Introduction to Databricks and Apache Spark

Surendra Panpaliya

Agenda

- Overview of Databricks and its features
- Introduction to Apache Spark and its ecosystem
- Understanding the advantages of using Databricks for big data processing
- Creating & Using Databrick Spark Cluster

Agenda

- Spark architecture
- Driver program,
- Cluster manager
- Executors
- Spark operations
- Number of executors
- Executor memory

Overview of Databricks and its features

Surendra Panpaliya

Learning Objectives

- What is Databricks ?
- Key features and components of Databricks
- Databricks Architecture
- Databricks Advantages
- Databricks limitations

What is Databricks ?

Databricks is a unified analytics platform

designed to help organizations

process, analyze, and

gain insights from large volumes of data.

What is Databricks ?

- Cloud-based Data Engineering tool
- Widely used by companies
- to process and transform
- large quantities of data and
- explore the data.

Databricks

- Founded by the creators of Apache Spark
- Matei Zaharia
- Powerful open-source
- Data processing framework.
- CTO of Databricks
- Spark's vice president at Apache.

Databricks Introduction

- Databricks provides a collaborative environment
- for data engineers,
- data scientists, and
- analysts to work together on data-related tasks,
- such as data preparation,
- exploration, machine learning

Databricks Introduction

- Used to process and transform
- Extensive amounts of data and
- Explore it through
- Machine Learning models.

Databricks Introduction

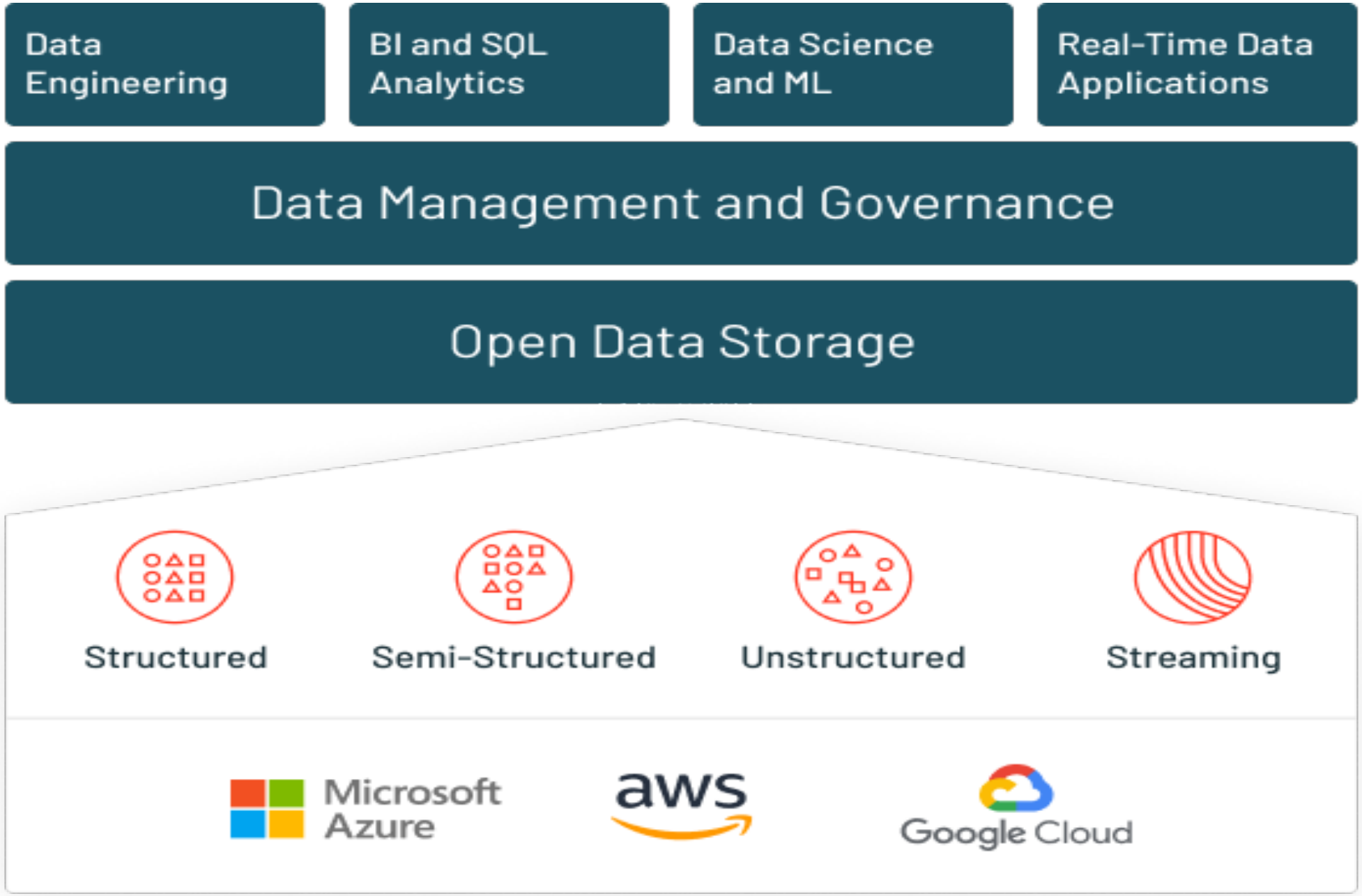
- Allows organizations
- to quickly achieve the full potential of
- combining their data,
- ETL processes, and
- Machine Learning.

