Turn Data into Insights with Data Lakes and Analytics on AWS

Osemeke Isibor

Partner Solutions Architect, AWS.

29th June 2020

Data is a strategic asset for every organization

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

The Economist

Theresa May v Brussels

Ten years on: banking after the crisis

South Korea’s unfinished revolution

Biology, but without the cells

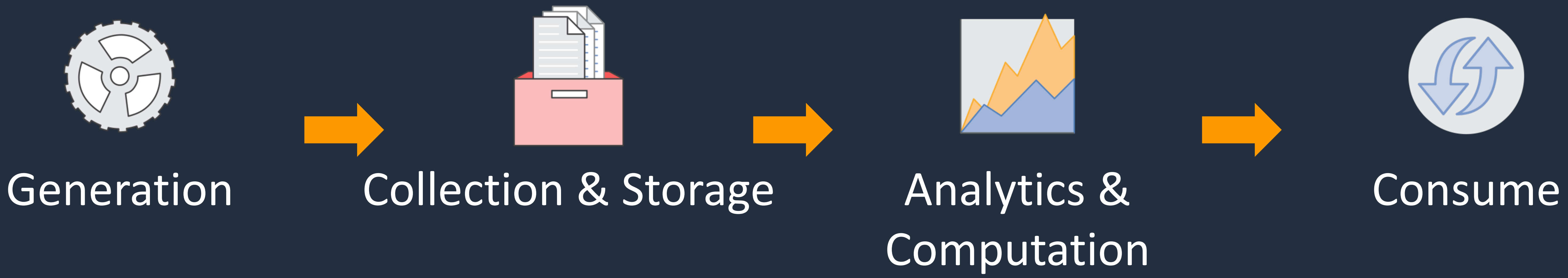
MAY 6TH–12TH 2017

The world’s most valuable resource

Data and the new rules of competition

*Copyright: The Economist, 2017, David Parkins

Data Analytics Workflow



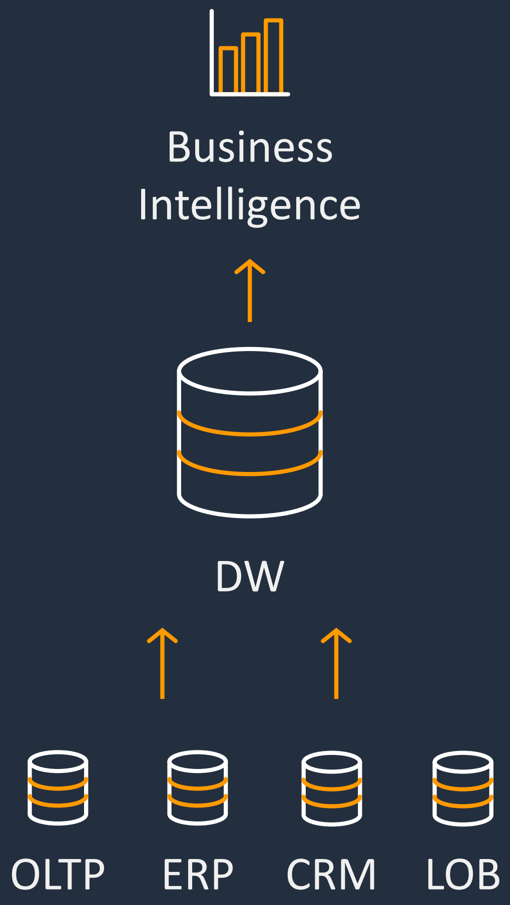
Generation

Collection & Storage

Analytics & Computation

Consume

Traditionally data Analytics revolved around data warehouses



Customers want more value from their data

Growing exponentially

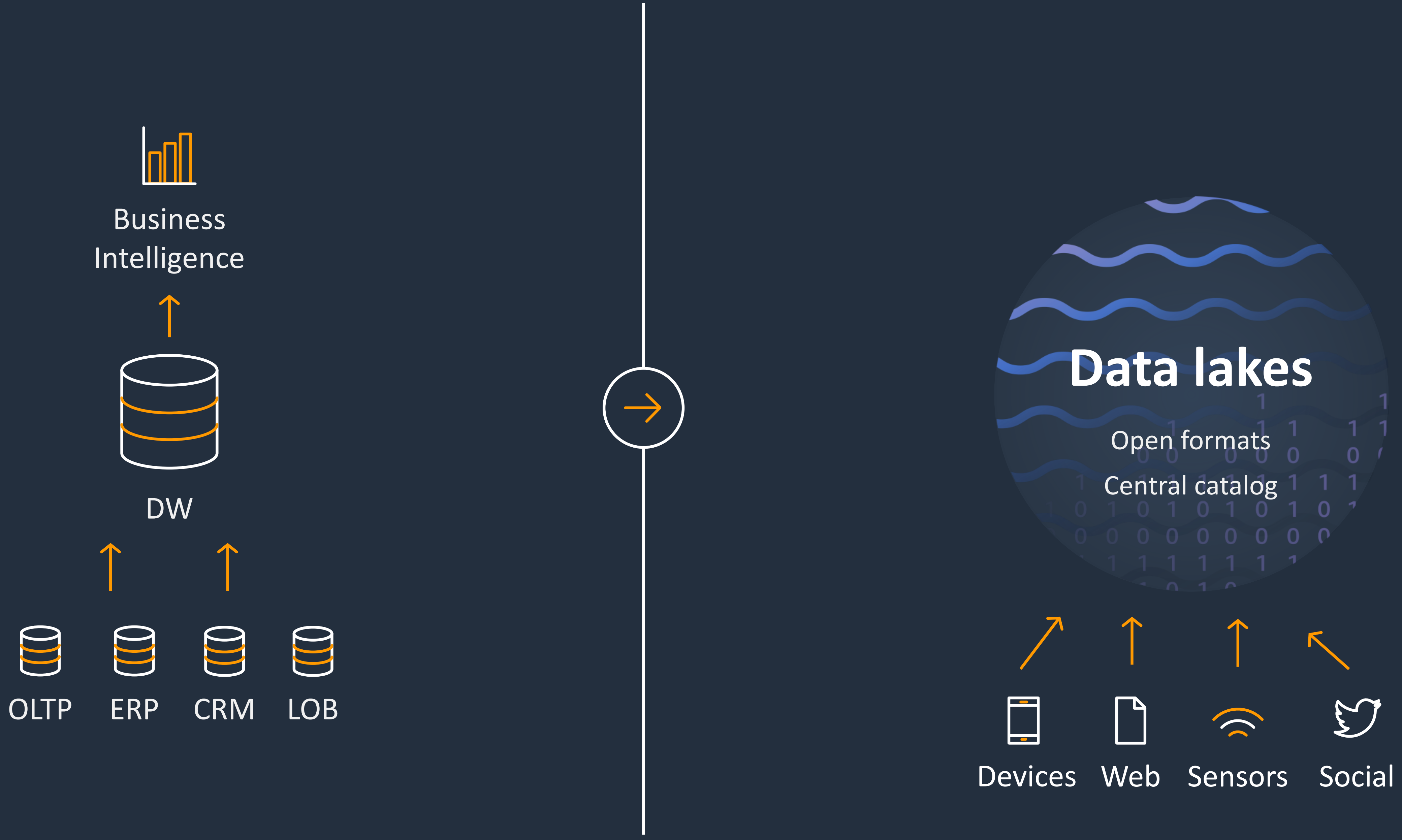
From new sources

Increasingly diverse

Used by many people

Analyzed by many applications

Customers moving to data lake architectures



Business Intelligence

DW

OLTP ERP CRM LOB

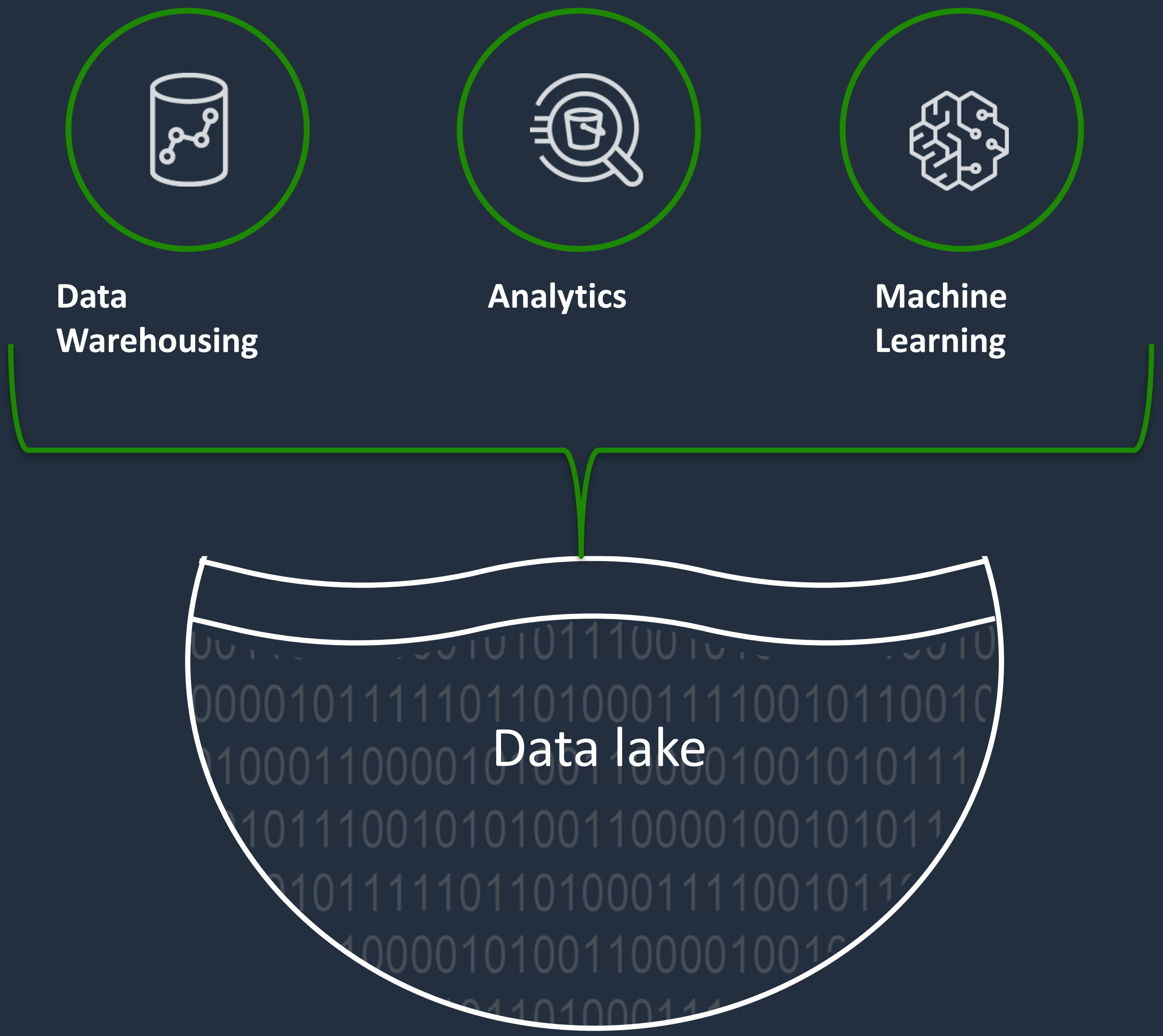
Data lakes

Open formats

Central catalog

Devices Web Sensors Social

Bringing together the best of both worlds



Extends or evolves DW architectures

Store any data in any format

Durable, available, and exabyte scale

Secure, compliant, auditable

Run any type of analytics from DW to Predictive

