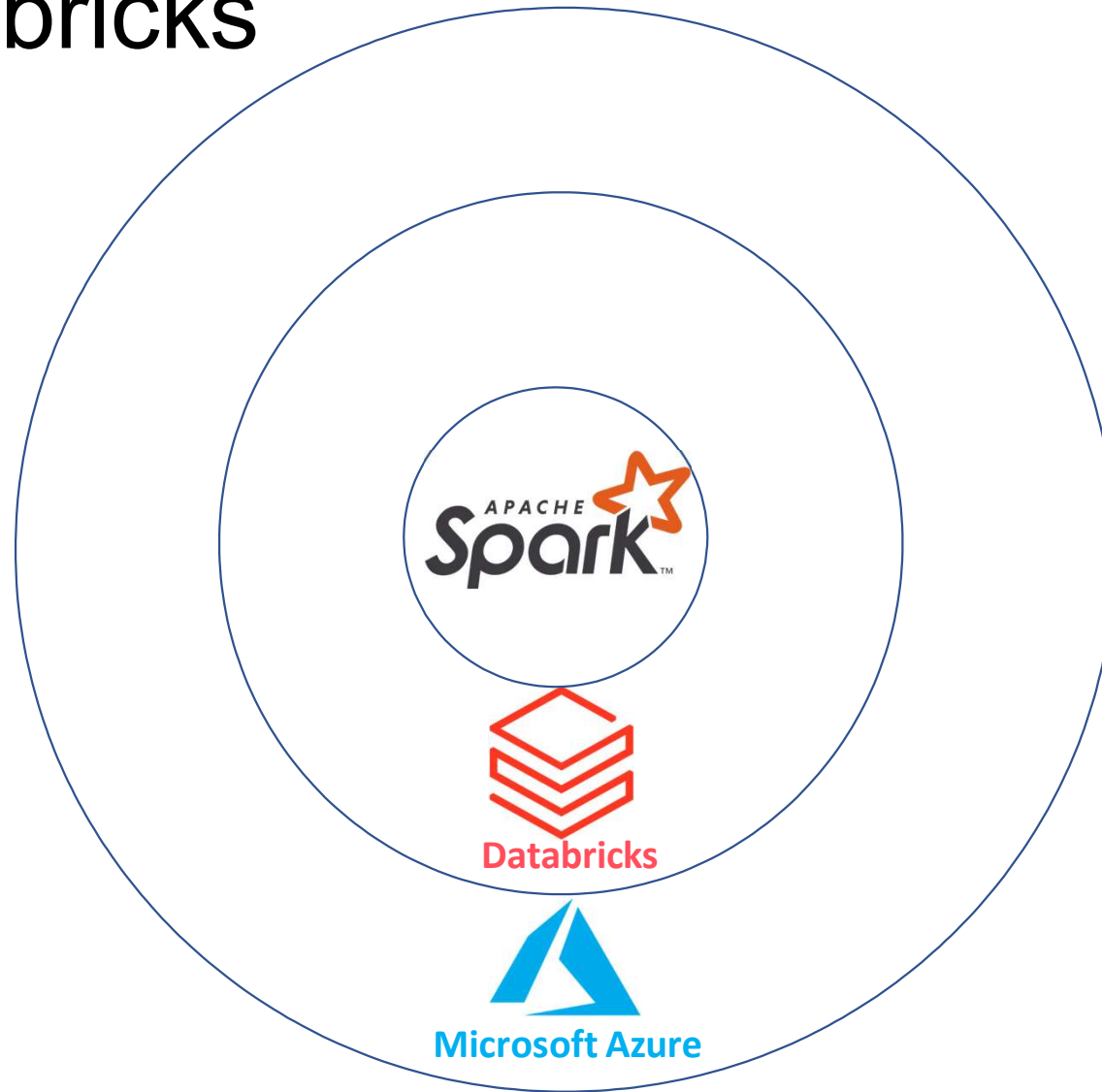


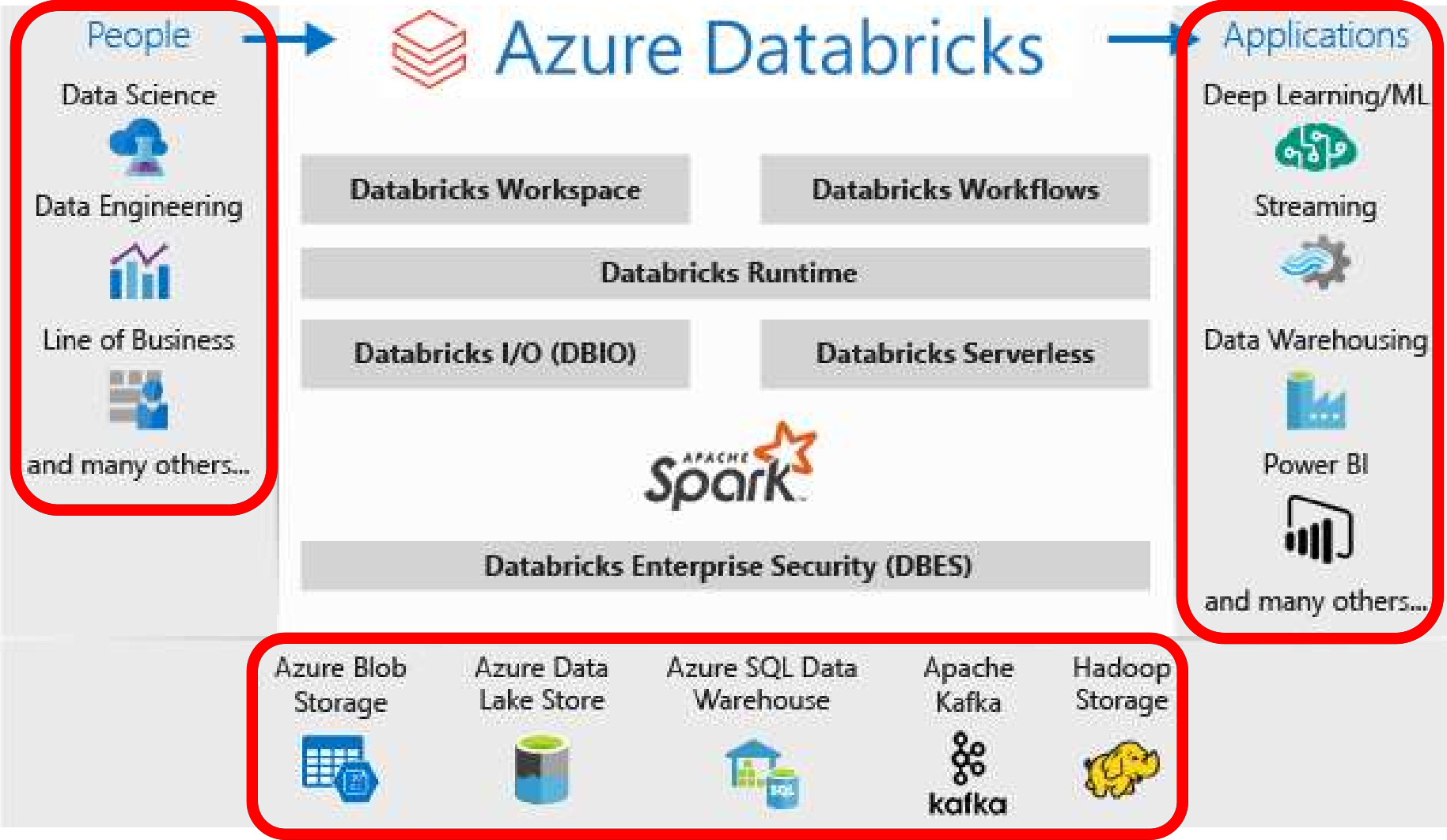
Introduction to Azure Databricks