Key features and components of Databricks

- Apache Spark Integration
- Unified Workspace
- Data Engineering
- Data Science and Machine Learning
- Notebooks

Key features and components of Databricks

- Collaboration
- Managed Services
- Scalability
- Integration


Data Engineering
BI and SQL Analytics
Data Science and ML
Real-Time Data Applications
Data Management and Governance
Open Data Storage
Structured
Semi-Structured
Unstructured
Streaming
Microsoft Azure
aws
Google Cloud

Apache Spark Integration

- Enables distributed data processing and analytics.
- Spark's capabilities cover batch processing
- Real-time stream processing
- Machine learning
- Graph processing

Unified Workspace

Databricks offers a collaborative workspace

Teams can write and execute code,

visualize data, and share insights.

Supports multiple programming languages

Python, Scala, R, and SQL.

Data Engineering

- Databricks provides tools
- to ingest, transform, and manage data.
- It supports various data sources and
- provides a structured way
- to perform ETL (Extract, Transform, Load) tasks.

Data Science and Machine Learning

- Building and deploying Machine Learning models.
- Offers libraries, tools, frameworks
- To streamline model development &
- Experimentation.

Notebooks

Databricks uses interactive notebooks,

Similar to Jupyter notebooks,

to create and share code,

Visualizations, and explanations.

Notebooks help in collaborative coding

documentation.

Collaboration

Multiple team members can collaborate

on the same projects within Databricks.

Promotes cross-functional collaboration and

knowledge sharing.

Managed Services

Databricks provides managed services,

meaning users don't need to worry about

infrastructure setup, scaling, or maintenance.

It offers automated cluster management,

resource optimization, and

security features.

Scalability

Databricks can handle

large-scale data processing and analytics.

Can scale up or down

based on the workload requirements.

