AWS SUMMIT

Building Your Data Lake on AWS

Ben Snively, Senior Solutions Architect, AWS

August 14, 2017

What is a Data Lake?

An architectural approach that allows you to store massive amounts of “raw” data into a central location

It’s readily available to be categorized, processed, analyzed, and consumed by diverse groups

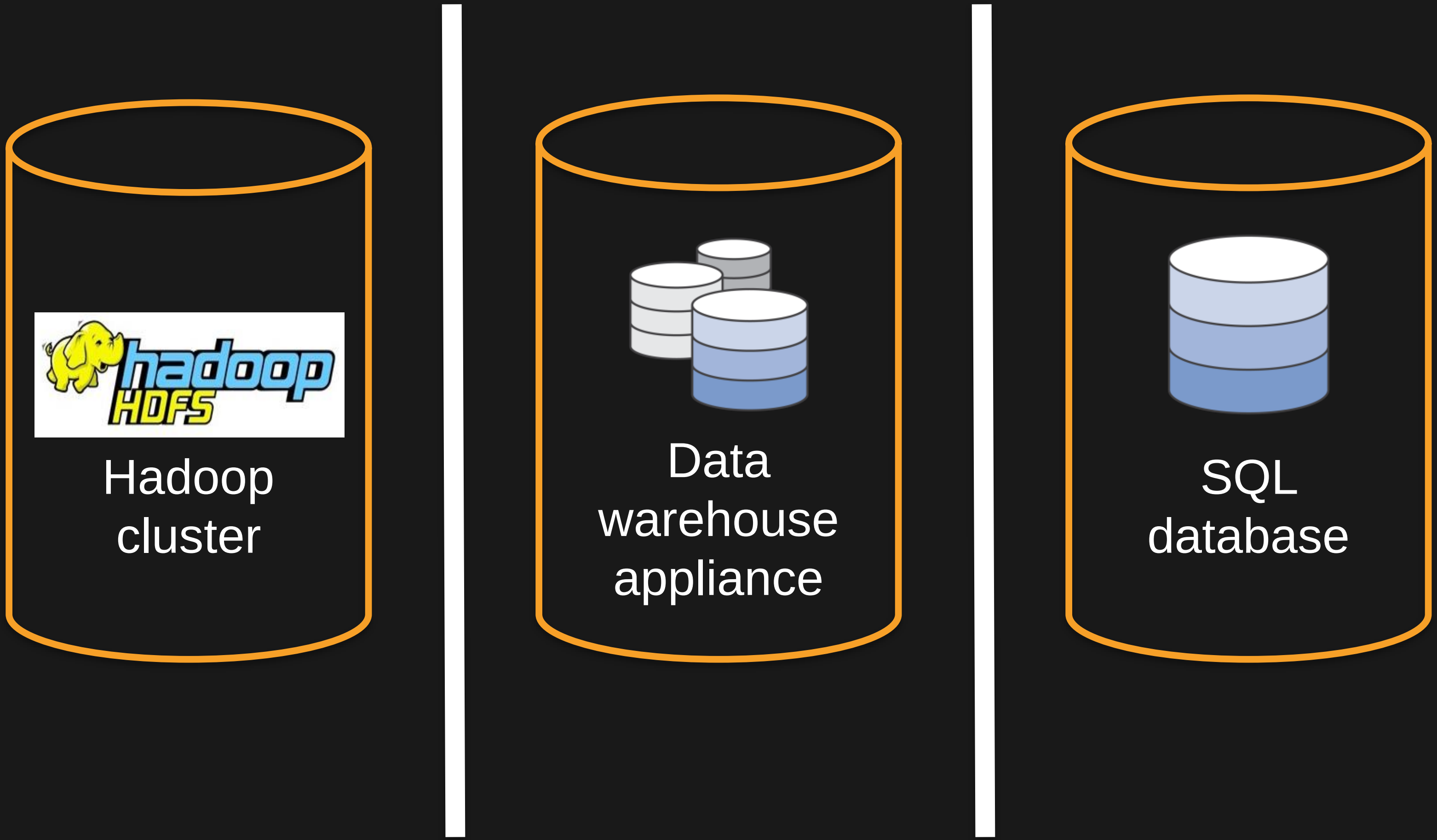
Why use a Data Lake?

Leverage all data within your organization

Leads to…

- Customer centricity
- Business agility
- Better predictions
- Competitive advantage

Legacy data architectures exist as isolated data silos



Hadoop cluster

Data warehouse appliance

SQL database

Navigating the Data Lake…

Data Lake is a new and increasingly popular architecture to store and analyze massive volumes and heterogenous types of data in a centralized repository

Building a Data Lake on AWS

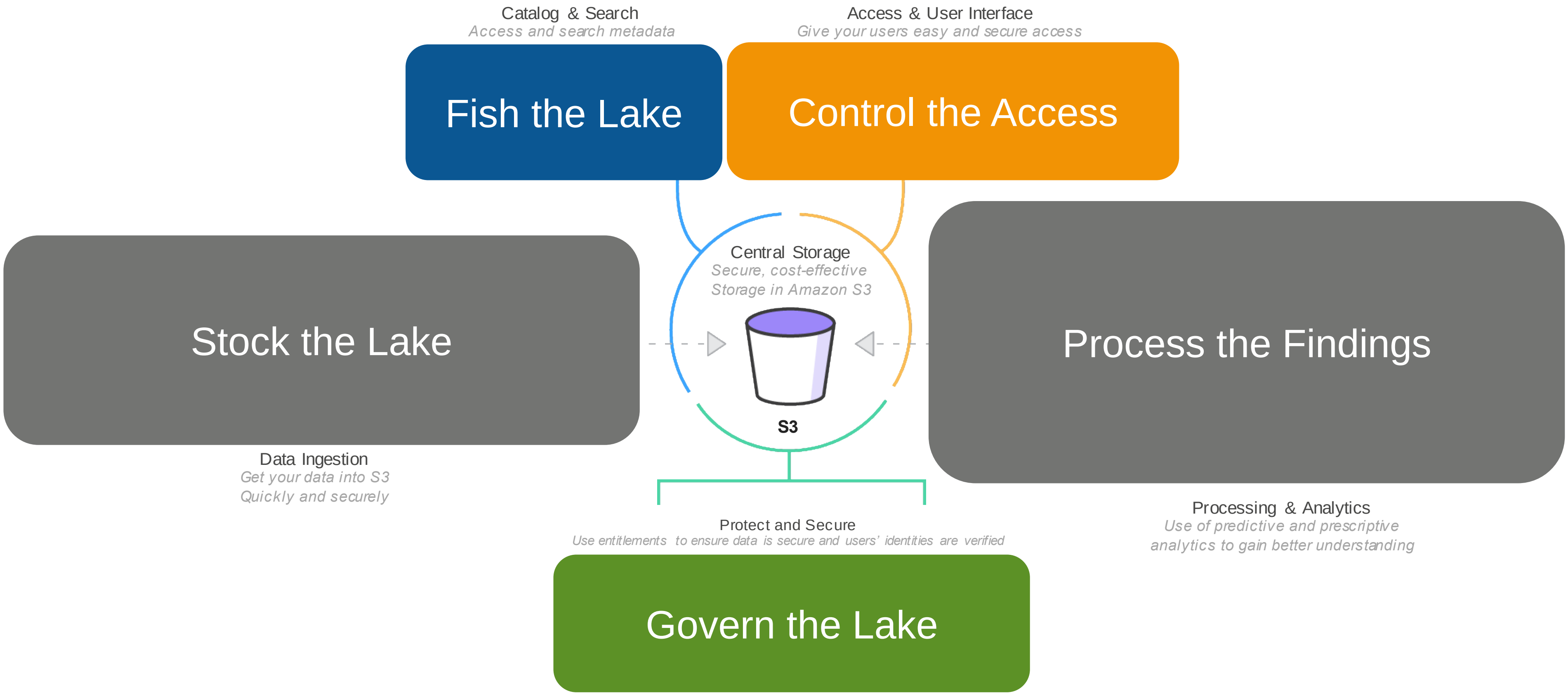
Why AWS?

