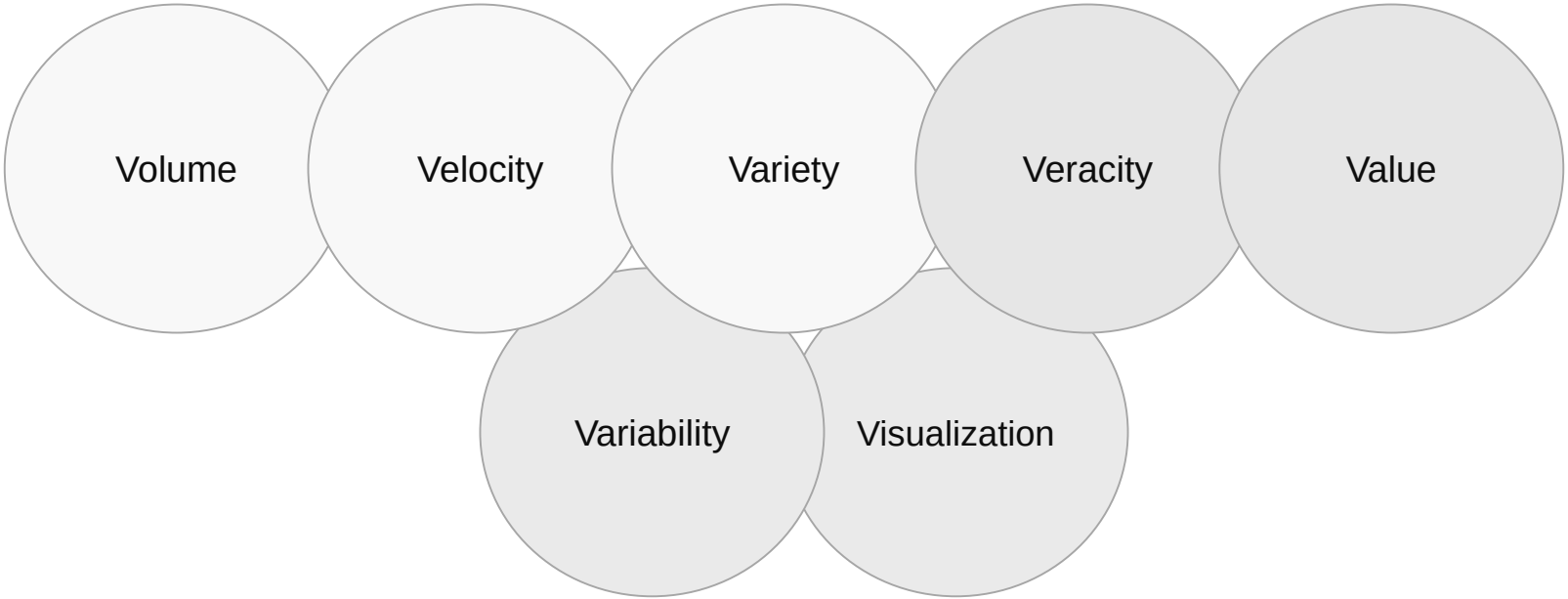
BDA305

Building Data Lakes and Analytics on AWS

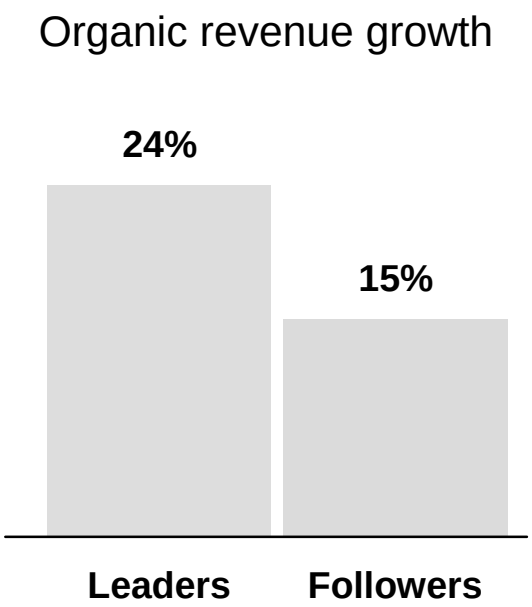
Ben Snively,

Specialist Solutions Architect, Amazon Web Services

Big Data Is Defined Many Different Ways



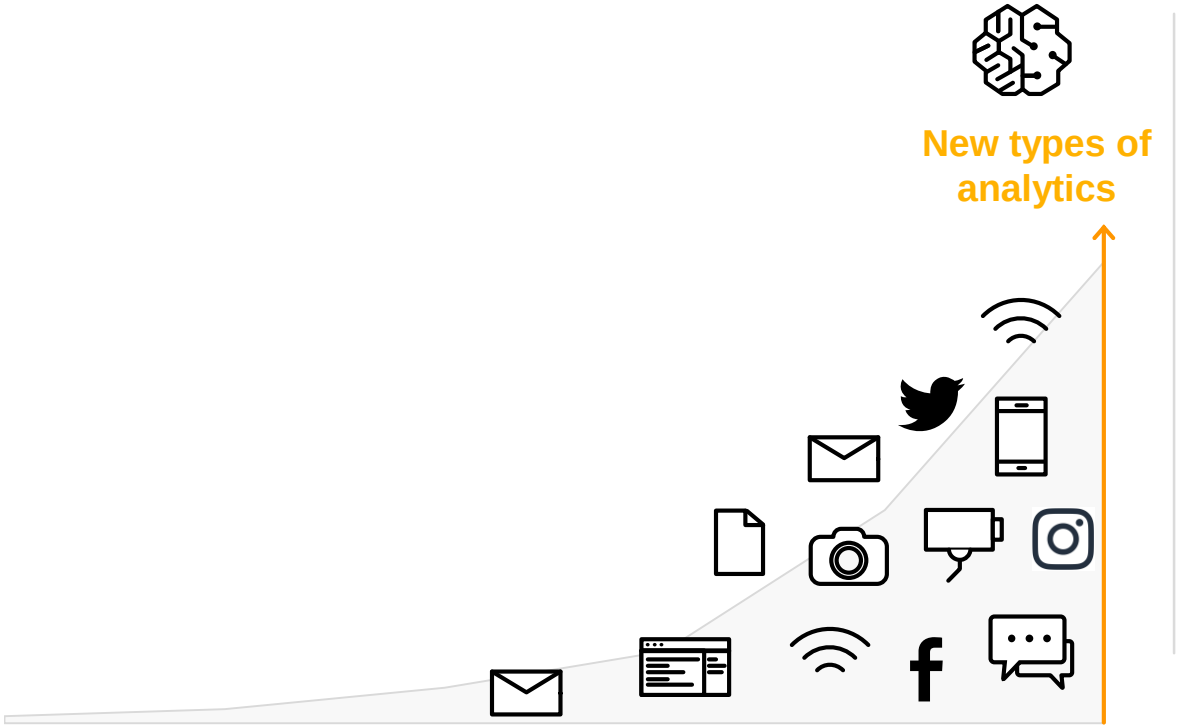
Most Important: Driving Value from Data



Organizations that successfully generate business value from their data will outperform their peers. An Aberdeen survey saw organizations who implemented a data lake outperforming similar companies by 9% in organic revenue growth.*

*Aberdeen: Angling for Insight in Today's Data Lake, Michael Lock, SVP Analytics and Business Intelligence

Data Is Changing → Analytics Are Adopting



Capture and store new data at PB-EB scale

Do new type of analytics in a cost effective way

- Machine learning
- Big data processing
- Real-time analytics
- Full-text search

Customers Are Doing This Today

FINRA oversees >3,000 securities firms doing business in the United States.