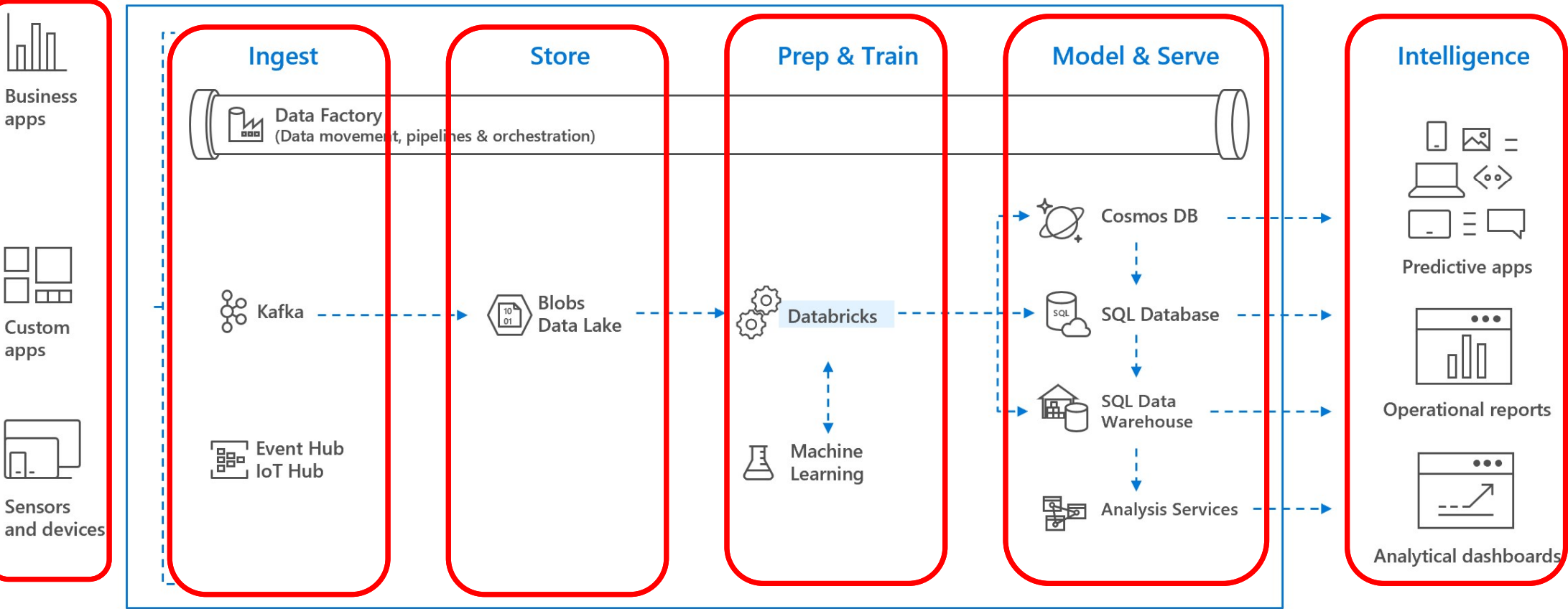
Azure Databricks



Azure Databricks



Azure Databricks



Apache Spark

Apache Spark is a lightning-fast unified analytics engine for big data processing and machine learning