Implementing a Data Lake architecture requires a broad set of tools and technologies to serve an increasingly diverse set of applications and use cases.

Boils down to:

Picking the right tool for the right job…on a consumption-based model…

Data Lake reference architecture



Catalog & Search

Access and search metadata

Fish the Lake

Access & User Interface

Give your users easy and secure access

Control the Access

Stock the Lake

Data Ingestion

Get your data into S3
Quickly and securely

Central Storage

Secure, cost-effective
Storage in Amazon S3

S3

Process the Findings

Processing & Analytics

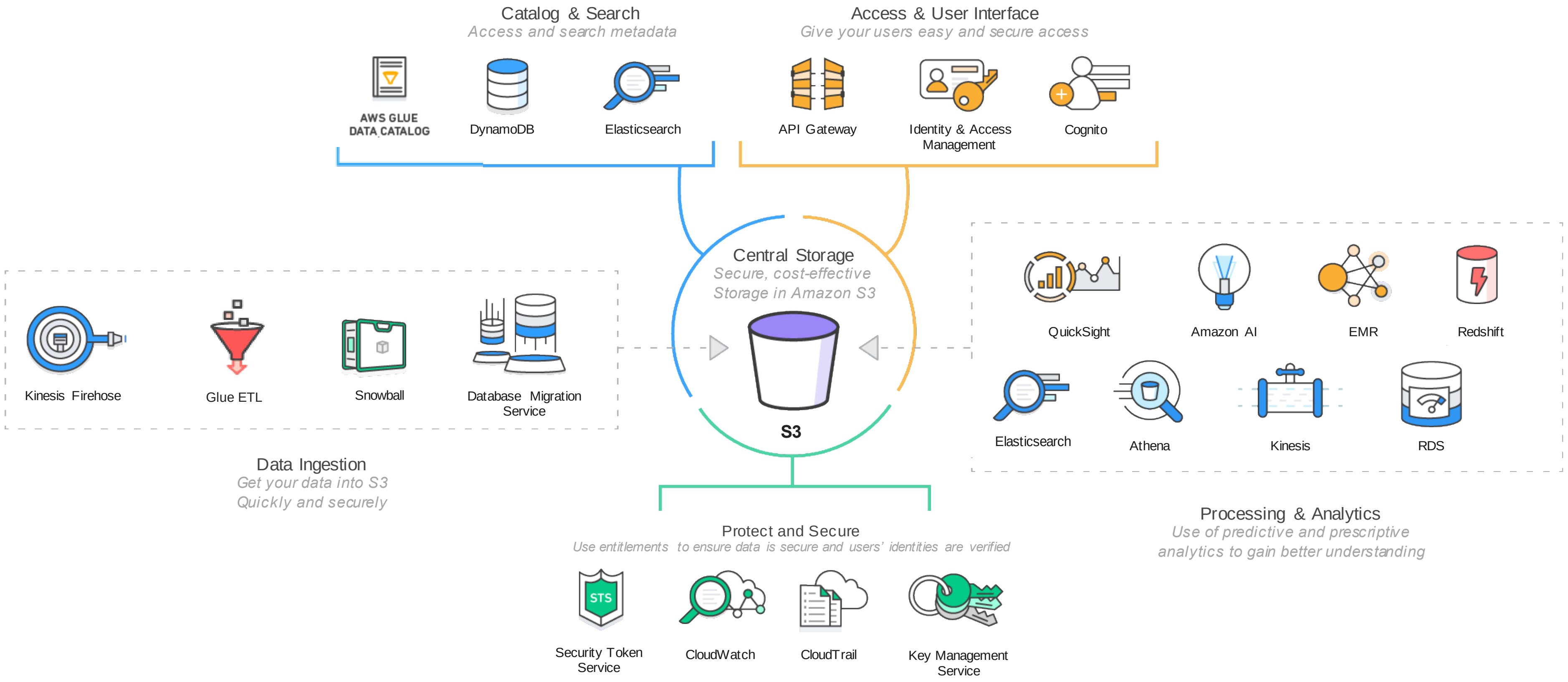
Use of predictive and prescriptive
analytics to gain better understanding

Protect and Secure

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

Govern the Lake

Data Lake reference architecture



S3 – Center of the Data Lake

Why Amazon S3 for Data Lake?

Durable

Designed for 11 9s
of durability

Available

Designed for
99.99% availability

High performance

- Multiple Upload
- Range GET

Easy to use

- Simple REST API
- AWS SDKs
- Read-after-create consistency
- Event notification
- Lifecycle policies

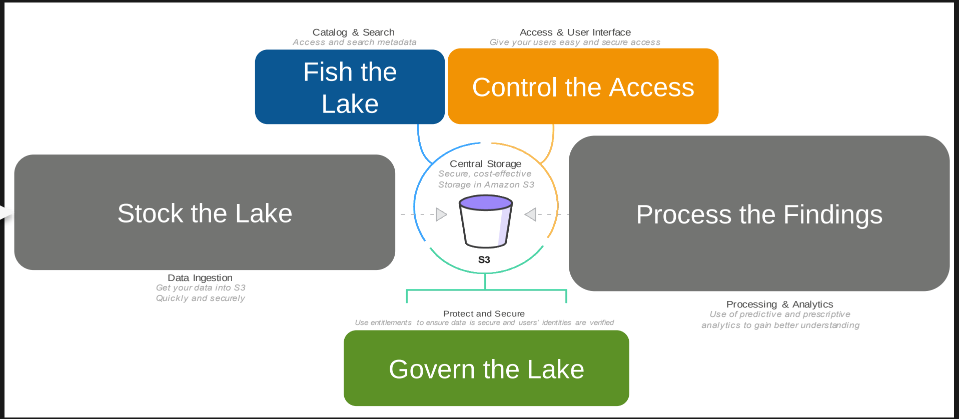
Scalable

- Store as much as you need
- Scale storage and compute independently
- No minimum usage commitments

Integrated

- Amazon EMR
- Amazon Redshift
- Amazon DynamoDB

Stock the Lake – Data Ingestion



Catalog & Search
Access and search metadata

Fish the Lake

Access & User Interface
Give your users easy and secure access

Control the Access

Stock the Lake

Data Ingestion
Get your data into S3
Quickly and securely

Central Storage
Secure, cost-effective
Storage in Amazon S3

S3

Process the Findings

Processing & Analytics
Use of predictive and prescriptive
analytics to gain better understanding

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

Govern the Lake

Amazon Glue ETL

- Serverless ETL engine generates Python code that is entirely customizable, reusable, and portable.