Why choose AWS for data lakes and analytics?

1. Easiest to build data lakes and analytics

Amazon S3: Highly durable, secure and scalable single storage layer for all analytics and ML

AWS Lake Formation: A service to build secure data lakes in days instead of months.

Deep integration across analytics & infrastructure(including federated queries)

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

2. Most secure infrastructure for analytics

Services for security and governance

Security

Amazon GuardDuty

AWS Shield

AWS WAF

Amazon Macie

VPC

Identity

AWS IAM

AWS SSO

Amazon Cloud Directory

AWS Directory Service

AWS Organizations

Encryption

AWS Certification Manager

AWS Key Management Service

Encryption at rest

Encryption in transit

Bring your own keys,
HSM support

Compliance

AWS Artifact

Amazon Inspector

Amazon Cloud HSM

Amazon Cognito

AWS CloudTrail

2. Most secure infrastructure: certifications

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

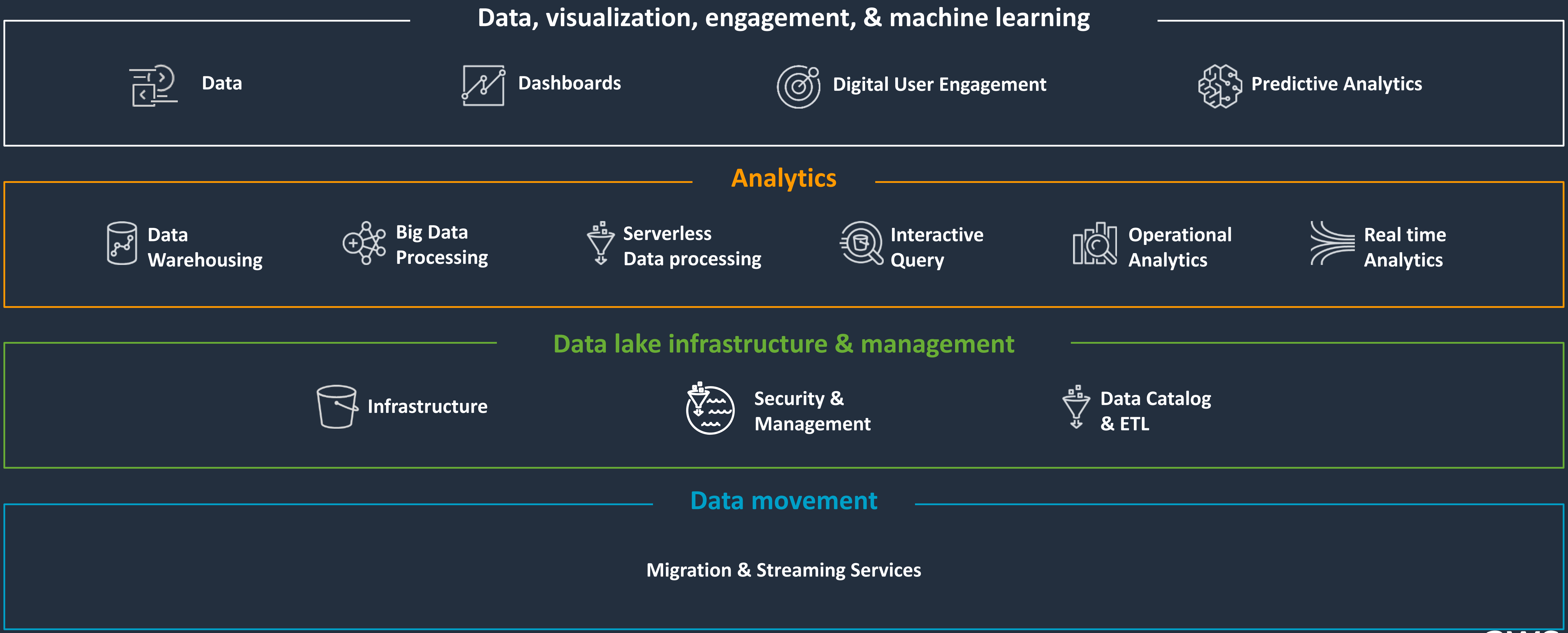
C5 [Germany]
Operational Security Attestation

Cyber Essentials Plus [UK]
Cyber Threat Protection

G-Cloud [UK]
UK Government Standards

IT-Grundschutz [Germany]
Baseline Protection Methodology

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

Data Exchange | QuickSight | Pinpoint | SageMaker | Comprehend | Lex | Polly | Rekognition | Translate

