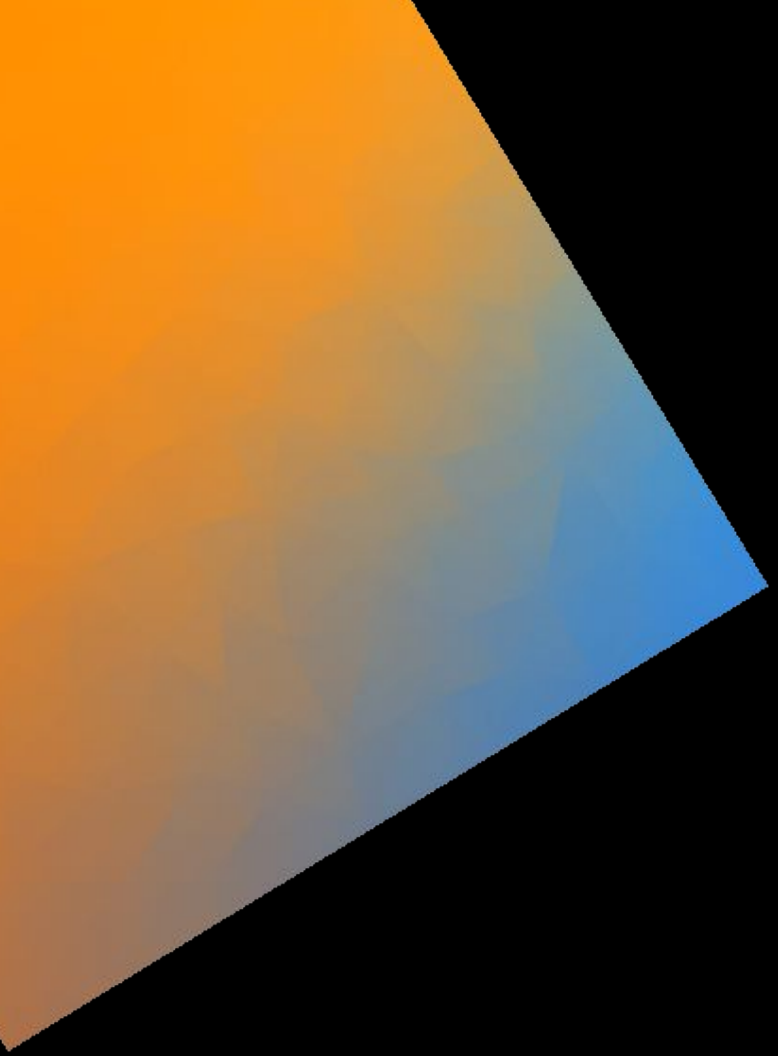


aws | databricks

Cloud Data Lake DevDay

Agenda

9:00-9:30 AM	Opening Remarks - Enabling Cloud Data Lakes for Analytics
9:30-9:45 AM	Customer Use Cases
9:45-9:55 AM	Break
9:55-10:40 AM	Architecture - Incorporating Roles; Scaling, Launching and Managing Clusters; Managing Pools and Containers
10:40-11:25 AM	Notebooks - Data Loading using Scala, Machine Learning Analytics using Python, and Redshift integration for data delivery to analysts
11:25-11:35 PM	Q&A

Current state of analytics

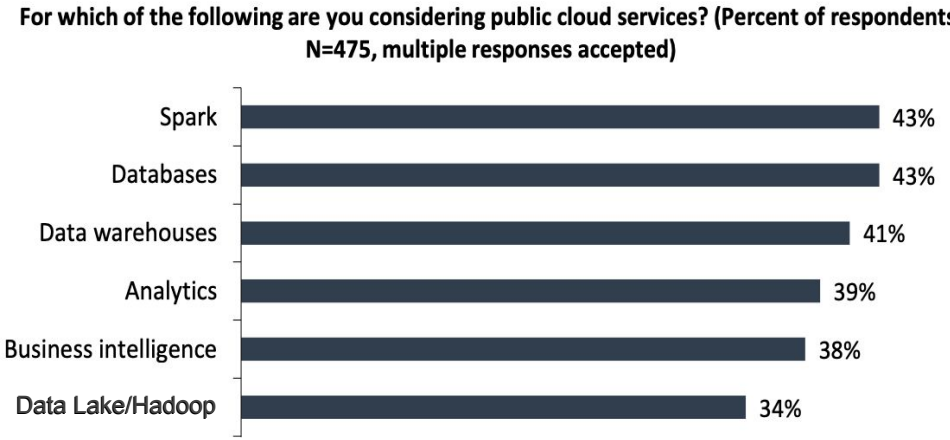
We have more data than ever before, often locked in systems and formats where we can't access it for analytics

On premises systems require excess capacity to process peak needs

Data security is costly and experts are hard to find

Analytic platforms are moving to the cloud

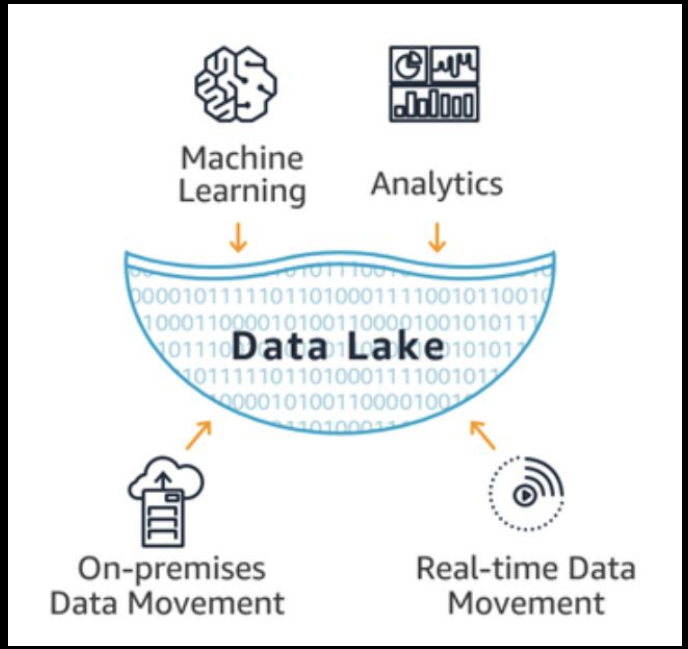
Figure 3. Migration to Cloud for Data Management Services



Source: AWS- Embracing a Data-centric Culture Anchored by a Cloud Data Lake http://bit.ly/CDLESG

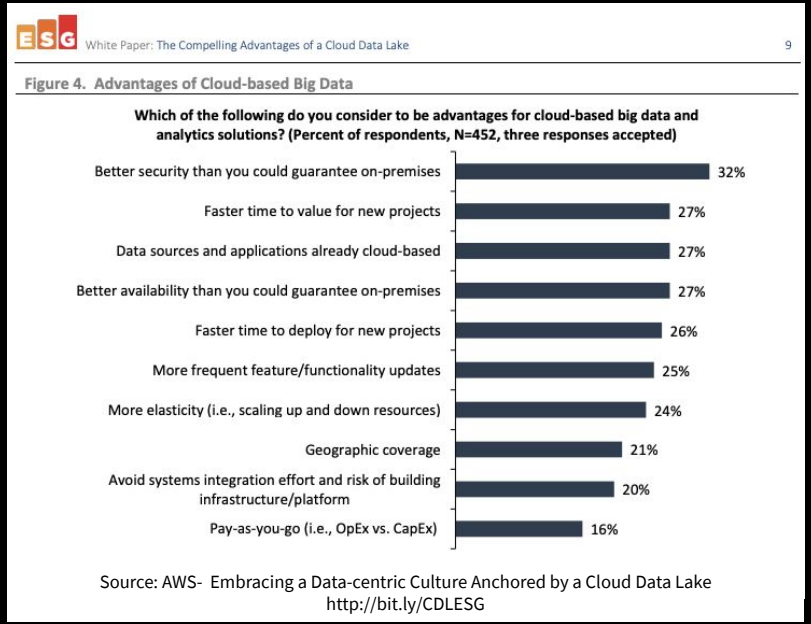
What is a data lake?

A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. Organizations see this as their golden repository.



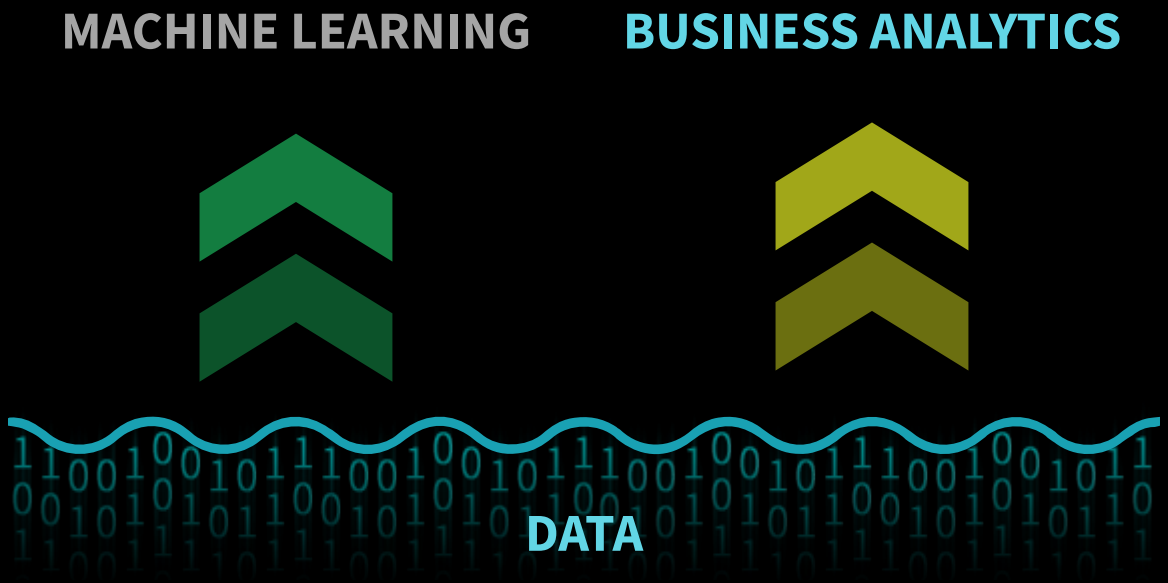
Benefits of cloud data lakes

- Pervasive security features
- Performance and scalability
- Reliability and availability
- Economics
- Integration
- Agility