Database Migration Service

Migrate your RDBMS into S3 (as well as other targets)

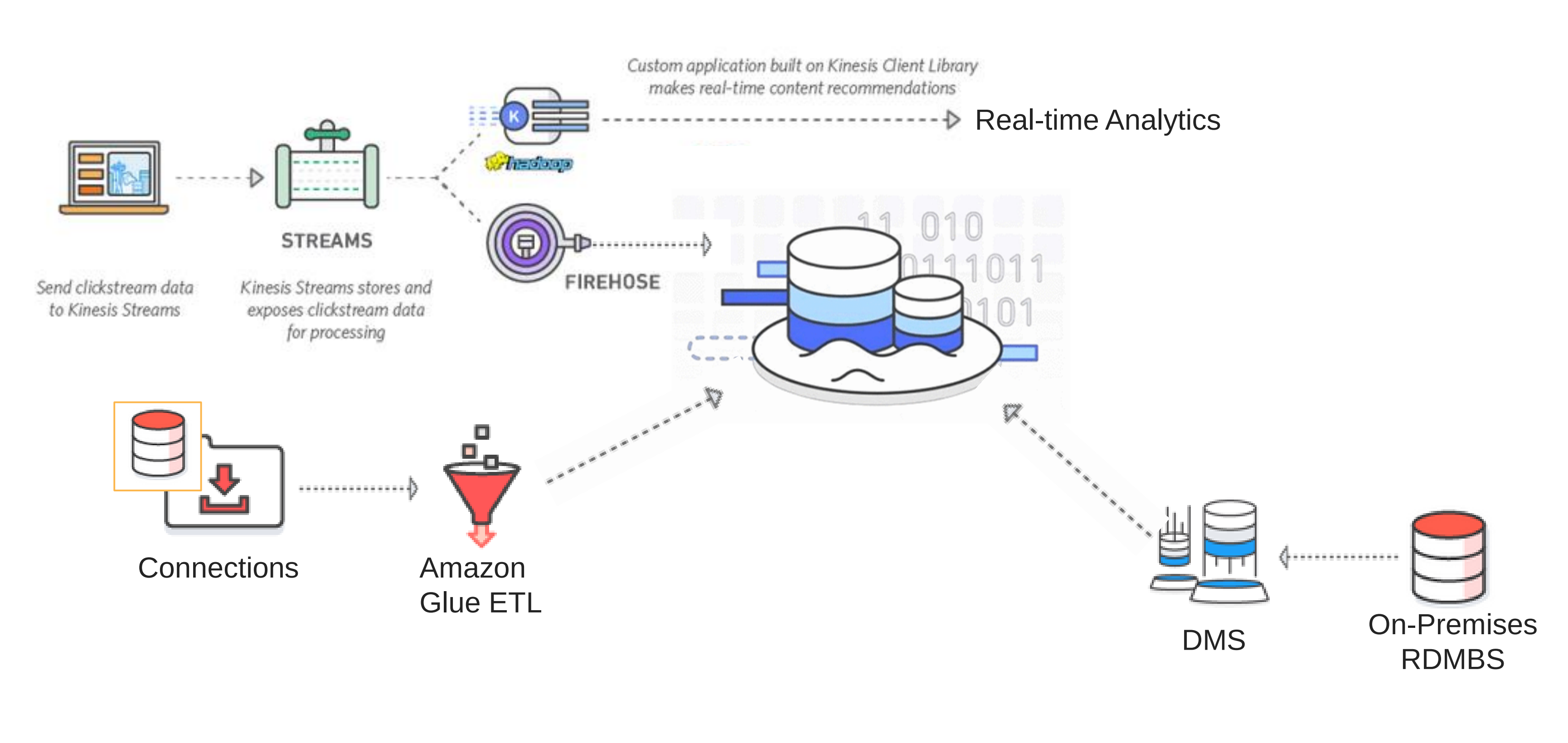
Amazon Kinesis Streams

- Build your own custom applications that process or analyze streaming data

Amazon Kinesis Firehose

- Easily load massive volumes of streaming data into S3, Amazon Redshift, and Amazon Elasticsearch Service

Stock the Lake – Data Ingestion

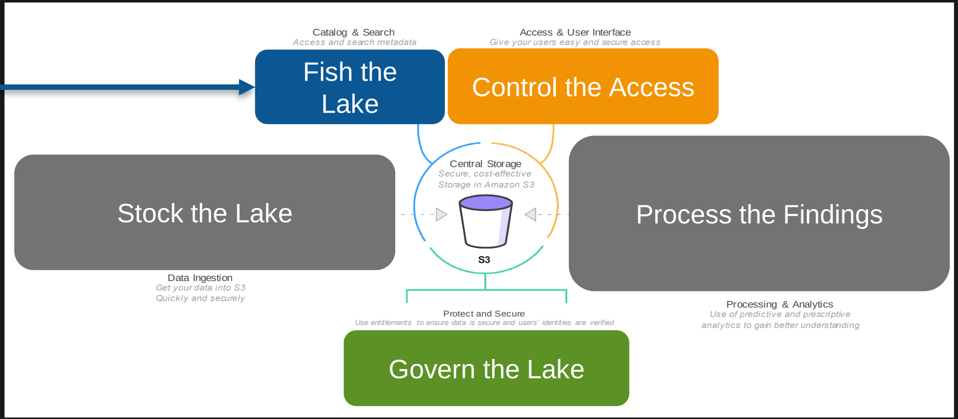

Custom application built on Kinesis Client Library makes real-time content recommendations
Real-time Analytics
K
hadoop
STREAMS
FIREHOSE
Send clickstream data to Kinesis Streams
Kinesis Streams stores and exposes clickstream data for processing
11 010
0111011
0101
Connections
Amazon Glue ETL
DMS
On-Premises RDMBS

Demonstration

What is the speed of the data?

What is the source of the data?

Fishing the Lake – Catalog/Search



Catalog & Search
Access and search metadata

Access & User Interface
Give your users easy and secure access

Fish the Lake

Control the Access

Stock the Lake

Central Storage
Secure, cost-effective Storage in Amazon S3

S3

Process the Findings

Data Ingestion
Get your data into S3 Quickly and securely

Processing & Analytics
Use of predictive and prescriptive analytics to gain better understanding

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

Govern the Lake

Glue Data Catalog

Structural and Operational metadata

Amazon DynamoDB

Metadata/Tag Lookup

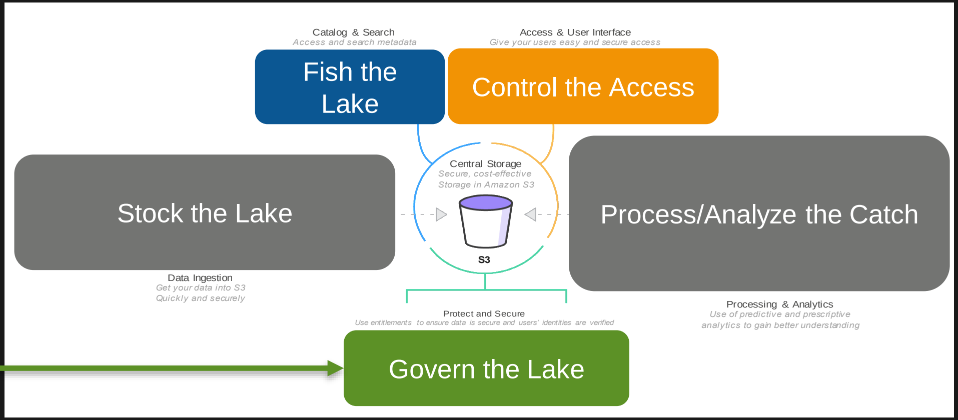
Search Content

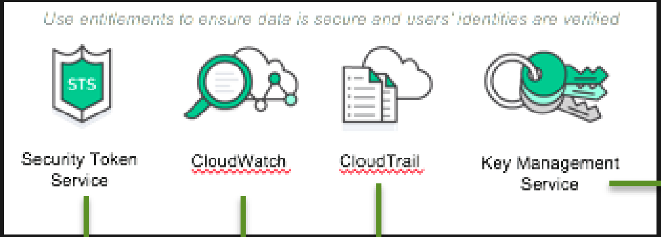
Demonstration

What type of data?