+ many more

Analytics

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

Data lake infrastructure & management

S3/Glacier | Lake Formation | AWS Glue

Data movement

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

3. Open standards, formats, and Apache open source

Flink
Ganglia
Hbase
HCatalog
HDFS
Hive
Hudi
Java
JupyterHub
Kafka
Livy

Mahout
MapReduce
MxNET
MySQL
Oozie
ORC
Parquet
Phoenix
Pig
Presto
Python

PyTorch
R
Scala
Spark
Sqoop
SQL
TensorFlow
Tez
YARN
Zeppelin
Zookeeper

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

4. Most scalable, cost-effective infrastructure for analytics

Some examples of advanced cost saving capabilities in analytics services

EMR

Autoscaling

57% less than on-premises per IDC report

Redshift

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

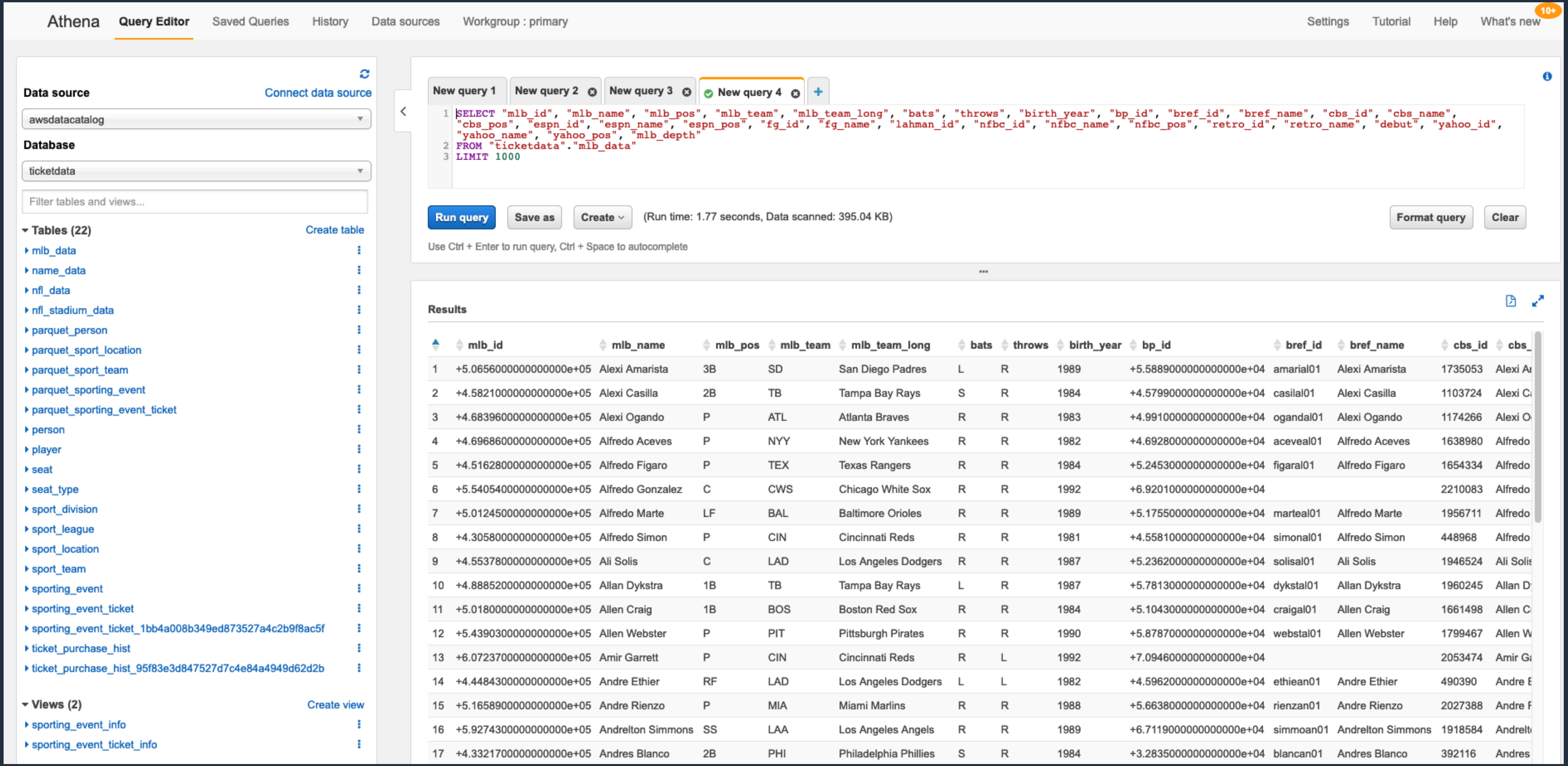
Athena & QuickSight

Serverless pay only for what is used

Pricing per session for visualization

Get answers from your data in AWS Data Lake

Query Directly with Amazon Athena



Data source — Connect data source

awsdatacatalog

Database

ticketdata

Filter tables and views...

Tables (22) — Create table

- mlb_data
- name_data
- nfl_data
- nfl_stadium_data
- parquet_person
- parquet_sport_location
- parquet_sport_team
- parquet_sporting_event
- parquet_sporting_event_ticket
- person
- player
- seat
- seat_type
- sport_division
- sport_league
- sport_location
- sport_team
- sporting_event
- sporting_event_ticket
- sporting_event_ticket_1bb4a008b349ed873527a4c2b9f8ac5f
- ticket_purchase_hist
- ticket_purchase_hist_95f83e3d847527d7c4e84a4949d62d2b

Views (2) — Create view

- sporting_event_info
- sporting_event_ticket_info

```sql
SELECT "mlb_id", "mlb_name", "mlb_pos", "mlb_team", "mlb_team_long", "bats", "throws", "birth_year", "bp_id", "bref_id", "bref_name", "cbs_id", "cbs_name",
"cbs_pos", "espn_id", "espn_name", "espn_pos", "fg_id", "fg_name", "lahman_id", "nfbc_id", "nfbc_name", "nfbc_pos", "retro_id", "retro_name", "debut", "yahoo_id",
"yahoo_name", "yahoo_pos", "mlb_depth"
FROM "ticketdata"."mlb_data"
LIMIT 1000
```

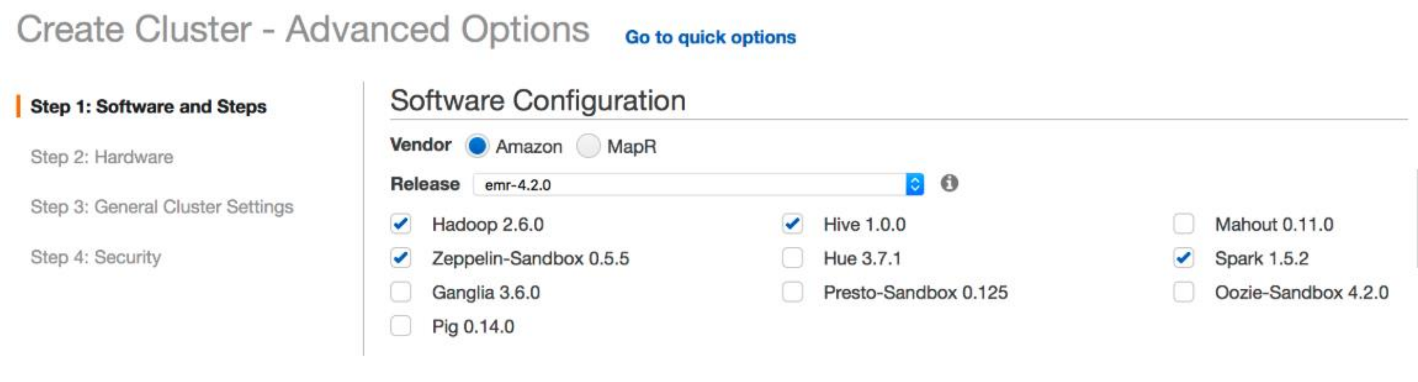
(Run time: 1.77 seconds, Data scanned: 395.04 KB)

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

Results

	mlb_id	mlb_name	mlb_pos	mlb_team	mlb_team_long	bats	throws	birth_year	bp_id	bref_id	bref_name	cbs_id	cbs_
1	+5.0656000000000000e+05	Alexi Amarista	3B	SD	San Diego Padres	L	R	1989	+5.5889000000000000e+04	amarial01	Alexi Amarista	1735053	Alexi Ar
2	+4.5821000000000000e+05	Alexi Casilla	2B	TB	Tampa Bay Rays	S	R	1984	+4.5799000000000000e+04	casilal01	Alexi Casilla	1103724	Alexi C
3	+4.6839600000000000e+05	Alexi Ogando	P	ATL	Atlanta Braves	R	R	1983	+4.9910000000000000e+04	ogandal01	Alexi Ogando	1174266	Alexi O
4	+4.6968600000000000e+05	Alfredo Aceves	P	NYY	New York Yankees	R	R	1982	+4.6928000000000000e+04	aceveal01	Alfredo Aceves	1638980	Alfredo
5	+4.5162800000000000e+05	Alfredo Figaro	P	TEX	Texas Rangers	R	R	1984	+5.2453000000000000e+04	figaral01	Alfredo Figaro	1654334	Alfredo
6	+5.5405400000000000e+05	Alfredo Gonzalez	C	CWS	Chicago White Sox	R	R	1992	+6.9201000000000000e+04			2210083	Alfredo
7	+5.0124500000000000e+05	Alfredo Marte	LF	BAL	Baltimore Orioles	R	R	1989	+5.1755000000000000e+04	marteal01	Alfredo Marte	1956711	Alfredo
8	+4.3058000000000000e+05	Alfredo Simon	P	CIN	Cincinnati Reds	R	R	1981	+4.5581000000000000e+04	simonal01	Alfredo Simon	448968	Alfredo
9	+4.5537800000000000e+05	Ali Solis	C	LAD	Los Angeles Dodgers	R	R	1987	+5.2362000000000000e+04	solisal01	Ali Solis	1946524	Ali Solis
10	+4.8885200000000000e+05	Allan Dykstra	1B	TB	Tampa Bay Rays	L	R	1987	+5.7813000000000000e+04	dykstal01	Allan Dykstra	1960245	Allan D
11	+5.0180000000000000e+05	Allen Craig	1B	BOS	Boston Red Sox	R	R	1984	+5.1043000000000000e+04	craigal01	Allen Craig	1661498	Allen C
12	+5.4390300000000000e+05	Allen Webster	P	PIT	Pittsburgh Pirates	R	R	1990	+5.8787000000000000e+04	webstal01	Allen Webster	1799467	Allen W
13	+6.0723700000000000e+05	Amir Garrett	P	CIN	Cincinnati Reds	R	L	1992	+7.0946000000000000e+04			2053474	Amir Ga
14	+4.4484300000000000e+05	Andre Ethier	RF	LAD	Los Angeles Dodgers	L	L	1982	+4.5962000000000000e+04	ethiean01	Andre Ethier	490390	Andre E
15	+5.1658900000000000e+05	Andre Rienzo	P	MIA	Miami Marlins	R	R	1988	+5.6638000000000000e+04	rienzan01	Andre Rienzo	2027388	Andre R
16	+5.9274300000000000e+05	Andrelton Simmons	SS	LAA	Los Angeles Angels	R	R	1989	+6.7119000000000000e+04	simmoan01	Andrelton Simmons	1918584	Andrelt
17	+4.3321700000000000e+05	Andres Blanco	2B	PHI	Philadelphia Phillies	S	R	1984	+3.2835000000000000e+04	blancan01	Andres Blanco	392116	Andres