Integration

Databricks can integrate with

Various data storage platforms,

Data warehouses

Other tools

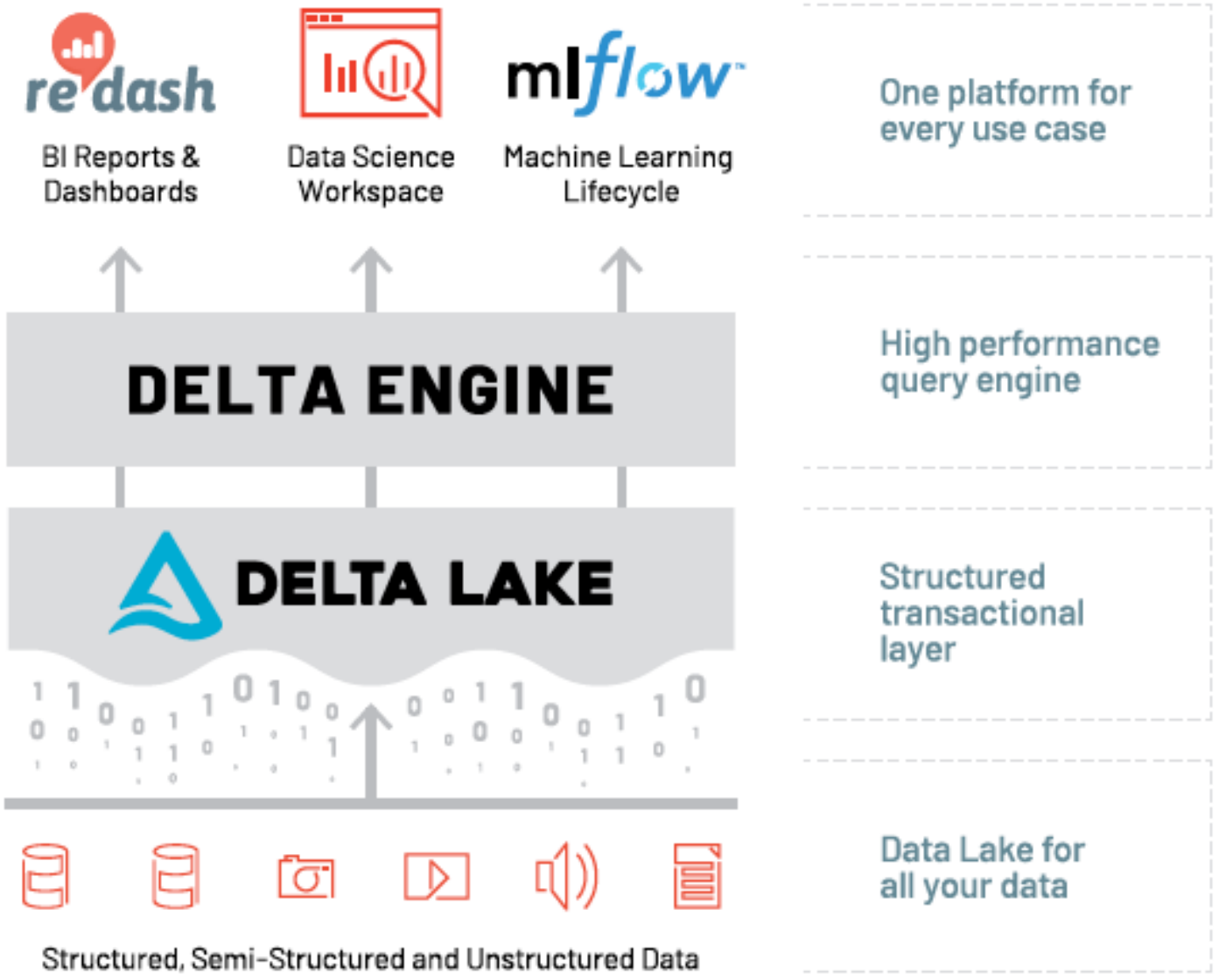
commonly used in the data ecosystem.

Data Source

- Connects with Cloud storage services provided by
- AWS, Azure, or Google Cloud
- connects to on-premise SQL servers, CSV, and JSON.
- extends connectivity to MongoDB,
- Avro files and others files