How is the data being queried?

Govern the Lake


Catalog & Search
Access and search metadata
Fish the Lake
Access & User Interface
Give your users easy and secure access
Control the Access
Central Storage
Secure, cost-effective Storage in Amazon S3
S3
Stock the Lake
Data Ingestion
Get your data into S3 Quickly and securely
Process/Analyze the Catch
Processing & Analytics
Use of predictive and prescriptive analytics to gain better understanding
Protect and Secure
Use entitlements to ensure data is secure and users' identities are verified
Govern the Lake


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

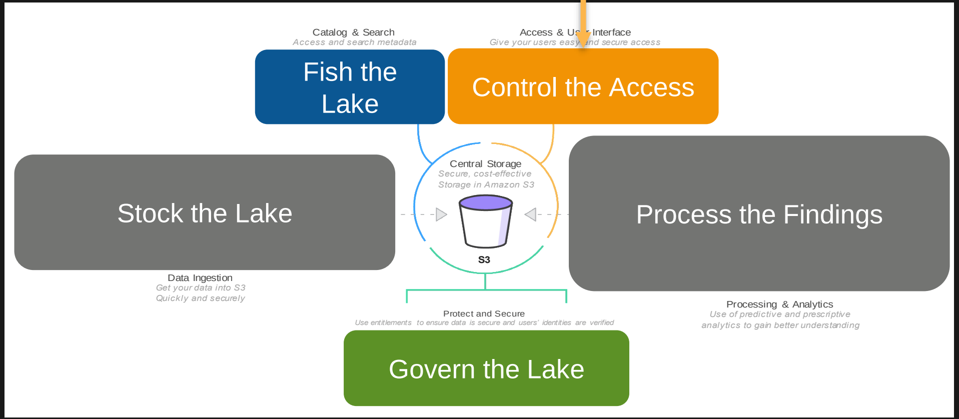
Temporary Tokens

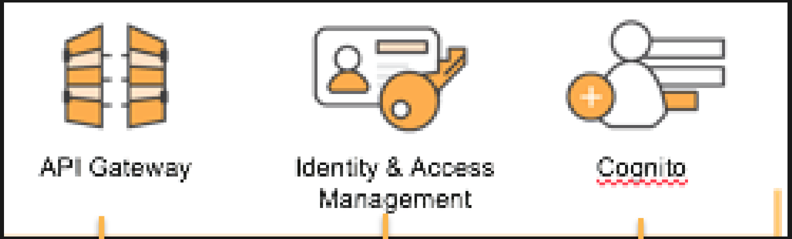
Performance

Auditing

Encryption

Control the Access


Catalog & Search
Access and search metadata
Fish the Lake
Access & User Interface
Give your users easy and secure access
Control the Access
Stock the Lake
Central Storage
Secure, cost-effective Storage in Amazon S3
S3
Process the Findings
Data Ingestion
Get your data into S3 Quickly and securely
Protect and Secure
Use entitlements to ensure data is secure and users' identities are verified
Processing & Analytics
Use of predictive and prescriptive analytics to gain better understanding
Govern the Lake



Interfaces

Identity/Access

User Authentication

Process/Analyze the Catch

Processing & Analytics

Real-time

- ElastiSearch Service
- Spark Streaming on EMR
- AWS Lambda
- Kinesis Analytics, Kinesis Streams
- Apache Flink on EMR
- Apache Storm on EMR