Analyze with Hadoop on Amazon EMR



Create Cluster - Advanced Options   Go to quick options

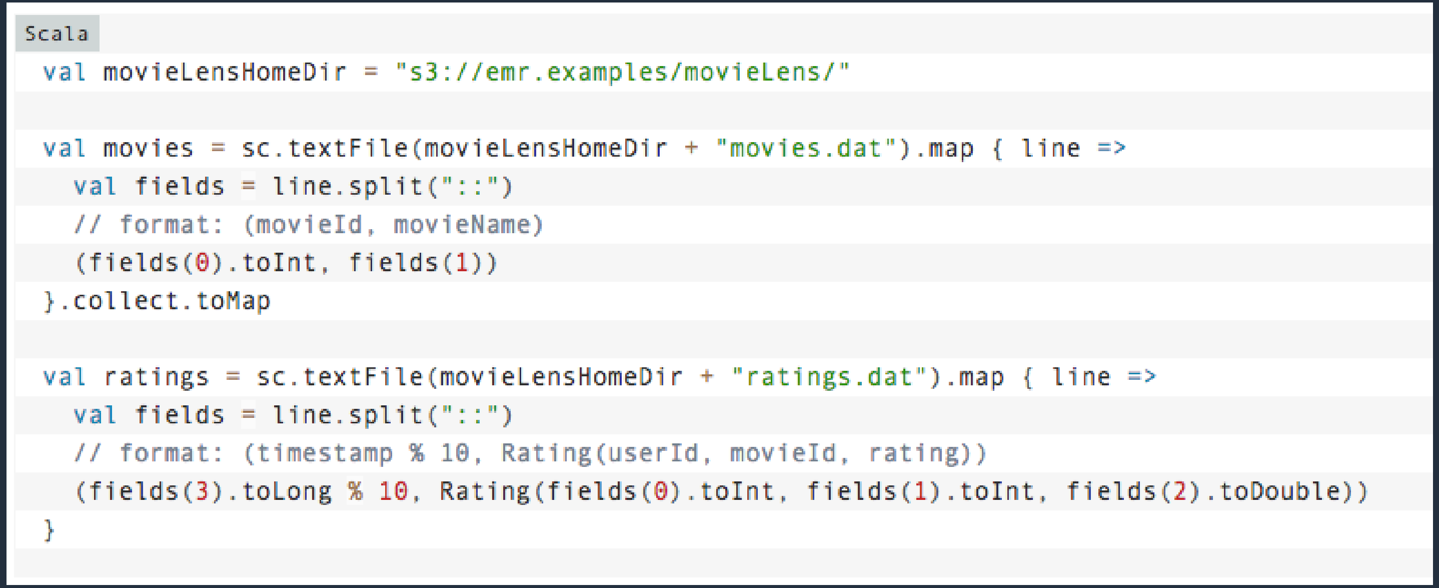
- **Step 1: Software and Steps**
- Step 2: Hardware
- Step 3: General Cluster Settings
- Step 4: Security

Software Configuration

Vendor ● Amazon ○ MapR

Release emr-4.2.0

☑ Hadoop 2.6.0
☑ Zeppelin-Sandbox 0.5.5
□ Ganglia 3.6.0
□ Pig 0.14.0
☑ Hive 1.0.0
□ Hue 3.7.1
□ Presto-Sandbox 0.125
□ Mahout 0.11.0
☑ Spark 1.5.2
□ Oozie-Sandbox 4.2.0

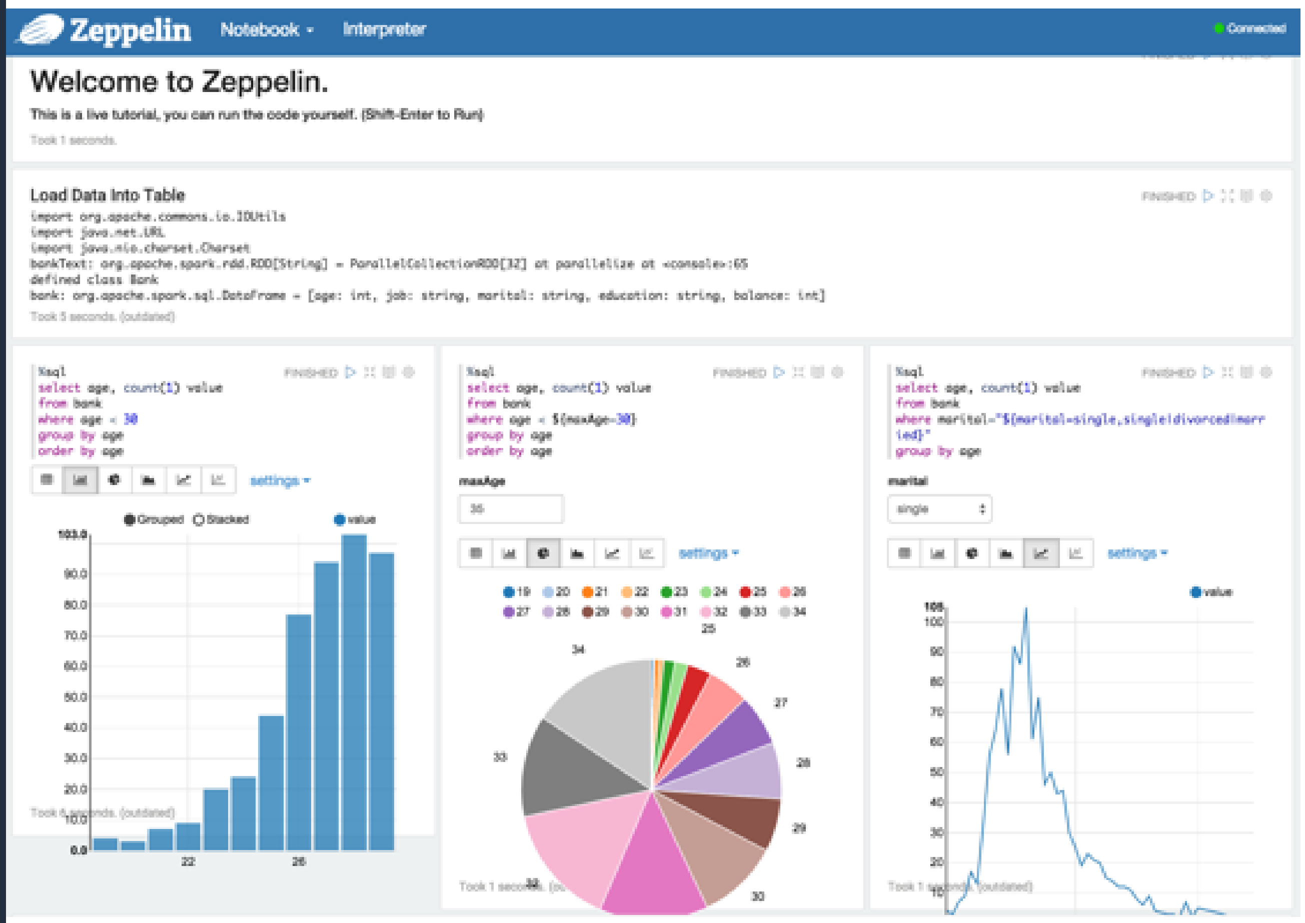


Scala

```scala
val movieLensHomeDir = "s3://emr.examples/movieLens/"

val movies = sc.textFile(movieLensHomeDir + "movies.dat").map { line =>
  val fields = line.split("::")
  // format: (movieId, movieName)
  (fields(0).toInt, fields(1))
}.collect.toMap

val ratings = sc.textFile(movieLensHomeDir + "ratings.dat").map { line =>
  val fields = line.split("::")
  // format: (timestamp % 10, Rating(userId, movieId, rating))
  (fields(3).toLong % 10, Rating(fields(0).toInt, fields(1).toInt, fields(2).toDouble))
}
```



Welcome to Zeppelin.

This is a live tutorial, you can run the code yourself. (Shift-Enter to Run)

Took 1 seconds.

Load Data Into Table

```
import org.apache.commons.io.IOUtils
import java.net.URL
import java.nio.charset.Charset
bankText: org.apache.spark.rdd.RDD[String] = ParallelCollectionRDD[32] at parallelize at <console>:65
defined class Bank
bank: org.apache.spark.sql.DataFrame = [age: int, job: string, marital: string, education: string, balance: int]
```

Took 5 seconds. (outdated)

```
%sql
select age, count(1) value
from bank
where age < 30
group by age
order by age
```

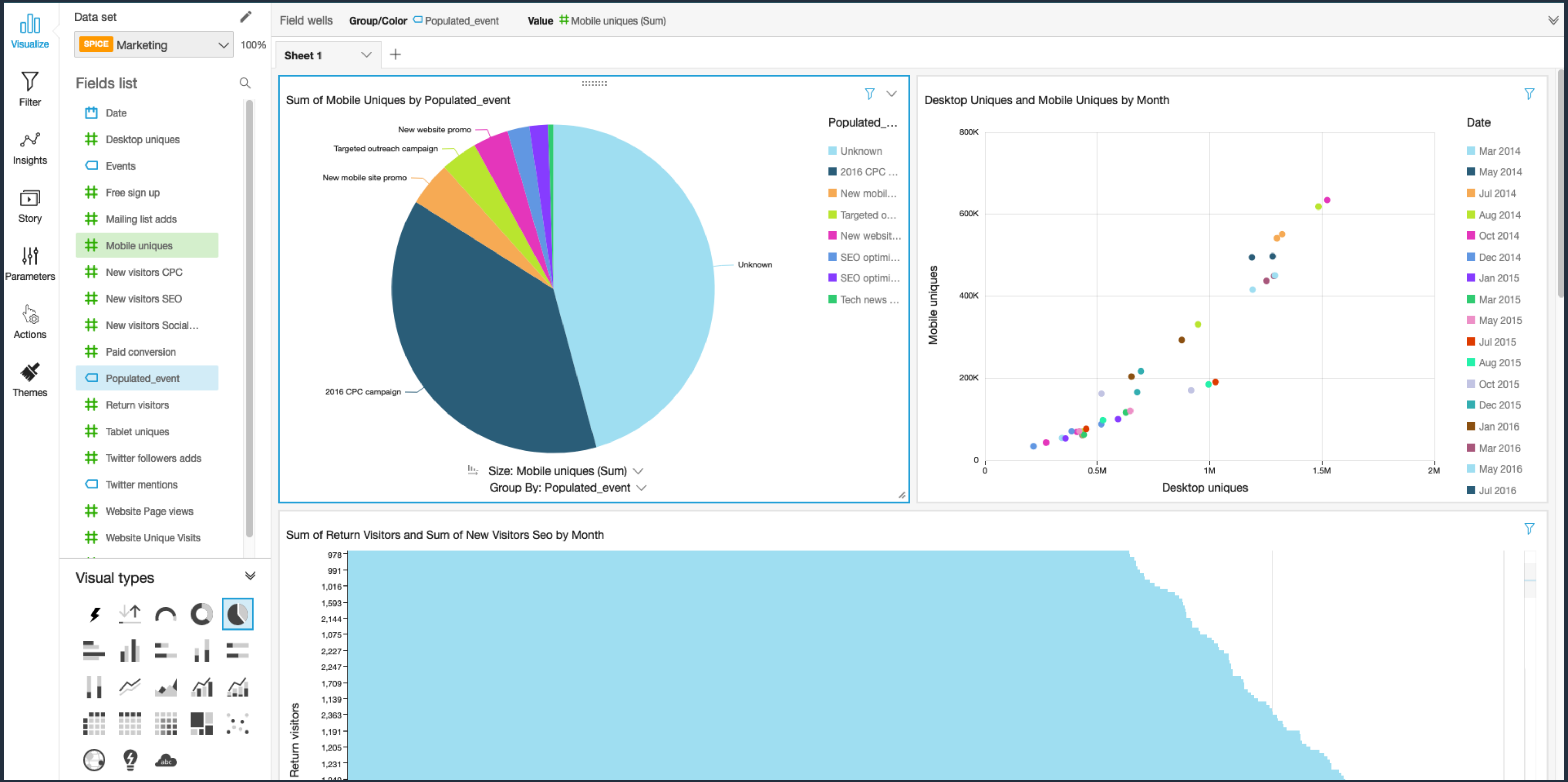
```
%sql
select age, count(1) value
from bank
where age < ${maxAge=30}
group by age
order by age
```

