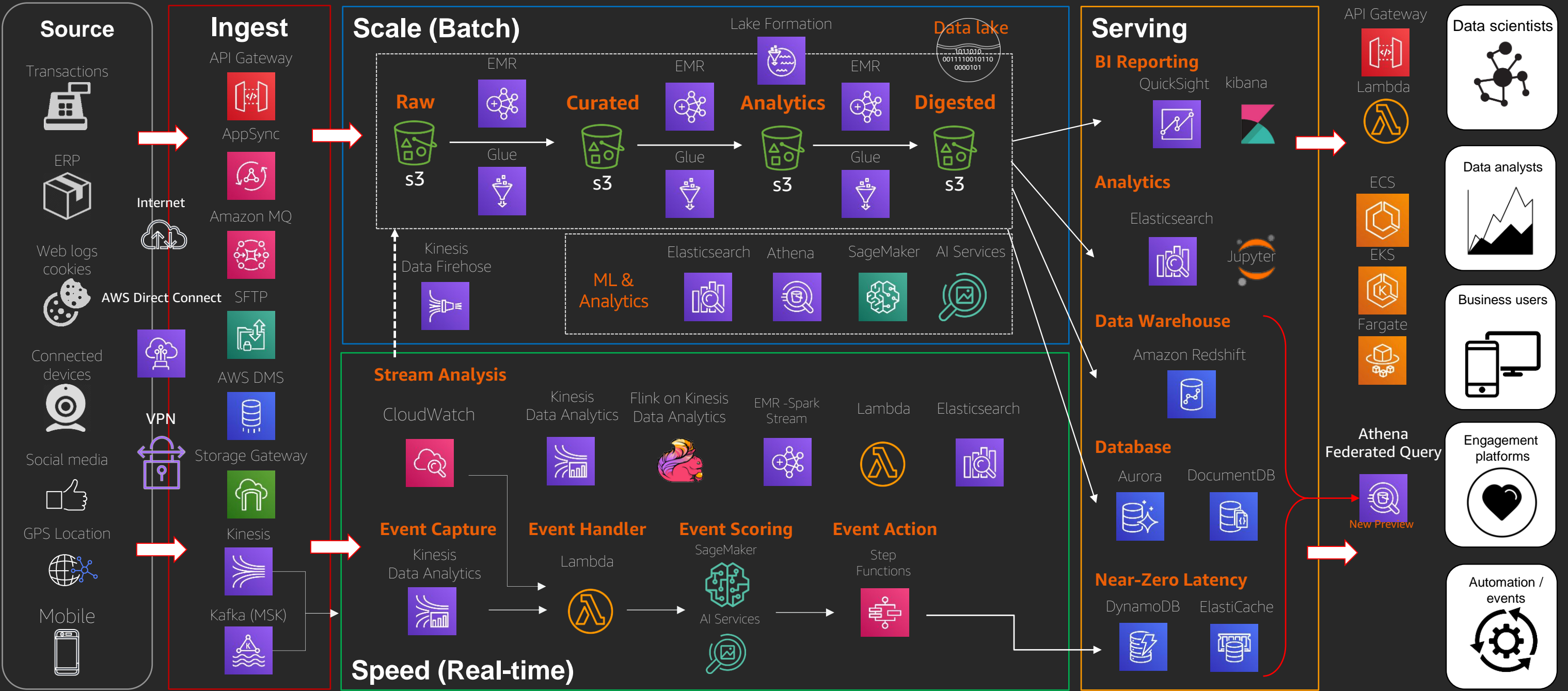


to >

Modern data architecture



After this session, what you will take away?



Let us start

Source

Ingest

Scale (Batch)

Serving

Speed (Real-time)