100% Open source under Apache License

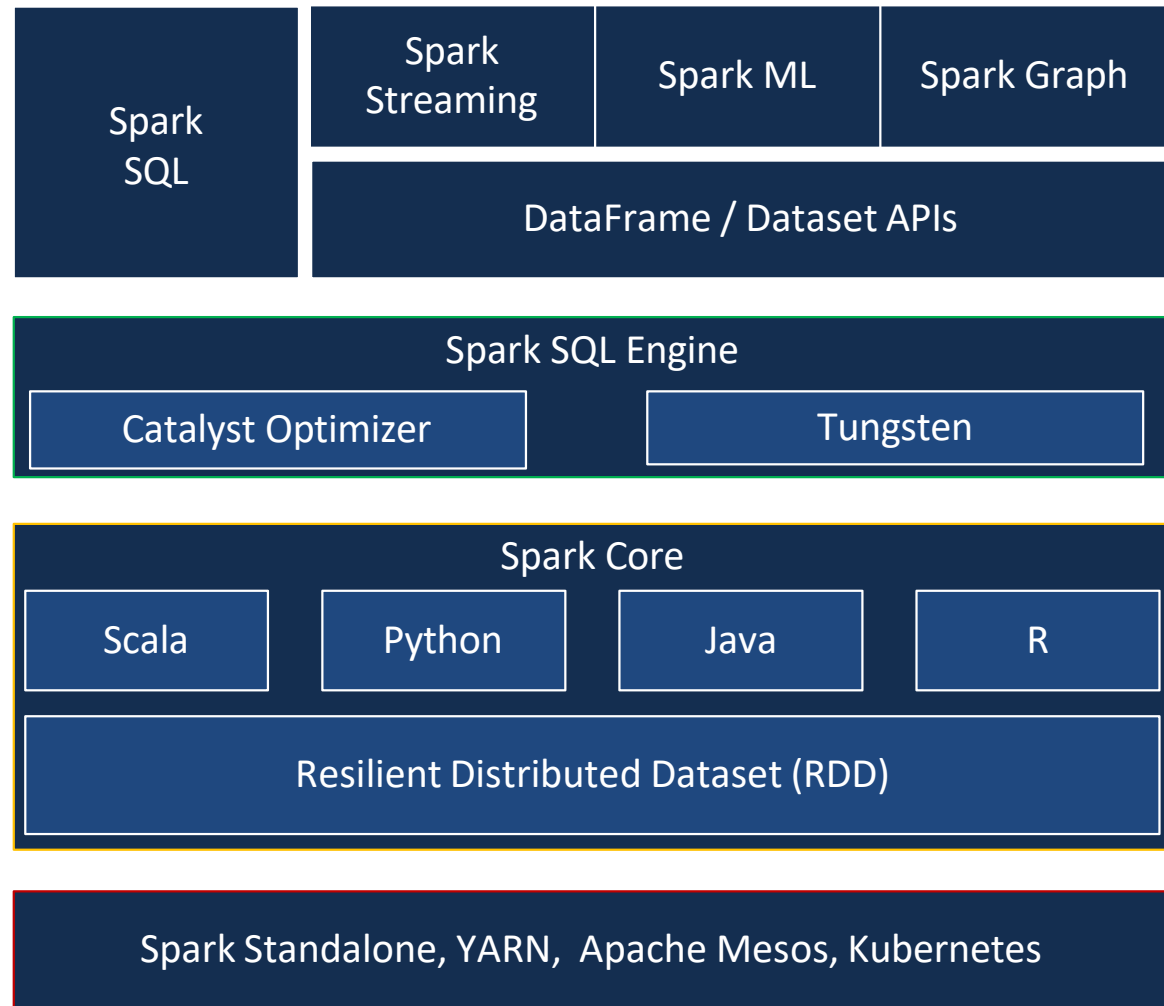
Simple and easy to use APIs

In-memory processing engine

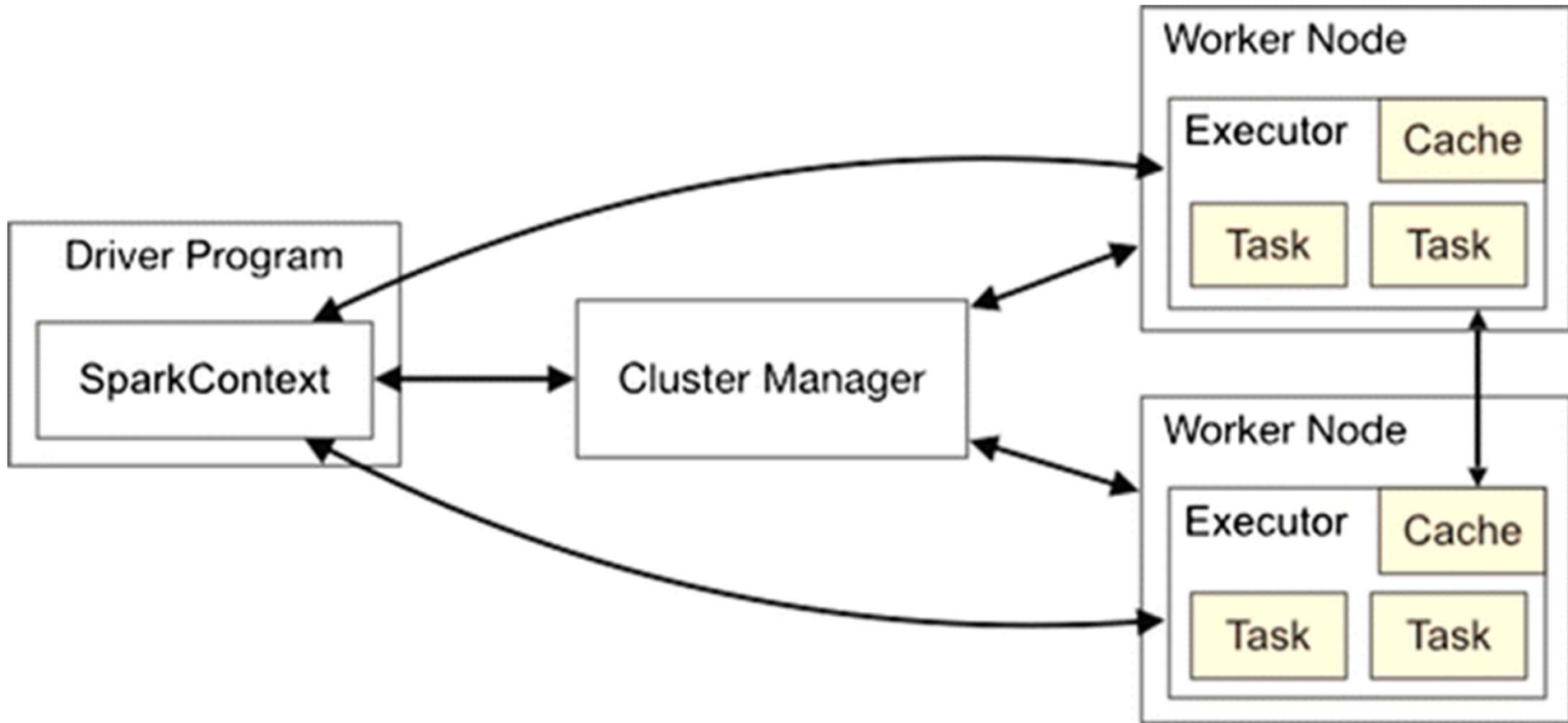
Distributed computing Platform

Unified engine which supports SQL, streaming, ML and graph processing

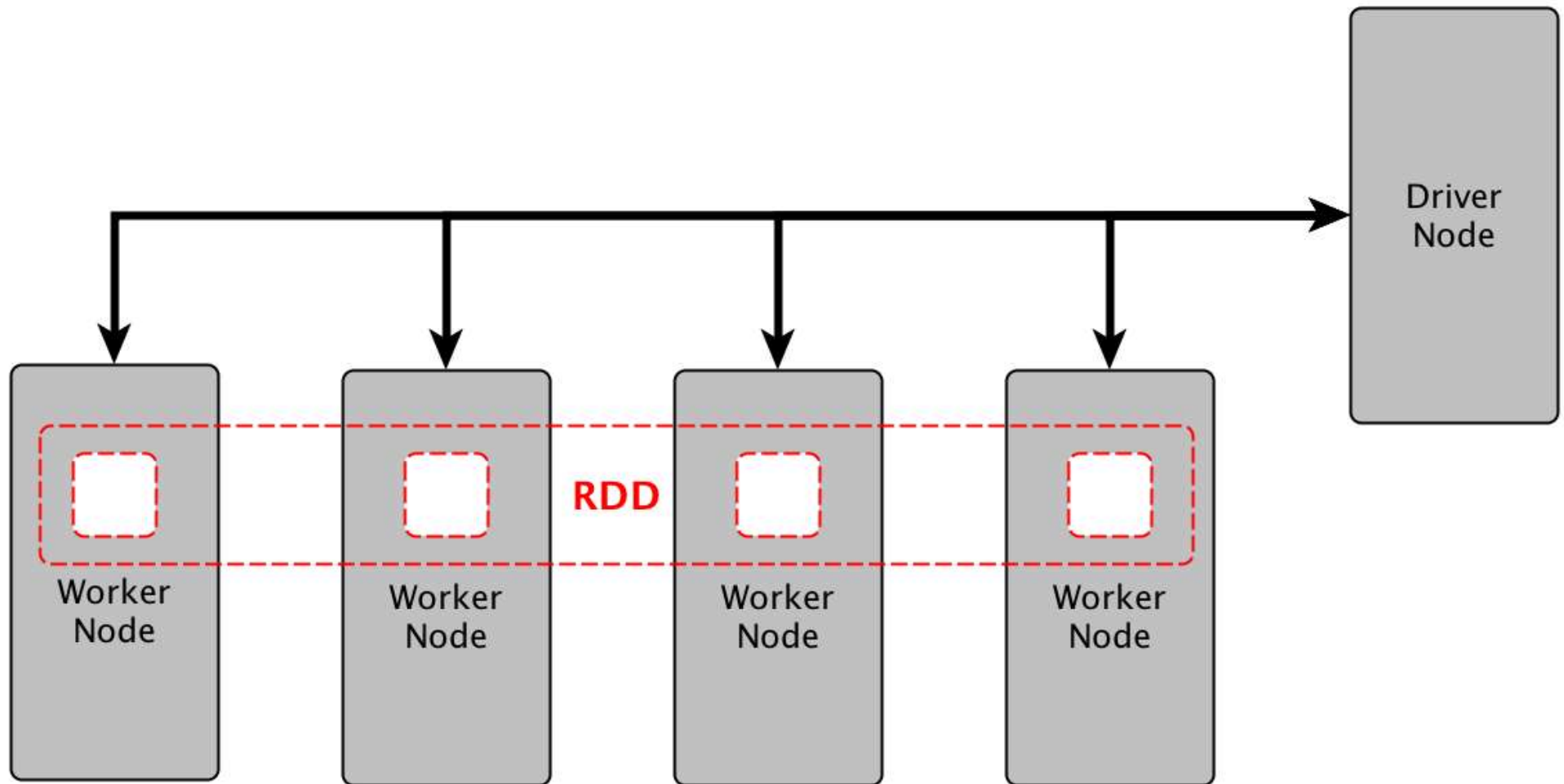
Apache Spark Architecture



Spark Architecture

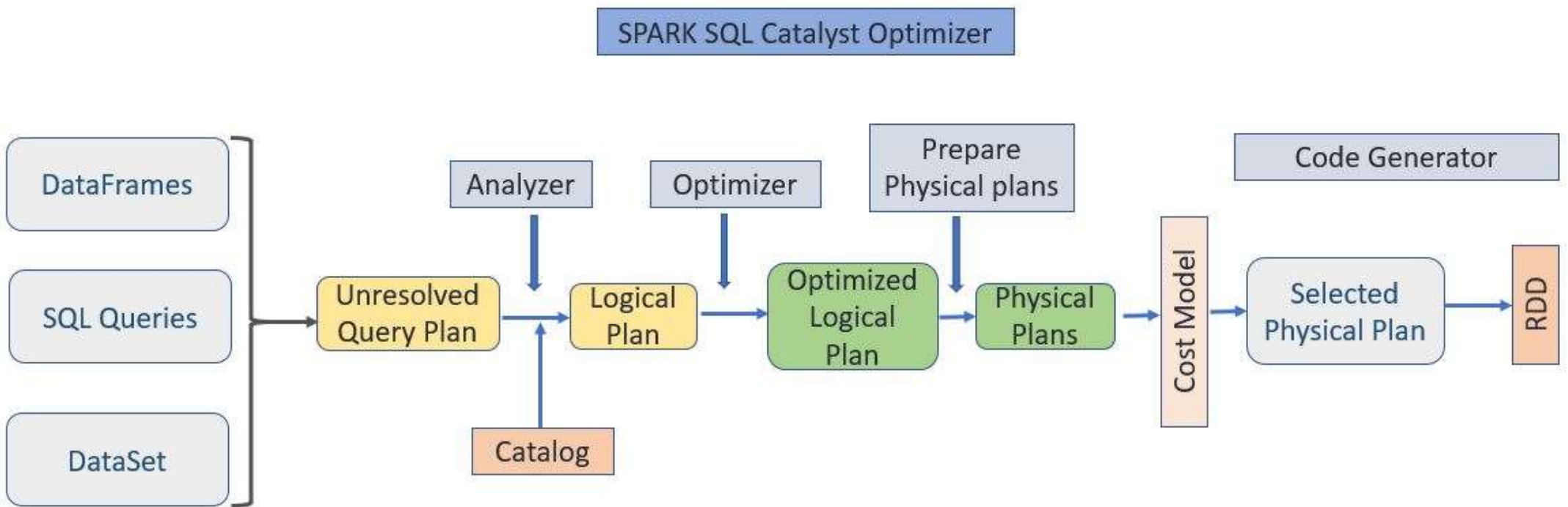