Databricks Architecture

Surendra Panpaliya

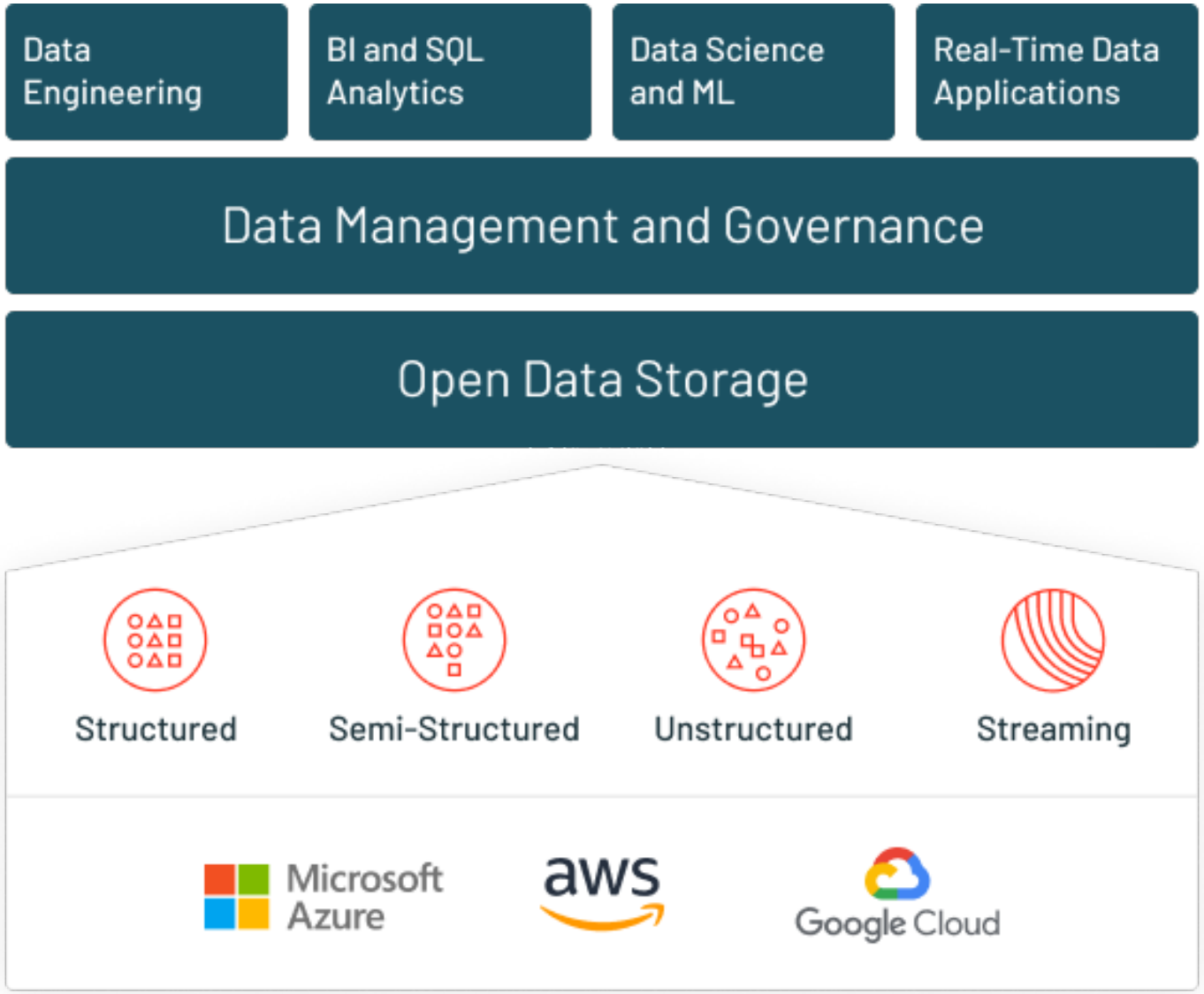

redash
BI Reports & Dashboards
Data Science Workspace
mlflow
Machine Learning Lifecycle
One platform for every use case
DELTA ENGINE
High performance query engine
DELTA LAKE
Structured transactional layer
Data Lake for all your data
Structured, Semi-Structured and Unstructured Data

DELTA LAKE

- Storage Layer
- Helps Data Lakes be more reliable.

Delta Lake

- Delta Lake integrates streaming and
- Batch data processing while providing
- ACID (Atomicity, Consistency, Isolation, and Durability)
- Transactions and scalable metadata handling.


Data Engineering
BI and SQL Analytics
Data Science and ML
Real-Time Data Applications
Data Management and Governance
Open Data Storage
Structured
Semi-Structured
Unstructured
Streaming
Microsoft Azure
aws
Google Cloud

Delta Engine

- Query engine
- Optimized for efficiently processing data
- stored in the Delta Lake.
- It also has other inbuilt tools that
- support Data Science,
- BI Reporting, MLOps.

Databricks Architecture and Components

Surendra Panpaliya

High-level architecture

Databricks architecture is built

to provide a unified platform for collaborative,

distributed data processing and analytics.