Source

Transactions



ERP



Web logs
cookies



Connected
devices



Social media



GPS Location



Mobile

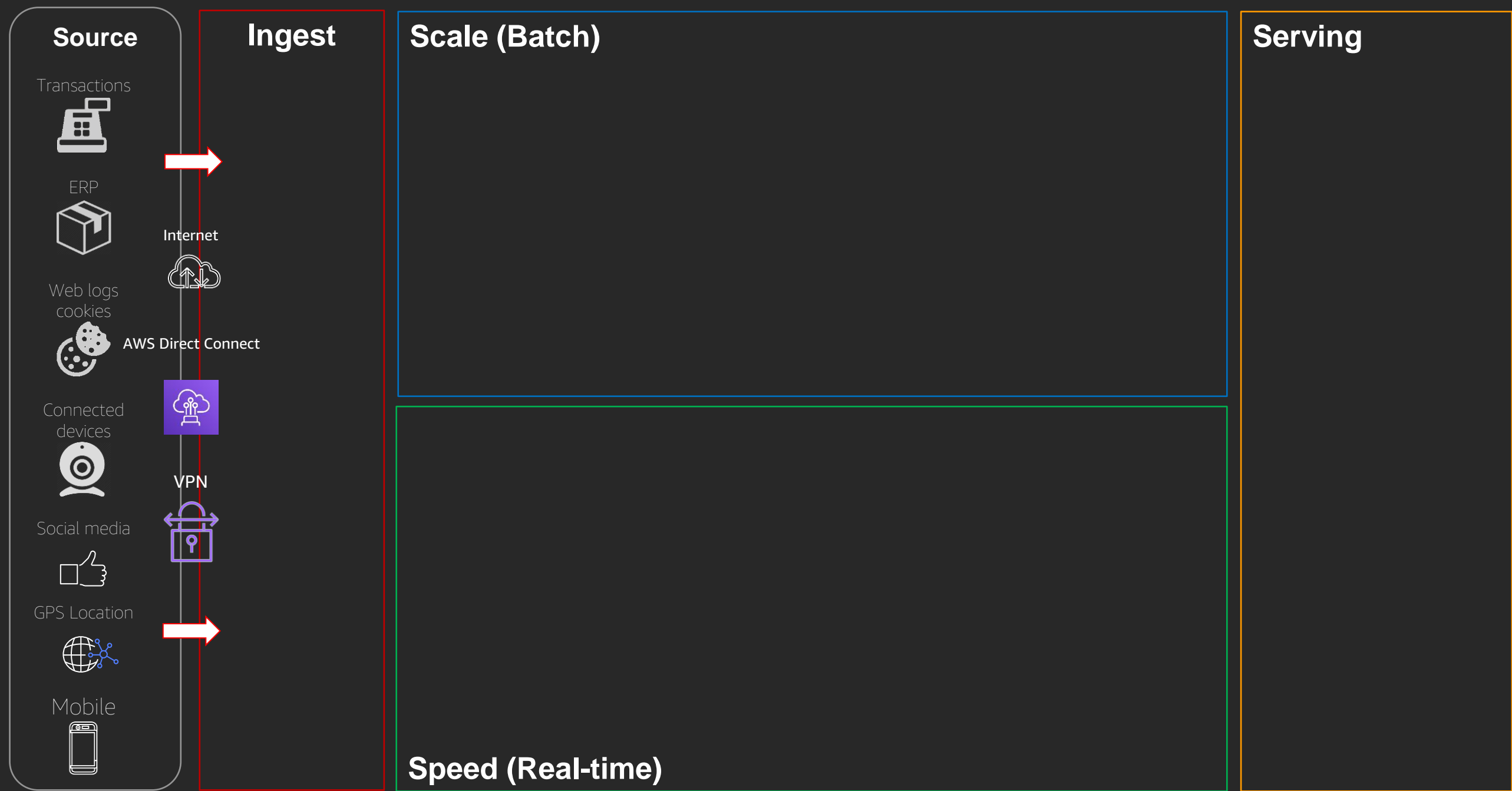


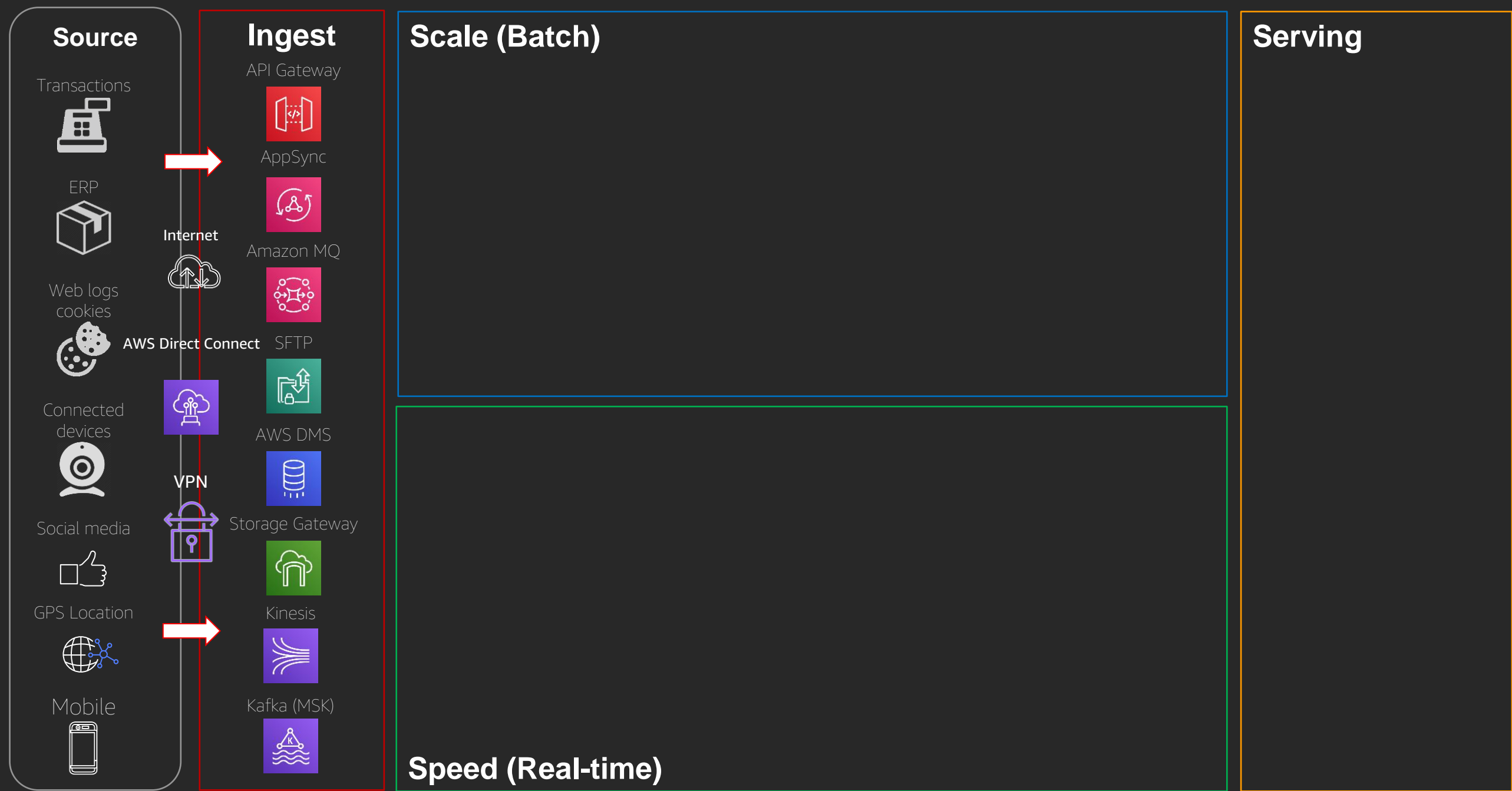
Ingest

Scale (Batch)

Serving

Speed (Real-time)





Source

Transactions



ERP



Web logs
cookies



AWS Direct Connect

Connected
devices



Social media



GPS Location



Mobile

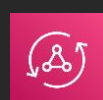


Ingest

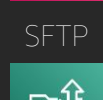
API Gateway



AppSync

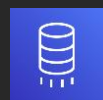


Amazon MQ



SFTP

AWS DMS



Storage Gateway



Kinesis



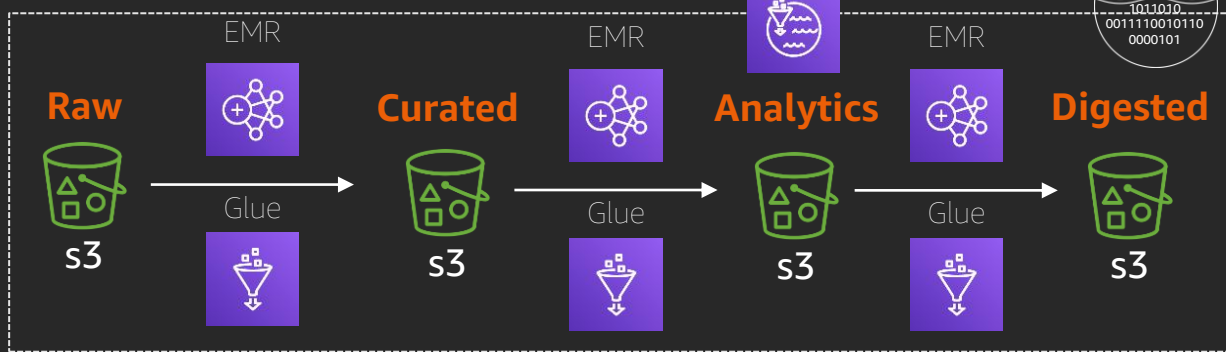
Kafka (MSK)



Scale (Batch)

Lake Formation

Data lake



Speed (Real-time)