Apache Spark Execution



Spark Catalyst Optimizer

Execution Model



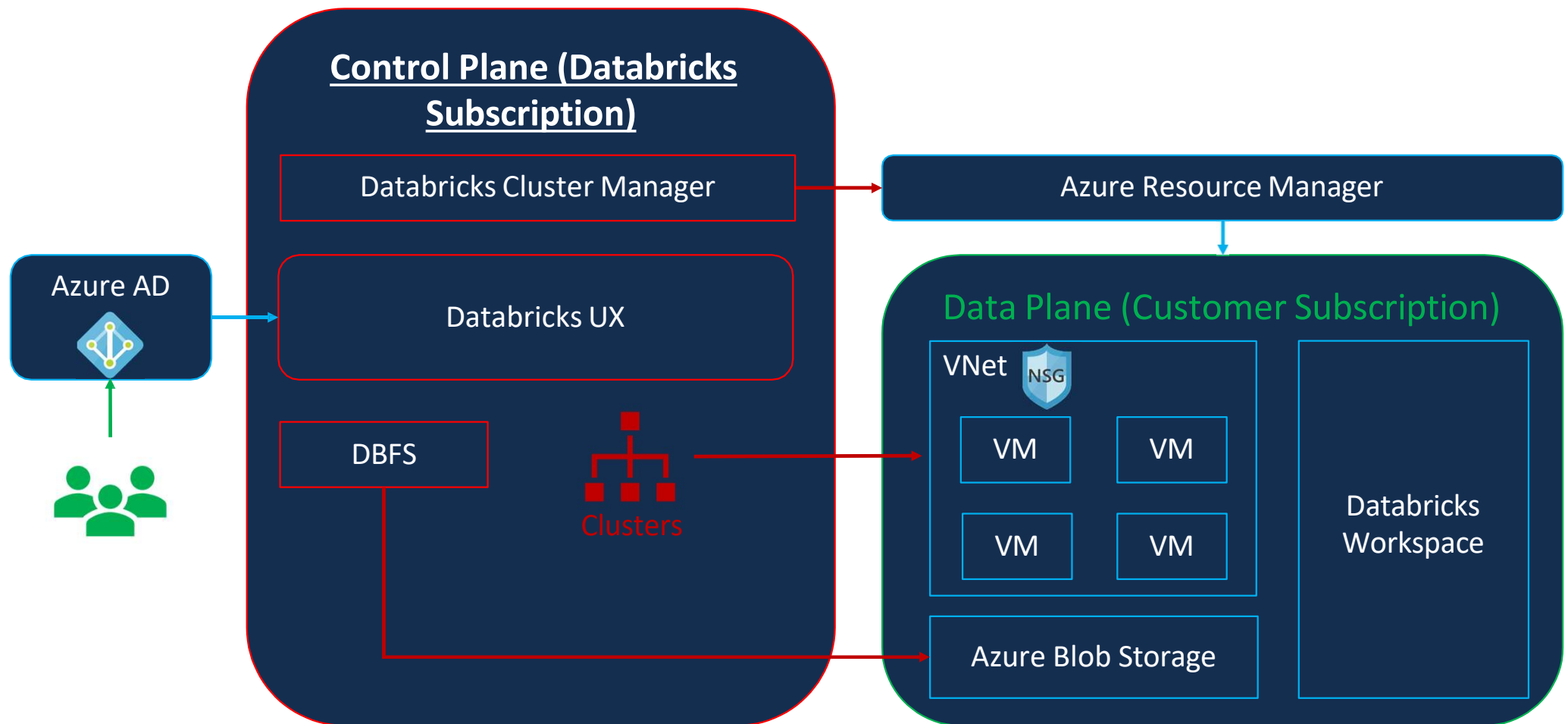
Creating Azure Databricks Service



Databricks Full vs Community edition

- Databricks Community Edition
 - Access to 15GB clusters
 - Not time-limited and users will not incur costs for their cluster usage.
- Full Databricks platform
 - Offers production-grade functionality