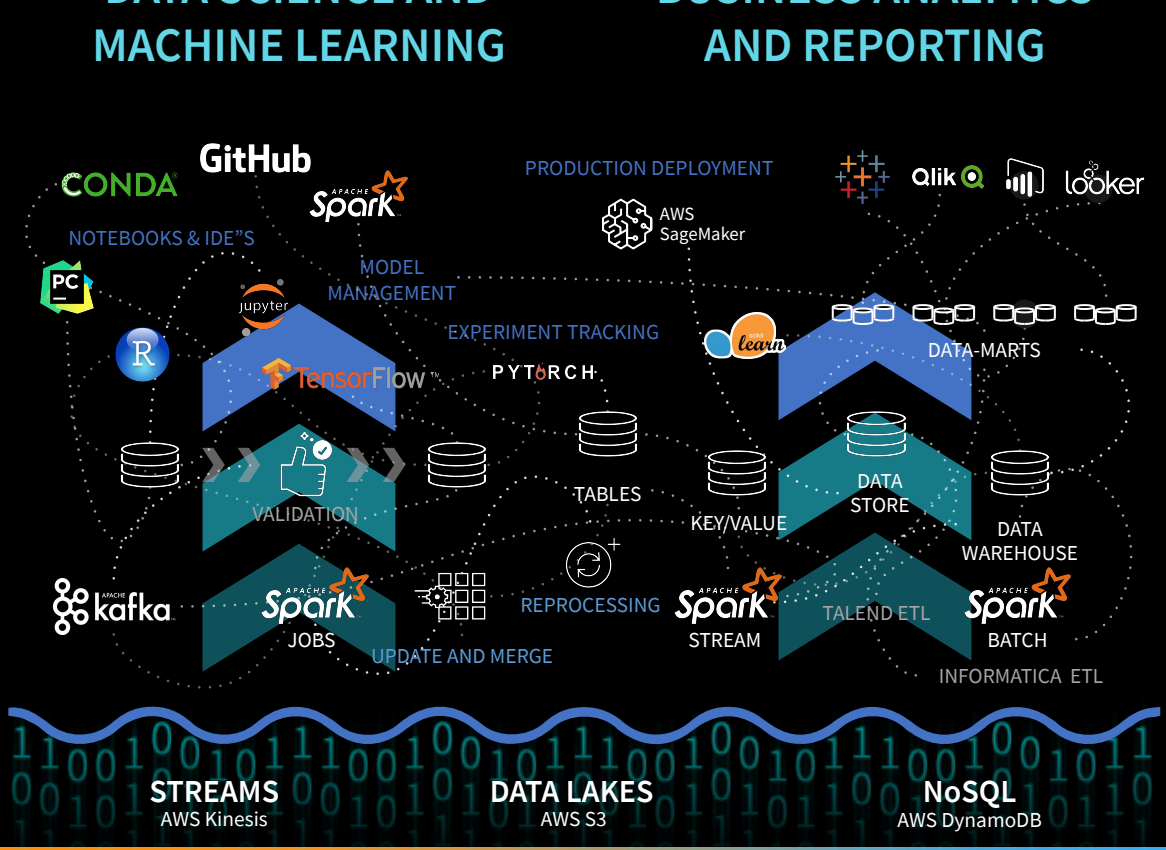


For organizations that are running analytics and machine learning workloads, the cloud makes the most sense, providing the balance of infinite scale and pay by use.

Organizations want to combine their massive **data lake** with **ML** and **BI** capabilities to unlock business value



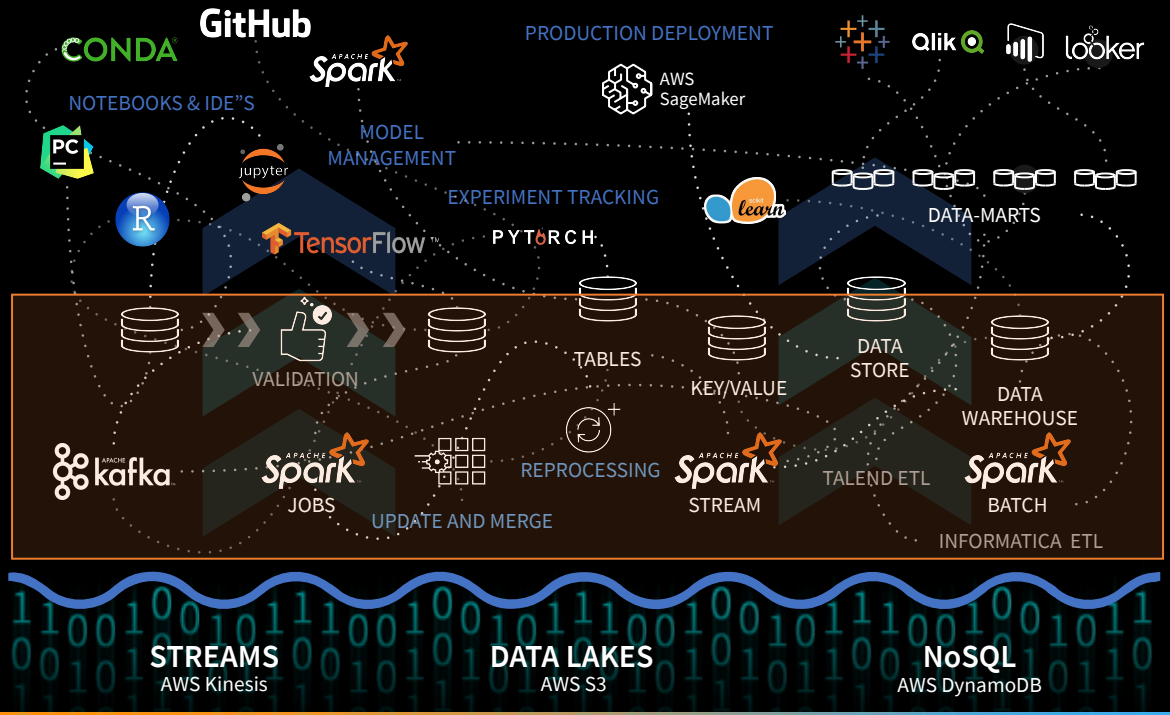
… but organizations fail to unlock business value due to data, technology and people silos



Why do organizations fail to unlock business value?

Data quality and reliability

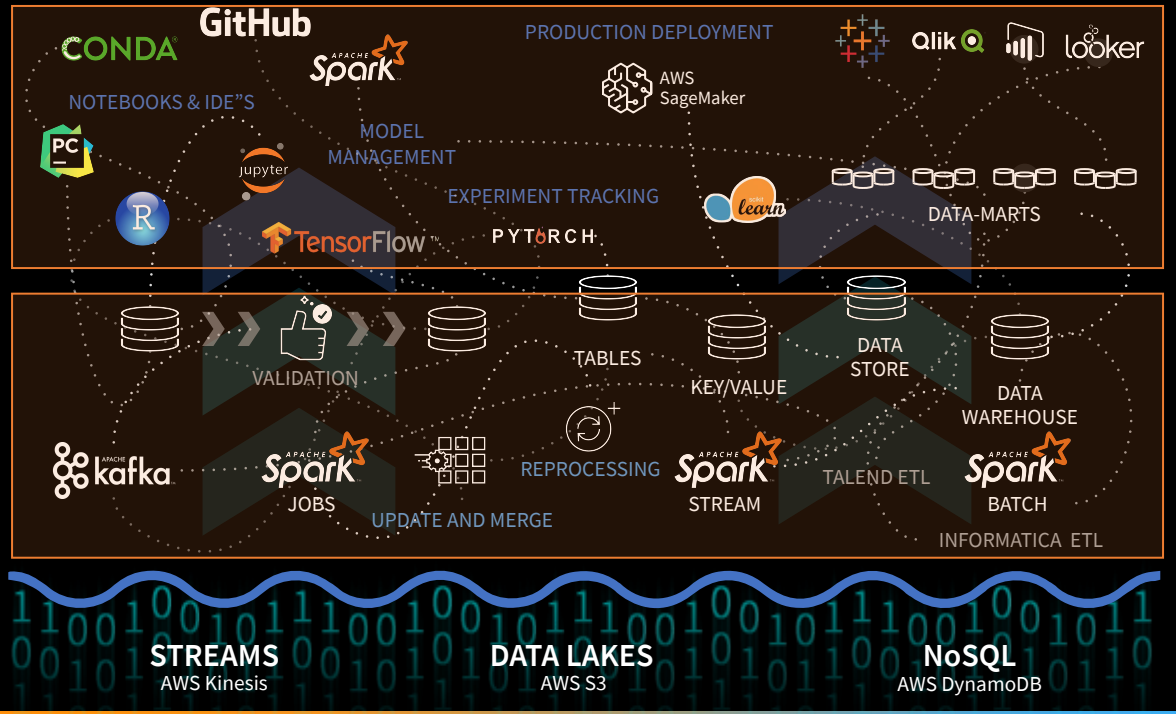
Multiple copies of inconsistent data built with unreliable data pipelines. Organizations want to guarantee data quality while retaining flexibility to ingest all data.



