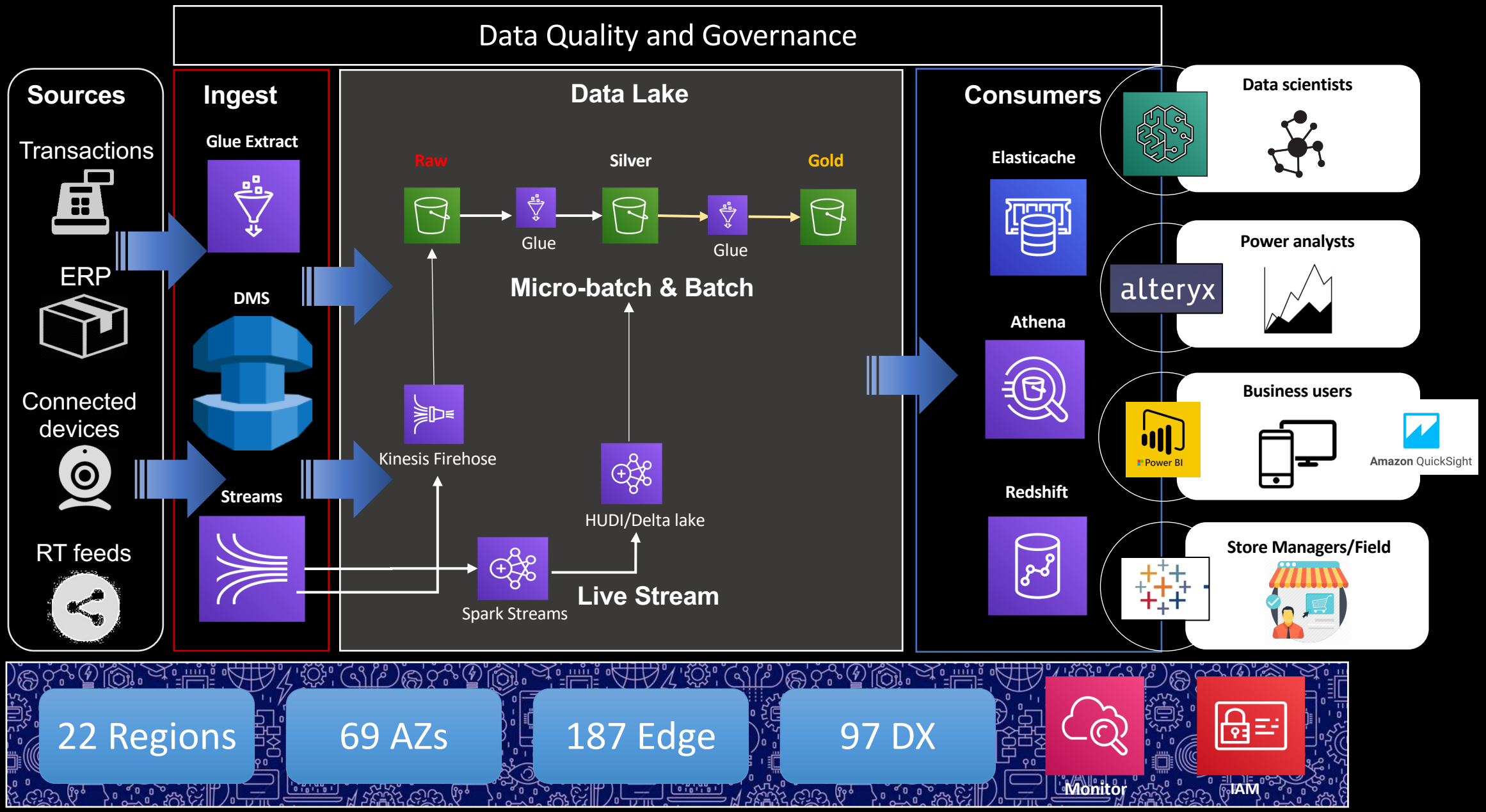