Challenge:

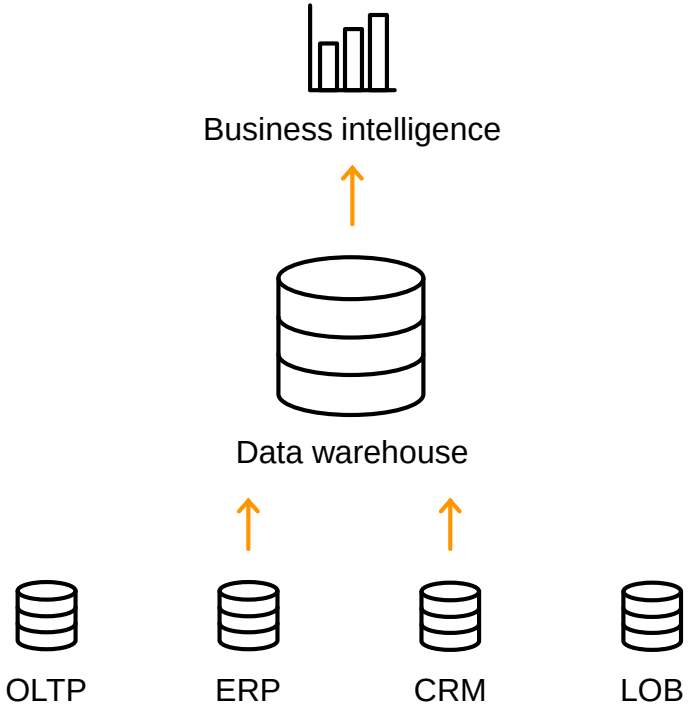
FINRA's legacy system did not scale well

- Up to 75 billion events per day
- Run complex surveillance queries over 20 PB of data

Solution:

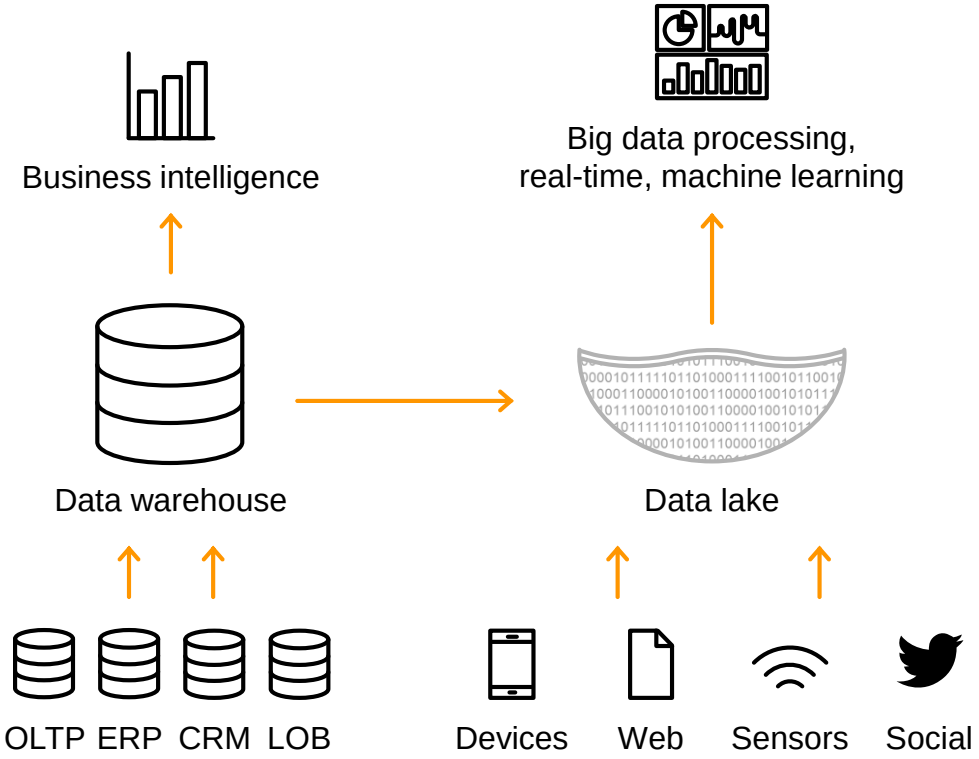
- Migrated their big data appliance to an S3 data lake and used EMR for ingestion and processing
- Migrated to RDS and testing Aurora

Traditionally, Analytics Used to Look Like This



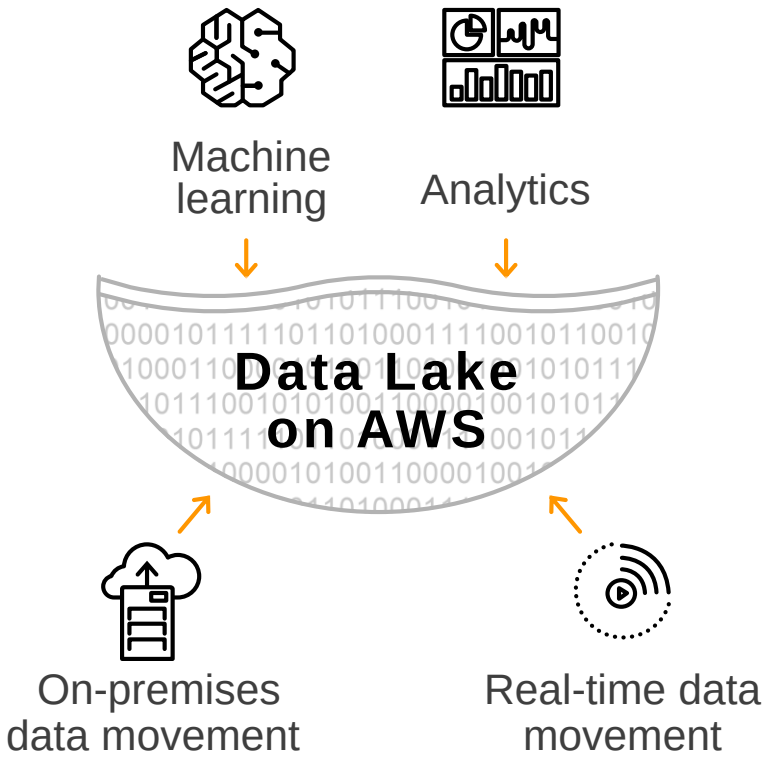
- Relational data
- TBs–PBs scale
- Schema defined prior to data load
- Operational reporting and ad hoc
- Large initial CAPEX + $10K–$50K/TB/year

Data Lakes Extend the Traditional Approach



- Relational and nonrelational data
- TBs–EBs scale
- Diverse analytical engines
- Low-cost storage & analytics

Data Lakes from AWS



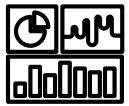
- Unmatched durability, and availability at EB scale
- Best security, compliance, and audit capabilities
- Object-level controls for fine-grain access
- Fastest performance by retrieving subsets of data
- The most ways to bring data in
- 2x as many integrations with partners
- Analyze with broadest set of analytics & ML services

Data Lakes, Analytics, and IoT Portfolio from AWS

Broadest, deepest set of analytic services

Machine learning

Managed ML Service
Deep Learning AMIs
Video and Image Recognition
Conversational Interfaces
Deep-Learning Video Camera
Natural Language Processing
Language Translation
Speech Recognition
Text-to-Speech



Analytics

Interactive Analysis
Hadoop & Spark
Data Warehousing
Full-text search
Real-time analytics
Dashboards & Visualizations

On-premises data movement

Dedicated Network connection
Secure appliances
Ruggedized Shipping Container
Database migration

Real-time data movement

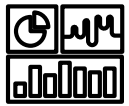
Connect Devices to AWS
Real-time Data Streams
Real-time Video Streams

Data Lakes, Analytics, and IoT Portfolio from AWS

Broadest, deepest set of analytic services

Machine learning

Amazon SageMaker
AWS Deep Learning AMIs
Amazon Rekognition
Amazon Lex
AWS DeepLens
Amazon Comprehend
Amazon Translate
Amazon Transcribe
Amazon Polly



Analytics

Amazon Athena
Amazon EMR
Amazon Redshift
Amazon Elasticsearch Service
Amazon Kinesis
Amazon QuickSight

On-premises data movement

AWS Direct Connect
AWS Snowball
AWS Snowmobile
AWS Database Migration Service

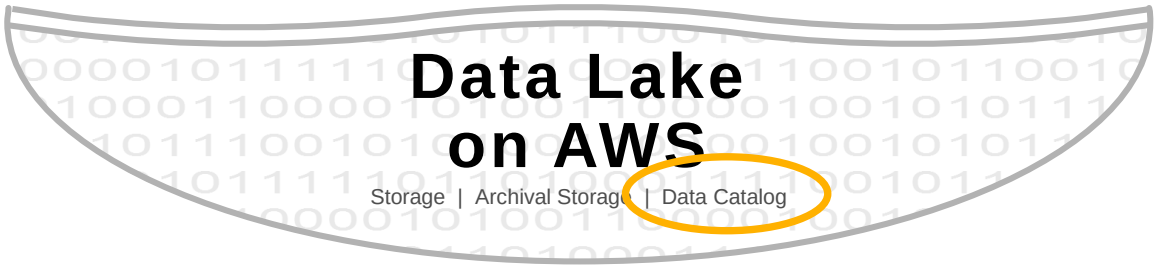
Real-time data movement

AWS IoT Core
Amazon Kinesis Data Firehose
Amazon Kinesis Data Streams
Amazon Kinesis Video Streams

What Data Do I Have?