Why do organizations fail to unlock business value?

Disparate technologies
Need to stitch different IDE's, ML frameworks and deployment tools together

Data quality and reliability
Multiple copies of inconsistent data built with unreliable data pipelines



Why do organizations fail to unlock business value?

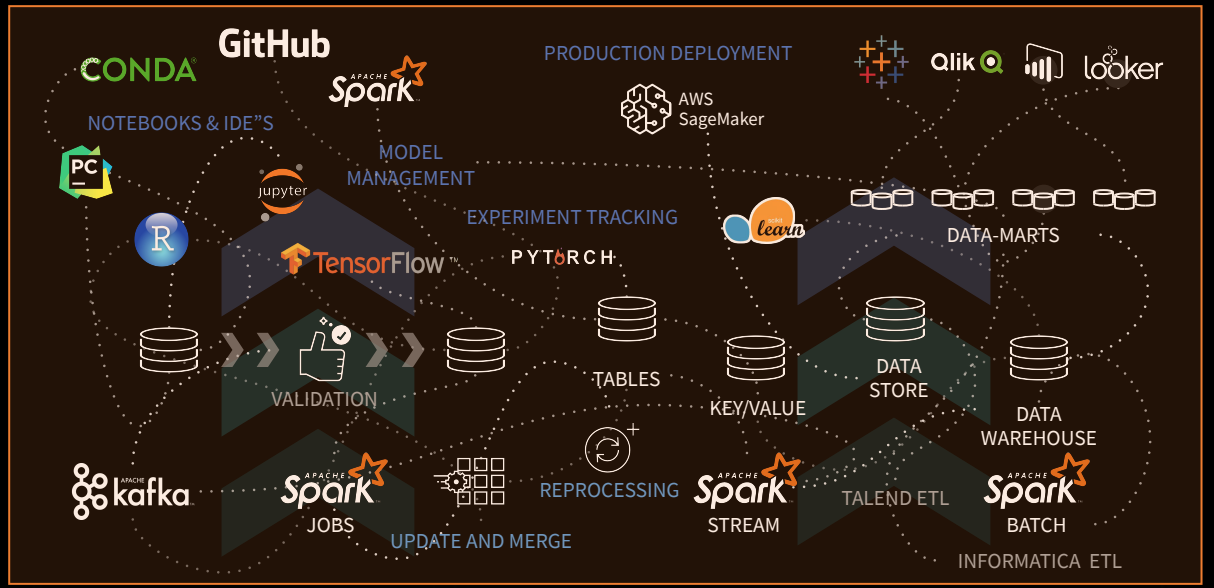
Disparate technologies
Need to stitch different IDE's, ML frameworks and deployment tools together

Data quality and reliability
Multiple copies of inconsistent data built with unreliable data pipelines

Fragmented security
End-to-end security with enterprise SLA's is a nightmare

DATA SCIENCE AND MACHINE LEARNING

BUSINESS ANALYTICS AND REPORTING



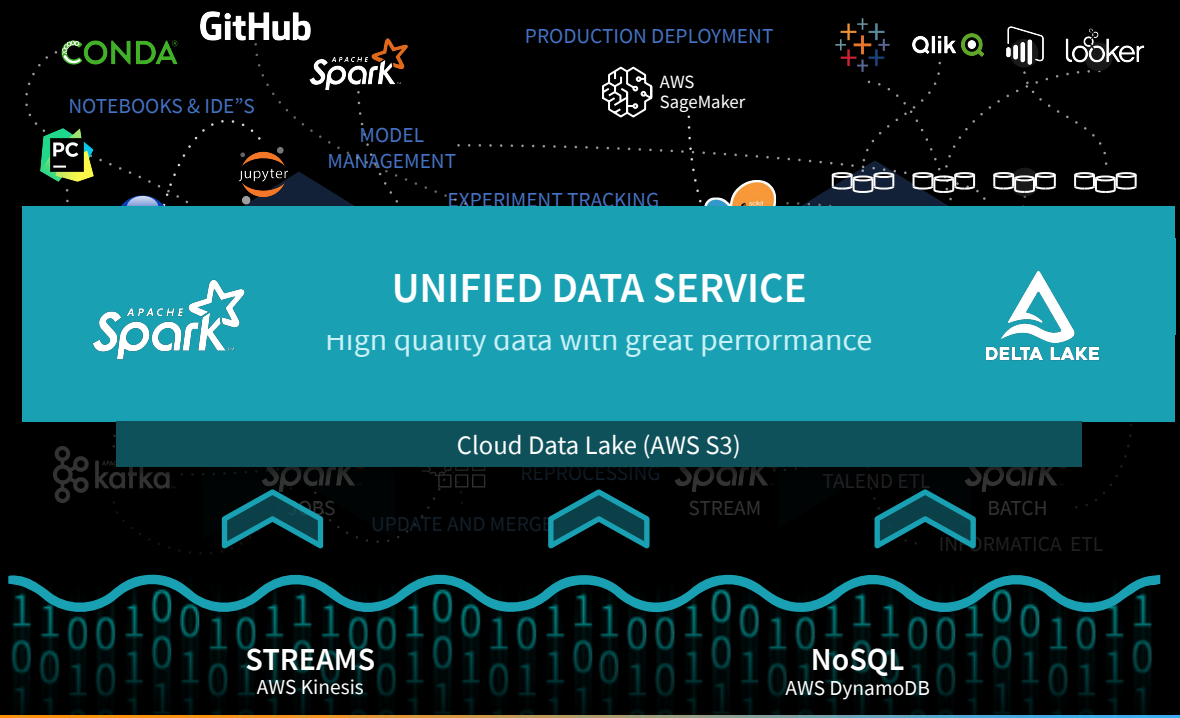
STREAMS
AWS Kinesis

DATA LAKES
AWS S3

NoSQL
AWS DynamoDB

Databricks accelerates data-driven innovation

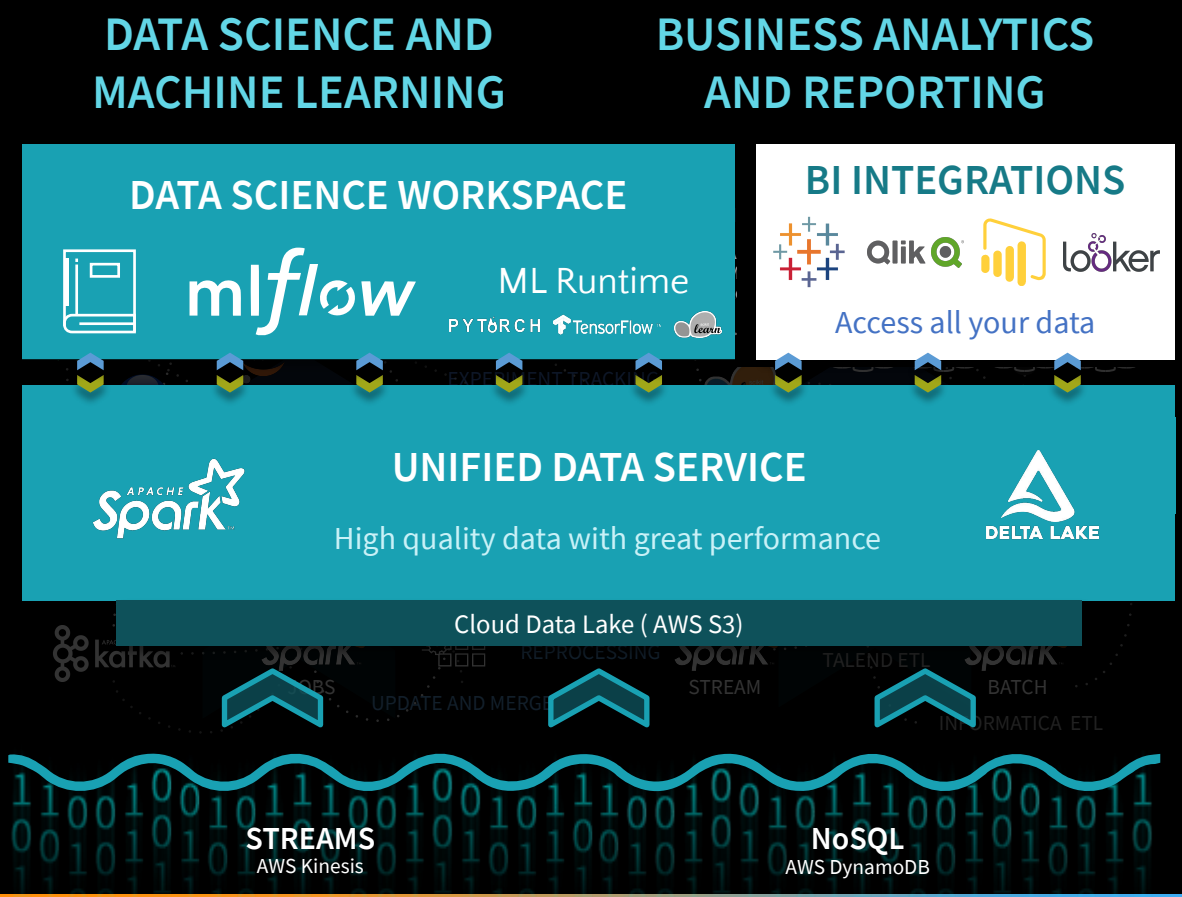
Scalable, reliable, and fast data - built on your existing data lake



