將數據轉化為商機

Turning Data into Insights July 30th

Jayson Hsieh Senior Solution Architect
Justine Peng Lead Development Representatives

Why Data Analytics ?

Forbes

“ Data is the fuel that powers AI & Digital Transformation. By making 10% more data accessible, a typical Fortune 1000 company will see a $65 million increase in net income. ”

Gartner®

$62.3B

MARKET OPPORTUNITY IN 2020

aws

Common analytics use cases – which do you need?

Manufacturing – Defect Detection / Performance Prediction

E commerce / Media – Recommendation

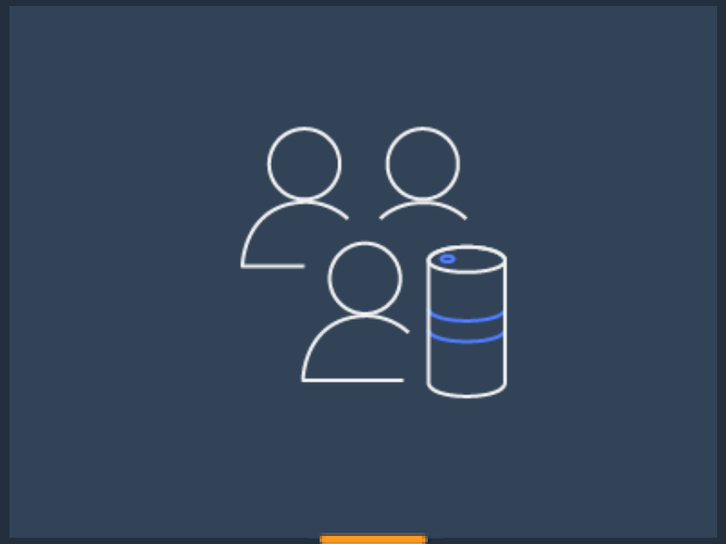
Retail – Data Warehouse Modernization, Logistic Management

Semiconductor – Machine Maintenance Prediction

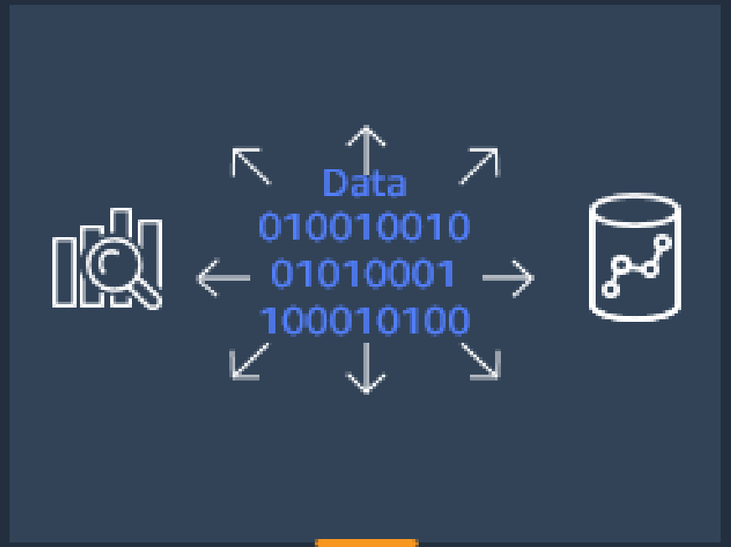
Gaming – User Behavior Analytics, Ad-hoc Customer Service

aws

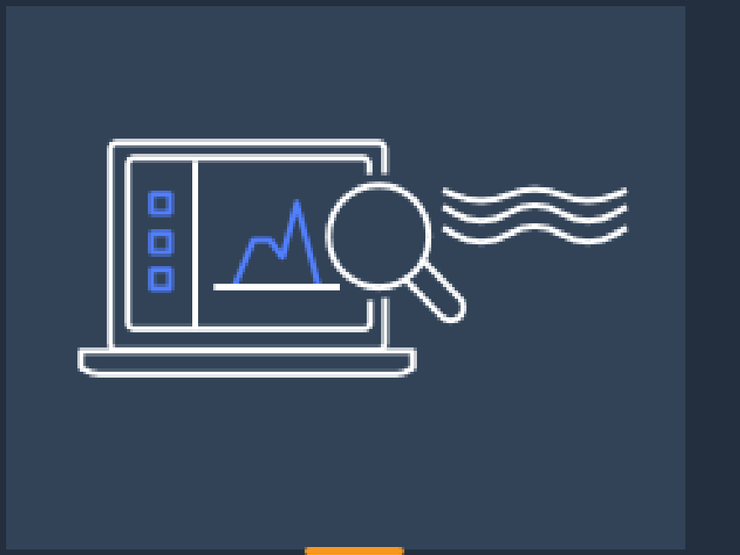
Why Data Analytics on cloud?