Gartner:

"*Through 2018, 80% of data lakes will not include effective metadata management capabilities, making them inefficient*."



AWS Glue

Data Catalog

Discover

Apache Hive Metastore compatible

Integrated with AWS services

Automatic crawling

Job Authoring

Develop

Auto-generates ETL code

Python and Apache Spark

Edit, debug, and share

Job Execution

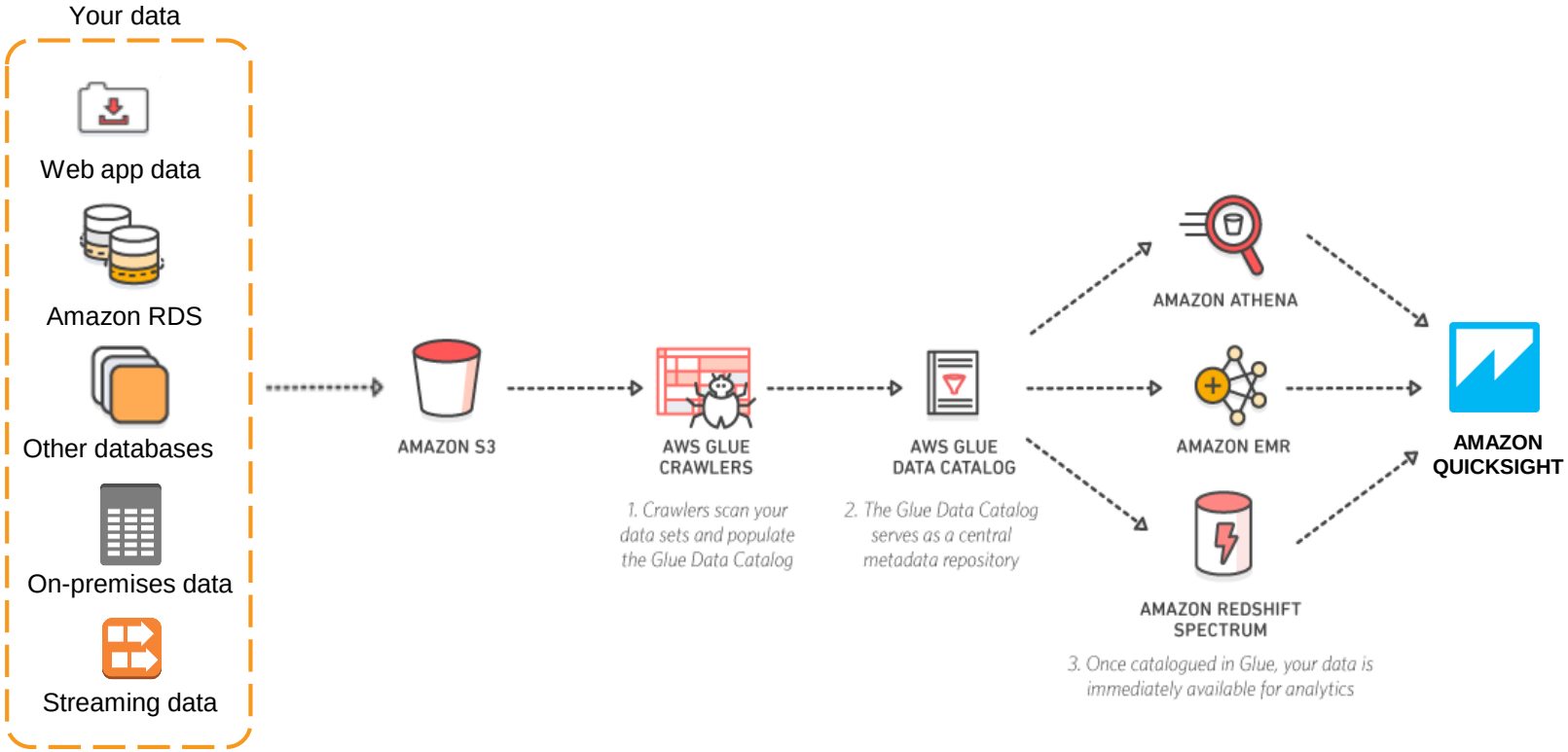
Deploy

Serverless execution

Flexible scheduling

Monitoring and alerting

Data Lake on Amazon S3 with AWS Glue



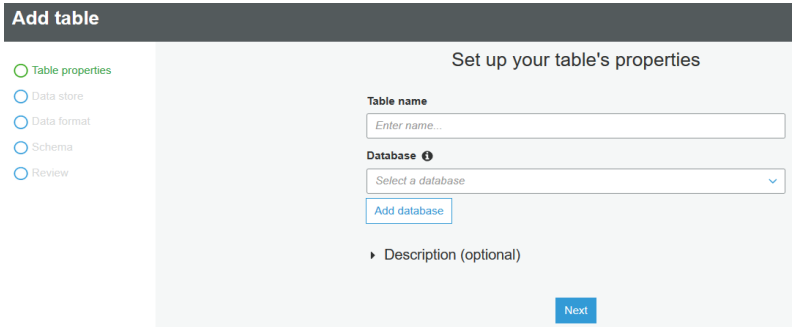
aws SUMMIT

Demonstration

Let’s Discover New Data in Our Data Lake

Other Ways of Populating the Catalog

Create table manually



Run Hive DDL statement

```
CREATE EXTERNAL TABLE IF NOT EXISTS elb_logs_raw_native_part (
  request_timestamp string,
  elb_name string,
  request_ip string,
  request_port int,
  backend_ip string,
  backend_port int,
  request_processing_time double,
  backend_processing_time double,
  client_response_time double,
  elb_response_code string,
  backend_response_code string,
  received_bytes bigint,
  sent_bytes bigint,
  request_verb string,
  url string,
  protocol string,
  user_agent string,
  ssl_cipher string,
  ssl_protocol string )
PARTITIONED BY(year string, month string, day string)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
WITH SERDEPROPERTIES (
        'serialization.format' = '1','input.regex' = '([^ ]*) ([^ ]*) ([^ ]*):([0-9]*) ([^ ]*):([0-9]*) ([.0-9]*) ([.0-
LOCATION 's3://athena-examples/elb/raw/';
```

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Run Query | Save As | Format Query | New Query (Run time: 2.03 seconds, Data scanned: 0KB)

Call the AWS Glue CreateTable API

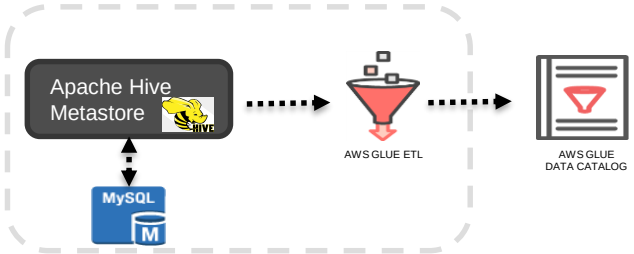
CreateTable

Creates a new table definition in the Data Catalog.