Batch

- EMR Hadoop, Spark, Presto
- Redshift Data Warehouse
- Athena Query Service

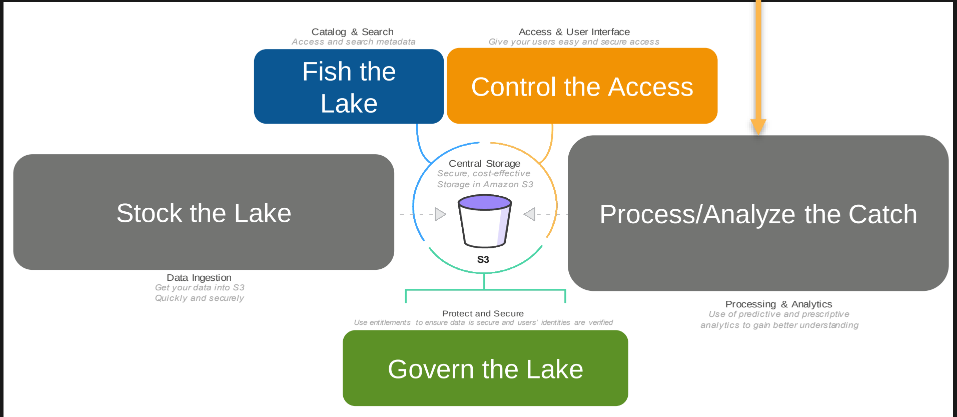
AI & Predictive

- Amazon Lex
- Amazon Rekognition
- Amazon Polly
- Machine Learning

Transactional & RDBMS

- DynamoDB, NoSQL DB
- Aurora Relational Database

BI & Data Visualization



Catalog & Search
Access and search metadata

Fish the Lake

Access & User Interface
Give your users easy and secure access

Control the Access

Stock the Lake

Data Ingestion
Get your data into S3
Quickly and securely

Central Storage
Secure, cost-effective
Storage in Amazon S3

S3

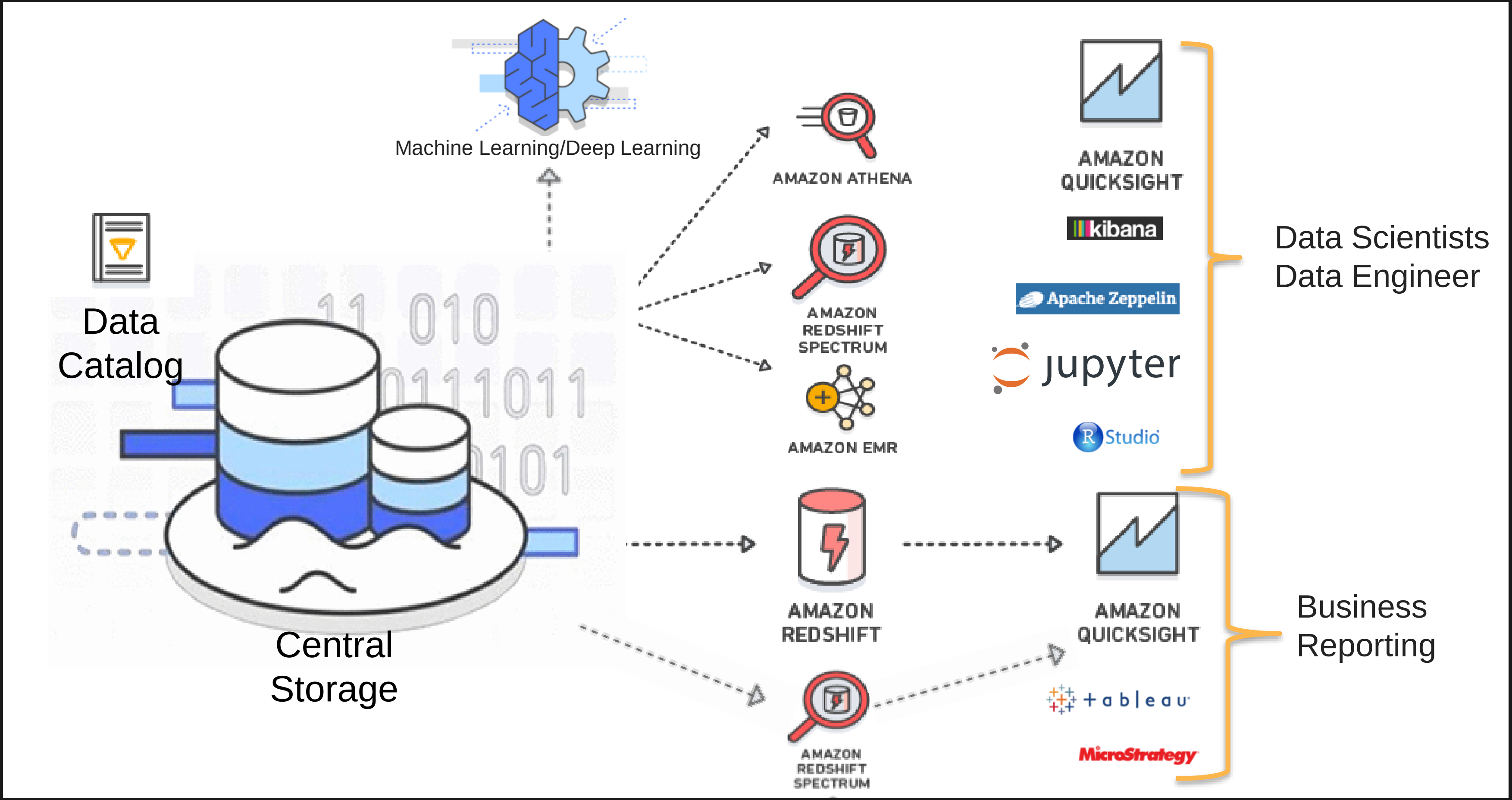
Process/Analyze the Catch

Processing & Analytics
Use of predictive and prescriptive
analytics to gain better understanding

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

Govern the Lake

Process/Analyze the Catch



Interactive query service

Amazon Athena

- Query directly from Amazon S3
- Use ANSI SQL
- Serverless
- Multiple data formats
- Pay per query

Amazon
Elastic
MapReduce

Hadoop/HDFS clusters

Hive, Pig, Impala, Hbase, Spark, Presto

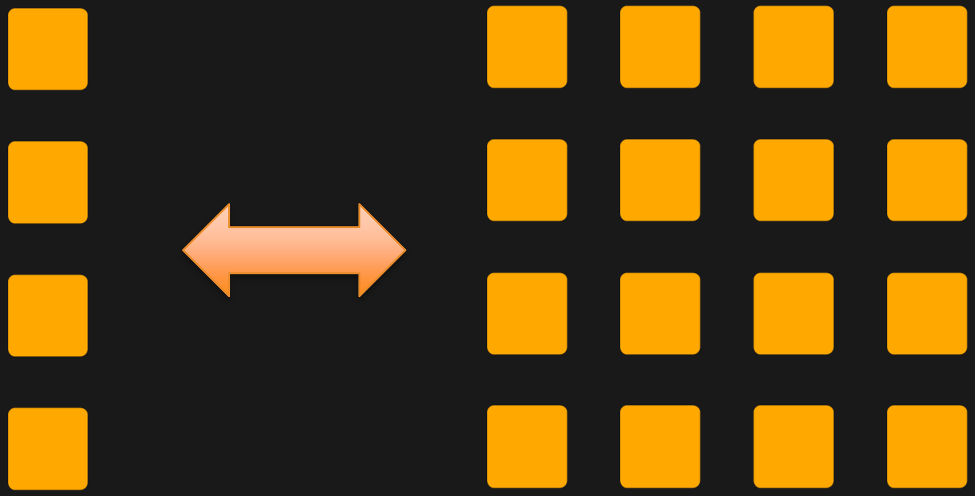
Easy to use, fully managed

On-demand, reserved instance, and Spot pricing

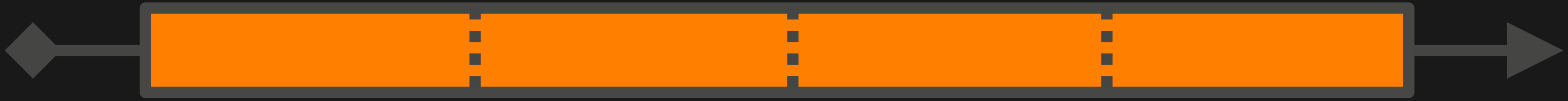
Tight integration with Amazon S3, DynamoDB, and Kinesis

Resizable clusters

Easy to add and remove compute capacity on your cluster.