maxAge: 35

```
%sql
select age, count(1) value
from bank
where marital="${marital=single,single|divorced|married}"
group by age
```

marital: single

Create Visualizations with Amazon QuickSight

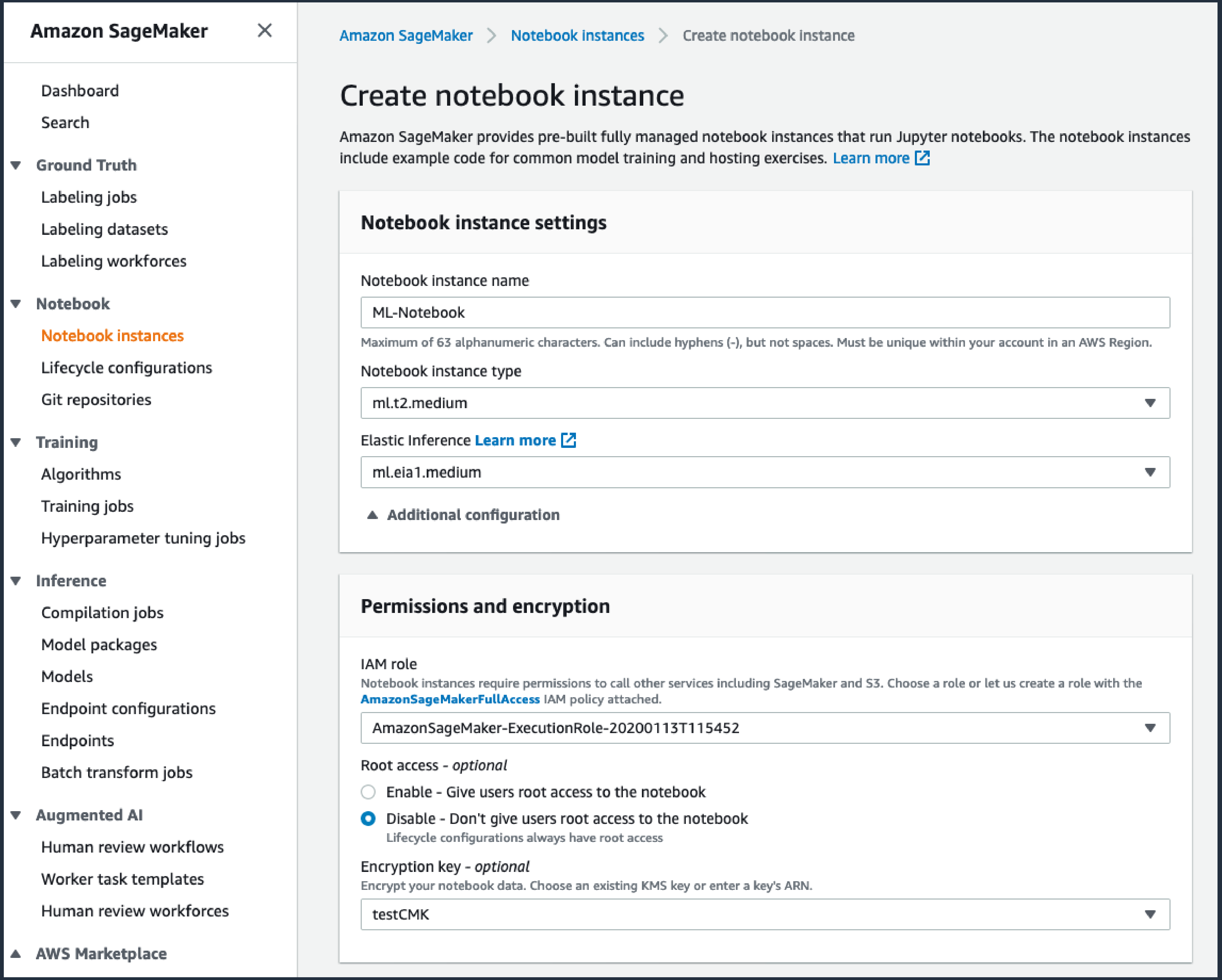


Data set: SPICE Marketing 100%

Field wells Group/Color Populated_event Value Mobile uniques (Sum)

Size: Mobile uniques (Sum)
Group By: Populated_event

Train ML Models with Amazon SageMaker



Create notebook instance

Amazon SageMaker provides pre-built fully managed notebook instances that run Jupyter notebooks. The notebook instances include example code for common model training and hosting exercises. Learn more

Notebook instance settings

Notebook instance name

ML-Notebook

Maximum of 63 alphanumeric characters. Can include hyphens (-), but not spaces. Must be unique within your account in an AWS Region.

Notebook instance type

ml.t2.medium ▼

Elastic Inference Learn more

ml.eia1.medium ▼

▲ Additional configuration

Permissions and encryption

IAM role

Notebook instances require permissions to call other services including SageMaker and S3. Choose a role or let us create a role with the **AmazonSageMakerFullAccess** IAM policy attached.

AmazonSageMaker-ExecutionRole-20200113T115452 ▼

Root access - *optional*

○ Enable - Give users root access to the notebook

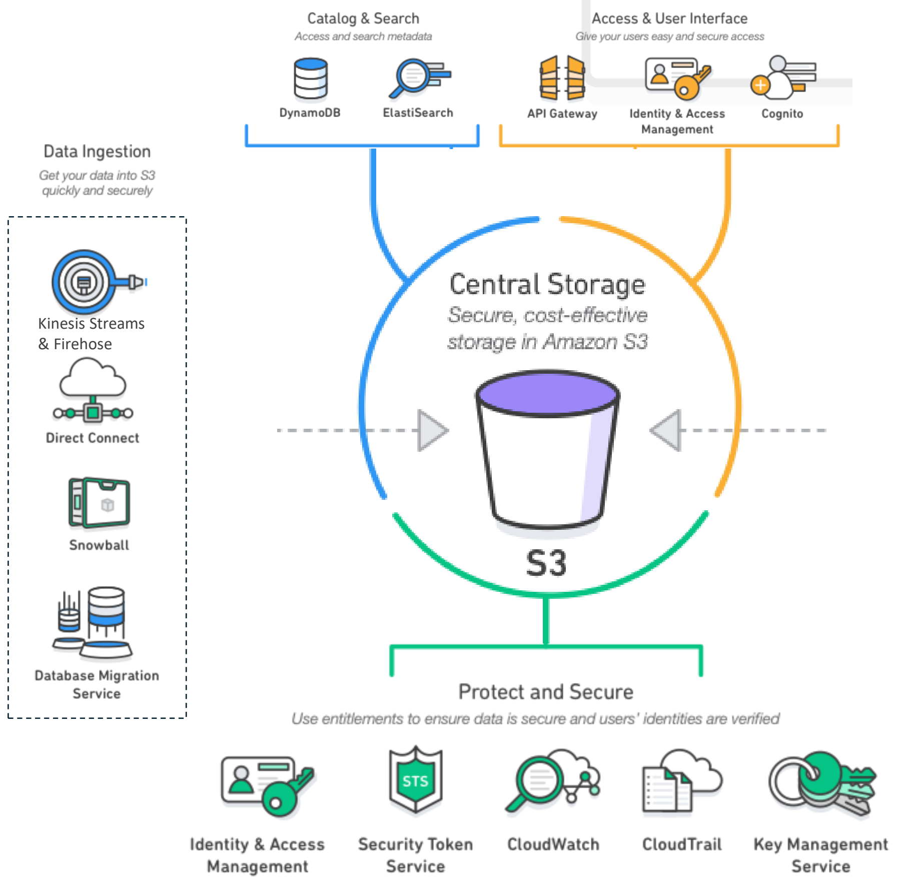
● Disable - Don't give users root access to the notebook
Lifecycle configurations always have root access

Encryption key - *optional*

Encrypt your notebook data. Choose an existing KMS key or enter a key's ARN.