Maintenance of large scale data center



Exponential growth of event data



End-to-end insights from analyzing all your data

aws

Why and Why Not for Analytic on Cloud

Lower Operation Effort

Faster Time to Market

Improve System Reliability

Security

Cost Efficiency

No Domain Expert

aws

Consider Moving your Analytics Workload to AWS

aws

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.

aws

amazon.com

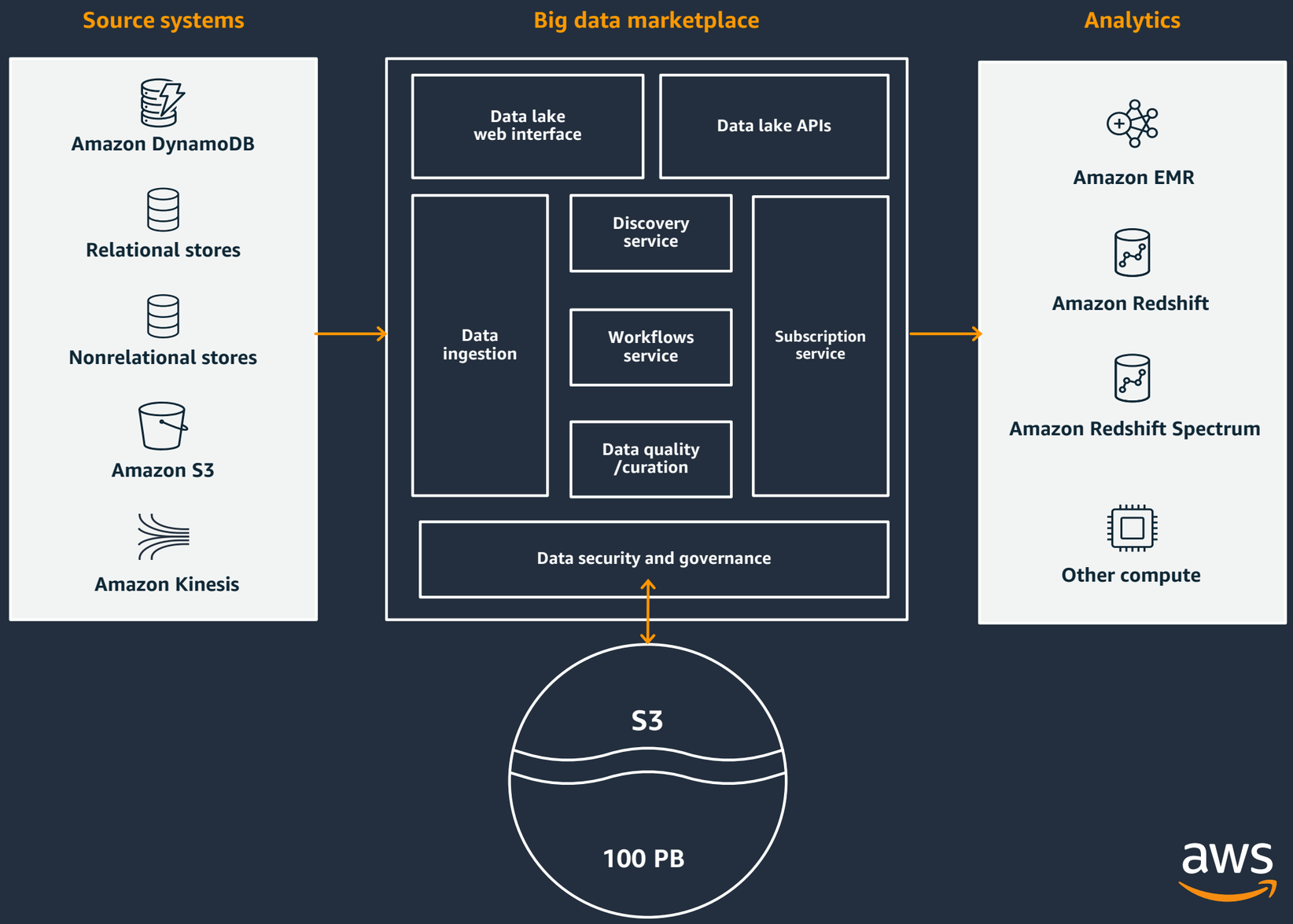