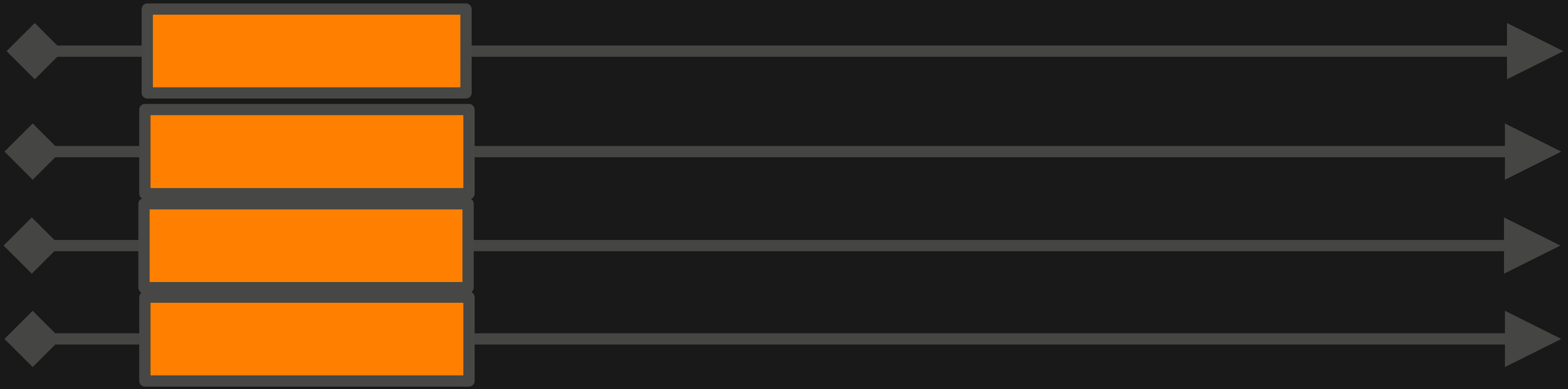


ON A SINGLE MACHINE



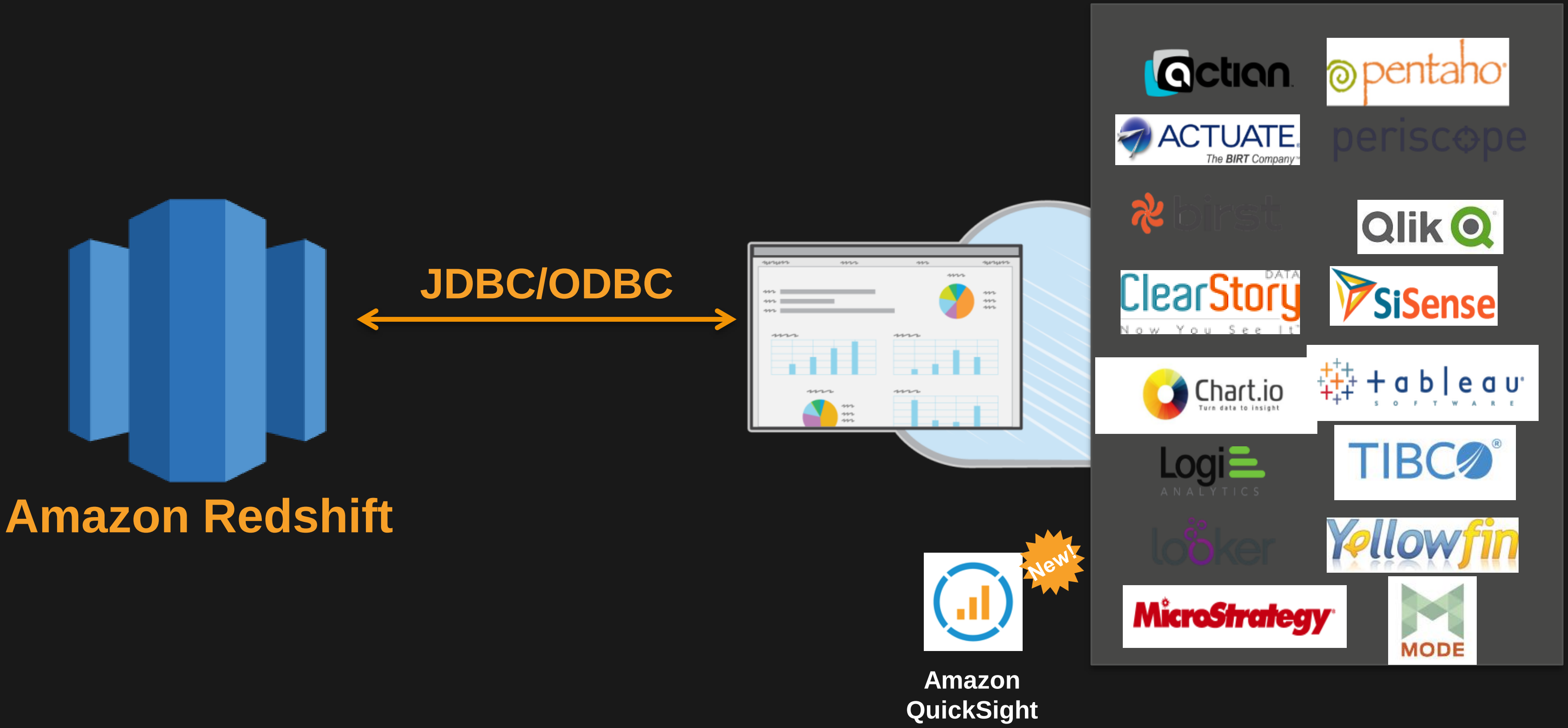
COST: 4h x $1.06 = **$4.24**
PROCESSING TIME: **4h**

ON MULTIPLE MACHINES



COST: 4 x 1h x $1.06 = **$4.24**
PROCESSING TIME: **1h**

Amazon Redshift works with third-party analysis tools



JDBC/ODBC

Amazon Redshift

New!

Amazon QuickSight

Amazon Redshift has security built in

SSL to secure data in transit

Encryption to secure data at rest

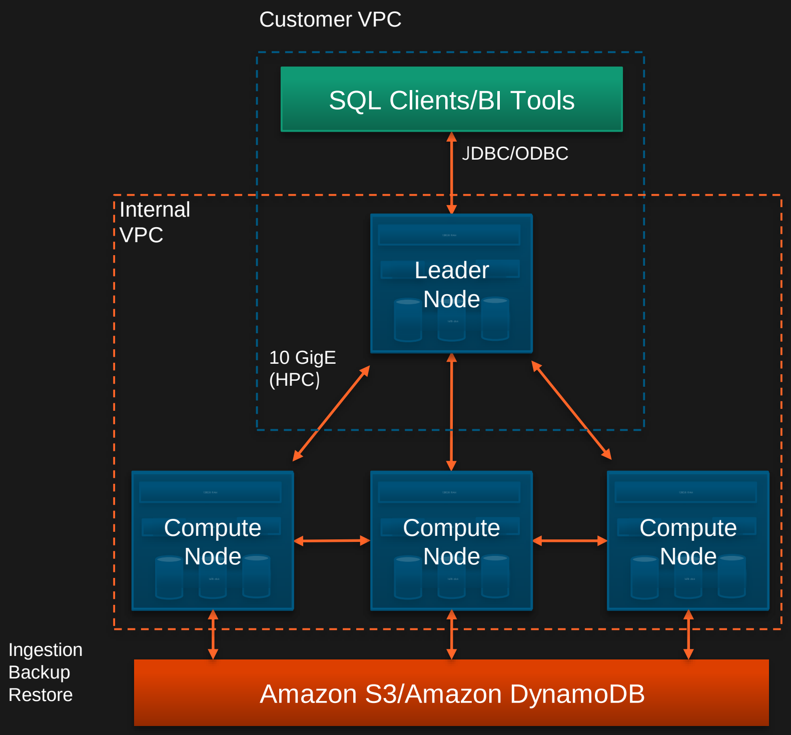
- AES-256; hardware accelerated
- All blocks on disks and in Amazon S3 encrypted
- HSM support