testCMK ▼

Fully integrate with other AWS Services



Catalog & Search

Access and search metadata

DynamoDB

ElastiSearch

Access & User Interface

Give your users easy and secure access

API Gateway

Identity & Access Management

Cognito

Data Ingestion

Get your data into S3 quickly and securely

Kinesis Streams & Firehose

Direct Connect

Snowball

Database Migration Service

Central Storage

Secure, cost-effective storage in Amazon S3

S3

Protect and Secure

Use entitlements to ensure data is secure and users' identities are verified

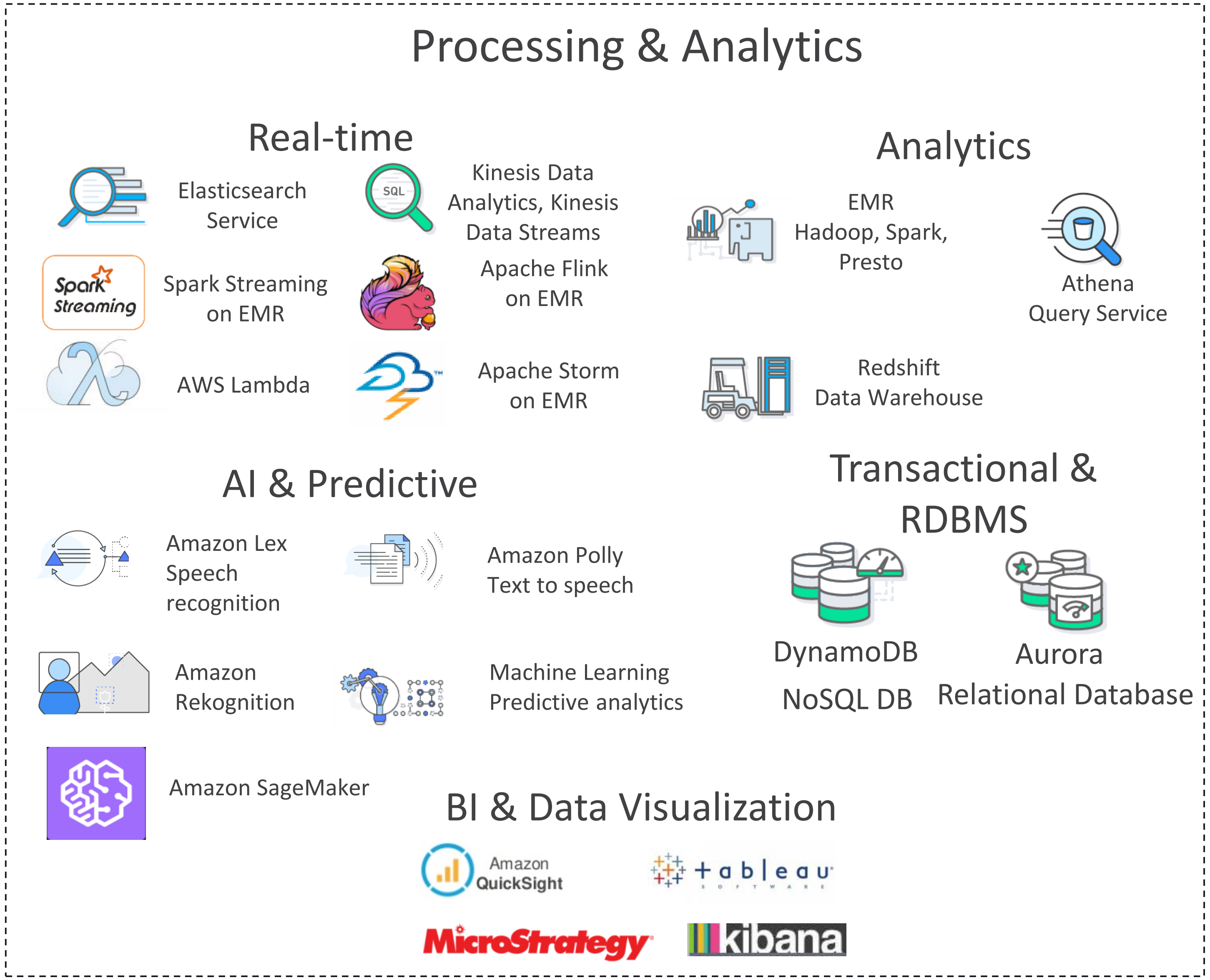
Identity & Access Management

Security Token Service

CloudWatch

CloudTrail

Key Management Service



Processing & Analytics

Real-time

Elasticsearch Service

Kinesis Data Analytics, Kinesis Data Streams

Spark Streaming on EMR

Apache Flink on EMR

AWS Lambda

Apache Storm on EMR

Analytics

EMR Hadoop, Spark, Presto

Athena Query Service

Redshift Data Warehouse

AI & Predictive

Amazon Lex Speech recognition

Amazon Polly Text to speech

Amazon Rekognition

Machine Learning Predictive analytics

Amazon SageMaker

Transactional & RDBMS

DynamoDB NoSQL DB

Aurora Relational Database

BI & Data Visualization

Amazon QuickSight

tableau

MicroStrategy

kibana

More data lakes and analytics than anywhere else

Tens of thousands of data lakes run on AWS across all industries

3M Science. Applied to Life.™ · 3VICTORS TRAVEL ai · ACQUIA · A+E · amazon · Adobe · AMGEN · ATLASSIAN

Agero · airbnb · BARCLAYS · Capital One · Chick-fil-A · CIVIS · CIVITAS LEARNING · CoffeeMeetsBagel

dataxu · CROWDSTRIKE · DOMO · EA · EPIC GAMES · FAST RETAILING · FINANCIAL TIMES · FOURSQUARE · FUNNY OR DIE

GP Georgia-Pacific · gogo · HEARST corporation · hudl · HYATT™ · instacart · JustGiving · lyft

MONSTER · mparticle · Nasdaq · NETFLIX · docomo · NU SKIN · Olive · OLX GROUP · Panasonic · Pfizer

Pinterest · pitney bowes · realnetworks · REDFIN · ROVIO · SCHOLASTIC · slack · SONOS · reddit

T · · Mobile · Sysco · THOMSON REUTERS · ticketmaster · TUNE · verizon · yelp · Zillow · ZUMBA

Complemented by AWS Partner Network (APN) Solutions providers

Collection & preparation

ATTUNITY

Informatica

MATILLION

Paxata

talend

TRIFACTA

Governance

collibra

DATAGUISE

DATAGUISE

DATA REPUBLIC

Informatica

ZALONI THE DATA LAKE COMPANY

Visualization

SISENSE

MicroStrategy

looker

tableau SOFTWARE

TIBCO

Get help from Data & analytics APN consulting partners

GLOBAL

accenture Deloitte.
wipro Infosys® Mindtree Welcome to possible

JAPAN

class method cloudpack TECHORUS

CHINA

博思云为 bosicloud
eCloud valley

LATAM

Morris & Opazo Business Solutions

NORTH AMERICA

1Strategy 47Lining A Hitachi Vantara Company 8K Miles
AGILISIUM BLUESENTRY CAMBRIDGE TECHNOLOGY
clearscale Cloudwick CORECOMPETE
VERTICAL TRAIL SOLUTIONS
MACTORES NorthBay
Provectus
ONICA PARIVEDA SOLUTIONS slalom
softserve iOLAP TEKsystems Our people make IT possible.

EMEA

BEEVA claranet CloudMas
D2SI by devoteam CLOUDREACH EMET E&M COMPUTING
IPPON Discovery to Delivery KCOM Reply storm
SOLITA dnm SEE CLEARER. ACT SMARTER. tecRACER Cloud Enabling Your Business
DoiT INTERNATIONAL keepler

APAC

blazeclan Cloud IT Better Powerup Cloud
TO THE NEW bluepi

aws

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/

Customer Success

Powered by AWS.

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

Challenge:

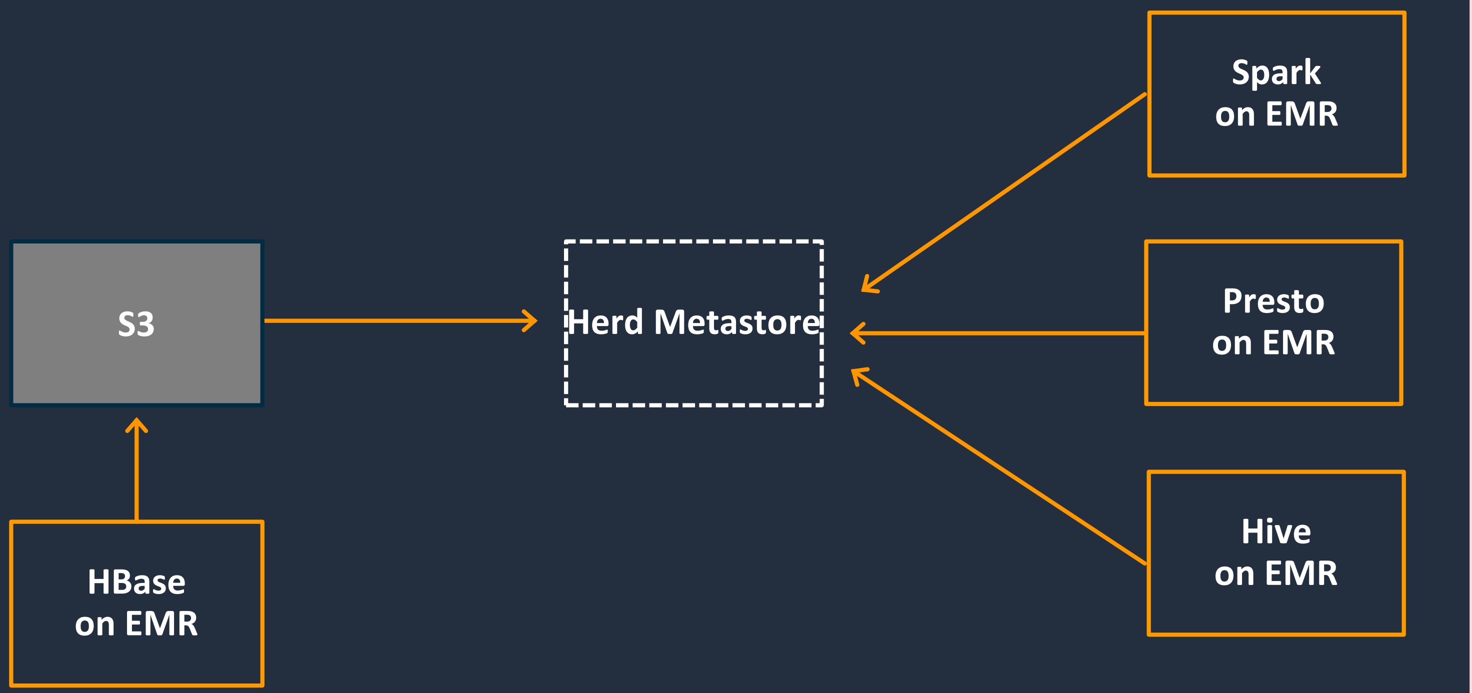
FINRA's legacy system did not scale well