Serving

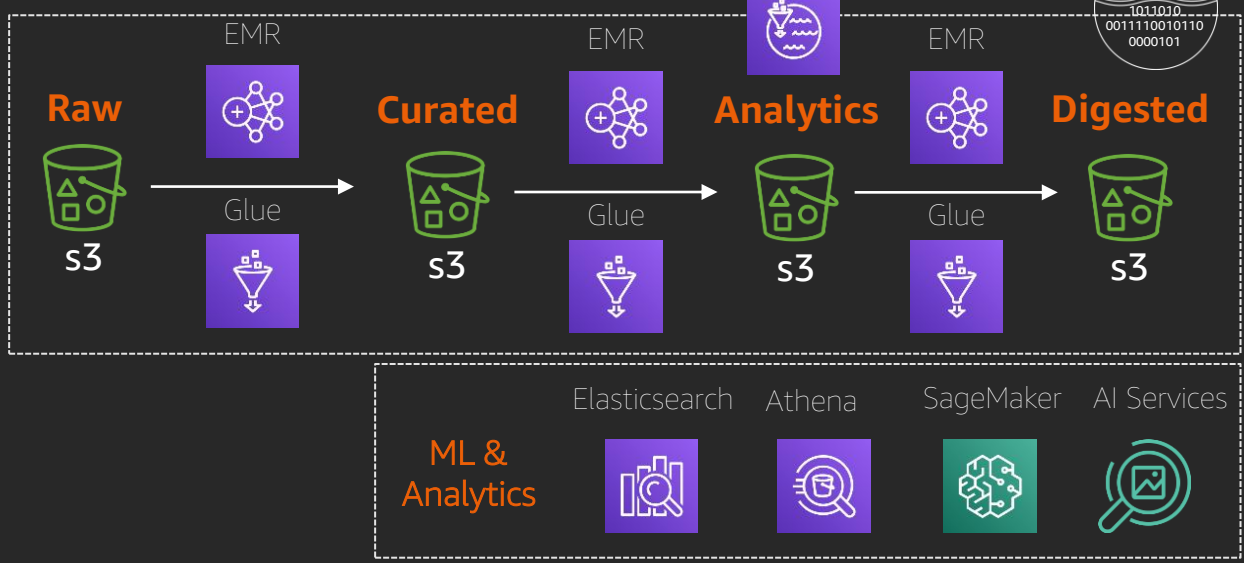
Source

- Transactions
- ERP
- Web logs cookies
- Connected devices
- Social media
- GPS Location
- Mobile

Ingest

- API Gateway
- AppSync
- Amazon MQ
- SFTP
- AWS DMS
- Storage Gateway
- Kinesis
- Kafka (MSK)

Scale (Batch)



Speed (Real-time)

Serving

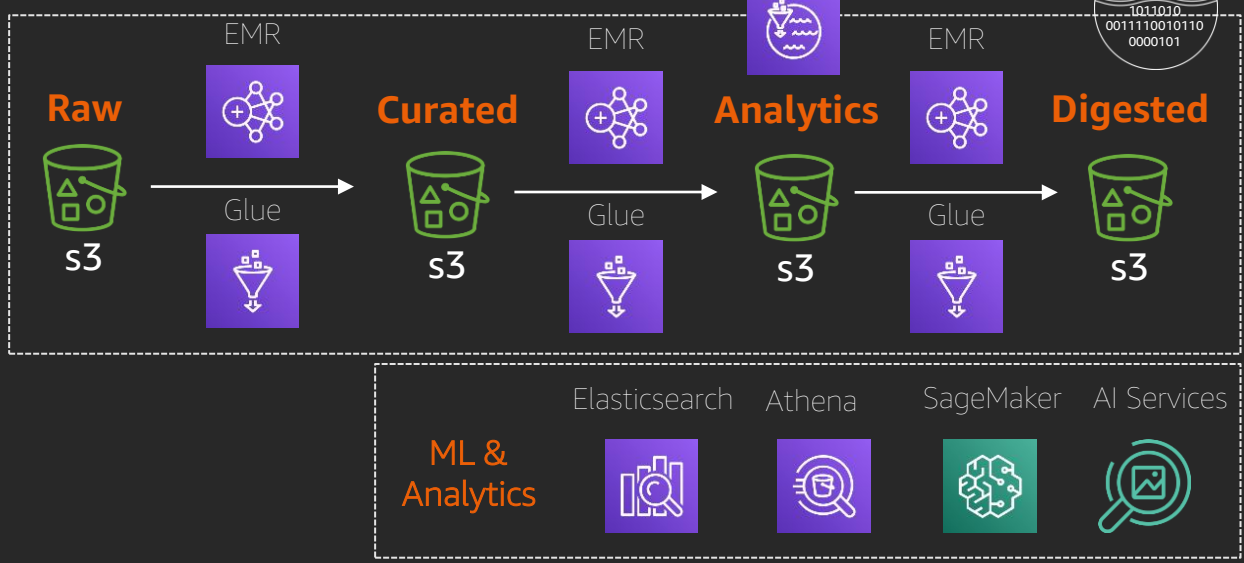
Source

- Transactions
- ERP
- Web logs cookies
- Connected devices
- Social media
- GPS Location
- Mobile

Ingest

- API Gateway
- AppSync
- Amazon MQ
- SFTP
- AWS DMS
- Storage Gateway
- Kinesis
- Kafka (MSK)

Scale (Batch)



Stream Analysis

- CloudWatch
- Kinesis Data Analytics
- Flink on Kinesis Data Analytics
- EMR -Spark Stream
- Lambda
- Elasticsearch

Speed (Real-time)