Amazon uses an AWS data lake

Solution

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

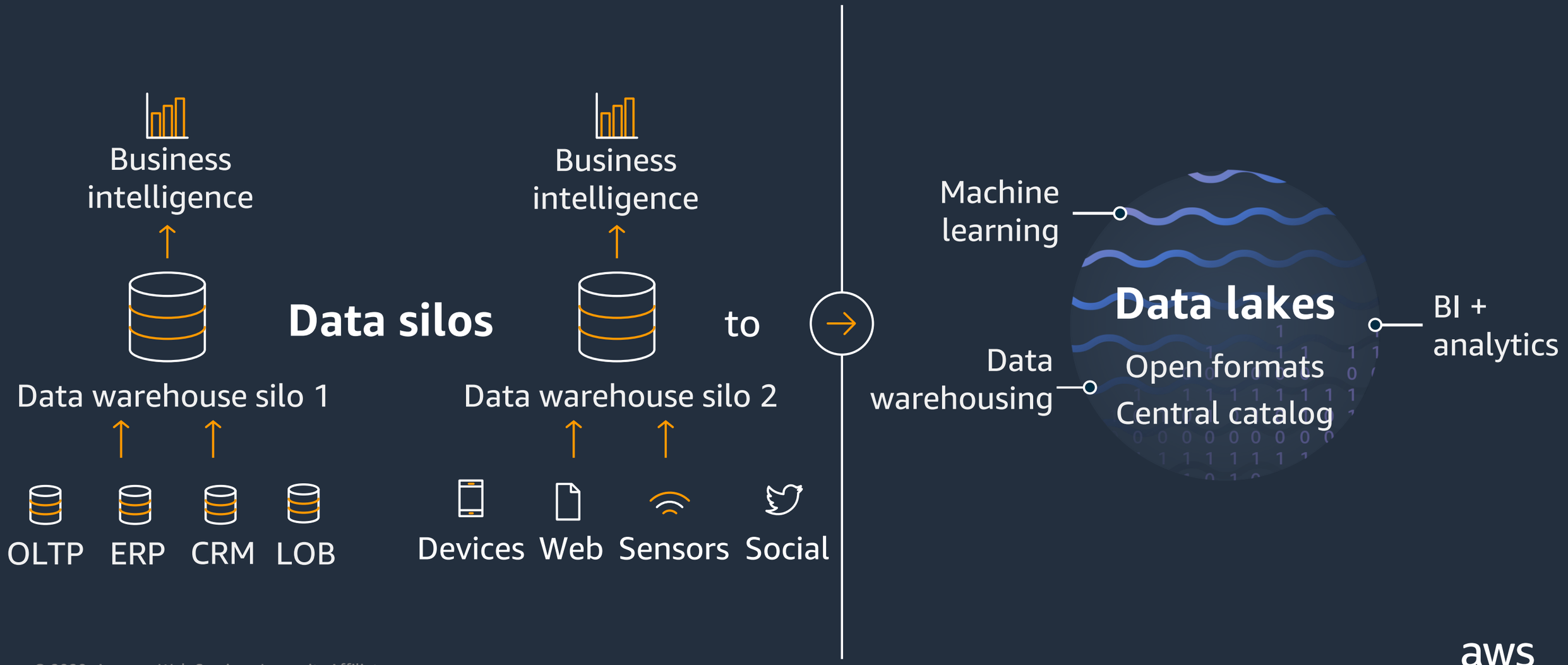
Benefits

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



aws

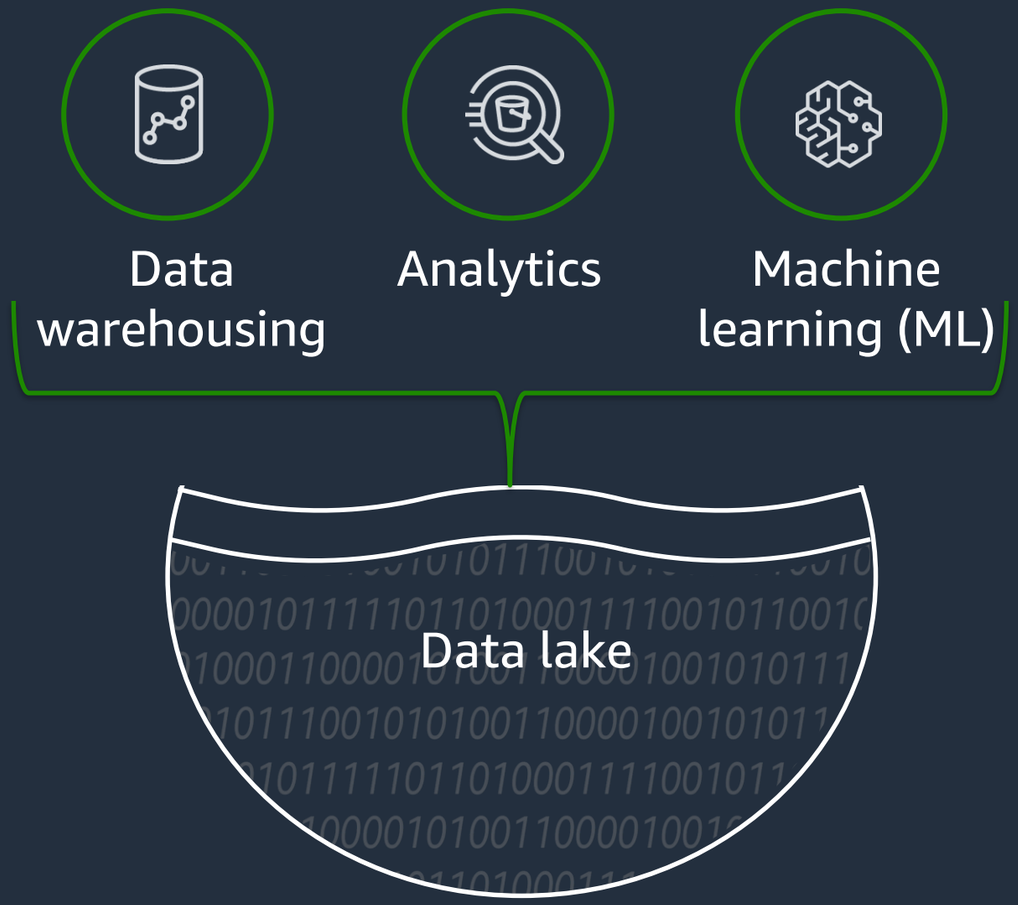
Traditional data warehousing approaches don’t scale



aws

Customers moving to data lake architectures

Bringing together the best of both worlds



aws

Why choose AWS for data lakes and analytics?