Databricks accelerates data-driven innovation

Collaborative workspace for data teams across the full lifecycle

Scalable, reliable, and fast data - built on your existing data lake

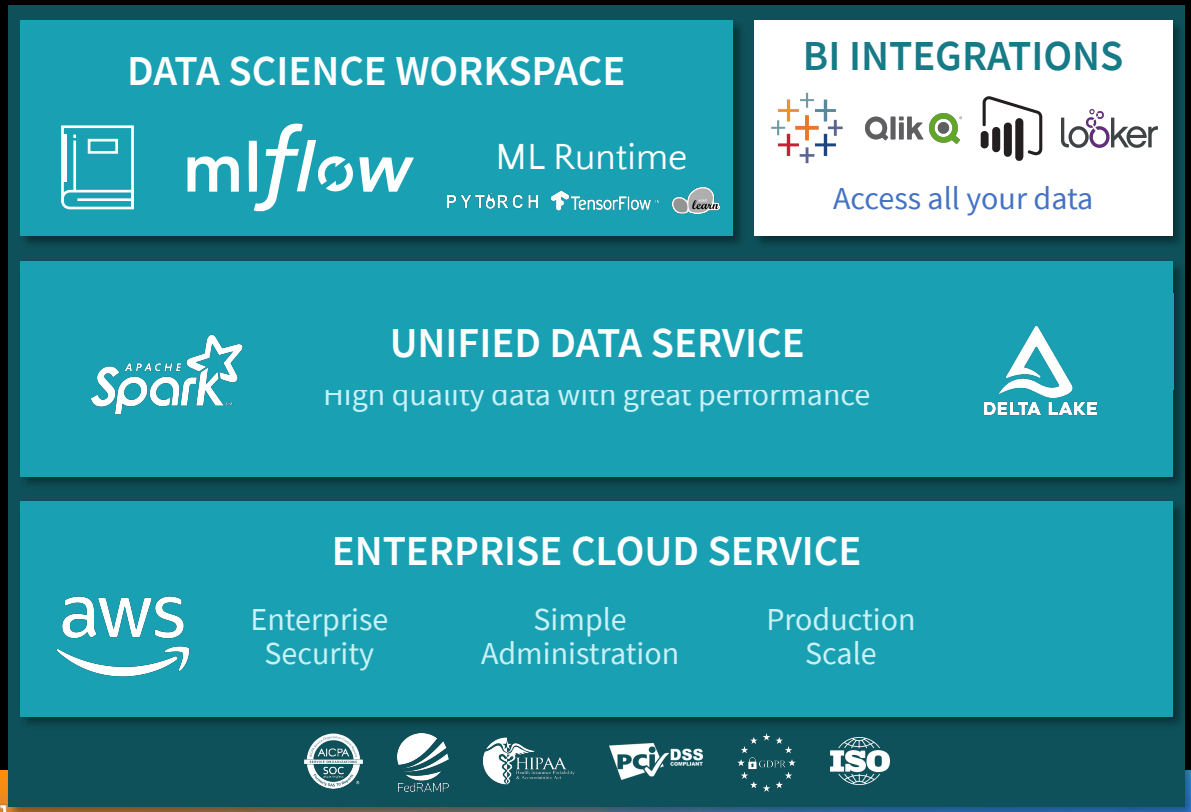


Databricks accelerates data-driven innovation

Collaborative workspace for data teams across the full lifecycle

Scalable, reliable, and fast data - built on your existing data lake

One fully-integrated security model for production infrastructure



UNIFIED DATA ANALYTICS PLATFORM

databricks

Accelerating data-driven innovation across data science, data engineering, and business analytics

mlflow **DATA SCIENCE WORKSPACE**
Collaboration across the lifecycle
PYTORCH learn TensorFlow

BI INTEGRATIONS
Qlik looker
Access all your data

APACHE Spark **UNIFIED DATA SERVICE**
High quality data with great performance
DELTA LAKE

aws **ENTERPRISE CLOUD SERVICE**
A simple, scalable, and secure managed service

RAW DATA LAKE

Successful cloud analytics implementations

Cloud Migration $100M Fin Srv Company	Pipeline Automation $1B Marketing Svcs Org	Retail ML $36B Retail Company	Fraud Detection $27B Fin Srv Company
• Previously running Hadoop on premises • Maintenance **cost $2M/yr** • Migrated to Databricks on AWS • Estimated costs savings **of almost $1M.** • Much more capable platform to drive their business	• Previous projects were **costing $65-70 per run** with Hadoop on premises • Databricks ran faster and provided expertise to optimize these jobs • Reduced cost to **under $1 per run**	• Enabled a DevOps approach to data science and productionized machine learning models • Now used by every data science team in the organization • Annual **savings of $15M** from a unified data analytics platform	• Established statistically-based fraud detection; running **successfully in production for several years** • Extending to detecting phone motion and key entry patterns to detect bots and other automated applications
Estimated cost savings up to $1M/year	Created automated pipelines that enable hundreds of machine learning models	Now powering every team and domain with data science and machine learning	Developing innovative AI applications with huge business implications

Today you will learn

- Keys to implement a cloud data lake
- How to prepare data for analytics
- How to create automated data pipelines
- How to create an infinite storage capacity that scales economically

These lessons will enable you to have a huge impact on your organization

The Role of Data Lakes

Cloud data lakes are great for data storage

Data Lake is a file system that supports

- High-velocity data
- Open storage
- Separate storage from compute
- Many structures and even unstructured data
- Lower-cost storage

Organizations want to operationalize

To operationalize data lakes, you need features you expect on a database

- Transactions
- Recovery
- Snapshots
- Indexes

But with the benefits of a data lake:

- No schema, flexible schema or fixed schema
- Integrated stream and batch processing
- Open format so your data is not locked and subject to vendor tax.

Cloud data lake blueprint

1. Manage your ingest process - handle batch and streaming data
2. Transform to open formats like Parquet
3. Utilize ACID transactions to prevent partial writes
4. Curate your data and refine it in a series of steps
5. Setup a data catalog for your data
6. Define centralized security, governance and audit procedures
7. Distribute results to LOB units for action

Introducing Delta Lake

A new standard for building data lakes

Open Format Based on Parquet

ACID Transactions

Apache Spark API's

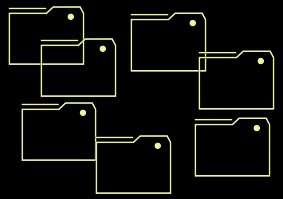
Data reliability challenges with data lakes

Lack of schema enforcement creates inconsistent and low quality data

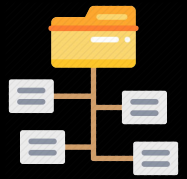
Failed production jobs leave data in corrupt state requiring tedious recovery

Lack of consistency makes it almost impossible to mix appends and reads, batch and streaming

Performance challenges with data lakes



Too many small or very big files - more time opening & closing files rather than reading contents (worse with streaming).

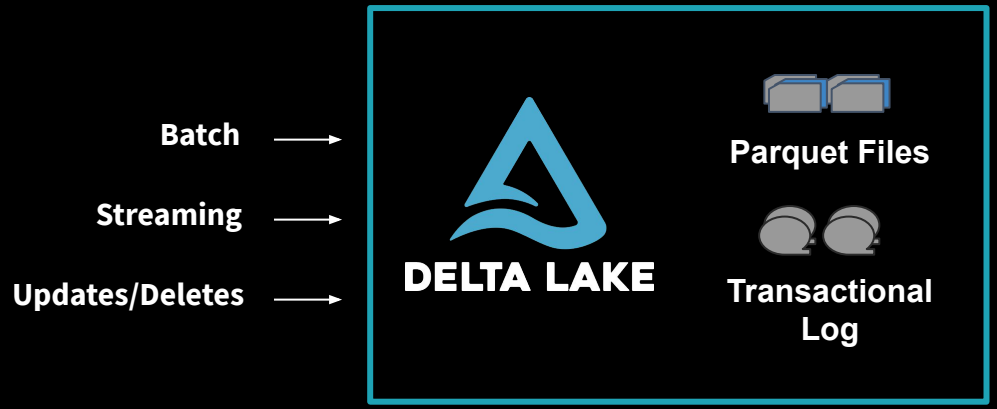


Partitioning aka “poor man’s indexing”- breaks down if you picked the wrong fields or when data has many dimensions, high cardinality columns.

No caching - storage throughput is low (storage is 20-50MB/s/core vs 300MB/s/core for local SSDs).

Delta Lake ensures data reliability

1. Manage your ingest process - handle batch and streaming data



2. Transform to open formats like Parquet

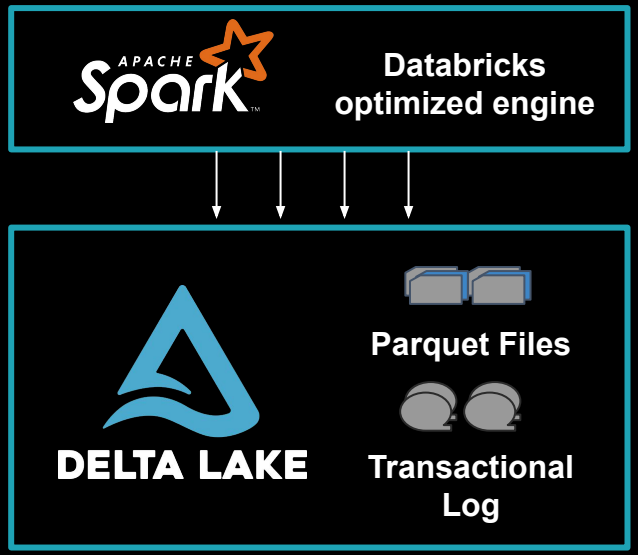
High Quality & Reliable Data always ready for analytics

3. Utilize ACID transactions to prevent partial writes

Key Features

- ACID Transactions
- Schema Enforcement
- Unified Batch & Streaming
- Time Travel/Data Snapshots