High-level architecture

It leverages Apache Spark as

its core processing engine.

Add additional components

to enable seamless data engineering,

machine learning, and collaboration.

High-level architecture

Databricks is structured to enable

Secure cross-functional team collaboration

Backend services managed by Databricks

High-level architecture

Focused on Data science,

Data analytics

Data engineering tasks

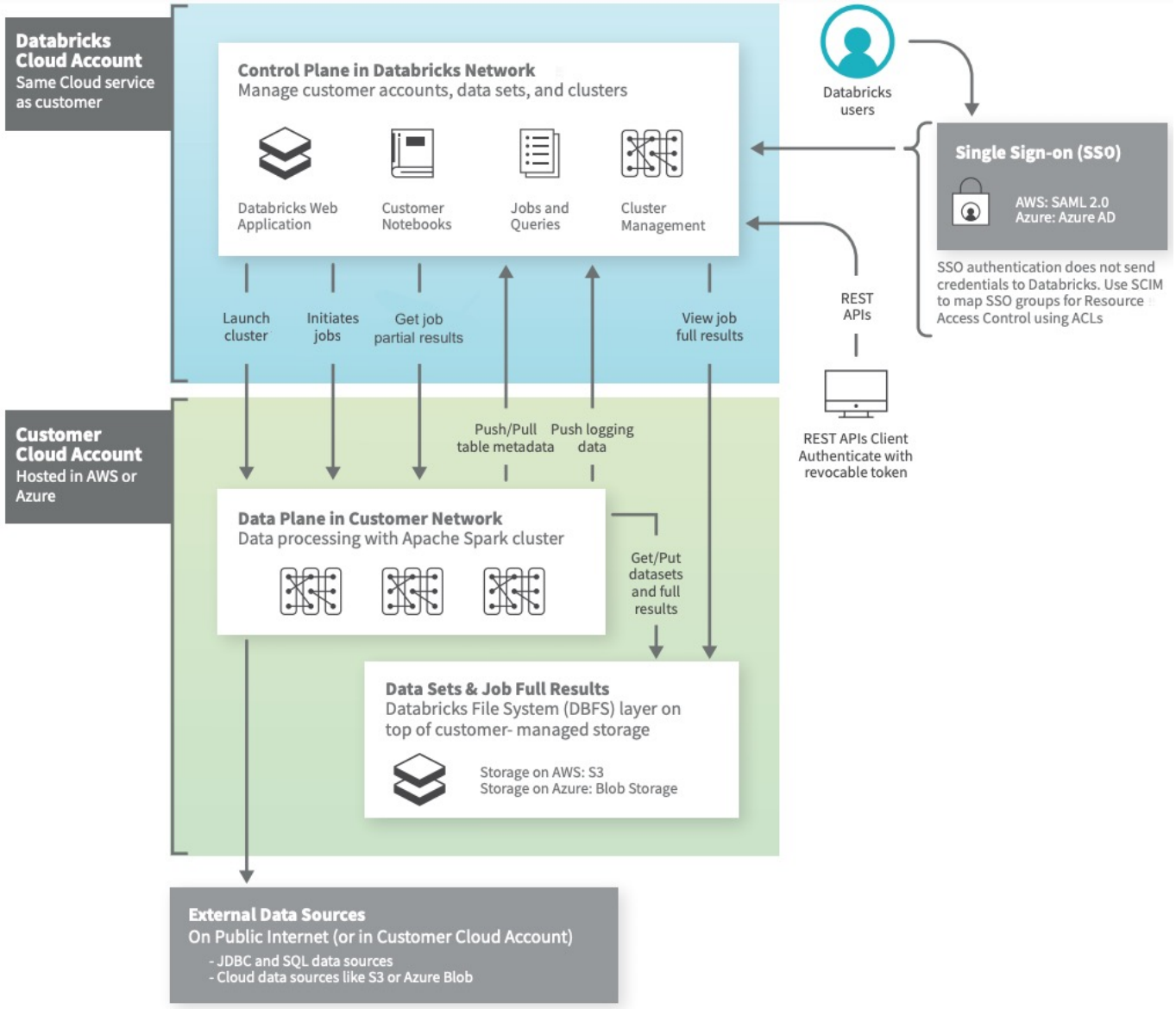
High-level architecture

Databricks operates out of

Control plane and a *Data plane*.