Request Syntax

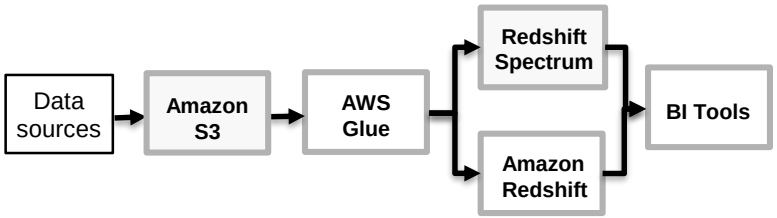
```
{
   "CatalogId": "string",
   "DatabaseName": "string",
   "TableInput": {
      "Description": "string",
```

Import from Apache Hive Metastore



NUVIAD—Data Lake Analytics with Redshift Spectrum

NUVIAD is a marketing platform that helps media buyers optimize their mobile bidding



Use AWS for marketing campaign and bidding analytics

Scale Amazon S3 storage for unlimited data capacity

Use Spectrum for unlimited scale and query concurrency

80% performance gain using parquet data format

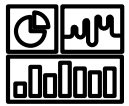
*"**Amazon Redshift Spectrum is a game changer for us.** Reports that took minutes to produce are now delivered in seconds. We like the ability scale compute on-demand to query petabytes of data in S3 in various open file formats."*

-- Rafi Ton, CEO, NUVIAD

How Do I Drive Value?

Machine learning

Amazon SageMaker
AWS Deep Learning AMIs
Amazon Rekognition
Amazon Lex
AWS DeepLens
Amazon Comprehend
Amazon Translate
Amazon Transcribe
Amazon Polly



Analytics

Amazon Athena
Amazon EMR
Amazon Redshift
Amazon Elasticsearch Service
Amazon Kinesis
Amazon QuickSight

Data Lake on AWS

Storage | Archival Storage | Data Catalog

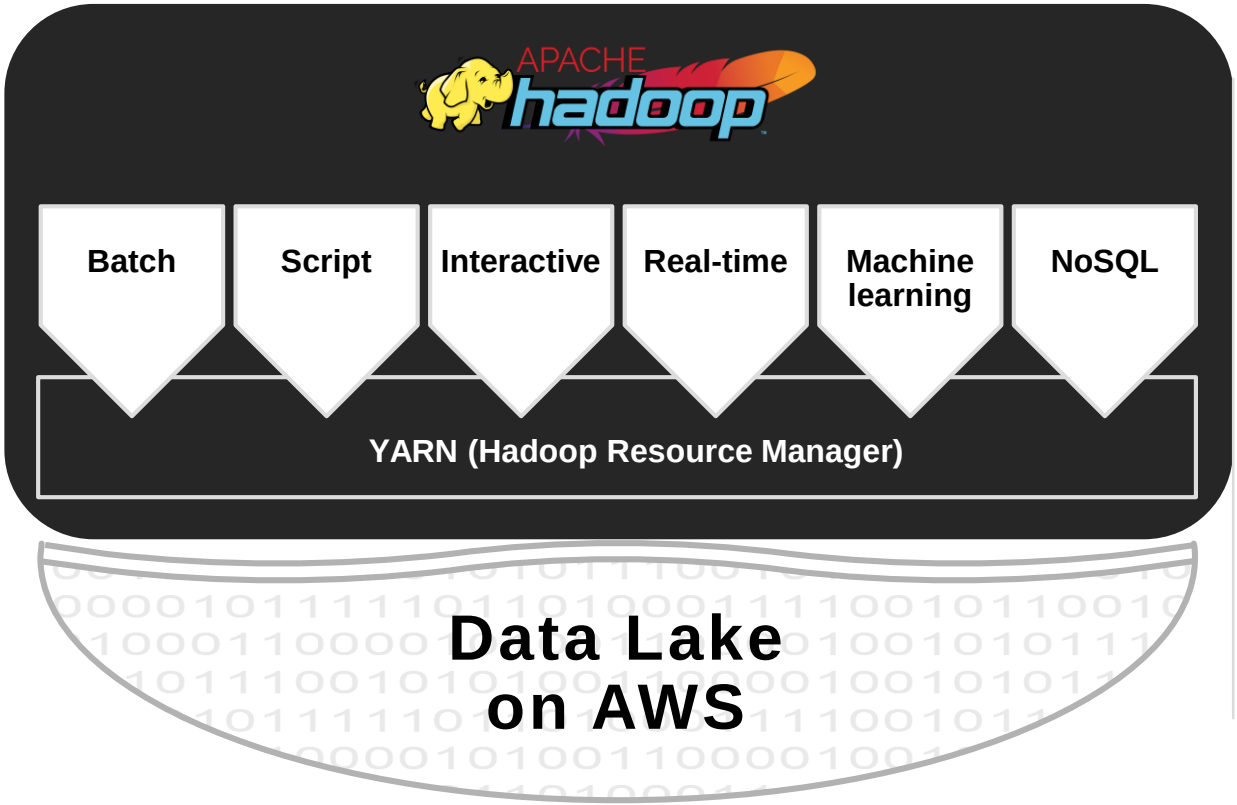
On-premises data movement

AWS Direct Connect
AWS Snowball
AWS Snowmobile
AWS Database Migration Service

Real-time data movement

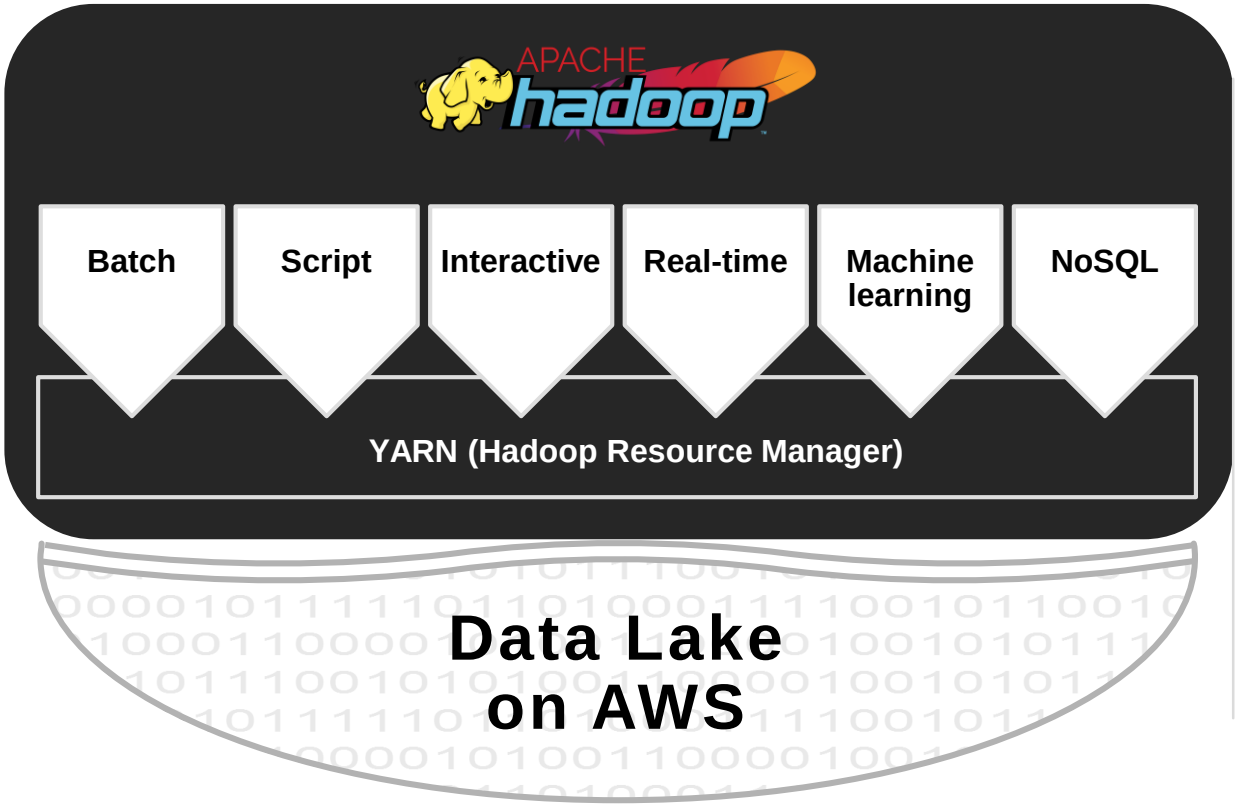
AWS IoT Core
Amazon Kinesis Data Firehose
Amazon Kinesis Data Streams
Amazon Kinesis Video Streams