Serving

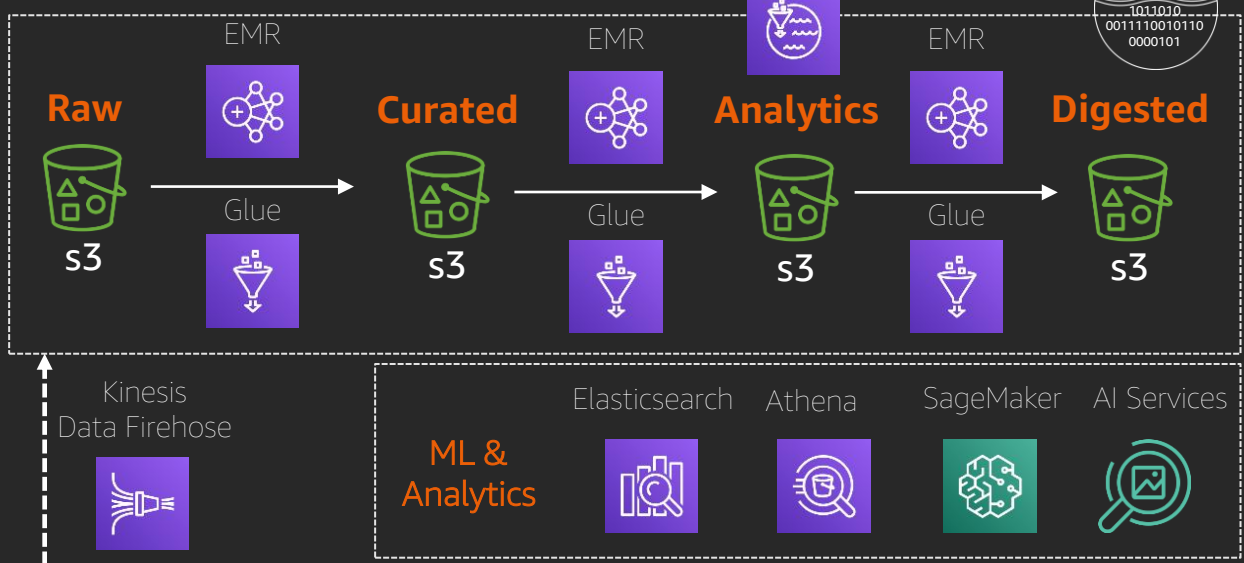
Source

- Transactions
- ERP
- Web logs cookies
- Connected devices
- Social media
- GPS Location
- Mobile

Ingest

- API Gateway
- AppSync
- Amazon MQ
- SFTP
- AWS DMS
- Storage Gateway
- Kinesis
- Kafka (MSK)

Scale (Batch)



Stream Analysis

- CloudWatch
- Kinesis Data Analytics
- Flink on Kinesis Data Analytics
- EMR -Spark Stream
- Lambda
- Elasticsearch

Speed (Real-time)

Serving

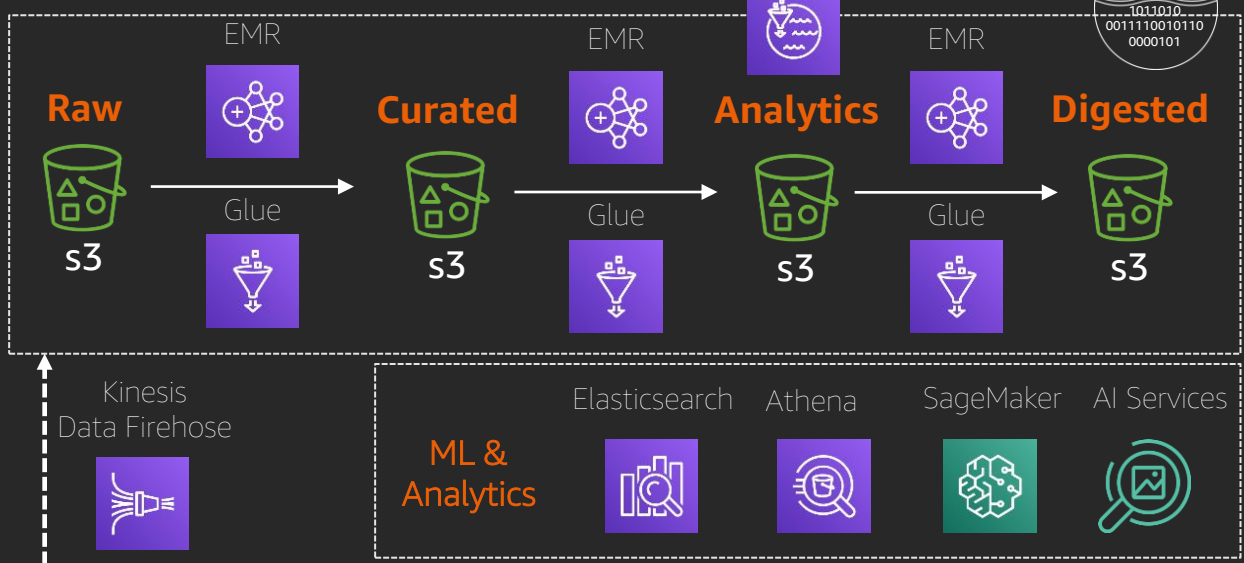
Source

- Transactions
- ERP
- Web logs cookies
- Connected devices
- Social media
- GPS Location
- Mobile

Ingest

- API Gateway
- AppSync
- Amazon MQ
- SFTP
- AWS DMS
- Storage Gateway
- Kinesis
- Kafka (MSK)

Scale (Batch)



Stream Analysis

- CloudWatch
- Kinesis Data Analytics
- Flink on Kinesis Data Analytics
- EMR -Spark Stream
- Lambda
- Elasticsearch

Event Capture

- Kinesis Data Analytics

Event Handler

- Lambda

Event Scoring

- SageMaker
- AI Services

Event Action

- Step Functions

Speed (Real-time)

Serving

Data lake

1011010
0011110010110
0000101

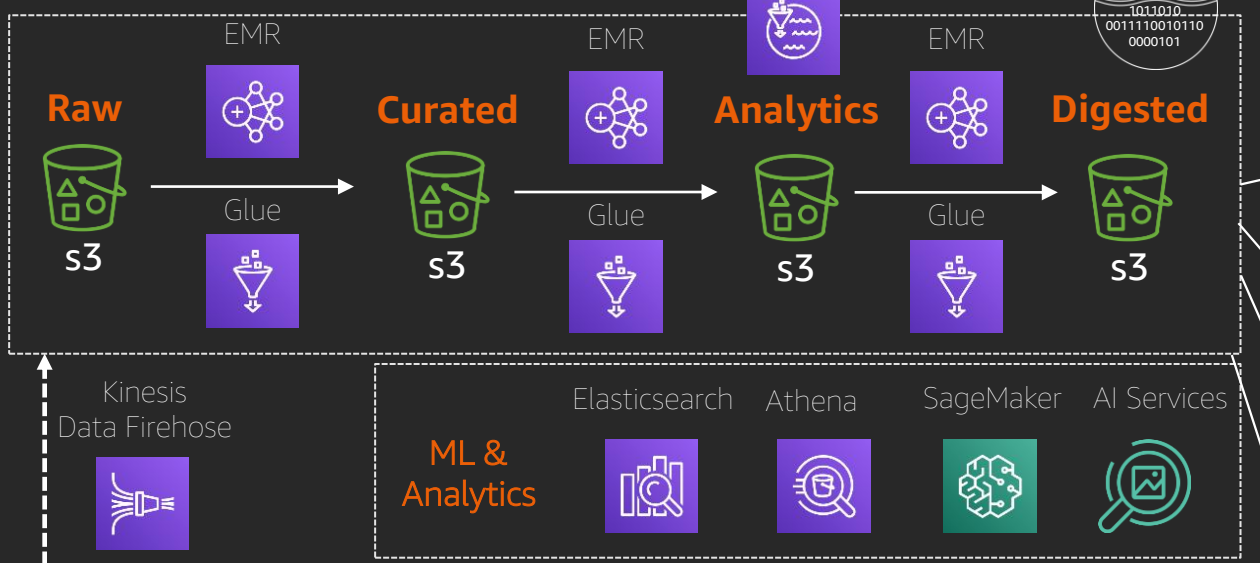
Source

- Transactions
- ERP
- Web logs cookies
- Connected devices
- Social media
- GPS Location
- Mobile