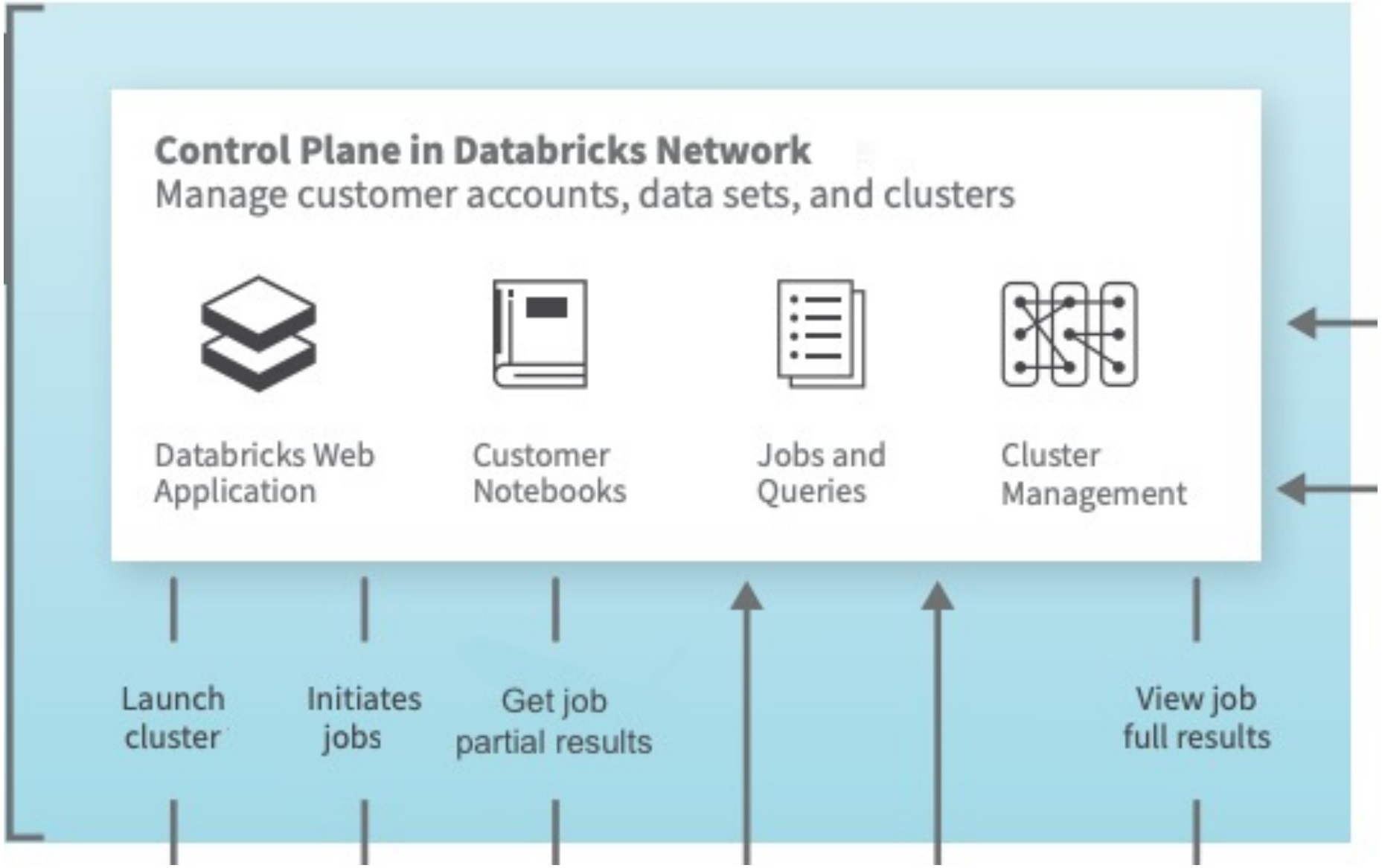
Architectures can vary

depending on custom configurations.