Hadoop/Spark Analytics



- Distributed processing
- Diverse analytics
  - Batch/Script (Hive/Pig)
  - Interactive (Spark, Presto)
  - Real-time (Spark)
  - Machine Learning (Spark)
  - NoSQL (HBase)
- For many use cases
  - Log and clickstream analysis
  - Machine learning
  - Real-time analytics
  - Large-scale analytics
  - Genomics
  - ETL

Hadoop/Spark Analytics on AWS



Amazon EMR

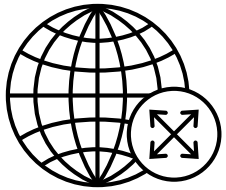
Managed Hadoop/Spark

Amazon S3

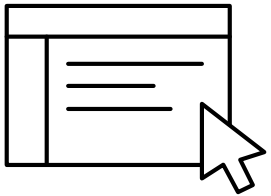
Object Storage

Introducing Amazon EMR

Managed Hadoop and Spark in the cloud at 1/8th the cost



Enterprise-grade



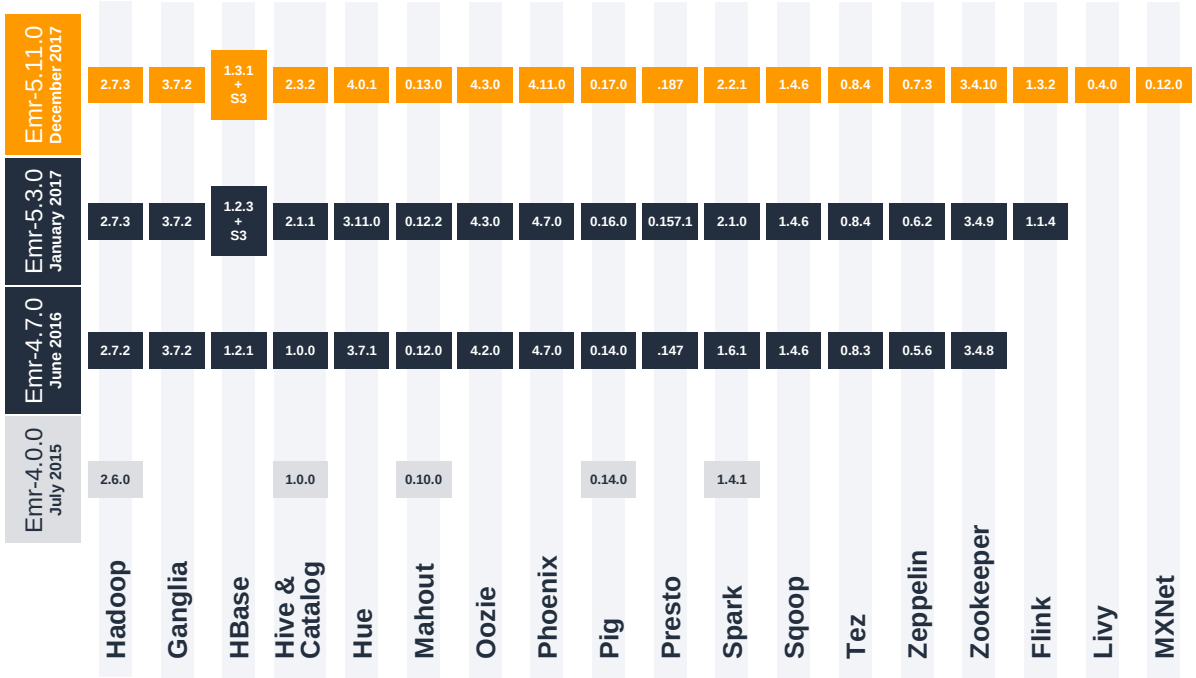
Easy

Lowest cost

EMR – Enterprise-grade Hadoop & Spark

Deploy latest releases in Hadoop and Spark ecosystems

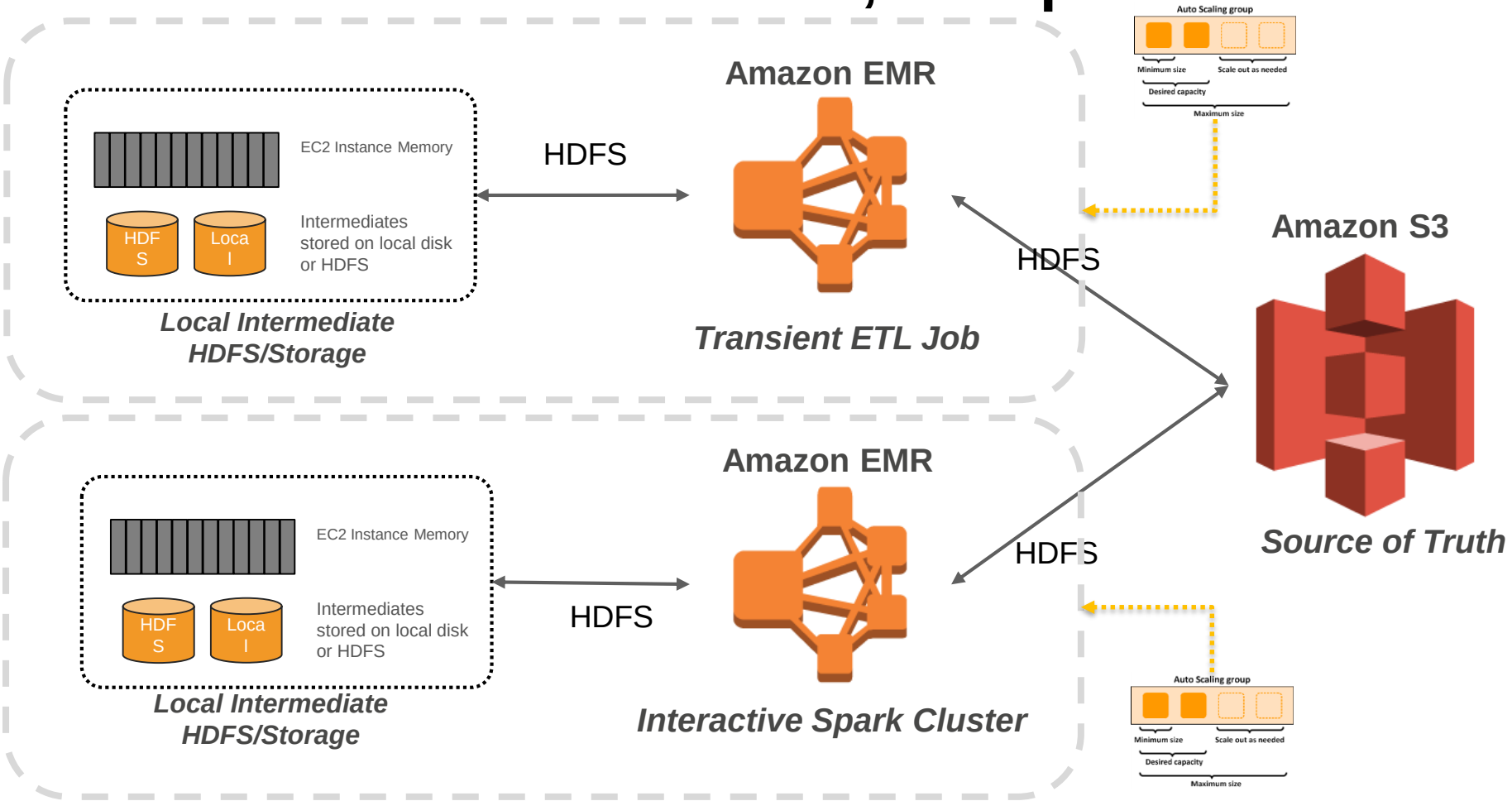
EMR releases



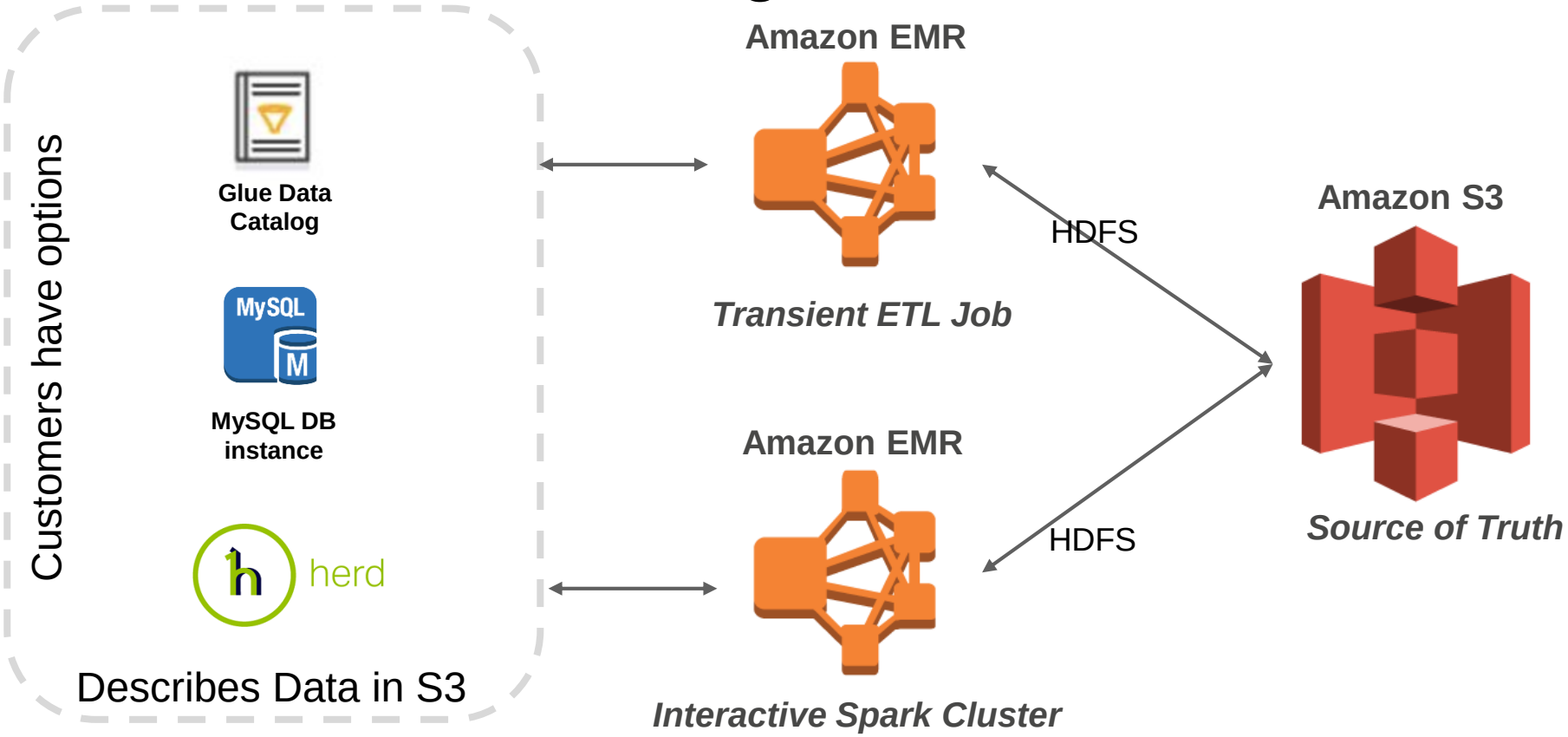
- Nineteen open-source projects: Apache Hadoop, Spark, HBase, Presto, and more
- Updated with the latest open source frameworks within 30 days of release

aws

Amazon S3 – Source of Truth, Multiple Clusters



External Metadata Management



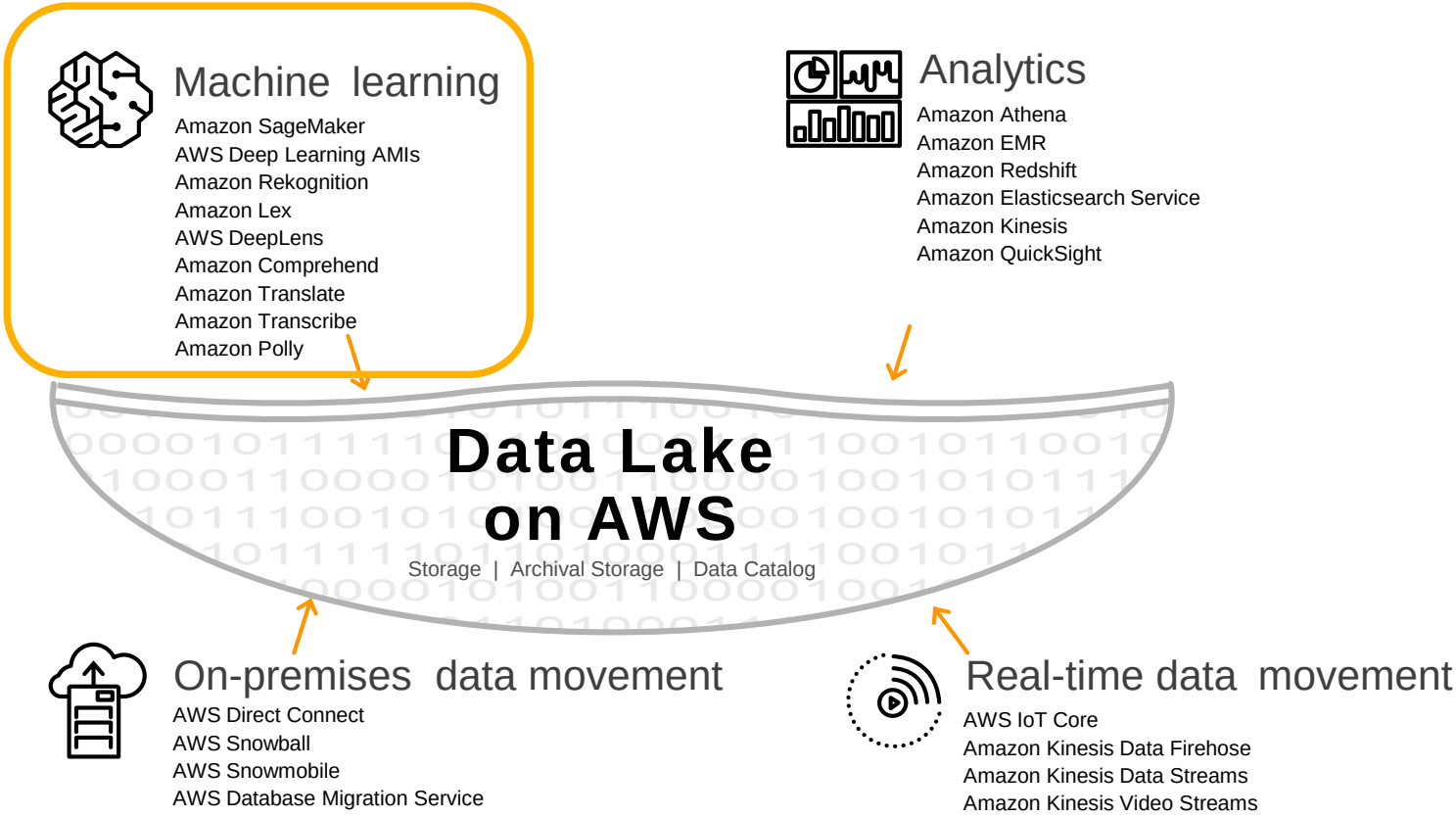
Amazon Athena is an interactive query service that makes it easy to analyze data directly from Amazon S3 using Standard SQL