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

Solution:

- Migrated their big data appliance to a S3 Data Lake and used EMR for ingestion and processing
- Migrated to RDS and testing Aurora

FINRA uses S3 to Build Data Lake with EMR



- Required fast access across trillions of trade records (20PB+)
- Migrated from on-premises system
- Use Apache HBase on Amazon EMR to store and serve this data
- Use EMR engines—Spark, Presto, and Hive to process data
- Lower costs by 60% over on-premises system

Nasdaq

Nasdaq operates financial exchanges around the world, and processes
large volumes of data.

Challenge:

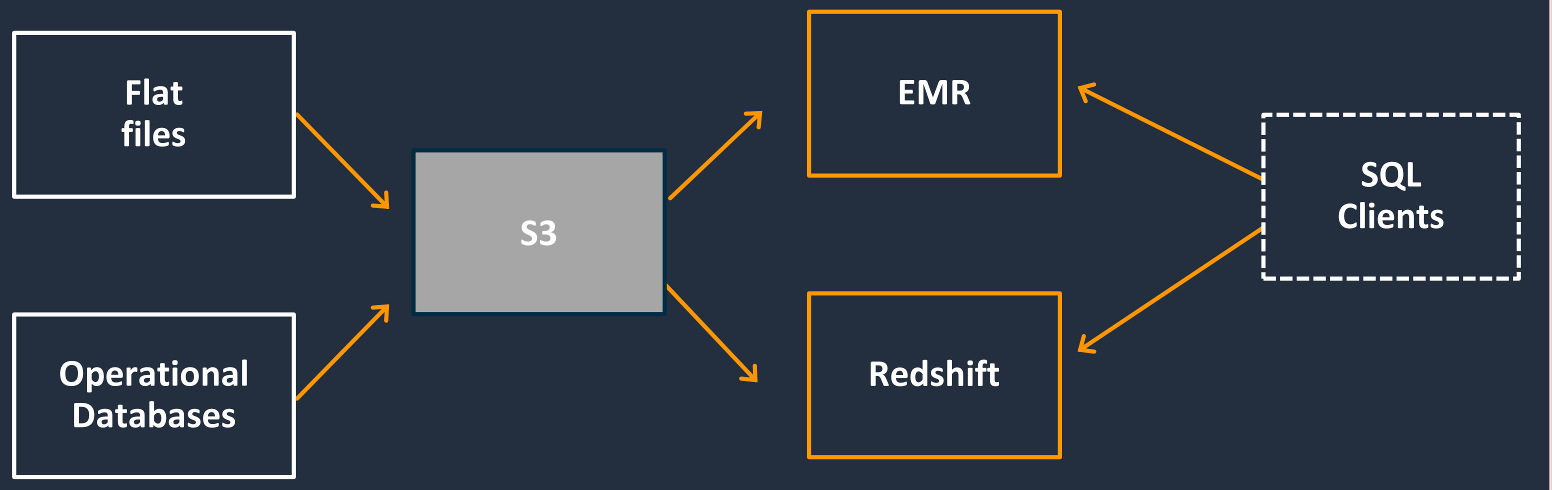
Nasdaq wanted to make their large historical data footprint available to analyze as a single dataset.

Solution:

- Use Amazon Redshift for interactive querying
- Use Amazon S3 as a Data Lake, and Presto on EMR to process historical data

Nasdaq Uses AWS to Build a Data Lake

Nasdaq



Data from all 7 exchanges operated by Nasdaq
(orders, quotes, trade executions)

- Migrate legacy on-premises warehouse to Amazon Redshift
- 4.8B rows inserted per trading day (orders, trades, quotes)
- Ingest data from multiple sources, validates, and stages in S3
- Redshift reads data out of S3 for fast queries
- Presto on EMR and S3 used for analysis of massive historical data set

amazon.com

Challenge:

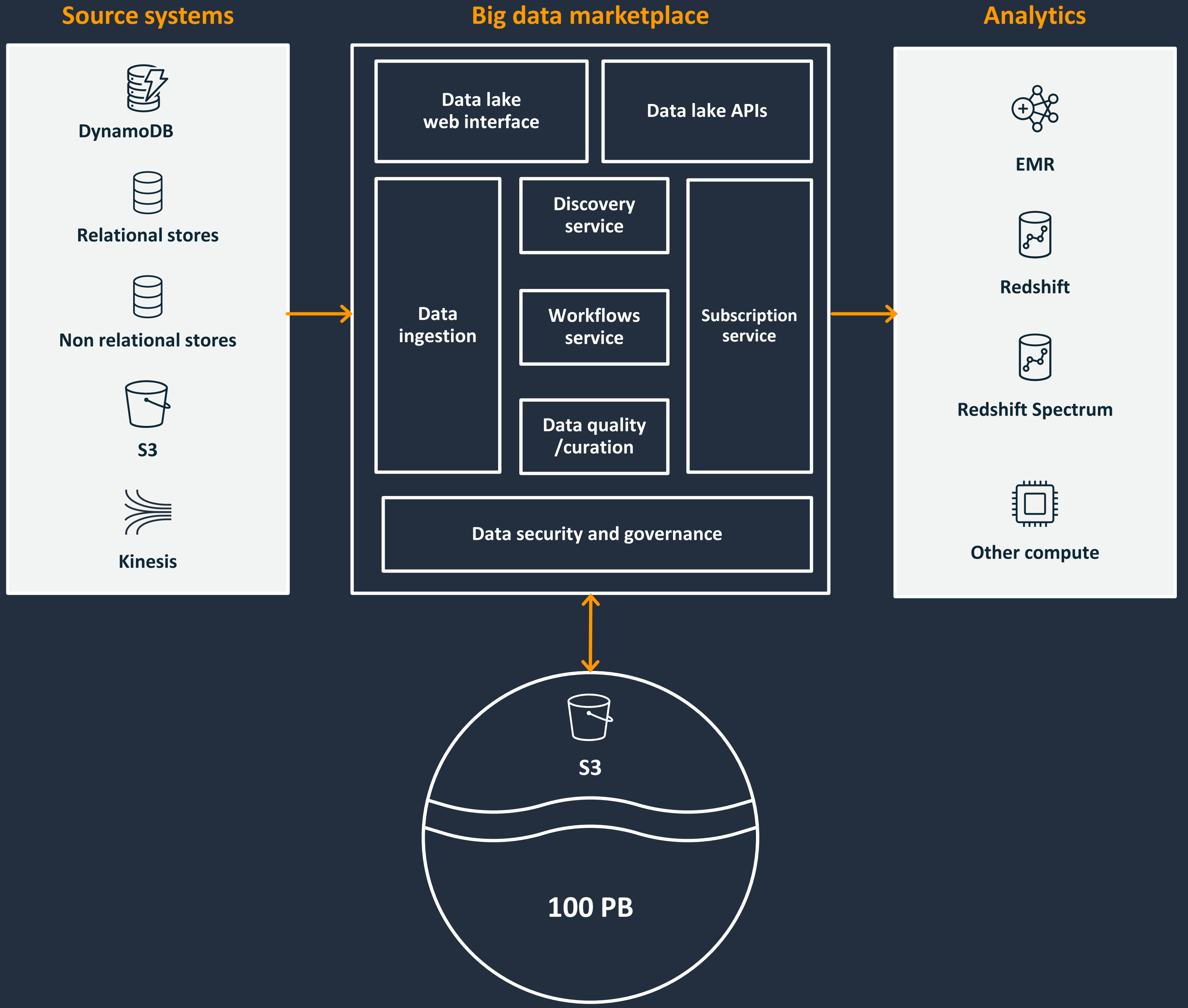
- Amazon needed to analyze a massive amount of data to find insights, identify opportunities, and evaluate business performance.

- The Oracle DW did not scale, was difficult to maintain, and costly.