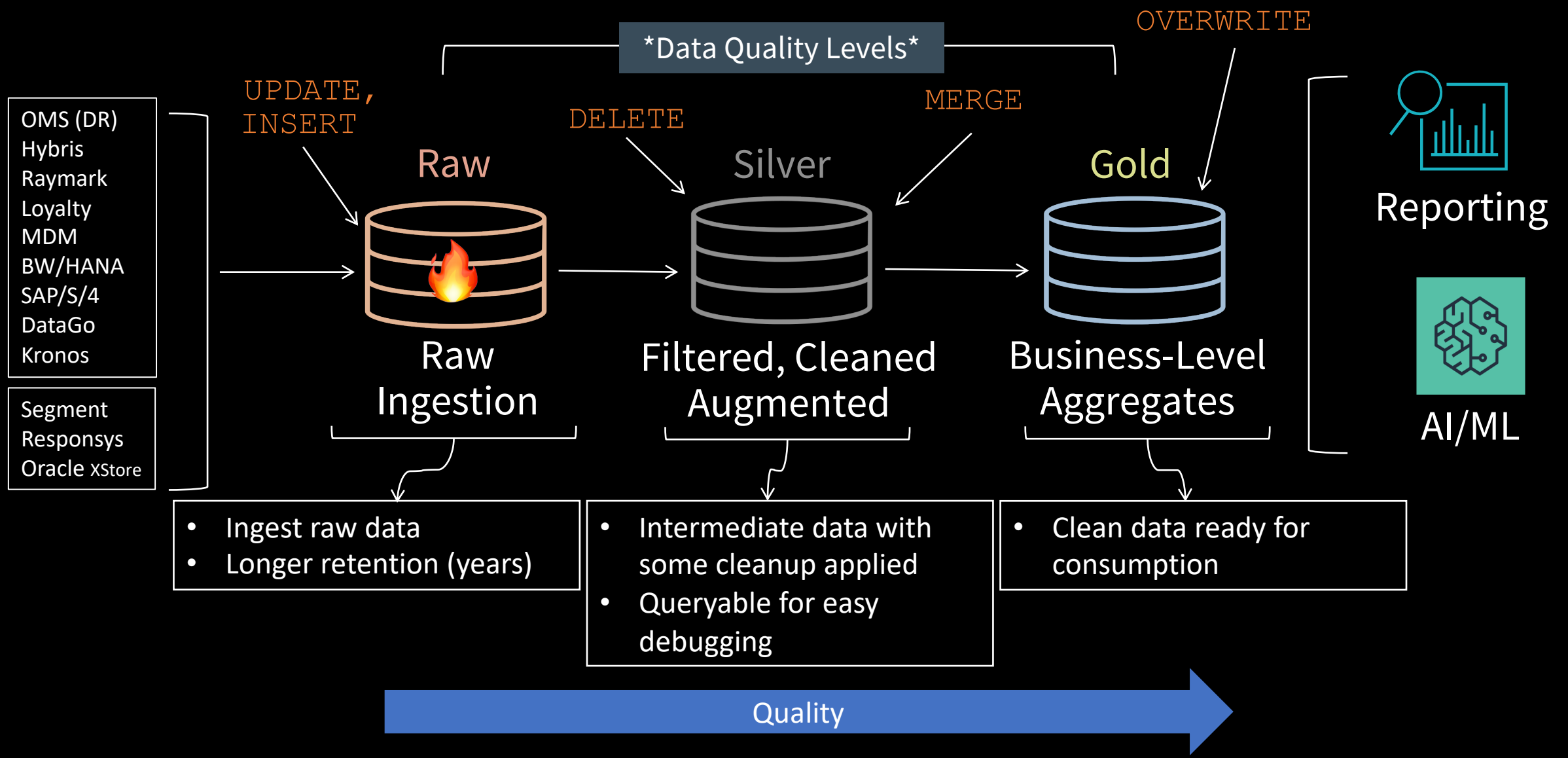
Data Lake and DS