Delta Lake optimizes performance



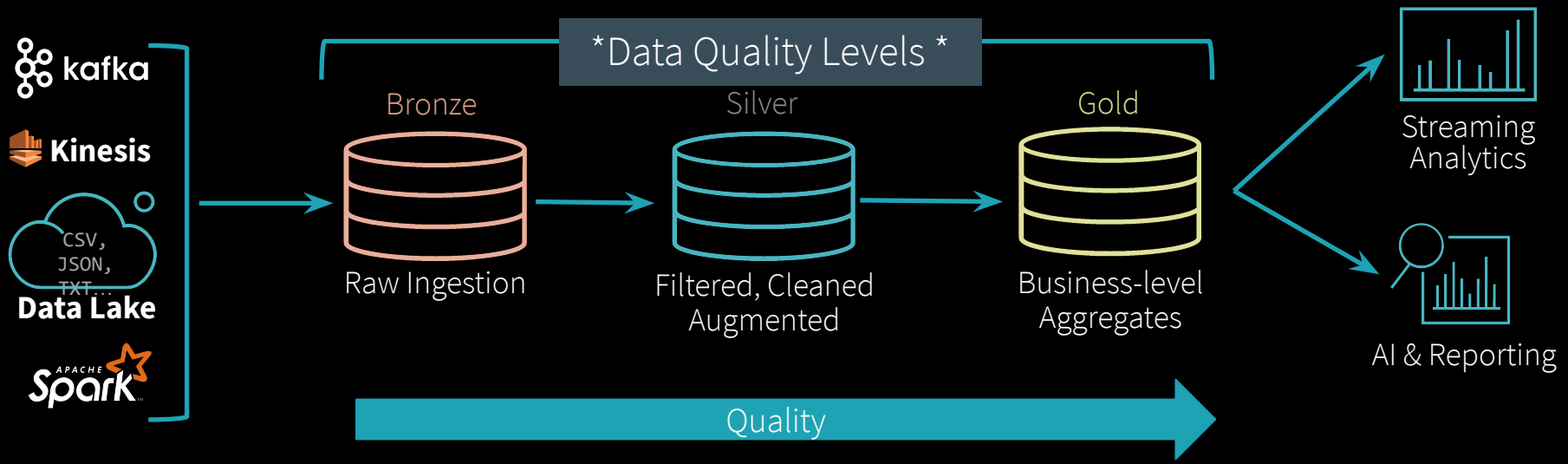
Highly Performant
queries at scale

Key Features

- Indexing
- Compaction
- Data skipping
- Caching

The DELTA LAKE

4 Curate your data and refine it in a series of steps



Delta Lake allows you to *incrementally* improve the quality of your data until it is ready for consumption.

Delta gives you the features you need

The features you expect on a database

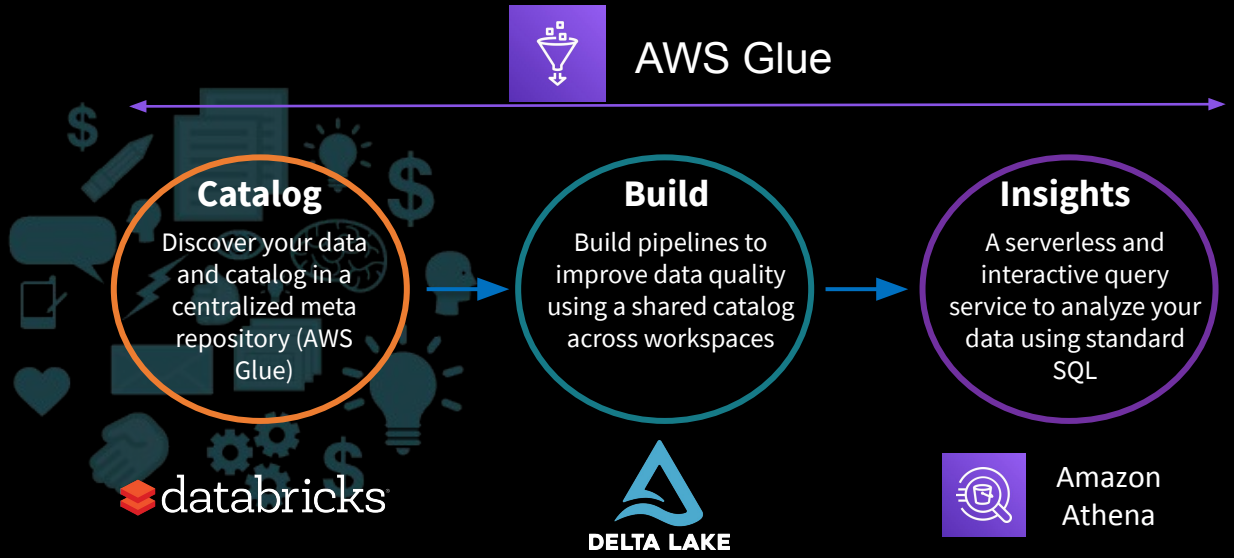
- Transactions
- Recovery
- Snapshots
- Indexes

With the benefits of a data lake:

- No schema, flexible schema or fixed schema
- Integrated stream and batch processing
- Open format so your data is not locked and subject to vendor tax.

Integration with Glue

5 Setup a data catalog for your data



Glue integration

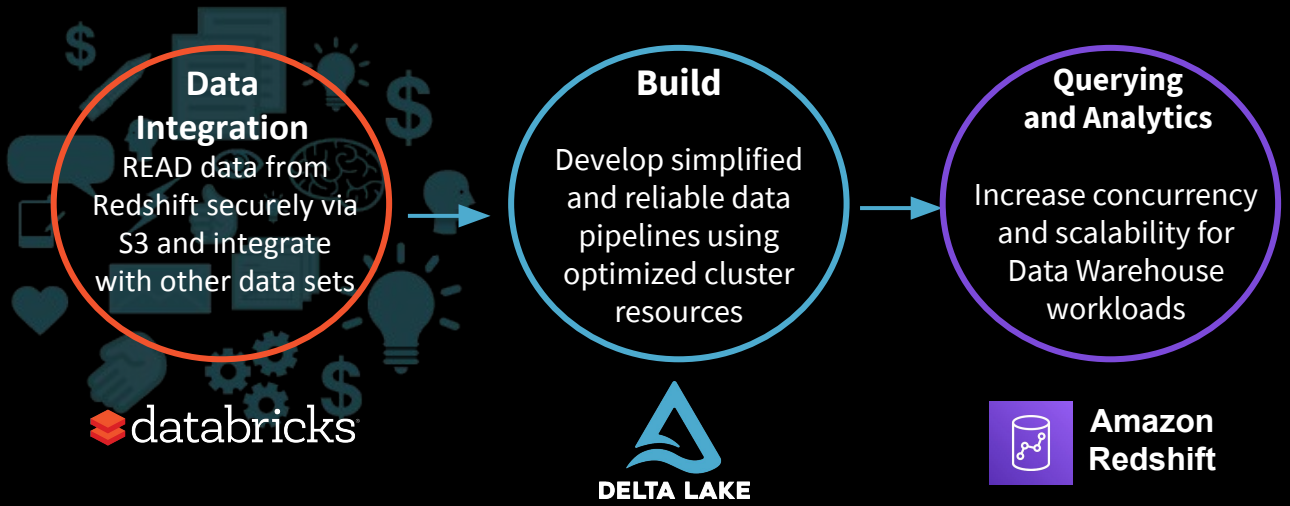
- Attach an IAM role to a Databricks cluster with the proper Glue Metastore permission
- Set the Spark Configuration, prior to launch the cluster
- Manage table definitions
- Simplifies manageability by using the same AWS Glue catalog across multiple workspace

Athena integration

- Create a manifest file and an external table definition
- Query the external table using Athena

Integration with Redshift

7 Distribute results to LOB units for action



Redshift integration

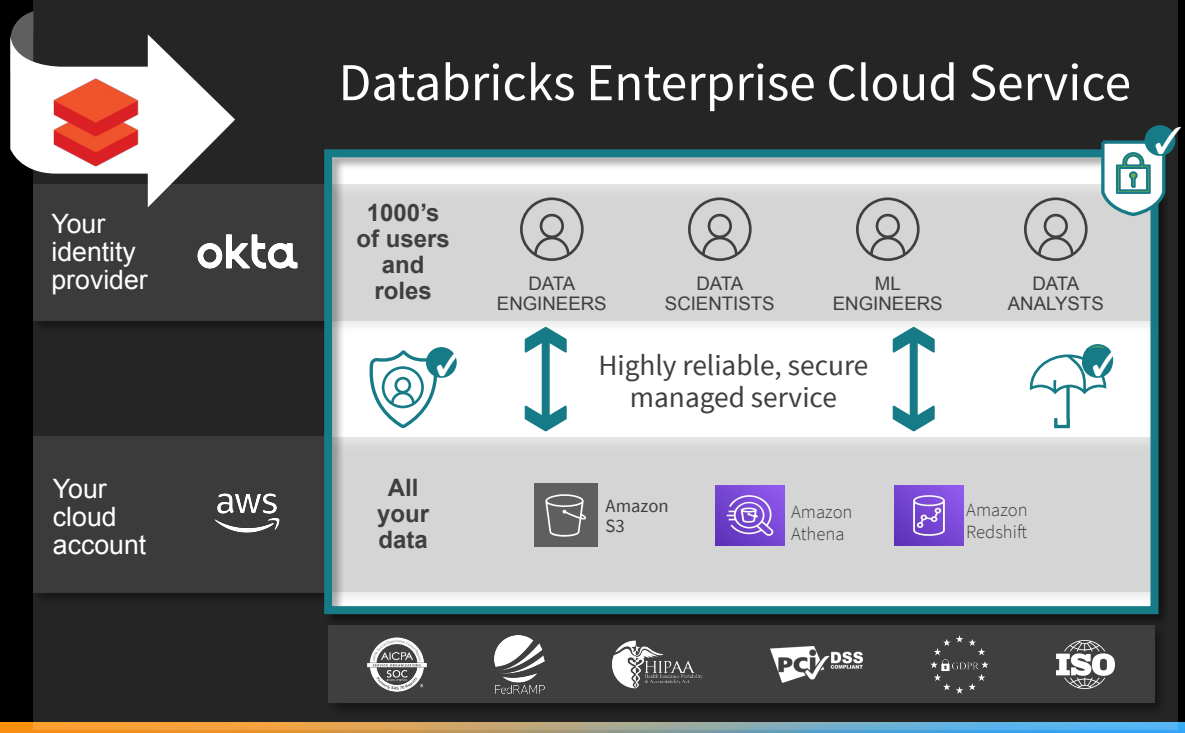
- Download and install the official Redshift JDBC driver on the cluster
- Authenticate to S3 and Redshift using IAM role
- S3 acts as an intermediary to store bulk data when reading from or writing to Redshift
- Secure your JDBC connection and encrypt the UNLOAD and COPY data operations following the Redshift documentation