Ingest

- API Gateway
- AppSync
- Amazon MQ
- SFTP
- AWS DMS
- Storage Gateway
- Kinesis
- Kafka (MSK)

Scale (Batch)



Stream Analysis

- CloudWatch
- Kinesis Data Analytics
- Flink on Kinesis Data Analytics
- EMR -Spark Stream
- Lambda
- Elasticsearch

Event Capture

- Kinesis Data Analytics

Event Handler

- Lambda

Event Scoring

- SageMaker
- AI Services

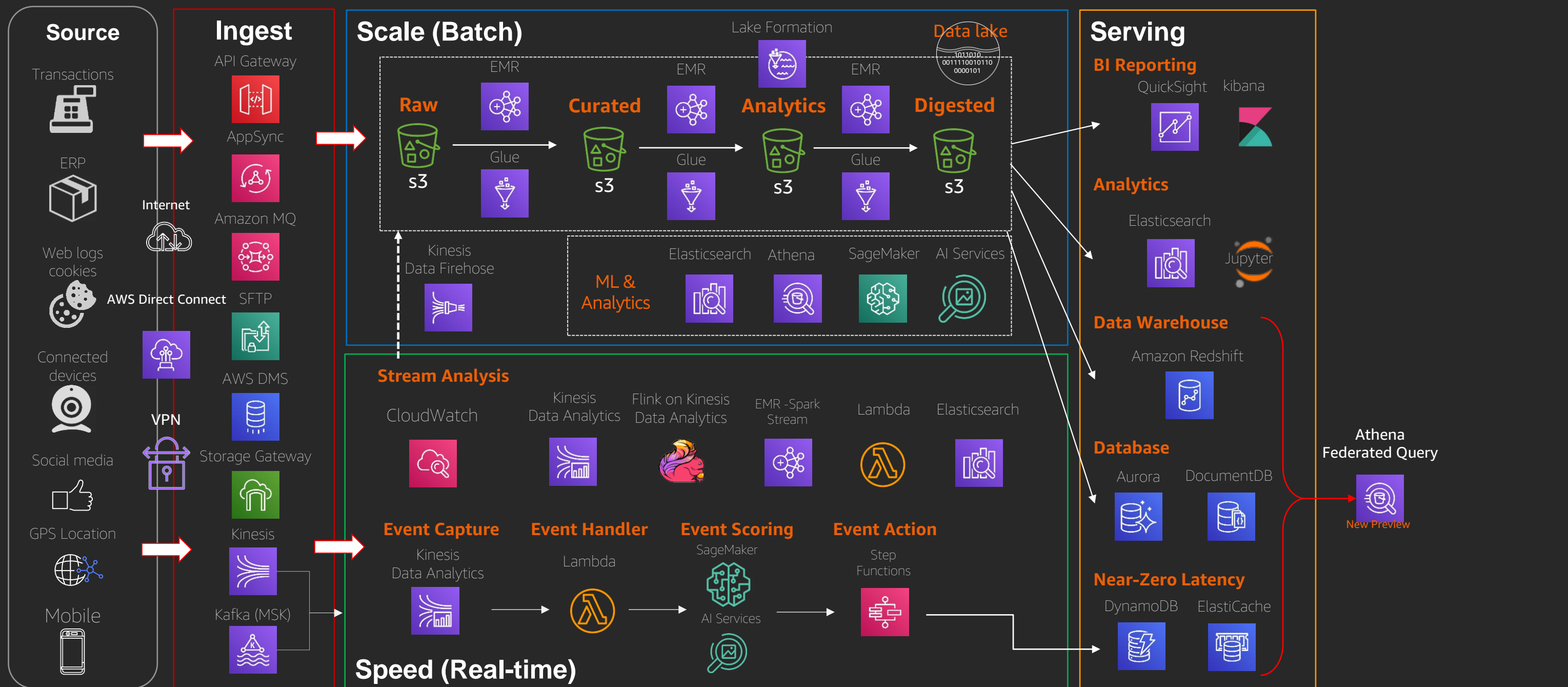
Event Action

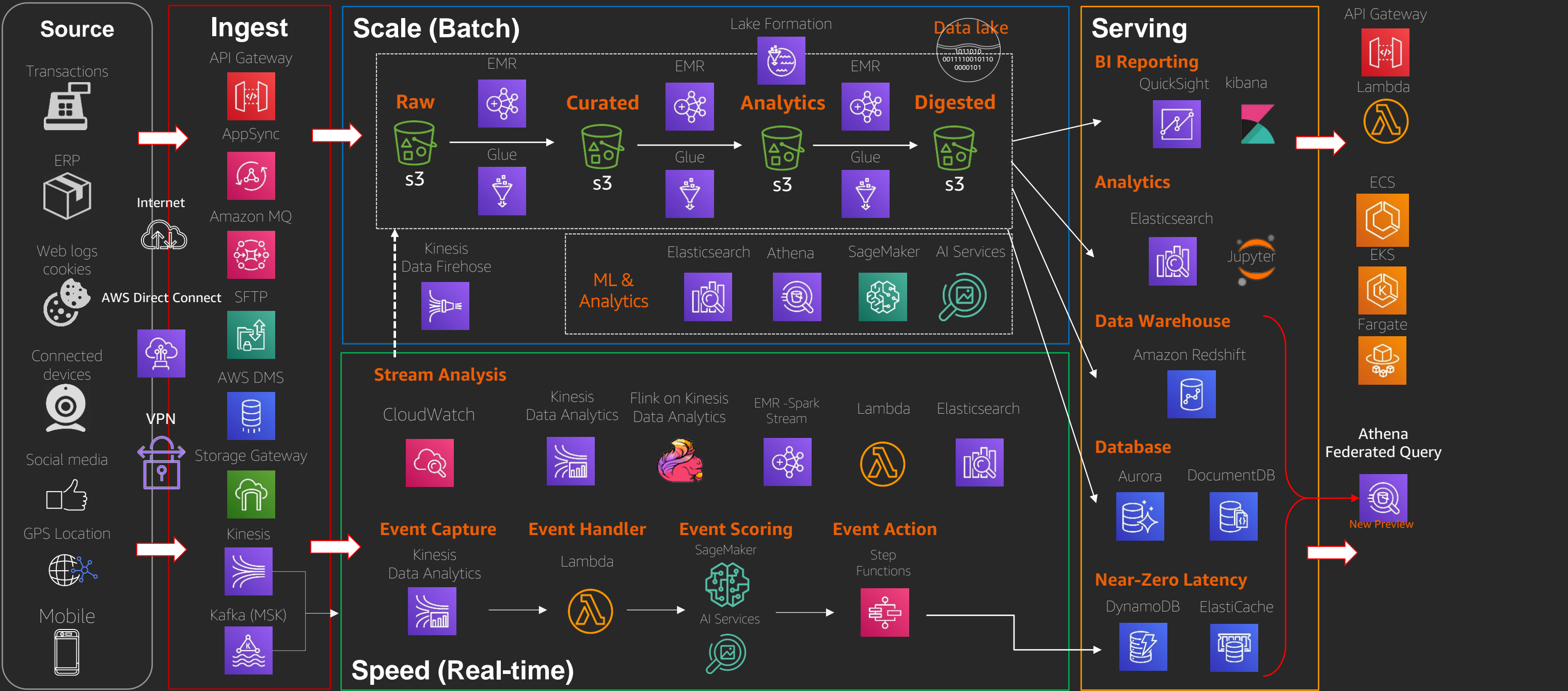
- Step Functions

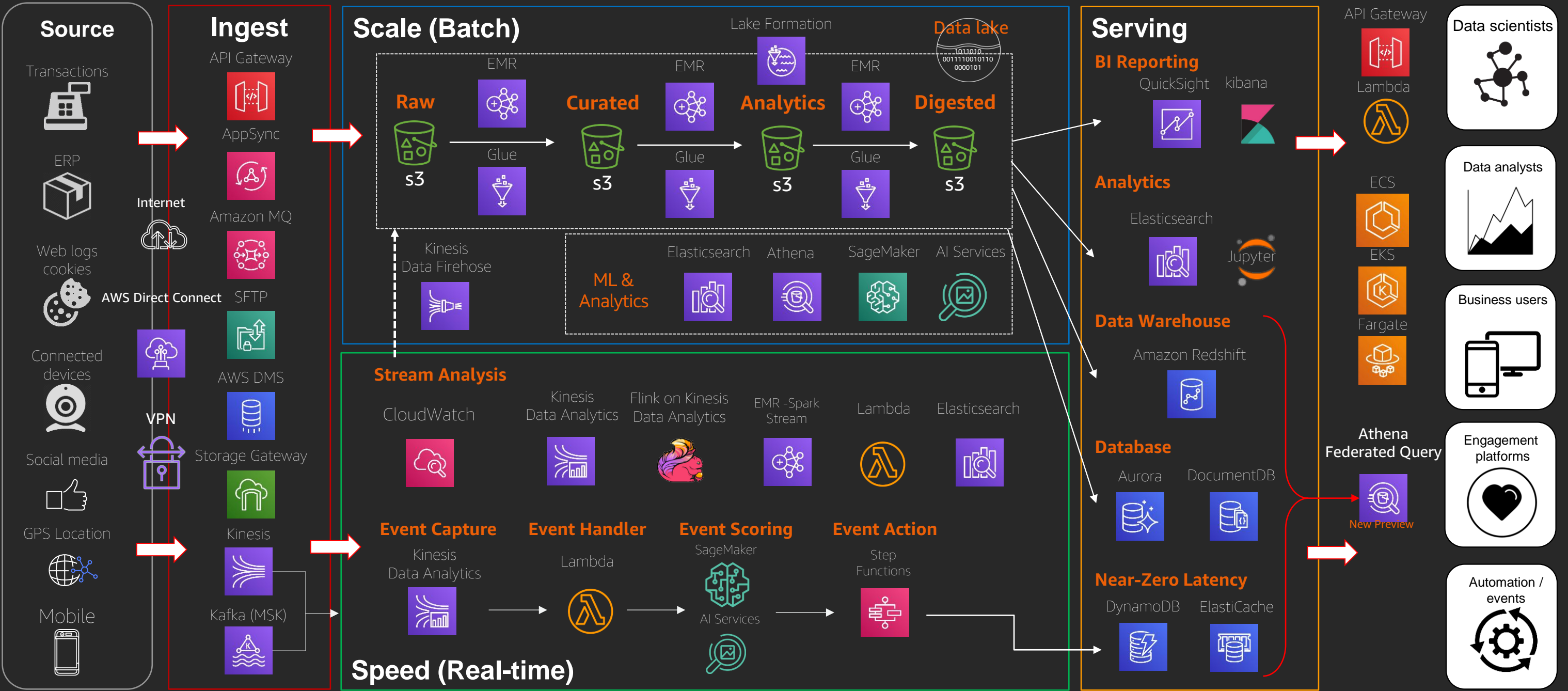
Speed (Real-time)

Serving

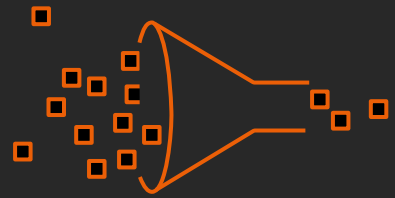
- BI Reporting: QuickSight, kibana
- Analytics: Elasticsearch, Jupyter
- Data Warehouse: Amazon Redshift
- Database: Aurora, DocumentDB
- Near-Zero Latency: DynamoDB, ElastiCache







Amazon S3 is the foundation of any data lake



Multiple data
input sources



Storage scales on
demand



Supports many
unique users and
teams



Analyzed by
many applications

Amazon S3 as the foundation for data lakes



Durable, available, exabyte-scalable

Secure, compliant, auditable

High performance

Low-cost storage and analytics

Broad network integration

AWS Lake Formation

Build a secure data lake in days

Build data lakes quickly



Move, store, catalog,
and clean your data faster

Transform to open formats like
Parquet and ORC

ML-based deduplication
and record matching

Simplify security management



Centrally define security, governance,
and auditing policies

Enforce policies consistently
across multiple services

Integrates with IAM and KMS

Provide self-service access to data



Build a data catalog that
describes your data

Enable analysts and data scientists
to easily find relevant data

Analyze with multiple analytics
services without moving data

Tier 1 Data Lake: Raw or Ingestion



Amazon S3

Single Source of Truth for Raw Data

Use Least Transformations

Use Lifecycle policies to S3-IA or Glacier

Tier 2 Data Lake: Curated



Amazon S3

Non-structured to structured Raw Data

Annotation

Data cleansing and transform

Uniform the data of encoding, format, types

(such as time format, string encoding, and etc)

Tier 3 Data Lake: Analytics



Amazon S3

Use columnar formats – Parquet/ORC

Organized into Partitions

Coalescing to Larger Partitions over time

Optimized for Analytics

Tier 4 Data Lake: Digested (Serving Stage)