Architectural overview

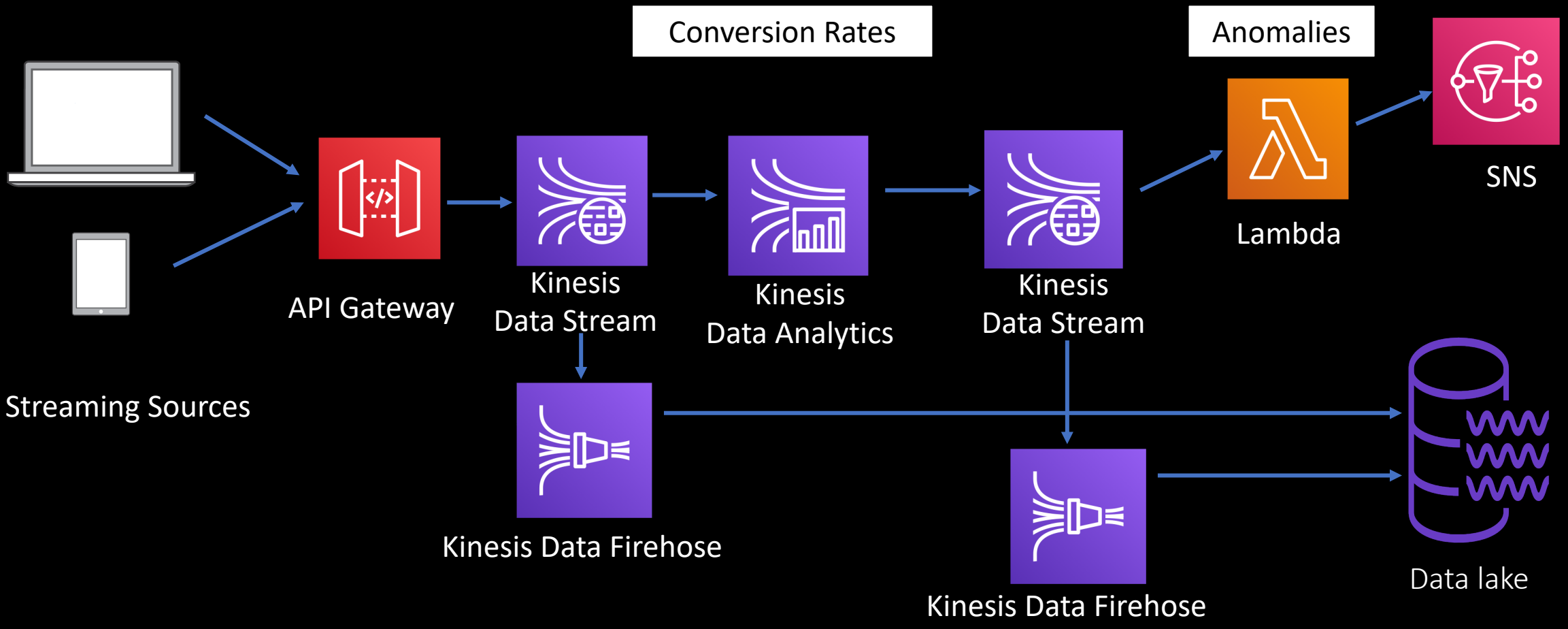
Global Data Lake Reference Architecture



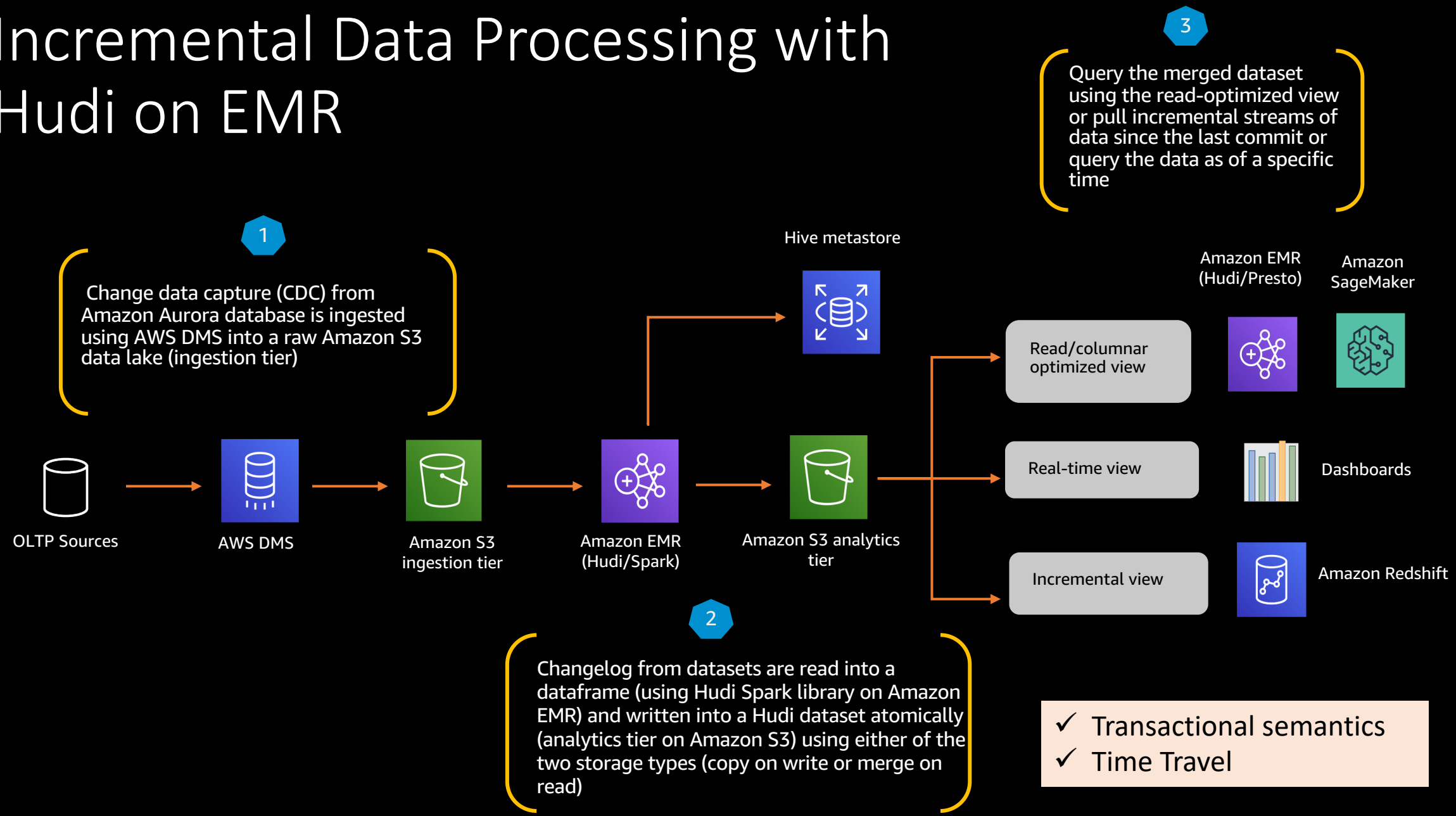
Incrementally improve quality of data until it is ready for consumption