aws SUMMIT

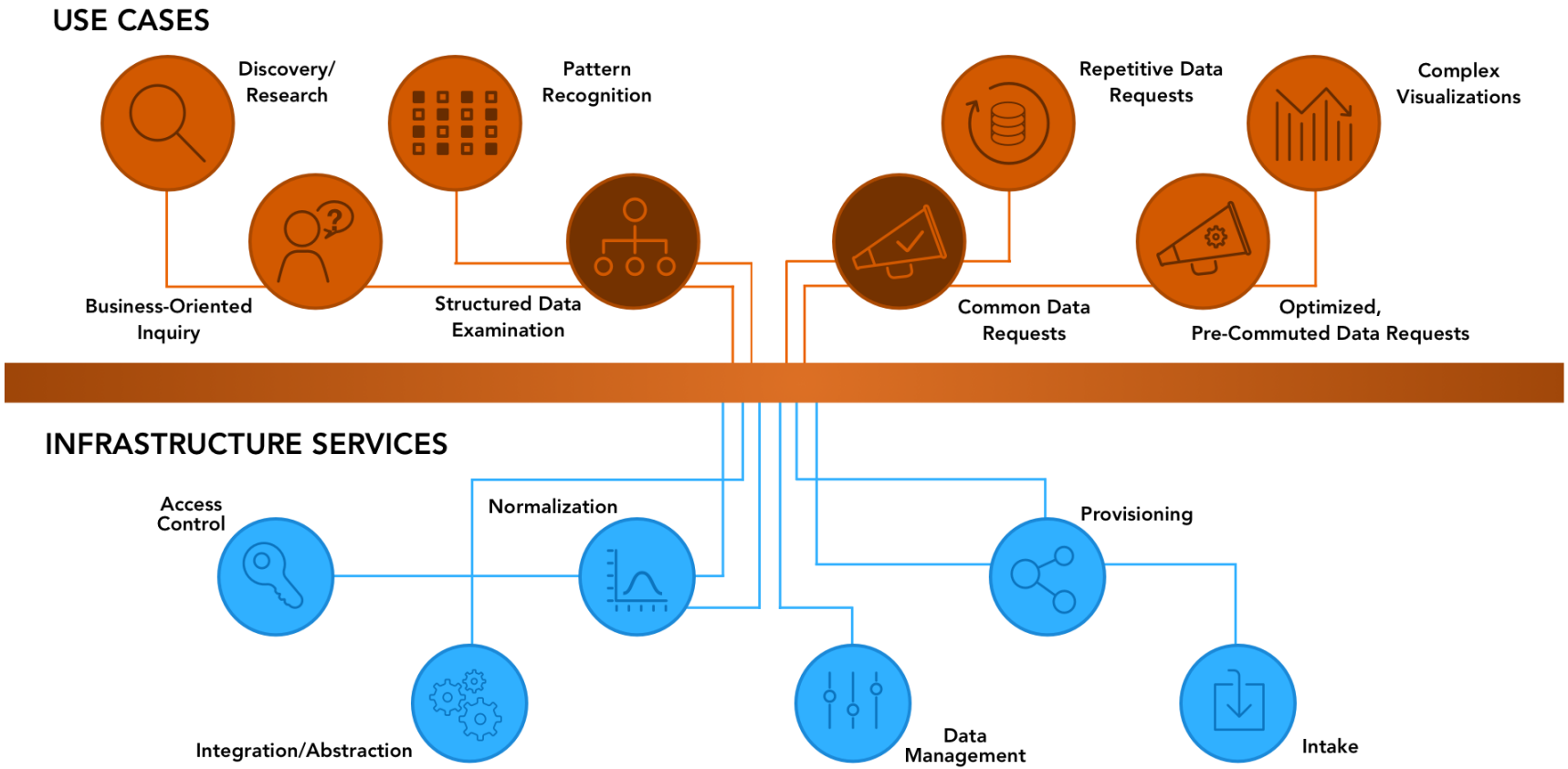
Demonstration
Running Hadoop/Spark Analytics