Databricks Cloud Account
Same Cloud service as customer
Control Plane in Databricks Network
Manage customer accounts, data sets, and clusters
Databricks Web Application
Customer Notebooks
Jobs and Queries
Cluster Management
Launch cluster
Initiates jobs
Get job partial results
View job full results
Push/Pull table metadata
Push logging data
Customer Cloud Account
Hosted in AWS or Azure
Data Plane in Customer Network
Data processing with Apache Spark cluster
Get/Put datasets and full results
Data Sets & Job Full Results
Databricks File System (DBFS) layer on top of customer- managed storage
Storage on AWS: S3
Storage on Azure: Blob Storage
External Data Sources
On Public Internet (or in Customer Cloud Account)
- JDBC and SQL data sources
- Cloud data sources like S3 or Azure Blob
Databricks users
Single Sign-on (SSO)
AWS: SAML 2.0
Azure: Azure AD
SSO authentication does not send credentials to Databricks. Use SCIM to map SSO groups for Resource Access Control using ACLs
REST APIs
REST APIs Client Authenticate with revocable token

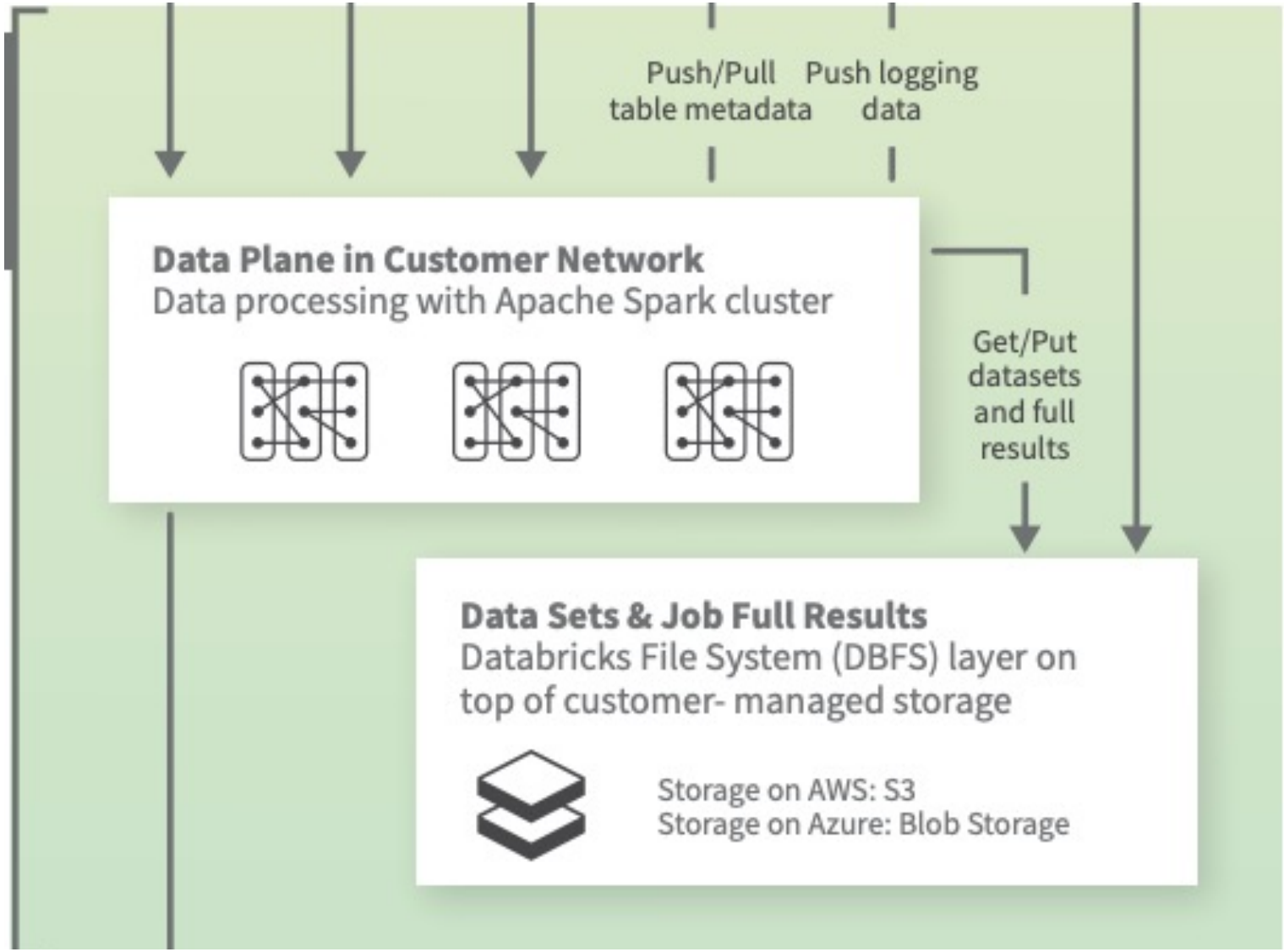
Control plane

- Backend services
- Notebook commands
- Other workspace configurations
- Stored in the control plane
- Encrypted at rest.


Control Plane in Databricks Network
Manage customer accounts, data sets, and clusters
Databricks Web Application
Customer Notebooks
Jobs and Queries
Cluster Management
Launch cluster
Initiates jobs
Get job partial results
View job full results

Data plane

Data is processed.



Key Components

- Workspace
- Notebooks
- Cluster Manager
- Databricks Runtime
- Job Scheduler
- Libraries and Dependencies

Key Components

- Data Import and Integration
- Collaboration and Sharing
- Security and Access Control
- Dashboards and Visualizations
- Machine Learning
- APIs and Integrations

Workspace

- Serves as the central hub
- Users to access, create
- Manage their data processing projects.
- Includes tools for creating notebooks,
- Organizing files
- Collaborating with team members.

Notebooks

Interactive environments

users can write and execute code,

visualize data

document work.

Notebooks

- Support multiple programming languages
- Python, Scala, R, and SQL.
- Conducting data analysis
- Building models within Databricks

Cluster Manager

- Create and manage clusters
- For running Spark workloads.