Clickstream with Real-Time Analytics

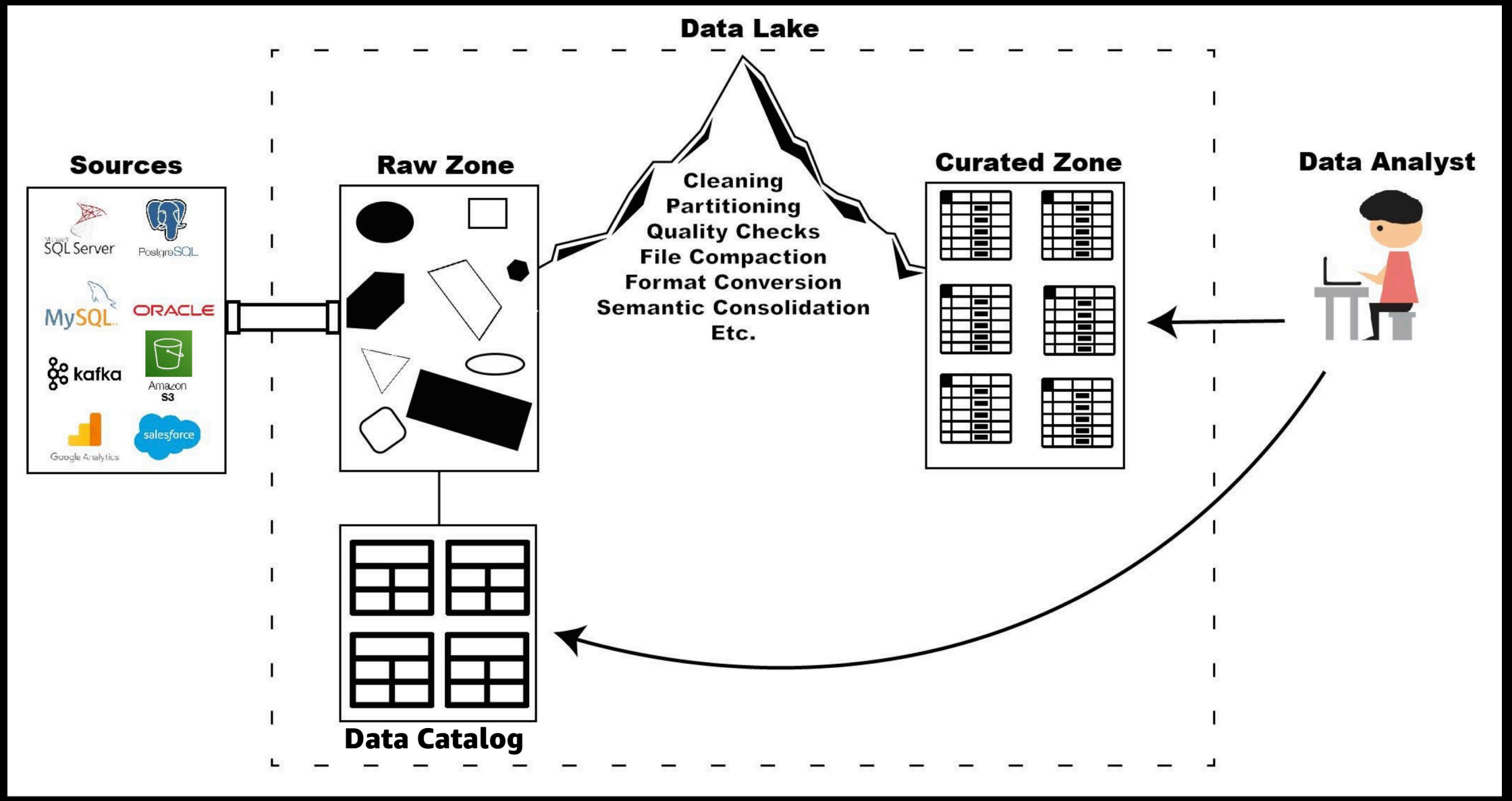


Incremental Data Processing with Hudi on EMR

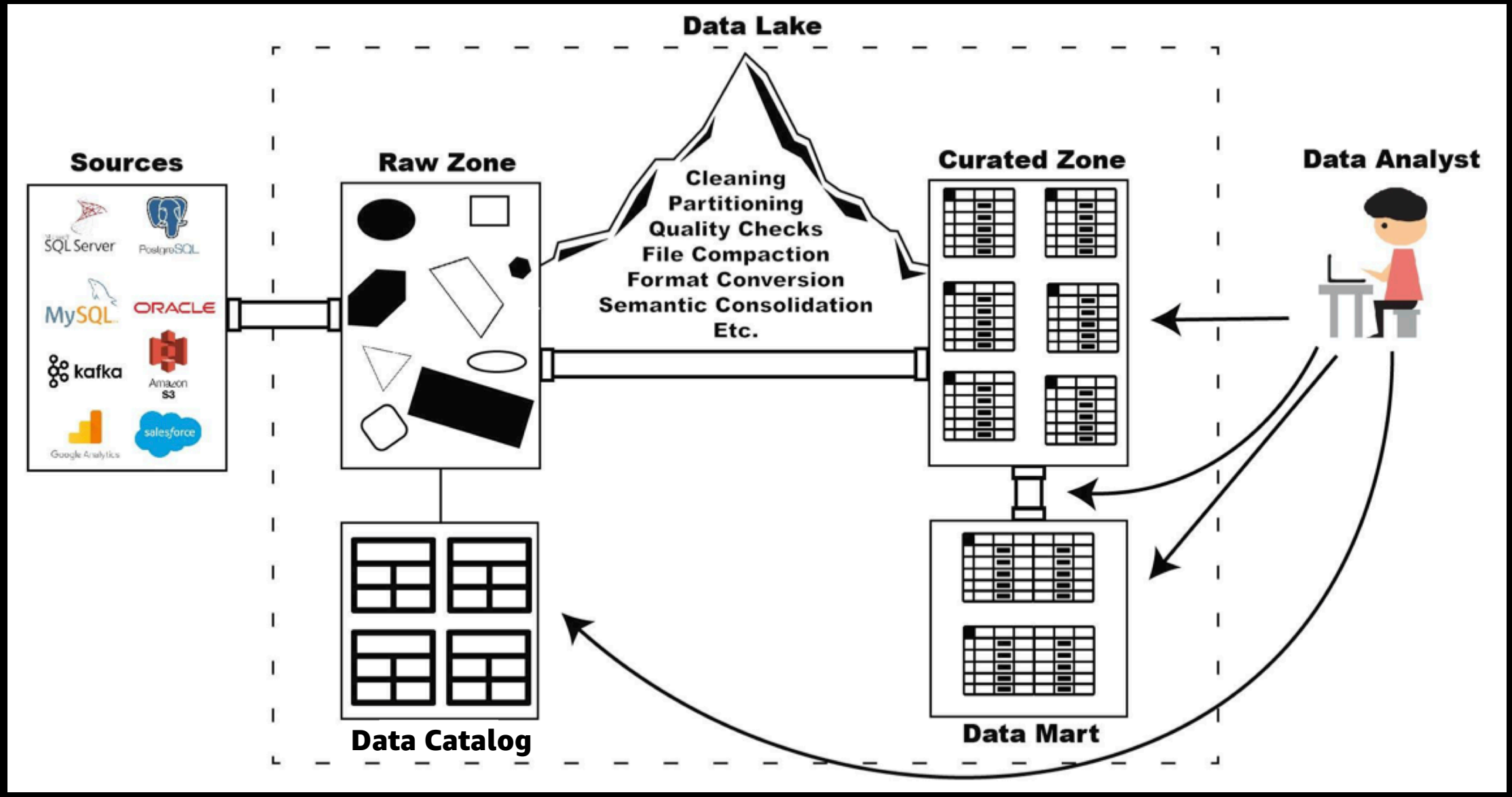


Data Lake - Best Practice

Give all data a minimal catalog for exploration



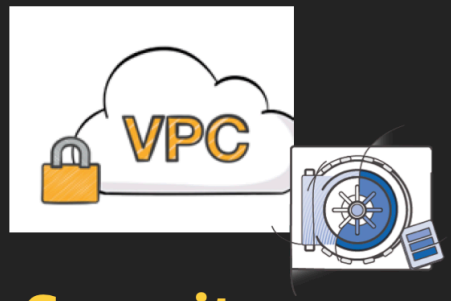
Self serve materialized views for Data Scientists



Security controls for your Data Lake

Encryption

- SSL endpoints
- Server-side encryption (SSE-S3)
- Amazon S3 server-side encryption with provided keys (SSE-C, SSE-KMS)
- Client-side encryption



Security

- Identity and Access Management (IAM) policies
- Bucket policies
- Access Control Lists (ACLs)
- Private VPC endpoints to Amazon S3

Compliance

- Buckets access logs
- Lifecycle management policies
- Access Control Lists (ACLs)
- Versioning & MFA deletes
- Certifications – HIPAA, PCI, SOC 1/2/3, etc.

Security & Governance

Security Overview

Region

VPC

Private subnet

Security group

Security group

Private subnet

Security group

Security group

AWS PrivateLink

Data Security

Amazon Macie

AWS Key Management Service

Identity

Permissions

AWS Identity & Access Management

Network

VPN connection

Peering connection

Corporate data center

Data Classification	Amazon Macie will detect and classify sensitive data using machine learning. Ideal for data lakes as it is integrated with S3. Fully integrated dashboards and alerting.
Infrastructure Zoning	VPC's provide network isolation within a region. Subnets provide network isolation within a VPC. Security Groups are firewall rulesets used to isolate compute instances. PrivateLink provides isolated connections to managed services like S3.
Data Syndication Security	Share data securely with partners through VPN connections to VPC, private peering into an AWS region, PrivateLink endpoints, SFTP, or SSL using your own certificates.
Cloud Zones for Sensitive Data	Use AWS regions for data locality. Data is never copied across regions unless initiated by the customer. Within regions use VPC to isolate infrastructure for processing sensitive data. Use KMS to manage keys used for encrypting sensitive data.
Decision Rights Management	Define policies and roles in IAM that are linked to LDAP. Federate authentication to identity providers like Okta. Use Organizations to put default policies and auditing in place.
Key Management	HSM for creating and storing your master keys and data encryption keys. Grant access to keys through IAM policies. Encrypt data at rest in all AWS storage and database services using cryptographic key material that only you control.