No direct access to compute nodes

Audit logging, AWS CloudTrail, AWS KMS integration

Amazon VPC support

SOC 1/2/3, PCI-DSS Level 1, FedRAMP, HIPAA



Redshift Spectrum

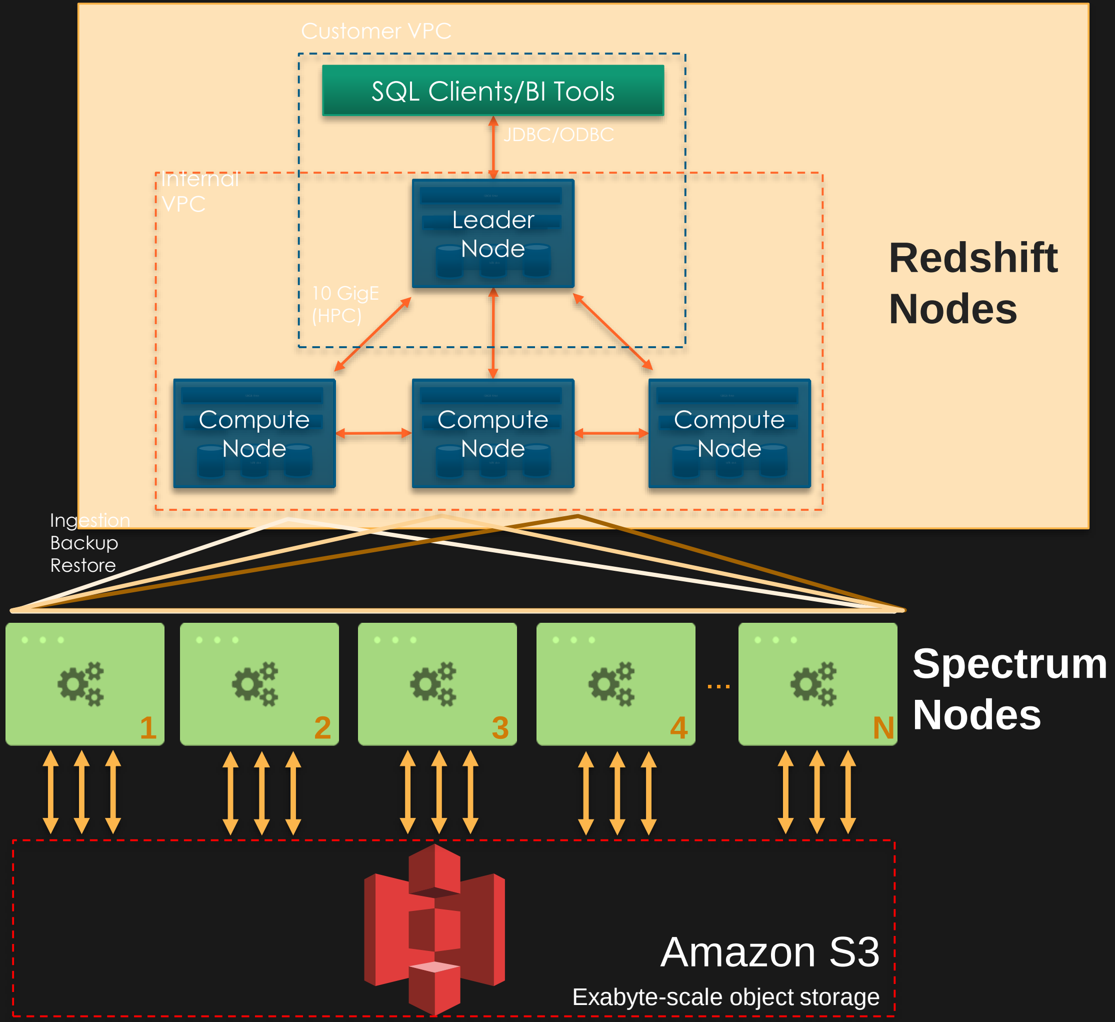
Leverages Amazon Redshift's advanced cost-based optimizer

Pushes down projections, filters, aggregations and join reduction

Dynamic partition pruning to minimize data processed

Automatic parallelization of query execution against Amazon S3 data

Efficient join processing within the Amazon Redshift cluster

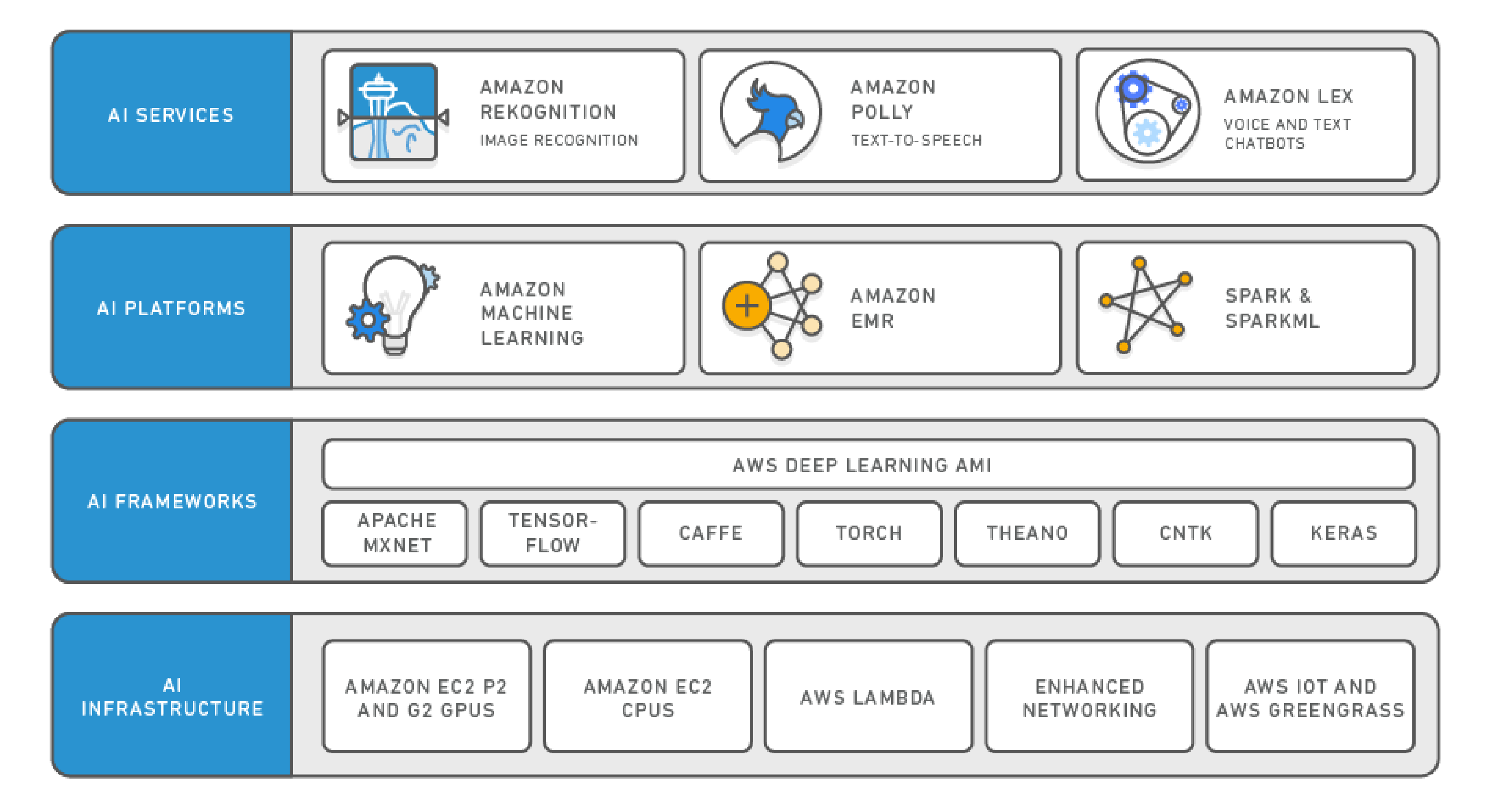


Data Catalog

Apache Hive Metastore

Amazon AI

Intelligent services powered by deep learning



AI SERVICES

- AMAZON REKOGNITION IMAGE RECOGNITION
- AMAZON POLLY TEXT-TO-SPEECH
- AMAZON LEX VOICE AND TEXT CHATBOTS

AI PLATFORMS

- AMAZON MACHINE LEARNING
- AMAZON EMR
- SPARK & SPARKML

AI FRAMEWORKS

- AWS DEEP LEARNING AMI
- APACHE MXNET
- TENSORFLOW
- CAFFE
- TORCH
- THEANO
- CNTK
- KERAS

AI INFRASTRUCTURE

- AMAZON EC2 P2 AND G2 GPUS
- AMAZON EC2 CPUS
- AWS LAMBDA
- ENHANCED NETWORKING
- AWS IOT AND AWS GREENGRASS

Proven customer success

The vast majority of big data use cases deployed in the cloud today run on AWS.