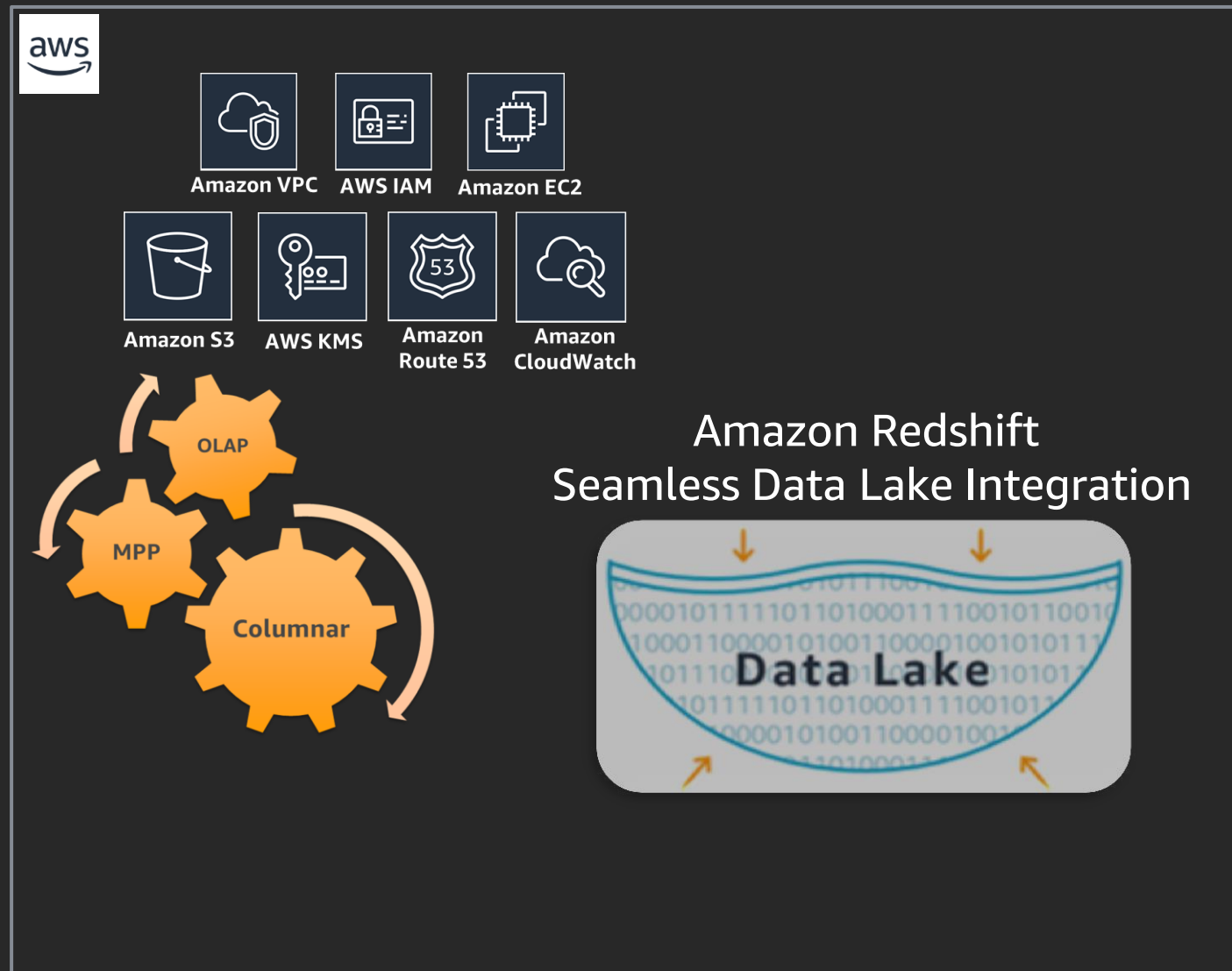
Amazon S3

Domain Level DataMart

Organized by use cases

Optimized for Specialized Analysis

Amazon Redshift: What's Under the Hood?



Amazon Redshift

Amazon Redshift is a fully managed data warehouse service that extends seamlessly to the data lake. It's highly performant, scalable, resilient, easy-to-use, cost-effective, & secure.