1

Easiest to build data lakes and analytics

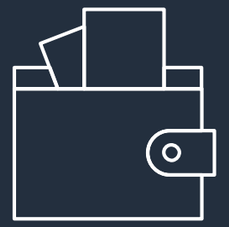
2

Most secure infrastructure for analytics

3

Most comprehensive and open

4



Most scalable and cost-effective

aws

1. Easiest to build data lakes and analytics

- 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)

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

aws

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

- Data
- Dashboards
- Digital user engagement
- Predictive analytics

Analytics

- Data warehousing
- Big data processing
- Serverless data processing
- Interactive query
- Operational analytics
- Real-time analytics

Data lake infrastructure & management

- Infrastructure
- Security & management
- Data catalog & ETL

Data movement

Migration & streaming services

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

aws

4. Most scalable, cost-effective infrastructure for analytics

Some examples of advanced capabilities in analytics services

Amazon EMR

Automatic scaling

57% less than on-premises per IDC report

Amazon Redshift

Less than 1/10 of the cost of traditional, on-premises solutions

Amazon Athena & Amazon QuickSight

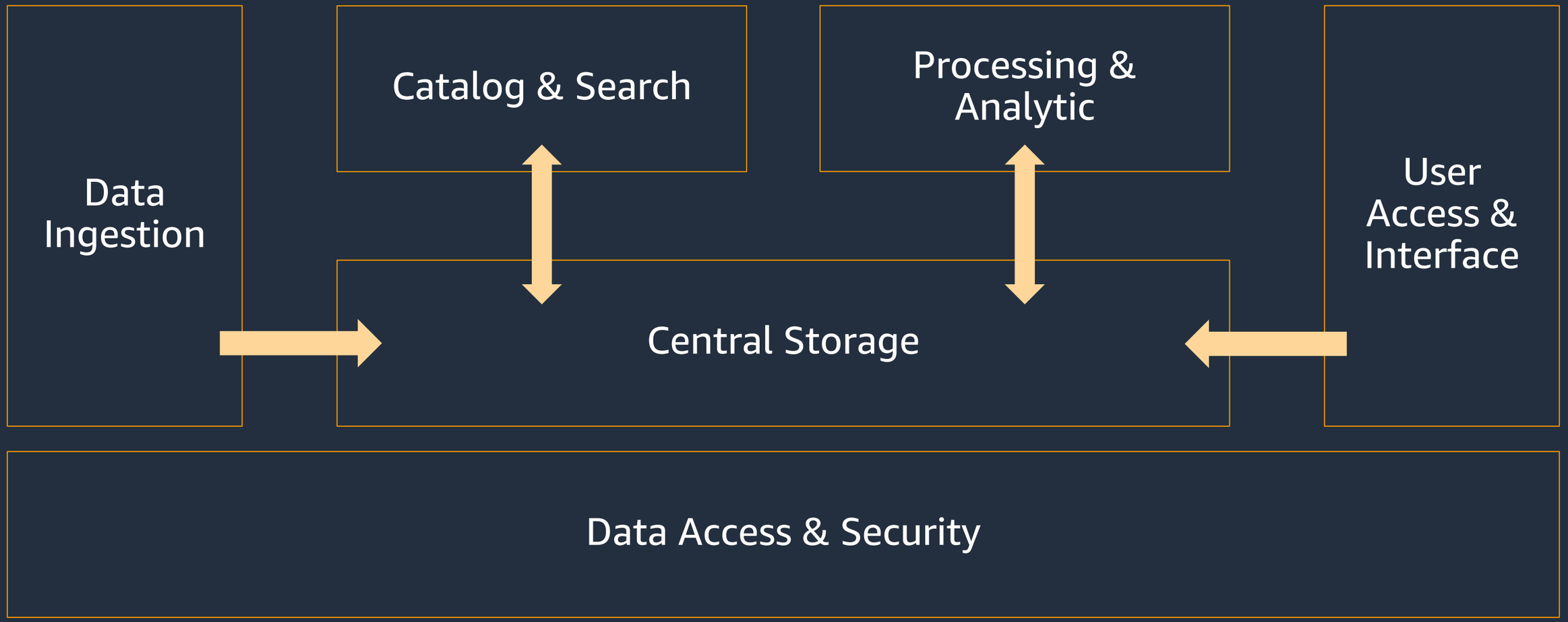
Serverless; pay only for what is used

Pricing per session for visualization

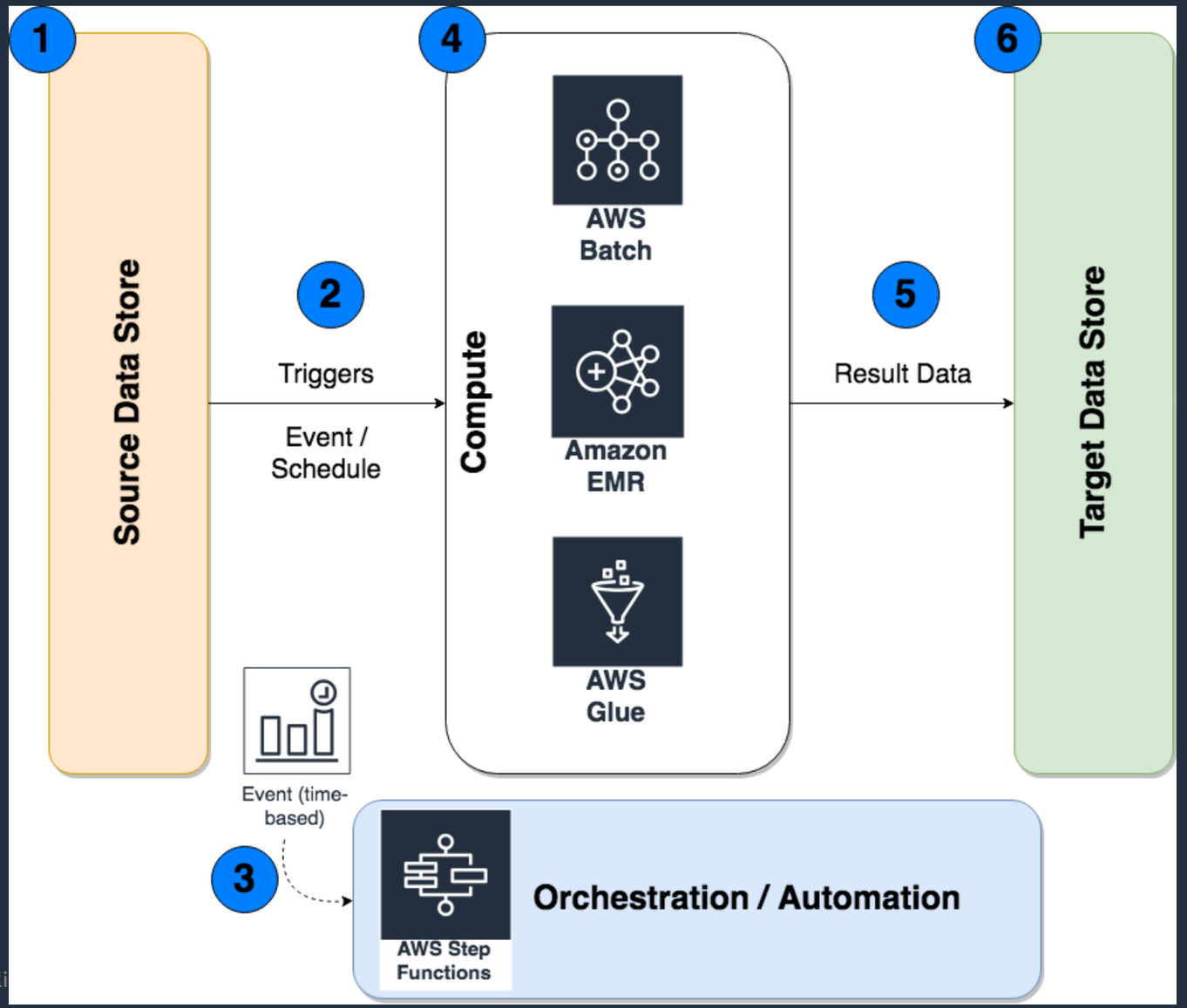
aws

Reference Architecture

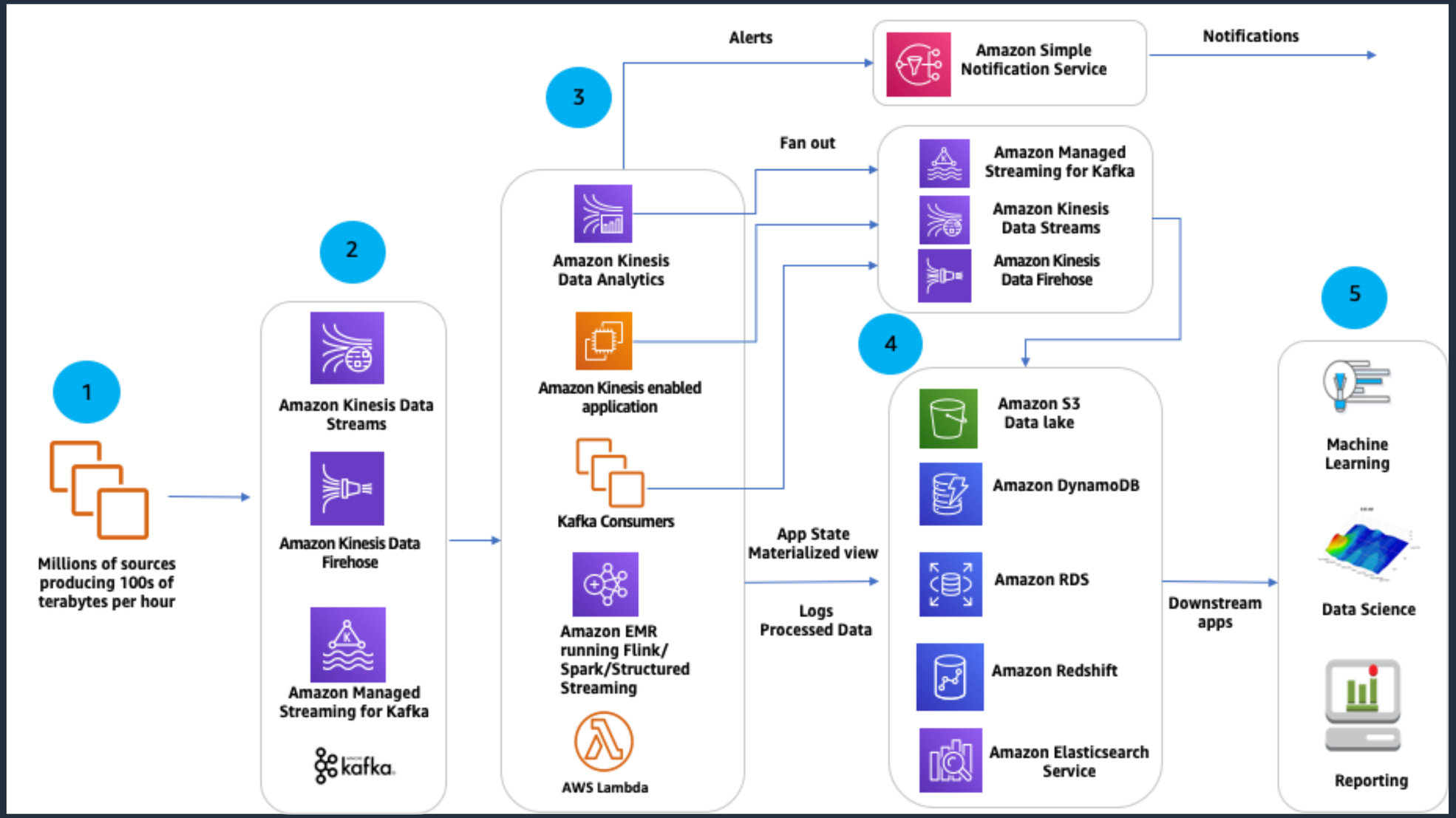