GE
airbnb
HEARST
FINRA
AdRoll
Johnson & Johnson
NETFLIX
Nasdaq
mlbam
Kellogg's
REDFIN
lyft
Pfizer
SIEMENS
PHILIPS
yelp
Zillow
BUILD FAX
Unilever

Case study: Re-architecting compliance

> *"For our market surveillance systems, we are looking at about 40% [savings with AWS], but the real benefits are the business benefits: We can do things that we physically weren't able to do before, and that is priceless."*
>
> \- Steve Randich, CIO

FINRA

What FINRA needed

- Infrastructure for its market surveillance platform
- Support of analysis and storage of approximately 75 billion market events every day

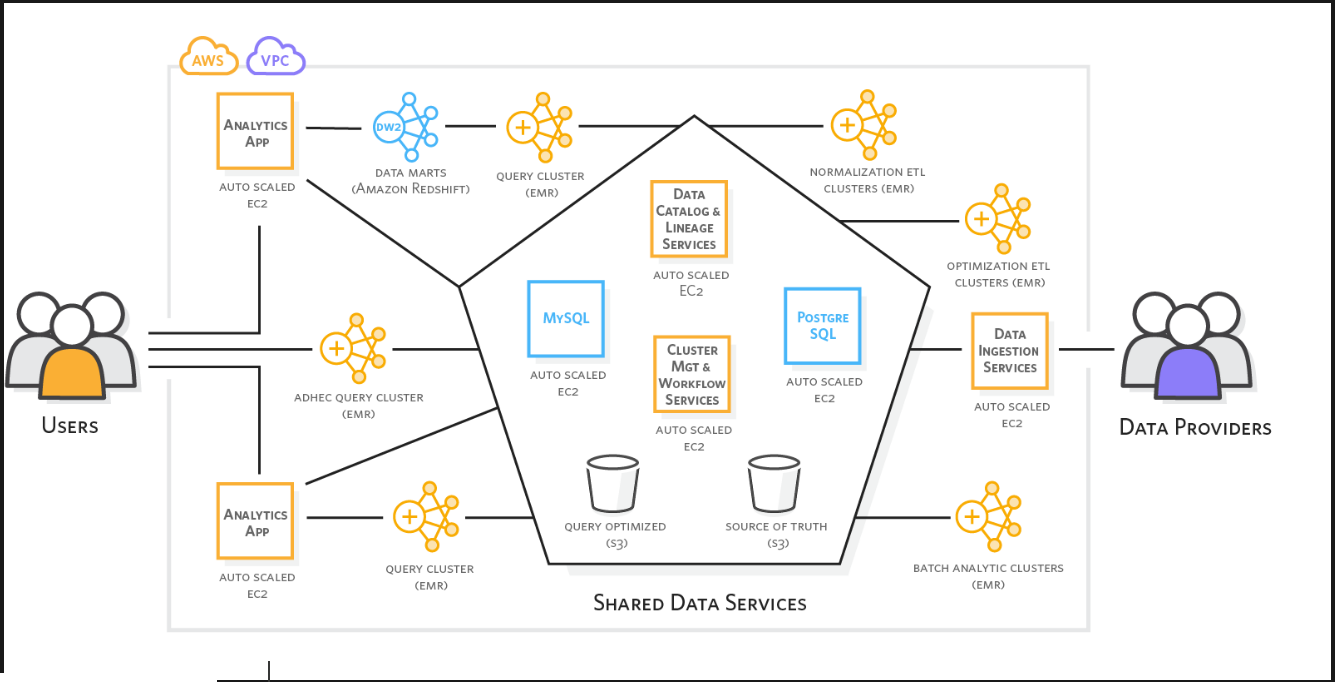
Why they chose AWS

- Fulfillment of FINRA's security requirements
- Ability to create a flexible platform using dynamic clusters (Hadoop, Hive, and HBase), Amazon EMR, and Amazon S3

Benefits realized

- Increased agility, speed, and cost savings
- Estimated savings of $10-20M annually by using AWS

Fraud detection



FINRA

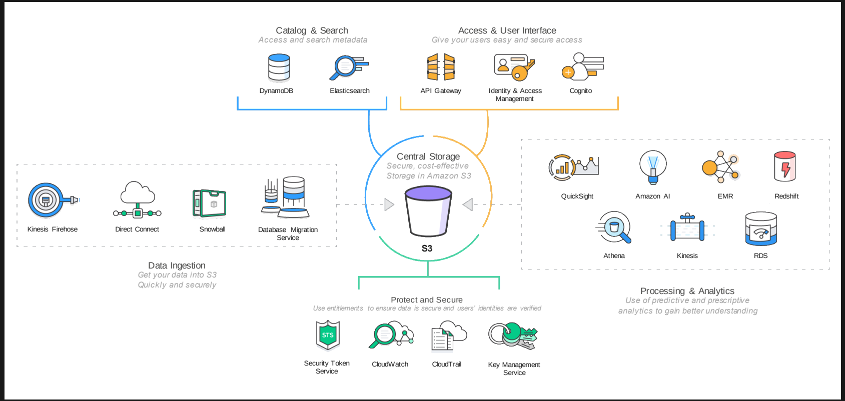
FINRA uses Amazon EMR and Amazon S3 to process up to 75 billion trading events per day and securely store over 5 petabytes of data, attaining savings of $10-20M per year.

Nasdaq

- Nasdaq implements an Amazon S3 data lake + Amazon Redshift data warehouse architecture
- Most recent two years of data is kept in the Amazon Redshift data warehouse and snapshotted into Amazon S3 for disaster recovery
- Data between two and five years old is kept in Amazon S3
- Presto on Amazon EMR is used to ad-hoc query data in Amazon S3
- Transitioned from an on-premises data warehouse to Amazon Redshift & Amazon S3 data lake architecture
- Over 1,000 tables migrated
- Average daily ingest of over 7B rows
- ***Migrated*** off legacy DW to AWS (start to finish) ***in 7 man-months***
- AWS costs were ***43%*** of legacy budget for the same data set (~1100 tables)

Building a Data Lake

An **architectural approach** that allows you to store massive amounts of “**raw**” data into a central location



It’s **readily available** to be **categorized, processed, analyzed, and consumed** by **diverse groups**

AWS

SUMMIT

Thank you!

amazon web services