Cloud native enterprise solution

6 Define centralized security, governance and audit procedures

Consistent security and roles enable easy enterprise roll-outs

- Adopt roles from SSO pass through
- Consolidated security
- Consistent governance

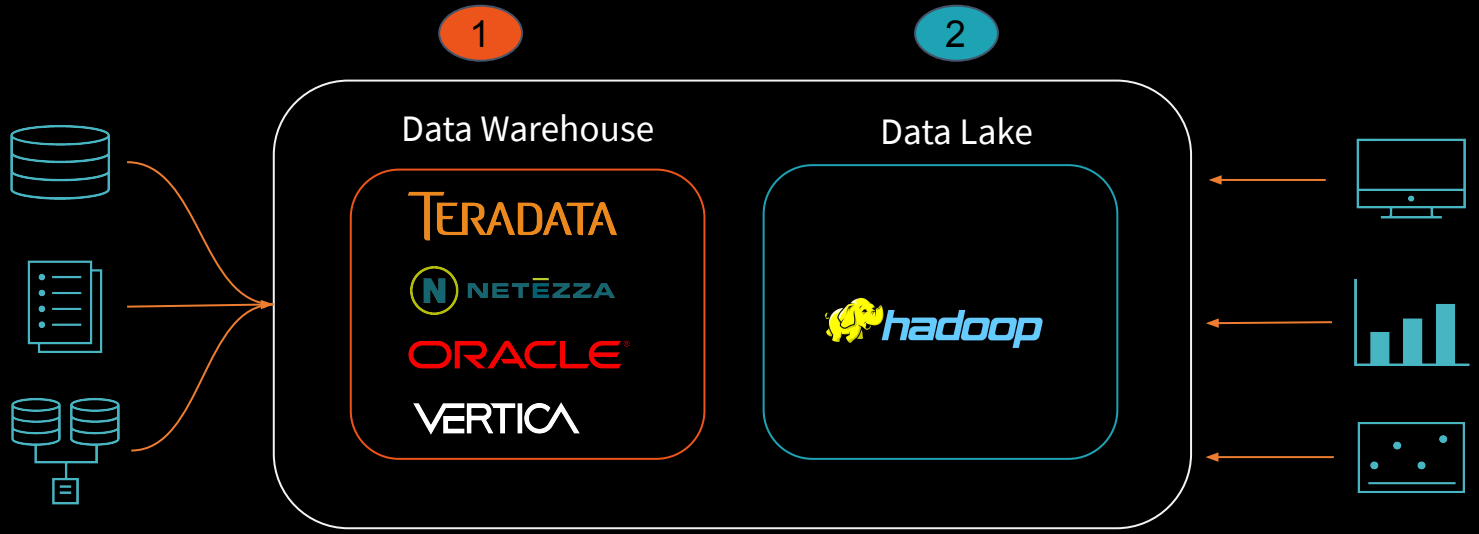


Best practices for building a cloud data lake

1. Manage your ingest process - handle batch and streaming data using Delta Lake
2. Transform to open formats like Parquet using Delta Lake
3. Utilize ACID transactions to prevent partial writes using Delta Lake
4. Curate your data and refine it in a series of steps using Delta Lake
5. Setup a data catalog for your data using Glue
6. Define centralized security, governance and audit procedures using AWS security
7. Distribute results to LOB units for action using Delta Lake or Redshift

On-premises analytics to cloud data lakes

Common Patterns



Data Sources

Legacy Data & Compute Architectures

BI, Analytics & ML

Hadoop: Complex, costly, risk-prone

1

Data reliability & performance Issues

- Hung jobs, corrupted-data
- Inconsistent schemas
- No roll-backs

2

Rigid, In-flexible clusters - cannot scale up/ down

- Unresponsive to business needs
- Cluster configurations rigid
- Hardware procurement cycles

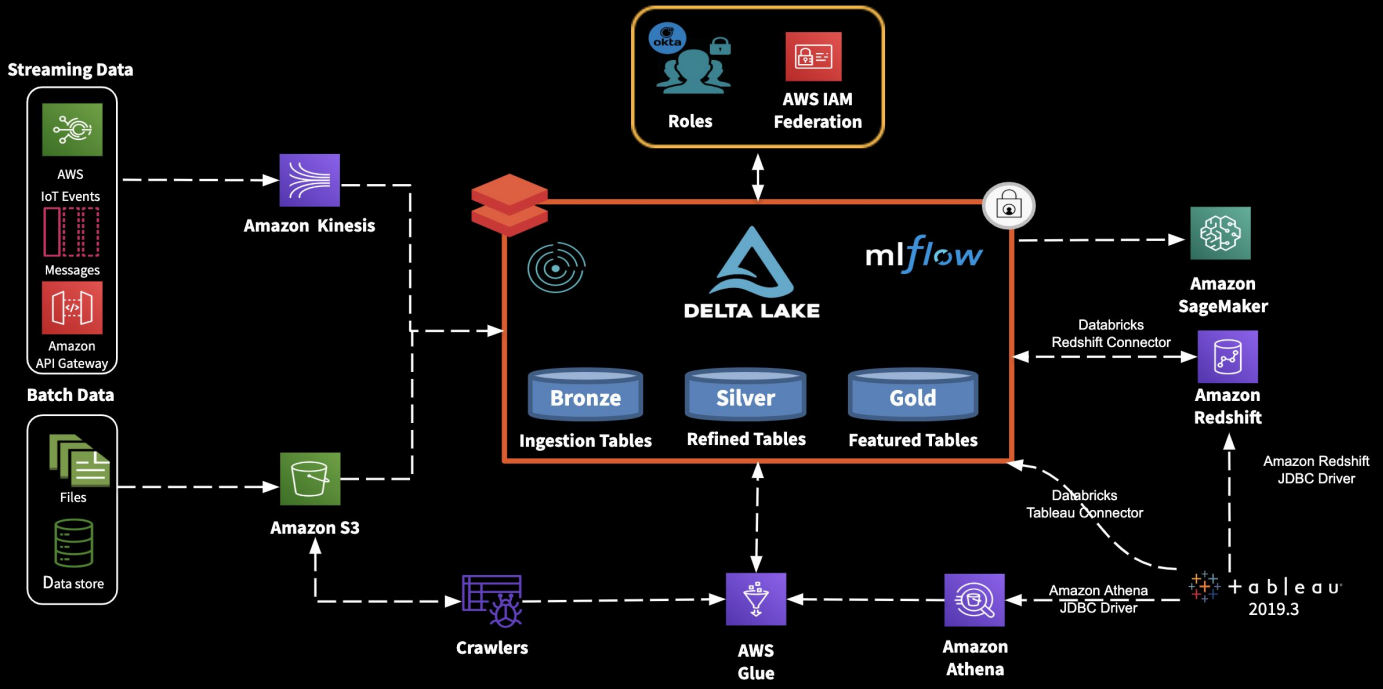
3

Devops Nightmare; Low productivity

- Silo'd AI/ Big Data environments
- Shared resource conflicts
- Upgrades, Maintenance delays

AWS data lake implementation

The Databricks Unified Data Analytics Platform



Kabbage migration

Lower costs, more capabilities

	Data Platform 2.0	**Data Platform 3.0**
Technology Stack - Major Components	MapR, Sql Server	Athena, Databricks, S3, Akka, Aurora
Technology Stack - Minor Components	Influx, Kafka, EMR, Druid, Nirvana, StreamSets, DCOS, Cloudera, Arrango, Vertica, S3	Influx, Kafka, EMR, Sql Server, Kubernetes
Cloud Environment	Hybrid	100% AWS
Maintainability	Labor Intensive	Auto Scalable, Auto Restart
Ease of Data Import/Export	Very Hard	Easy
Yearly Cost (approximate)	$2,000,000	$1,000,000 - $1,400,000

Data lake pipeline modernization case study

AWS ML workloads managed by Databricks since 2017
Data lake modernization project Q4 2019: Modernized native S3 data lake with Databricks Delta

Before: 72 person-days

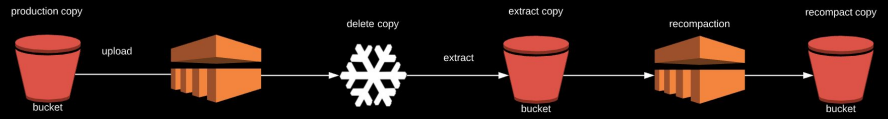
- Multiple (4) Copies of data; operational overhead, low data fidelity
- Manual Recompaction - operational overhead, costs
- Total = $397,000 / Quarter (incl $180K Snowflake + $160K EMR costs/quarter)