Our portfolio

Broad and deep portfolio, purpose-built for builders

Business Intelligence & Machine Learning

**Data Exchange**
Data exchange

**QuickSight**
Visualizations

**SageMaker**
ML

**Comprehend**
NLP

**Transcribe**
Speech-to-text

**Textract**
Extract text

**Personalize**
Recommendation

**Forecast**
Forecasts

**Translate**
Translation

**CodeGuru**
Code reviews

**Kendra**
Enterprise search


NEW


NEW


NEW

NEW

Analytics

**Redshift**
Data warehousing

**EMR**
Hadoop + Spark


**Athena**
Interactive analytics


AQUA

NEW

EMR on Outposts

NEW


**Elasticsearch Service**
Operational Analytics


**Kinesis Data Analytics**
Real time


UltraWarm


NEW


Databases


**Aurora**
MySQL, PostgreSQL


**DynamoDB**
Key value, Document


**Neptune**
Graph


**RDS**
MySQL, PostgreSQL, MariaDB, Oracle, SQL Server, RDS on VMware

**DocumentDB**
Document

**QLDB**
Ledger Database

**ElastiCache**
Redis, Memcached

**Timestream**
Time Series

**Managed Apache Cassandra Service**
Wide column

RDS Proxy


NEW


RDS on Outposts

NEW


NEW


Blockchain


**Managed Blockchain**

**Blockchain Templates**

Data Lake

**S3/Glacier**

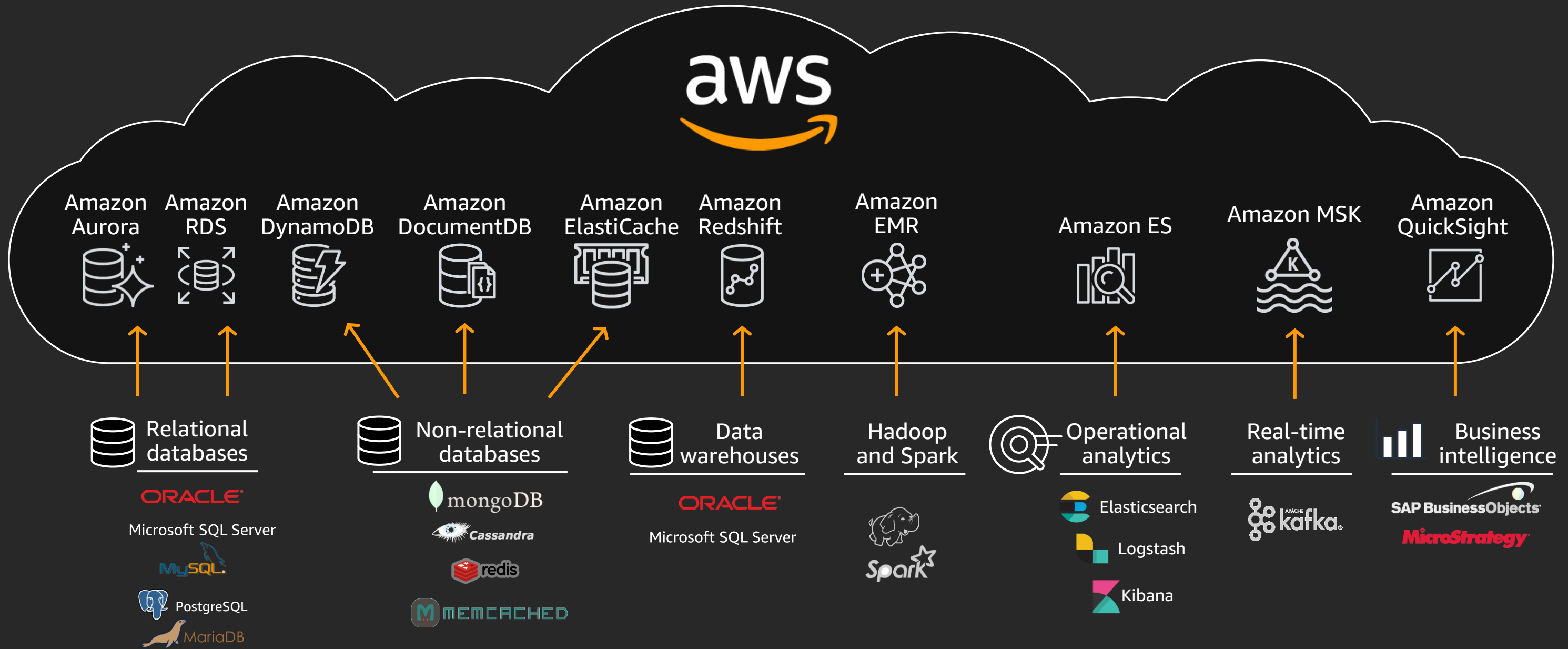
**Lake Formation**
Data Lakes

**Glue**
ETL & Data Catalog

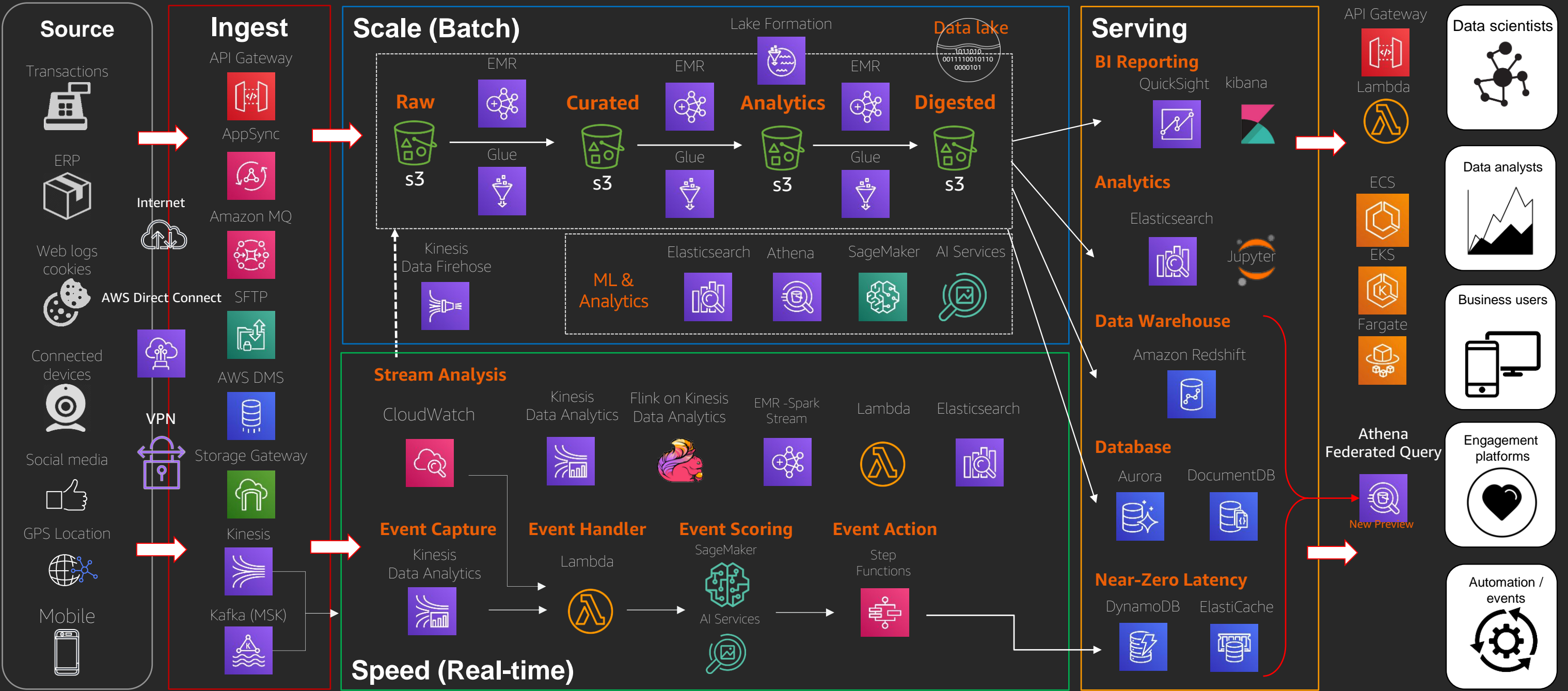
Data Movement

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

Broad database and analytics services portfolio



Take Away



Complemented by AWS Partner Network (APN)

Collection & preparation



Governance



Visualization



Data and analytics strategic & competency partners

Global



Japan



China



LATAM



North America



EMEA

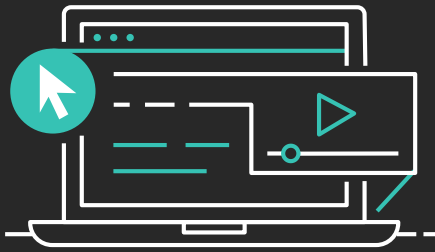


APAC



Learn storage with AWS Training and Certification

Resources created by the experts at AWS to help you build cloud storage skills



45+ free digital courses cover topics related to cloud storage, including:

- Amazon S3
- AWS Storage Gateway
- Amazon S3 Glacier
- Amazon Elastic File System (Amazon EFS)
- Amazon Elastic Block Store (Amazon EBS)



Classroom offerings, such as Architecting on AWS, feature AWS expert instructors and hands-on activities

Visit the storage learning path at <https://aws.training/storage>