Solution:

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

Amazon uses an AWS Data Lake



- Migrated system with 50 PB of analytics data
- Optimized scale and cost
- Eliminated Oracle licensing cost
- Expanded analytics toolset
- Doubled dataset used in analytics from 50PB to 100PB

EQUINOX

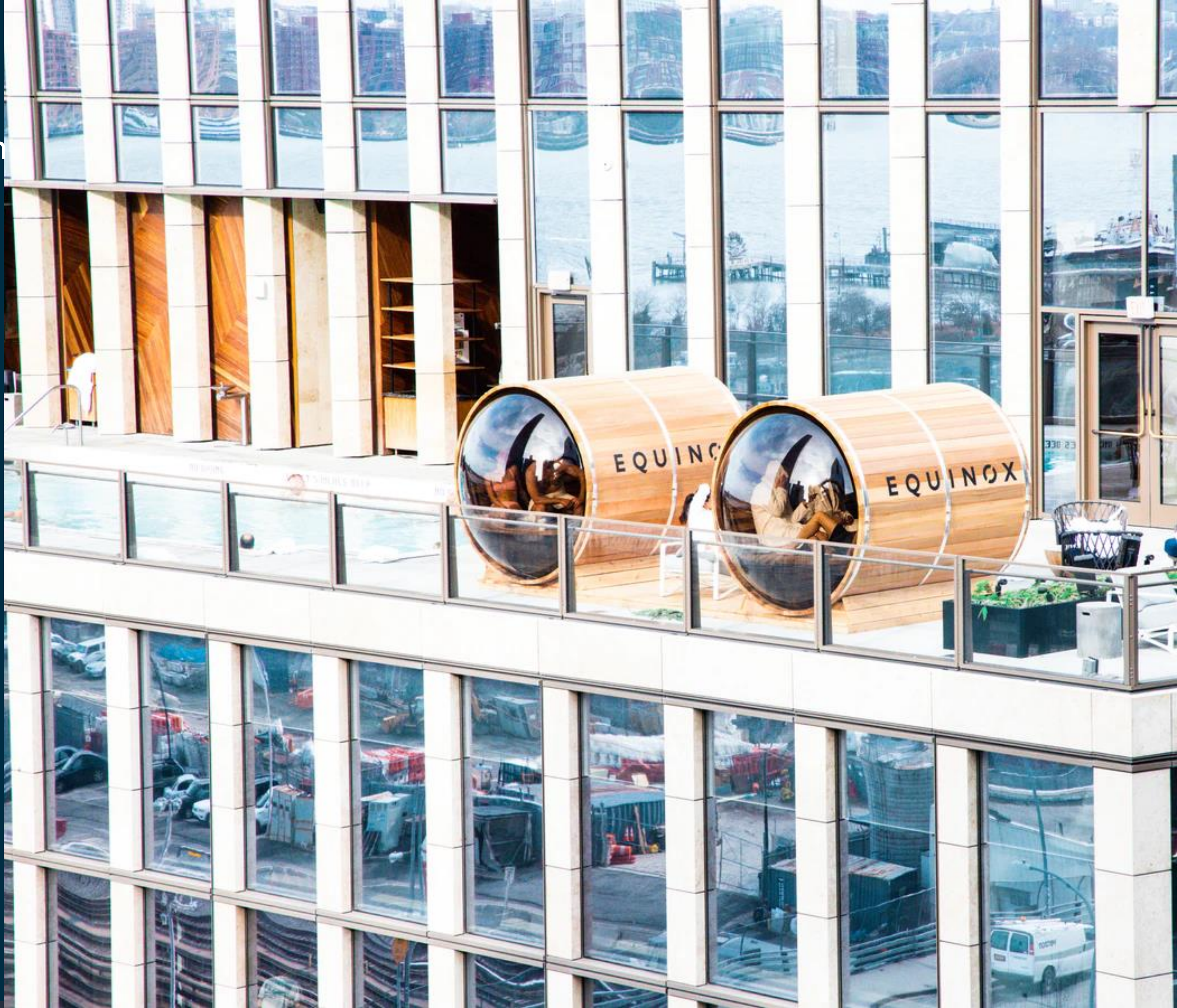
Equinox Fitness has a number of health and fitness brands

Challenge:

- Their data warehouse had limited integration.
- They needed to reduce administration and costs, blend structured and semi-structured data for analytics, and evolve into a data lake strategy.

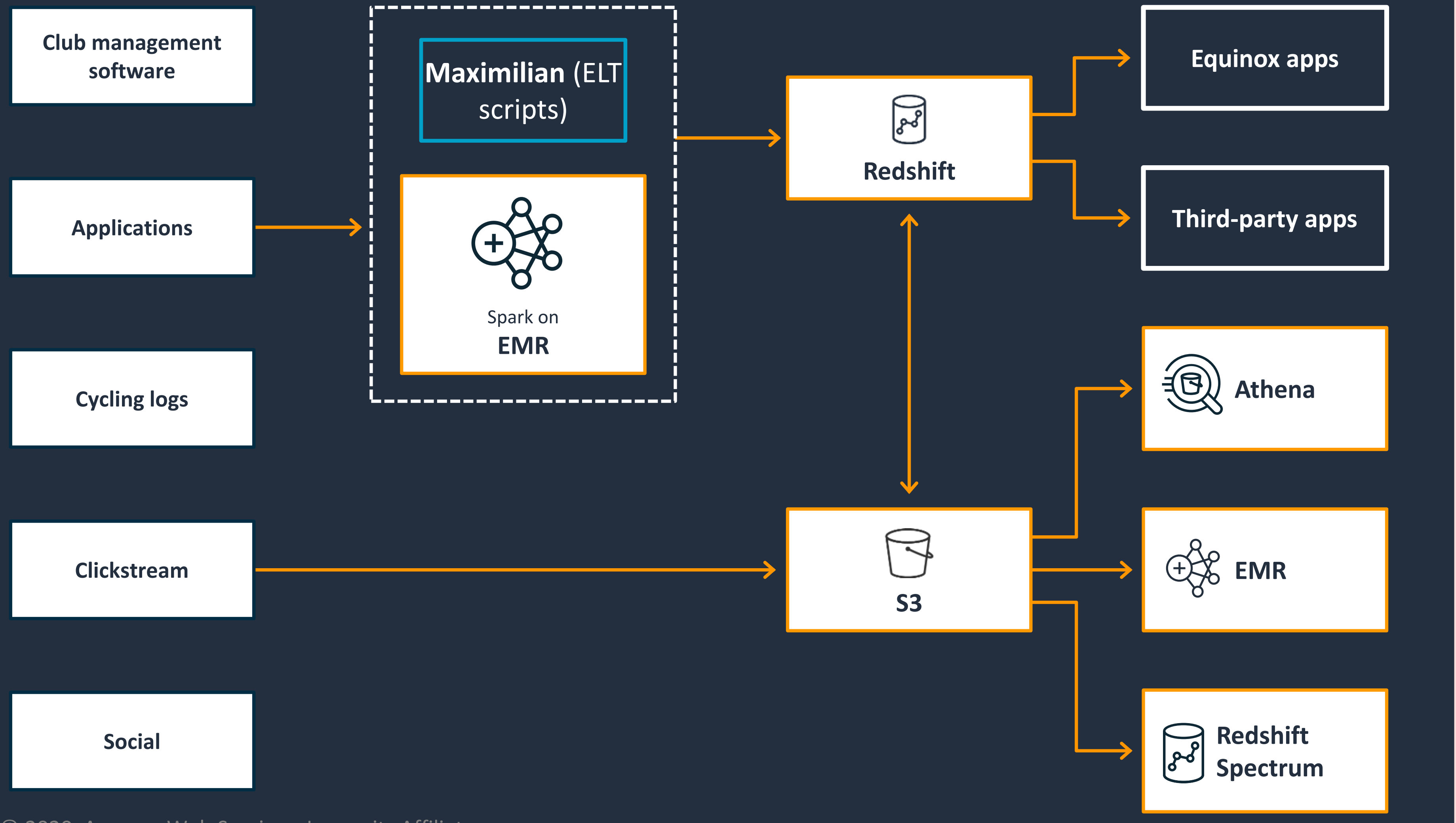
Solution:

- Migrated to Amazon Redshift to combine data from disparate sources.
- They land data directly in an Amazon S3 data lake and perform analytics using Amazon Redshift, Redshift Spectrum, and Amazon EMR.



80% cost savings by migrating to Amazon Redshift

EQUINOX



- Migrated from Teradata data warehouse
- Built a DW with Redshift and data lake with S3
- Analytics on data lake with Amazon Athena, Amazon Redshift Spectrum, and Amazon EMR
- Increased user productivity to move faster
- Amazon Redshift costs ~20% of its original Teradata maintenance and support
- Report time reduced from months to days

aws

Next steps…

1

Sign up for an AWS account

Instantly get access
to the AWS Free Tier

2

Learn with 10-minute tutorials

Explore and learn with
simple tutorials

3

Start building with AWS

Begin building with step-by-step
guide to help you launch
your AWS project

AWS Data Exchange

Easily find and subscribe to 3rd-party data in the cloud

Quickly find diverse data in one place

>1,000 data products

>80 data providers including include Dow Jones, Change Healthcare, Foursquare, Dun & Bradstreet, Thomson Reuters, Pitney Bowes, Lexis Nexis, and Deloitte

Easily analyze data

Download or copy data to S3

Combine, analyze, and model with existing data

Analyze data with EMR, Redshift, Athena, and AWS Glue

Efficiently access 3rd party data

Simplifies access to data: No need to receive physical media, manage FTP credentials, or integrate with different APIs

Minimize legal reviews and negotiations

Feedback Survey

Please let us know what you think about the session.

Get help from AWS with your analytics project.

Scan QR code to complete survey

As a token of appreciation, we’ll mail you an AWS souvenir

Thank you!

iosemeke@amazon.com