Amazon EMR and Amazon Athena

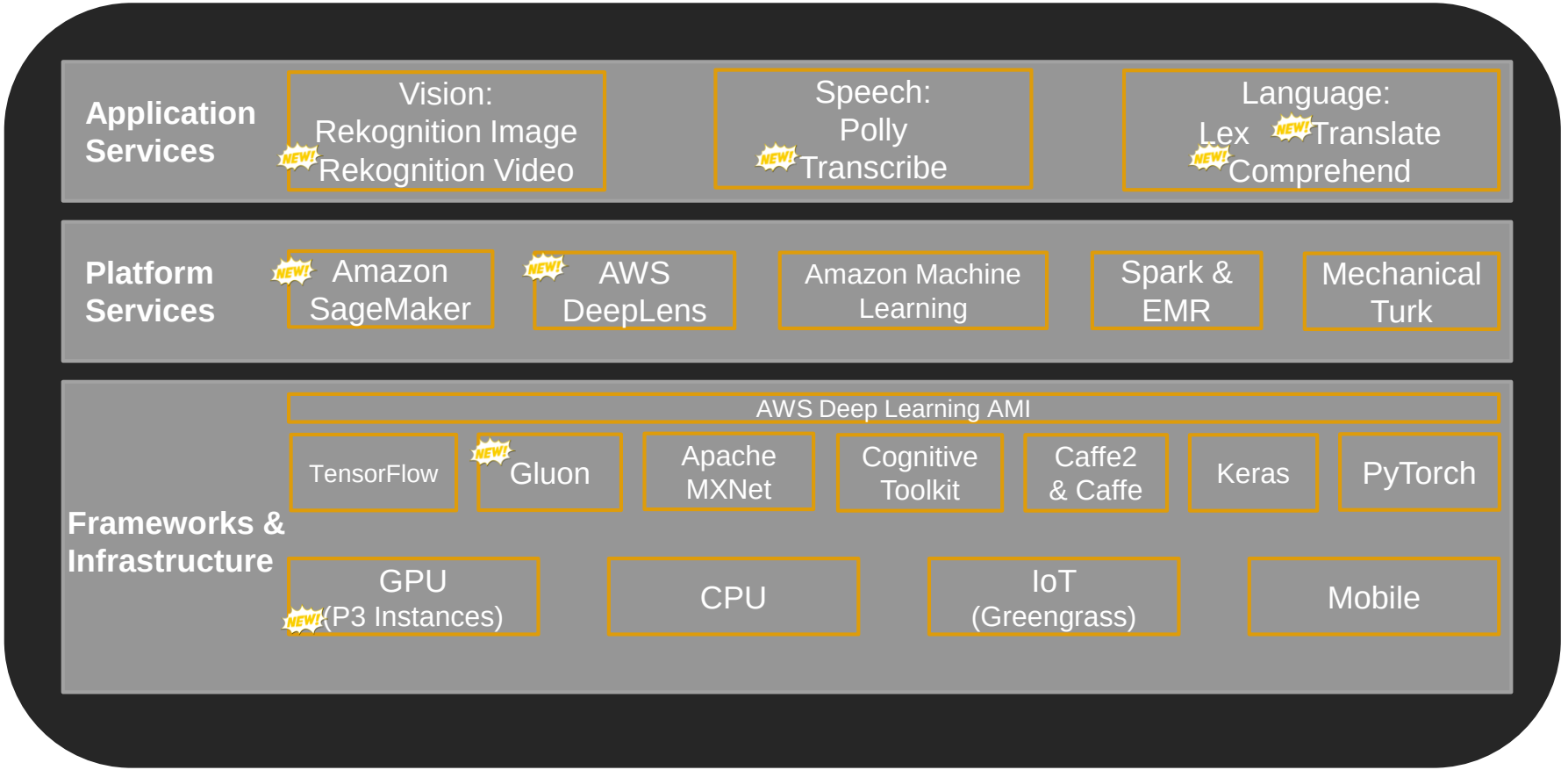
Machine Learning on Your Data Lake



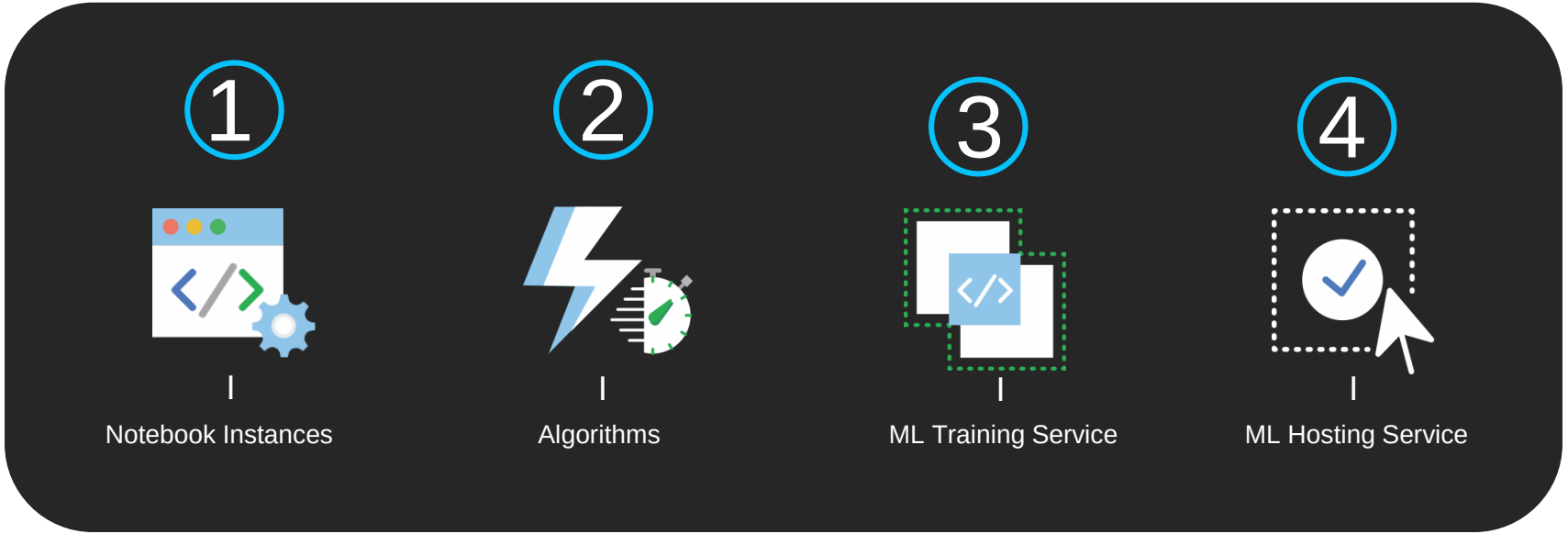
FINRA: Varied Analytic Use Cases



ML in the Hands of Every Developer



Amazon SageMaker



Digital Globe – Using ML to Find the Right Data

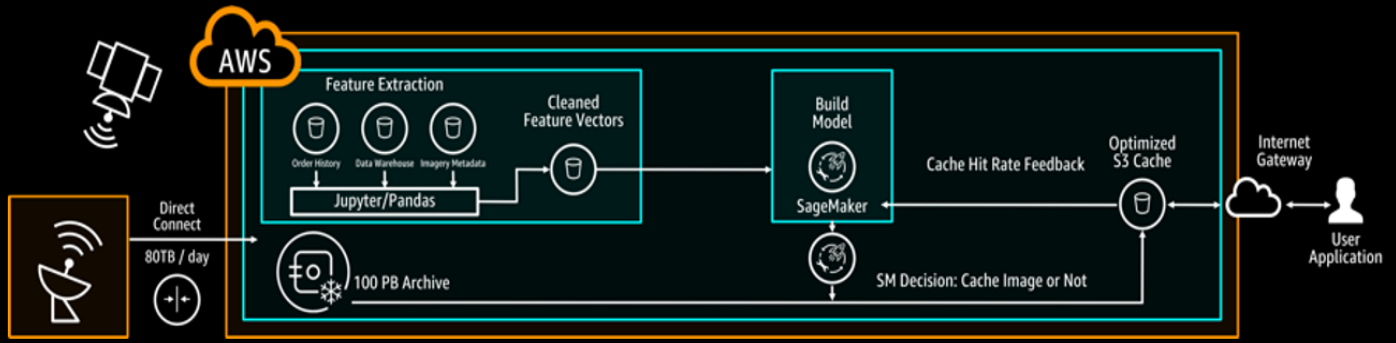
Data lake:

- 100 PB of data in cloud
- Optimize storage tiers

Solution:

- Optimize their data lake storage, cut costs in half

USING AMAZON SAGEMAKER TO CUT CLOUD STORAGE COSTS IN HALF



aws SUMMIT

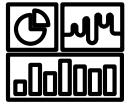
Demonstration:

Running Machine Learning on Your Data Lake

Agility and Innovation Are Key

Machine learning

Amazon SageMaker
AWS Deep Learning AMIs
Amazon Rekognition
Amazon Lex
AWS DeepLens
Amazon Comprehend
Amazon Translate
Amazon Transcribe
Amazon Polly



Analytics

Amazon Athena
Amazon EMR
Amazon Redshift
Amazon Elasticsearch Service
Amazon Kinesis
Amazon QuickSight

On-premises data movement

AWS Direct Connect
AWS Snowball
AWS Snowmobile
AWS Database Migration Service

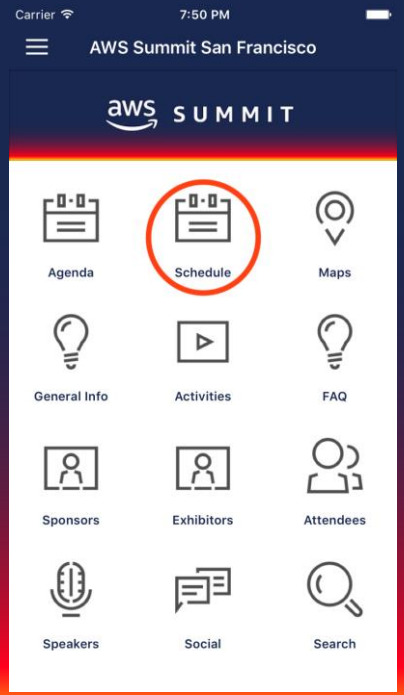
Real-time data movement

AWS IoT Core
Amazon Kinesis Data Firehose
Amazon Kinesis Data Streams
Amazon Kinesis Video Streams

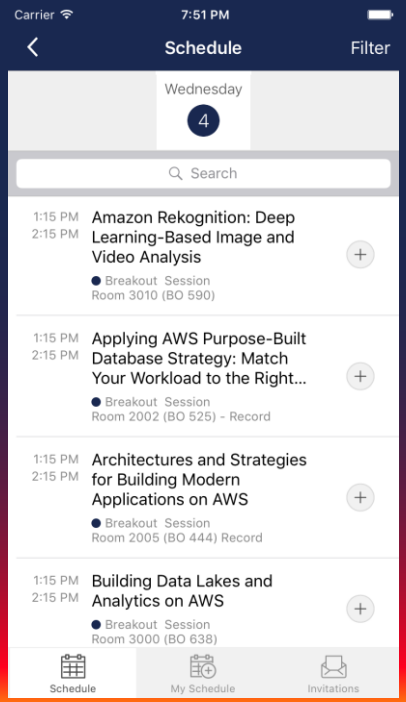
Please complete the session survey in the summit mobile app.

Submit Session Feedback

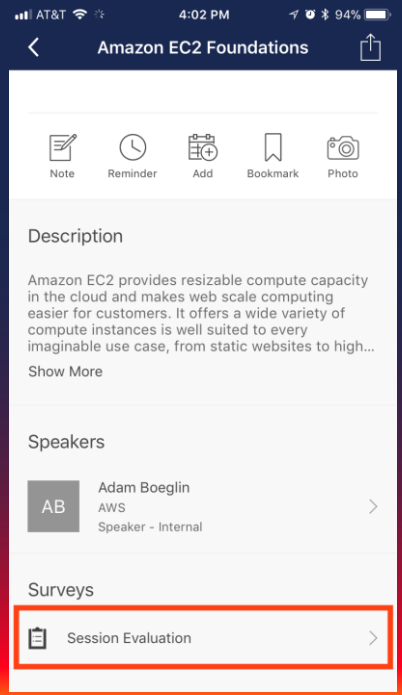
1. Tap the **Schedule** icon.



2. Select the session you attended.



3. Tap **Session Evaluation** to submit your feedback.



Thank you!