Security and Governance

	Athena	EMR	Glue	Redshift
Authentication	IAM/EC2 Key pair	Kerberos/LDAP/ EC2 Key pair/IAM	IAM Role	IAM/Native
Authorization	S3 Bucket Policies	S3 Bucket Policies/ Hive Grants/ EMRFS Auth	S3 Bucket Policies/ Fine Grained	S3 Bucket Policies/ Native Grants
Encryption of data at-rest	SSE-S3/ SSE-KMS/ CSE-KMS	SSE-S3/ SSE-KMS/ CSE-KMS/ CSE-CMK	SSE-S3	Database Encryption/ SSE-S3/ SSE-KMS/ CSE-CMK
Encryption of data in-transit	SSL	Yes, through Security Config	SSL	SSL
Audit	CloudTrail	Application Logs	CloudTrail	Database Audit
Compliance	HIPAA	FedRAMP/HIPAA	HIPAA	FedRAMP/HIPAA

AWS Services integrated with KMS - https://aws.amazon.com/kms/features/#AWS_Service_Integration

Compliance: Virtually every regulatory agency

Global

CSA
Cloud Security Alliance Controls

ISO 9001
Global Quality Standard

ISO 27001
Security Management Controls

ISO 27017
Cloud Specific Controls

ISO 27018
Personal Data Protection

PCI DSS Level 1
Payment Card Standards

SOC 1
Audit Controls Report

SOC 2
Security, Availability, & Confidentiality Report

SOC 3
General Controls Report

United States

CJIS
Criminal Justice Information Services

DoD SRG
DoD Data Processing

FedRAMP
Government Data Standards

FERPA
Educational Privacy Act

ISO FFIEC
Financial Institutions Regulation

FIPS
Government Security Standards

FISMA
Federal Information Security Management

GxP
Quality Guidelines and Regulations

HIPPA
Protected Health Information

ITAR
International Arms Regulations

MPAA
Protected Media Content

NIST
National Institute of Standards and Technology

SEC Rule 17a-4(f)
Financial Data Standards

VPAT/Section 508
Accountability Standards

Asia Pacific

FISC [Japan]
Financial Industry Information Systems

IRAP [Australia]
Australian Security Standards

K-ISMS [Korea]
Korean Information Security

MTCS Tier 3 [Singapore]
Multi-Tier Cloud Security Standard

My Number Act [Japan]
Personal Information Protection

Europe

C5 [Germany]
Operational Security Attestation

Cyber Essentials Plus [UK]
Cyber Threat Protection

G-Cloud [UK]
UK Government Standards

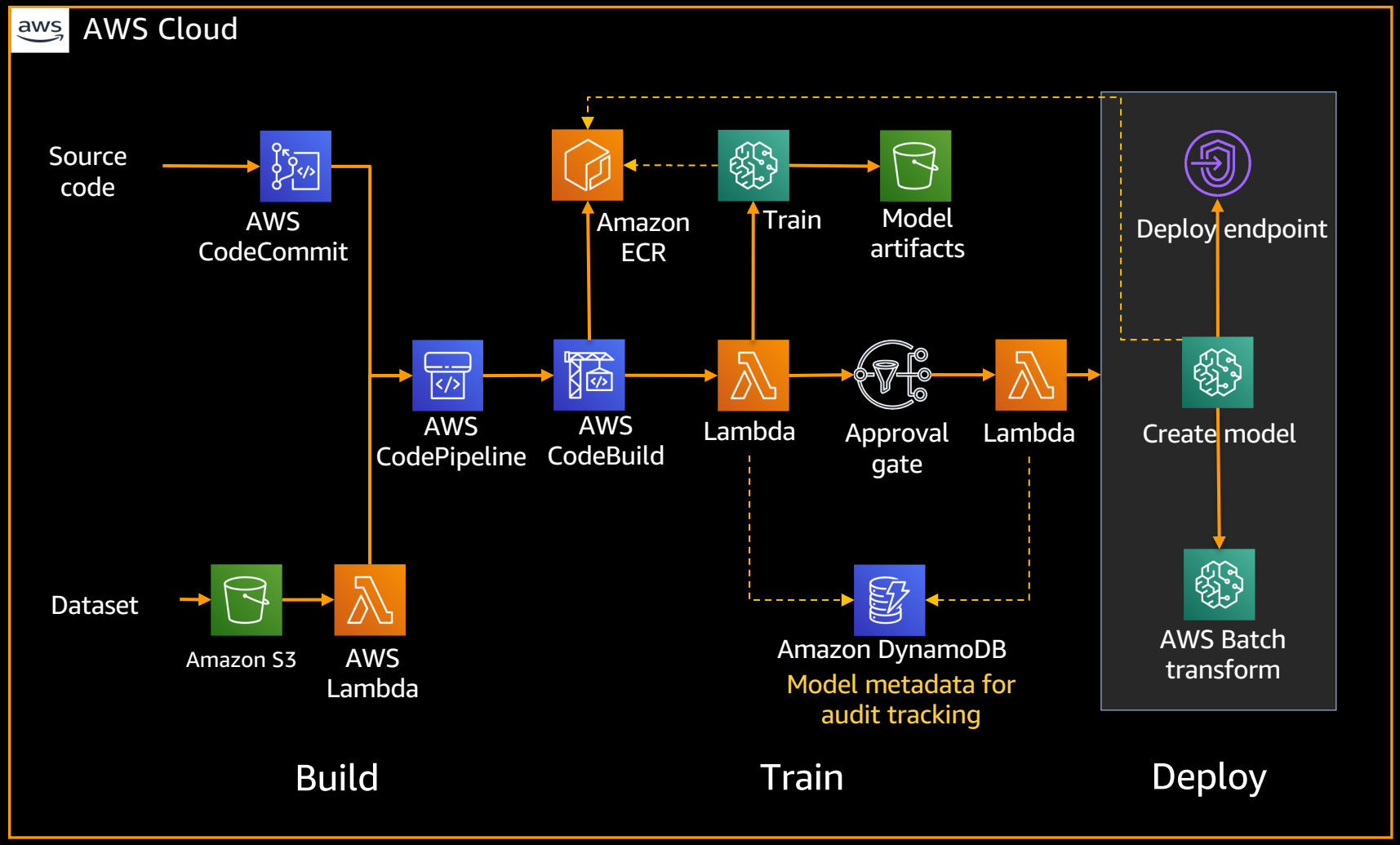
IT-Grundschutz [Germany]
Baseline Protection Methodology

HIPAA Eligible Services - https://aws.amazon.com/compliance/hipaa-eligible-services-reference/