After: 6 person days

- No copies. Deletes occur in-place.
- New Cost Estimate $23,000 per quarter

Total Yearly Savings: $1,496,000

Before: Data lake architecture



Databricks' migration solution

On-premises

NETEZZA

hadoop

VERTICA

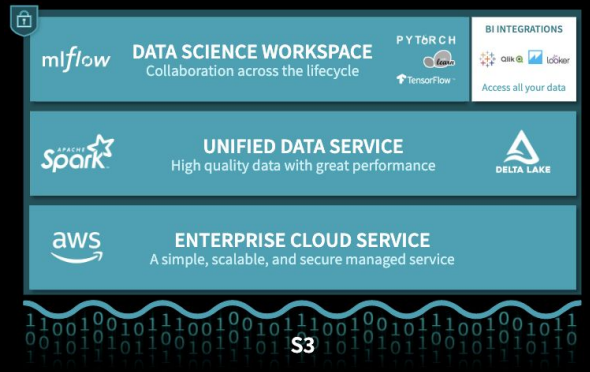
Other EDWs

Data
Metastore
Apps/ Jobs/ Queries
Security Governance

Migration

Reduce risk and accelerate outcomes
Automate migration data + code
Proven tools and methodologies
Working with your preferred partners

Databricks



ISV Partners

Migration: WanDisco, StreamSets, Talend, Infoworks
Security: Privacera, Immuta
CDC - Attunity, Syncsort

Program for Hadoop Migration

- Assessment
- Data replication
- Frameworks

www.databricks.com/migration

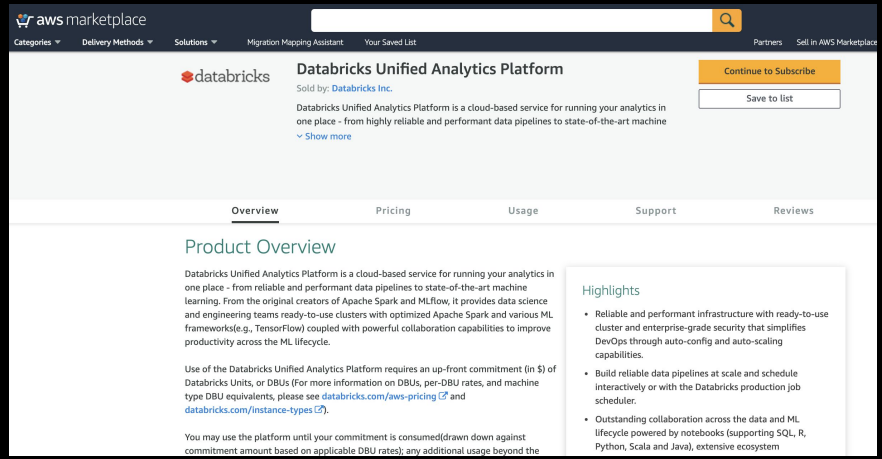
Consulting & SI Partners

KnowledgeLens, Slalom,Clarity Insights, CapGemini, MindTree, Cognizant, TCS

Now available on AWS Marketplace

- Simplified licensing
- One billing for Databricks usage

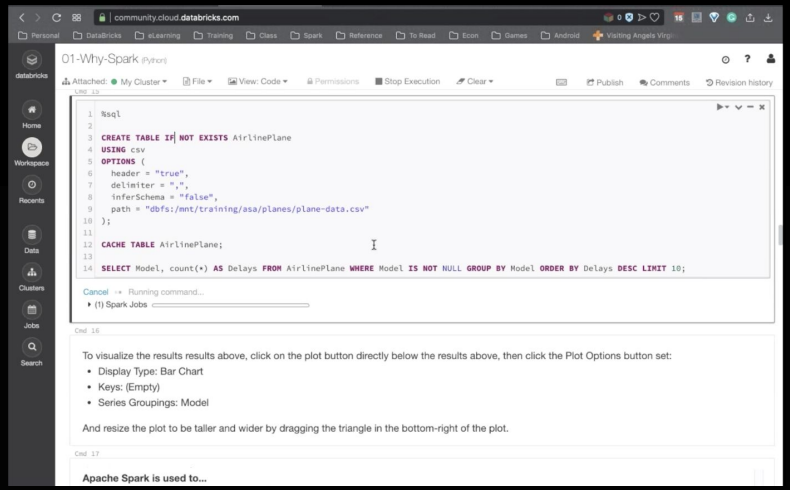
http://bit.ly/DATABRICKS



Free Training

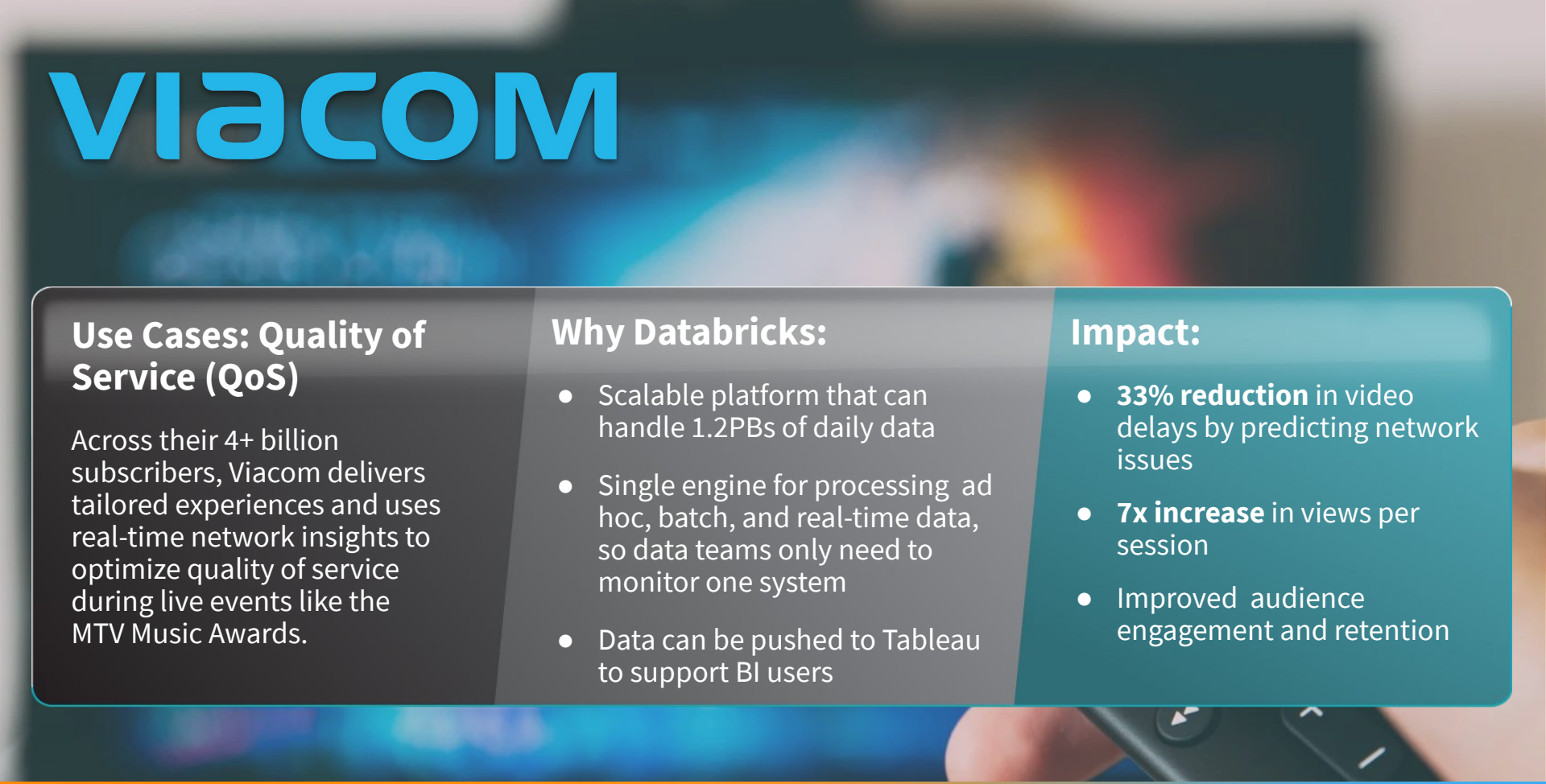
Three virtual 2-hour training sessions using Databricks on AWS:

- Getting Started with Apache Spark
- Data Engineering and Streaming Analytics
- Machine Learning



Sign up now:
http://bit.ly/TrainingAWS

Customer Profile Slides



Use Cases: Quality of Service (QoS)

Across their 4+ billion subscribers, Viacom delivers tailored experiences and uses real-time network insights to optimize quality of service during live events like the MTV Music Awards.

Why Databricks:

- Scalable platform that can handle 1.2PBs of daily data
- Single engine for processing ad hoc, batch, and real-time data, so data teams only need to monitor one system
- Data can be pushed to Tableau to support BI users

Impact:

- **33% reduction** in video delays by predicting network issues
- **7x increase** in views per session
- Improved audience engagement and retention

Analyzes 100 billion stock market events per day to identify fraud and wrong doing.