Data Architectures



Reference Architecture – Batch Data Processing

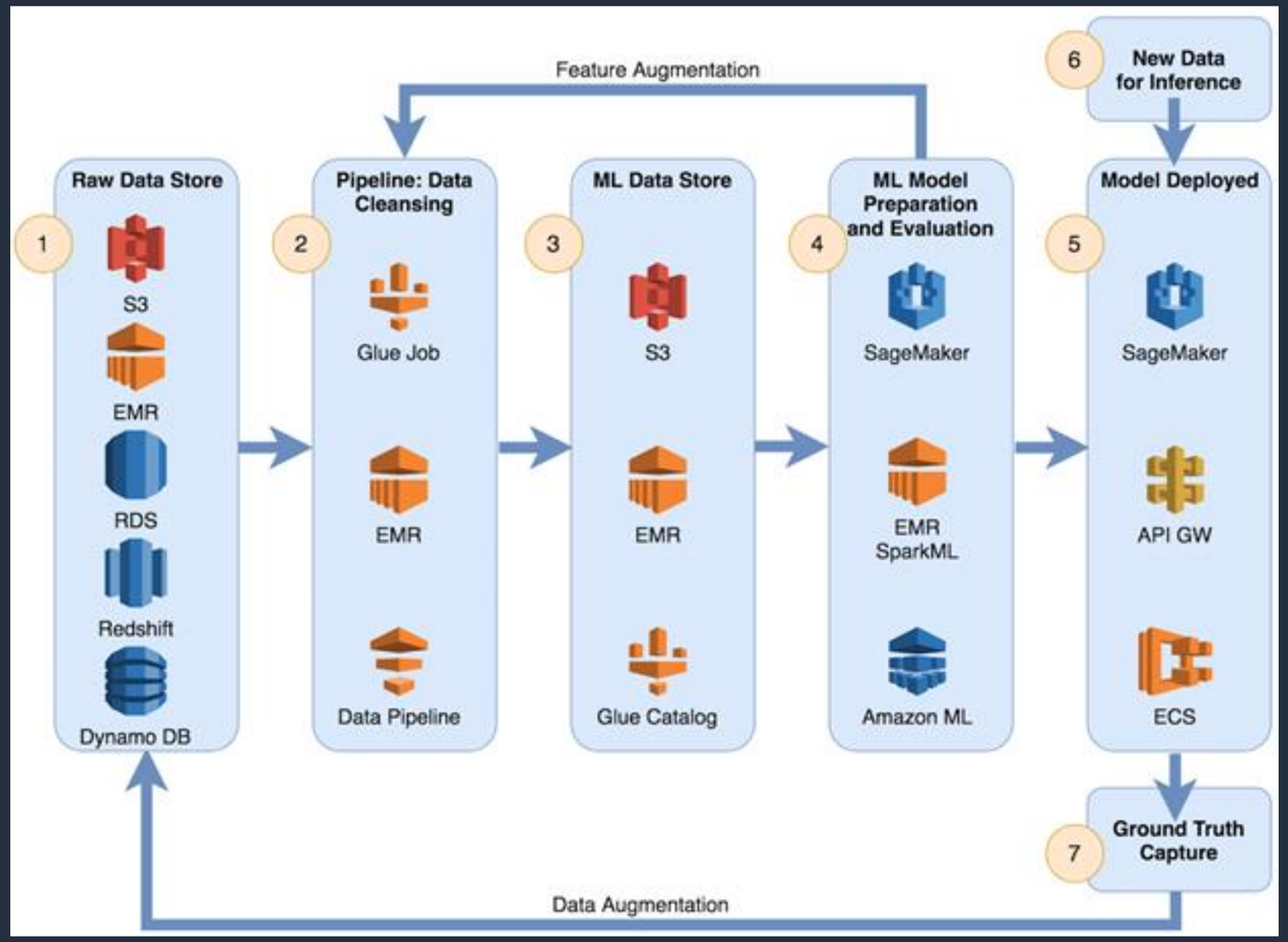


Reference Architecture – Streaming ingest & processing



aws

Reference Architecture – Data Science



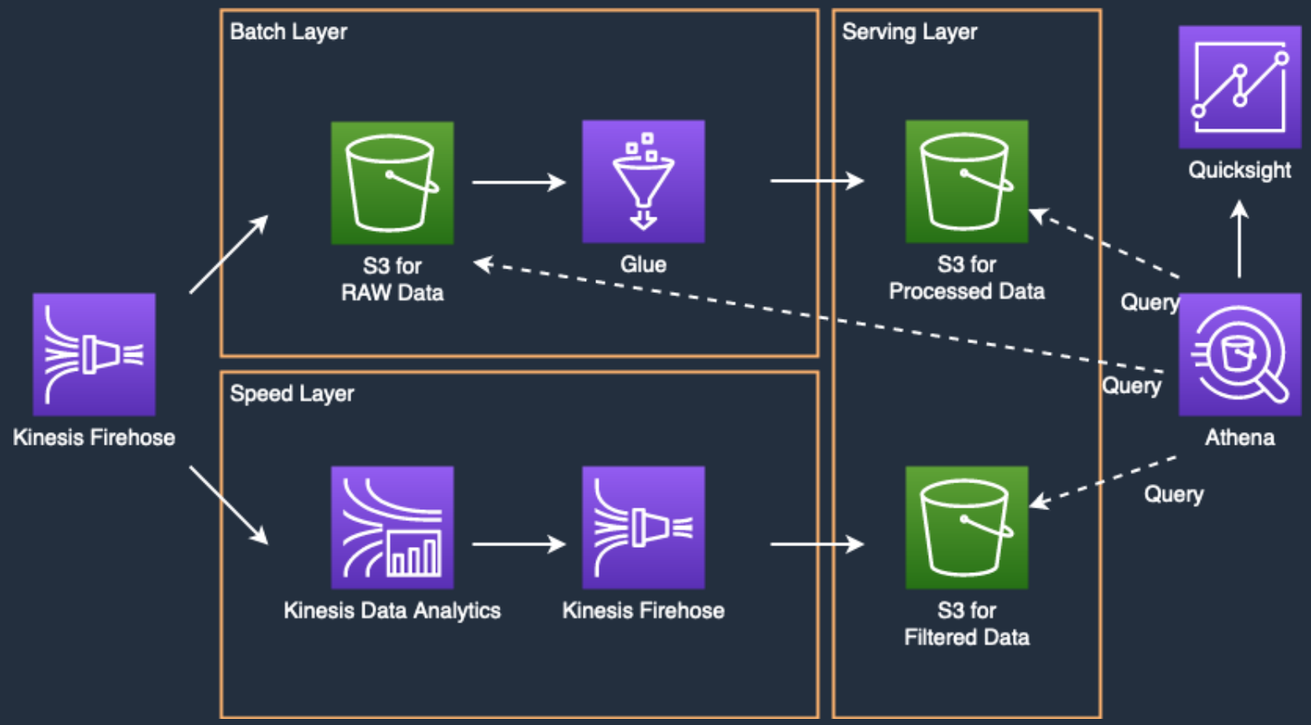
aws

Architecture Best Practices

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

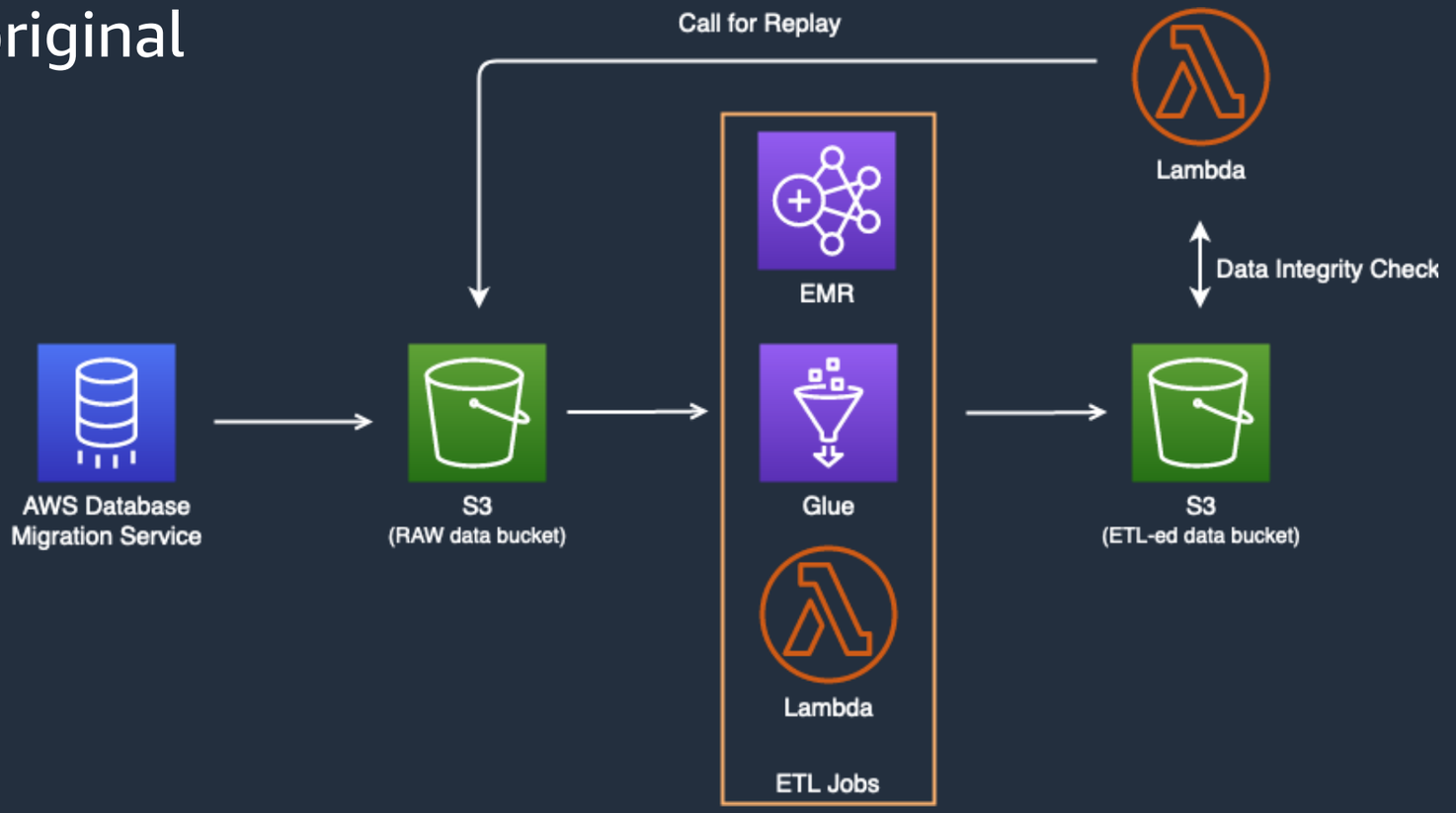


aws

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