- Provisions and allocates resources
- To clusters based on
- user-defined configurations.

Cluster Manager

It supports various cluster managers

Apache Hadoop YARN,

Apache Mesos

Databricks standalone

Databricks Runtime

Databricks provides

Own distribution of Apache Spark,

Known as Databricks Runtime.

Databricks Runtime

- Runtime includes enhancements
- Optimizations
- Additional libraries
- Suited for cloud-based analytics.

Databricks Runtime

- Databricks regularly updates
- Optimizes the runtime
- For better performance
- Compatibility.

Job Scheduler

- Schedule and automate
- Execution of Notebooks and Scripts
- Schedule tasks
- To run at specific intervals
- Response to triggers.

Libraries and Dependencies

Databricks allows you

To install external libraries and dependencies

Required for notebooks and tasks.

Libraries and Dependencies

Libraries can include

Machine learning frameworks

visualization libraries

other third-party tools.

Data Import and Integration

- Databricks supports integrations
- Various data sources and connectors
- To ingest data from databases
- Cloud storage
- Streaming platforms

Collaboration and Sharing

Facilitates collaboration

Among team Members

To share notebooks,

Visualizations

Insights

Collaboration and Sharing

- Supports Version control
- Commenting features
- To enhance teamwork
- Knowledge sharing

Security and Access Control

- Provides security features
- User authentication,
- Role-based access control
- Data encryption

Dashboards and Visualizations

Enables the creation of

Interactive dashboards

Visualizations using tools

Databricks Delta,

Enabling data exploration and

Communication of insights.

Machine Learning

- Supports machine learning
- MLflow library
- Helps with tracking and managing
- Machine learning experiments and models.

APIs and Integrations

Databricks offers APIs

For programmatic interactions

with the platform,

enabling automation,

integration with other tools, and

custom development.

Summary

Databricks architecture is

Designed to streamline the entire data lifecycle,

from data ingestion and exploration

to advanced analytics and model deployment.

Summary

Its unified approach allows data engineers,

data scientists, and analysts

to collaborate effectively and

derive meaningful insights from their data.