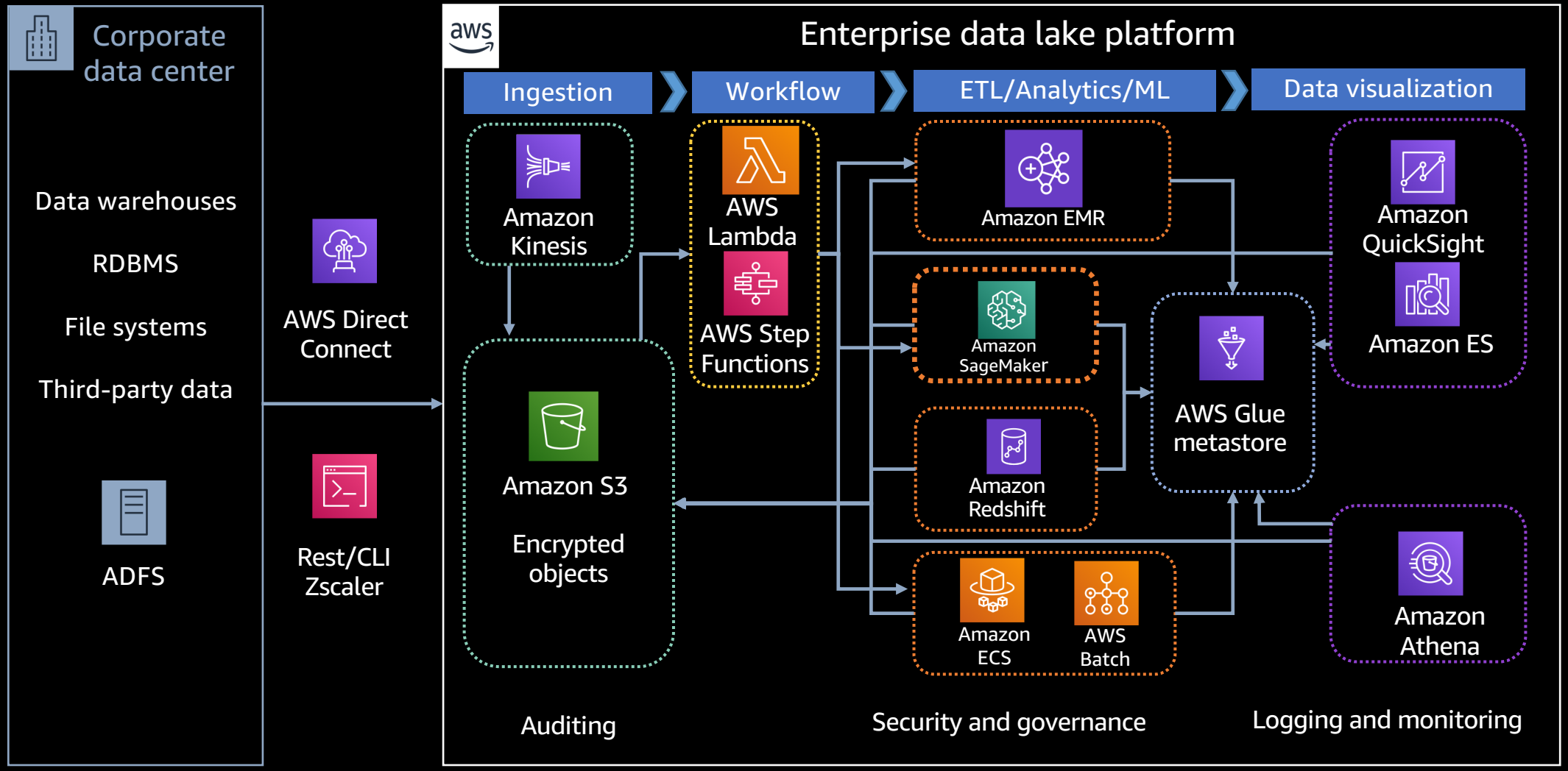
Consumption

Example: Machine learning orchestration with auditing

+ Reproducible and reusable pipeline
+ Built-in audit tracking capability
+ Other options: AWS Step Functions, Apache Airflow