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

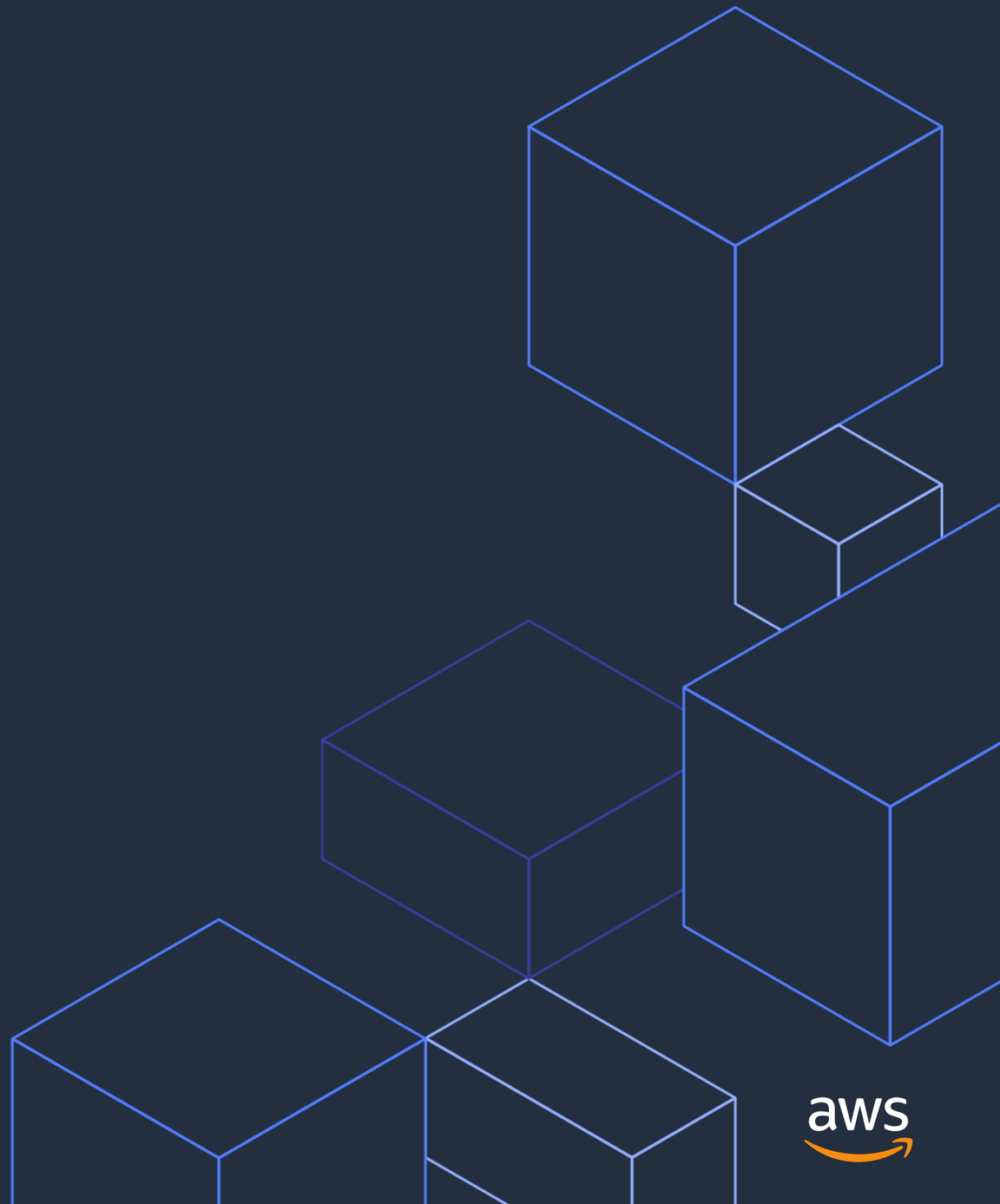
Use the right ETL tool for the job

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.

aws

AWS Can Help



What's Next?

Free Resources

- Webinars
- Workshop
- Best Practices

Training Courses

- Digital Training
- Hands on Lab

Programs

- Customer Innovation Program
- ML Labs
- Data Labs

aws

Q&A

Remember to complete your evaluations!

Thank You

aws