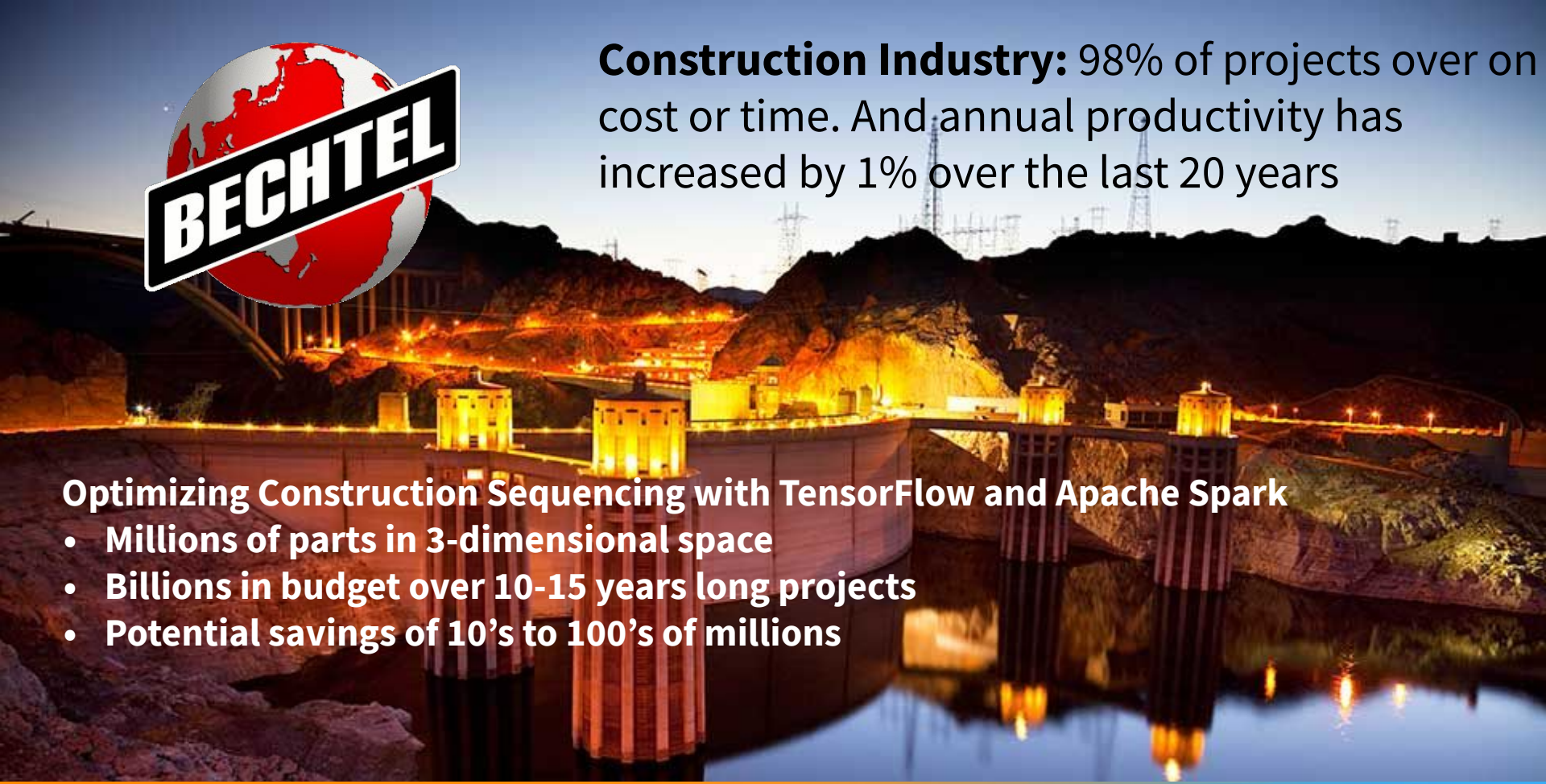
With Databricks we have created an ecosystem that allows data scientists self-service access to data, a direct path from prototyping to production and access to a common data lake from a suite of analytic technologies

REGENERON

"Our early large-scale data strategy has already paid off with a new drug target for chronic liver disease."

Saving Lives by Scaling Data Science

- **Data Science on** 250K genomes and growing to 500K genomes (10TB)
- **Increase model effectiveness by reducing** full data set query runtimes from 30 minutes to 5 seconds
- **Increased productivity of** Bioinformaticians and clinical researchers by automating devops



Construction Industry: 98% of projects over on cost or time. And annual productivity has increased by 1% over the last 20 years

Optimizing Construction Sequencing with TensorFlow and Apache Spark

- **Millions of parts in 3-dimensional space**
- **Billions in budget over 10-15 years long projects**
- **Potential savings of 10’s to 100’s of millions**

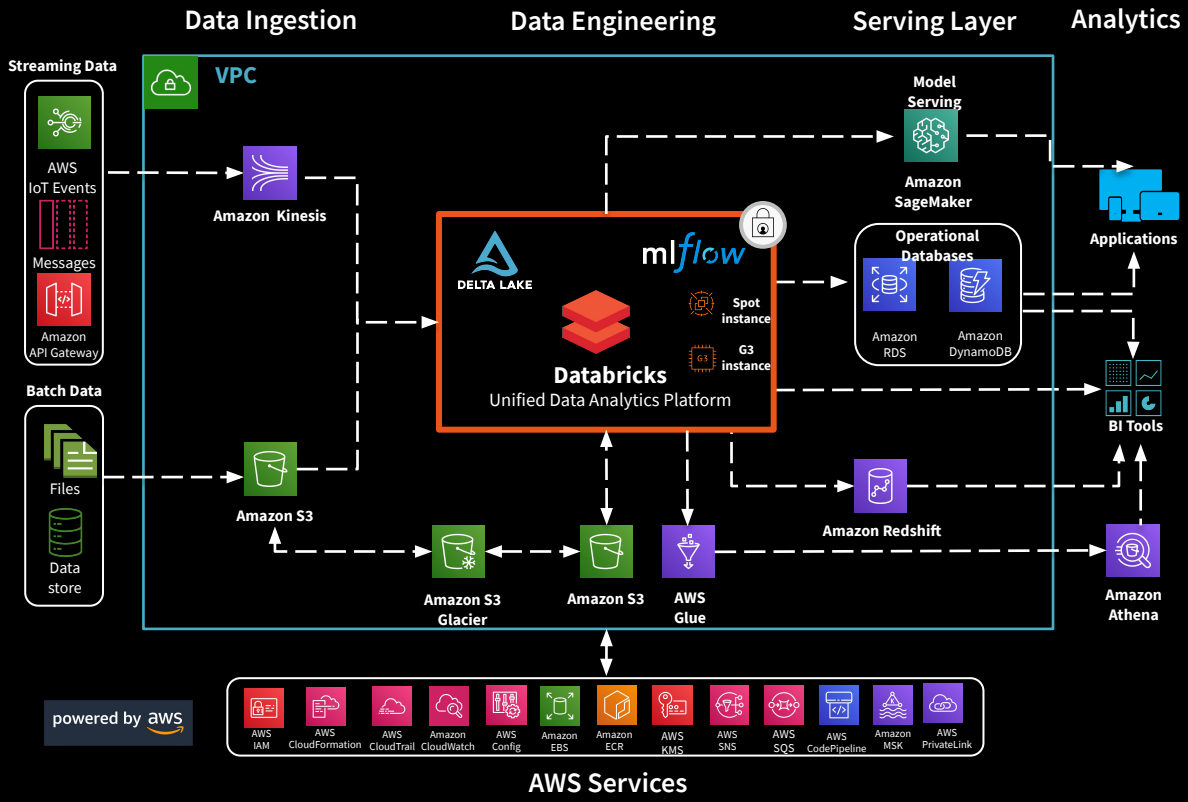

SPARK+AI
SUMMIT 2020
JUNE 22-25, 2020 | SAN FRANCISCO
ORGANIZED BY databricks
SAVE $450
REGISTER BY MARCH 31
SAVE YOUR SPOT
EXPANDED
TECHNICAL TRAINING
LEARN MORE

Agenda

9:00-9:30 AM	Opening Remarks - Enabling Cloud Data Lakes for Analytics
9:30-9:45 AM	Customer Use Cases
9:45-9:55 AM	Break
9:55-10:40 AM	Architecture - Incorporating Roles; Scaling, Launching and Managing Clusters; Managing Pools and Containers
10:40-11:25 AM	Notebooks - Data Loading using Scala, Machine Learning Analytics using Python, and Redshift integration for data delivery to analysts
11:25-11:35 PM	Q&A

Databricks Unified Data Analytics Technical Interactive Demo

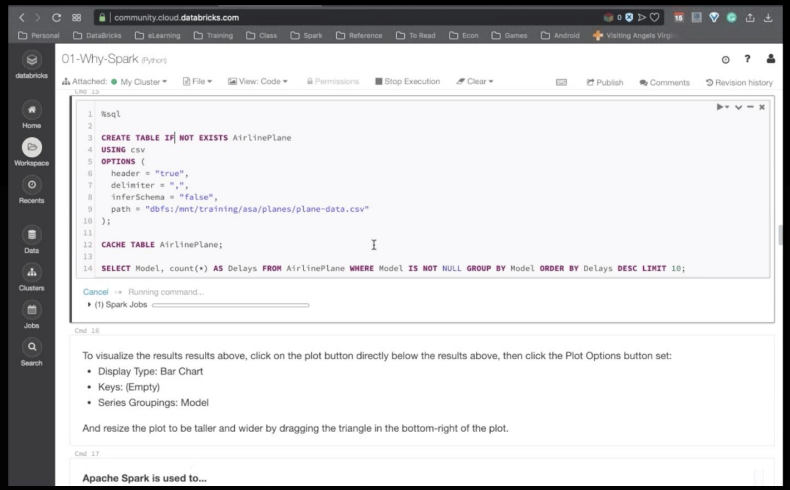
The Databricks Unified Data Analytics Platform on AWS



Free Training

Three virtual 2-hour training sessions using Databricks on AWS:

- Getting Started with Apache Spark
- Data Engineering and Streaming Analytics
- Machine Learning



Sign up now:
http://bit.ly/TrainingAWS

Thank you!