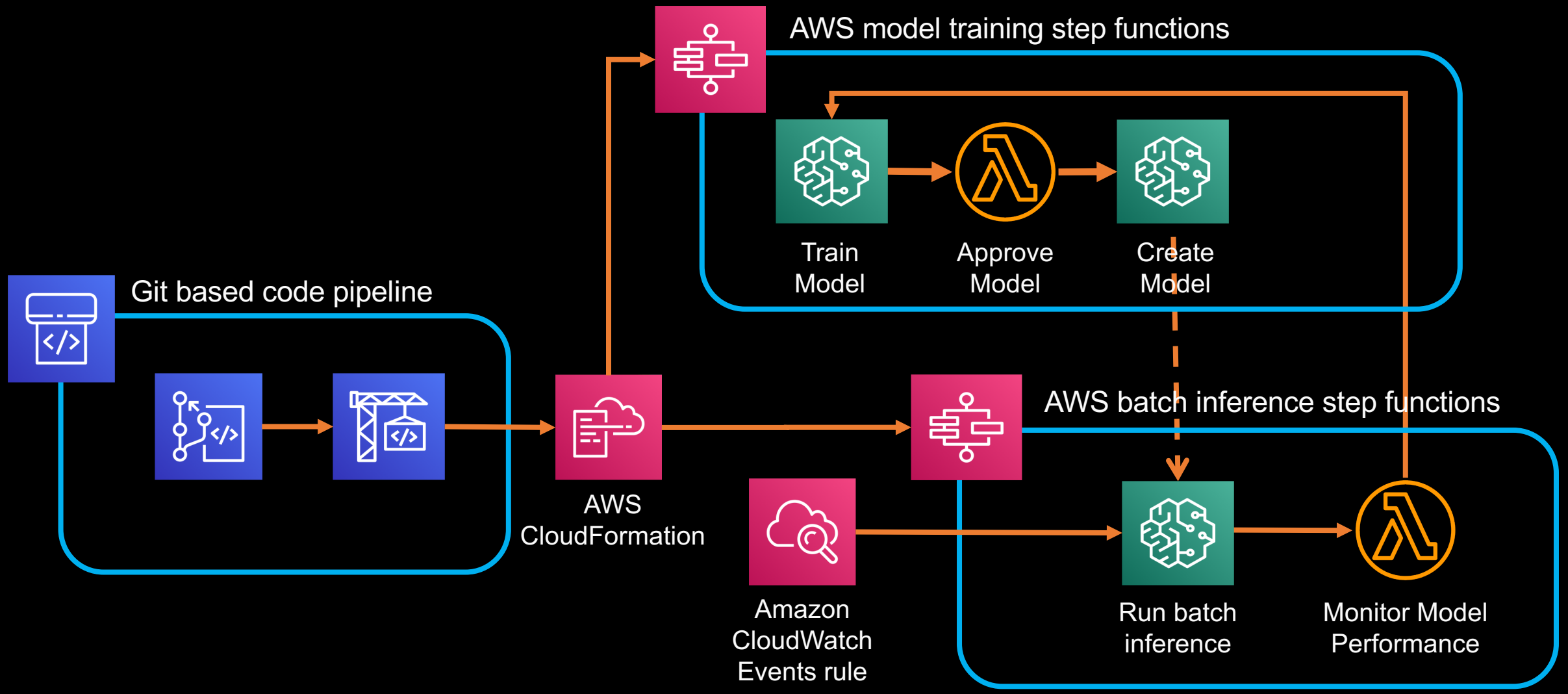
Example: SageMaker in EDL- Reference architecture



Platform built with 100% native AWS services => less integration challenges

AI/ML - MDLC

MDLC batch (offline) inference model workflow



Real-time (online) inference workflow