 - Unlimited number of clusters that easily scale up or down
 - Job launcher
 - Collaboration
 - Advanced security controls

Azure Databricks Architecture



Databricks Workspace Components

Notebooks

Data

Clusters

Jobs

Models


Cluster Configuration

Cluster Configuration

Policy 

Unrestricted

☒ Multi node ☐ Single node

Access mode 

Single user

Single user access 

Ramesh Retnasamy (az.adm1...)

Performance

Databricks runtime version 

Runtime: 11.3 LTS (Scala 2.12, Spark 3.3.0)

☐ Use Photon Acceleration 

Worker type 

Standard_DS3_v2



14 GB Memory, 4 Cores

Min workers

2

Max workers

8

 ☐ Spot instances 

Driver type

Same as worker

14 GB Memory, 4 Cores

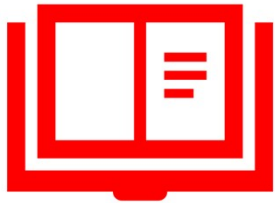
☒ Enable autoscaling 

☒ Terminate after 120 minutes of inactivity 

Creating Databricks Cluster



Databricks Notebooks



What's a notebook

Creating a notebook

Magic Commands

Databricks Utilities

Databricks Mounts



What is DBFS

What are Databricks mounts

Mount ADLS container to databricks

Create Service Principal

Create Azure Data Lake Storage Gen2

Creating Azure Key Vault

Creating Databricks secret scope

Databricks File System (DBFS)

Databricks Notebooks

Databricks CLI

Databricks API

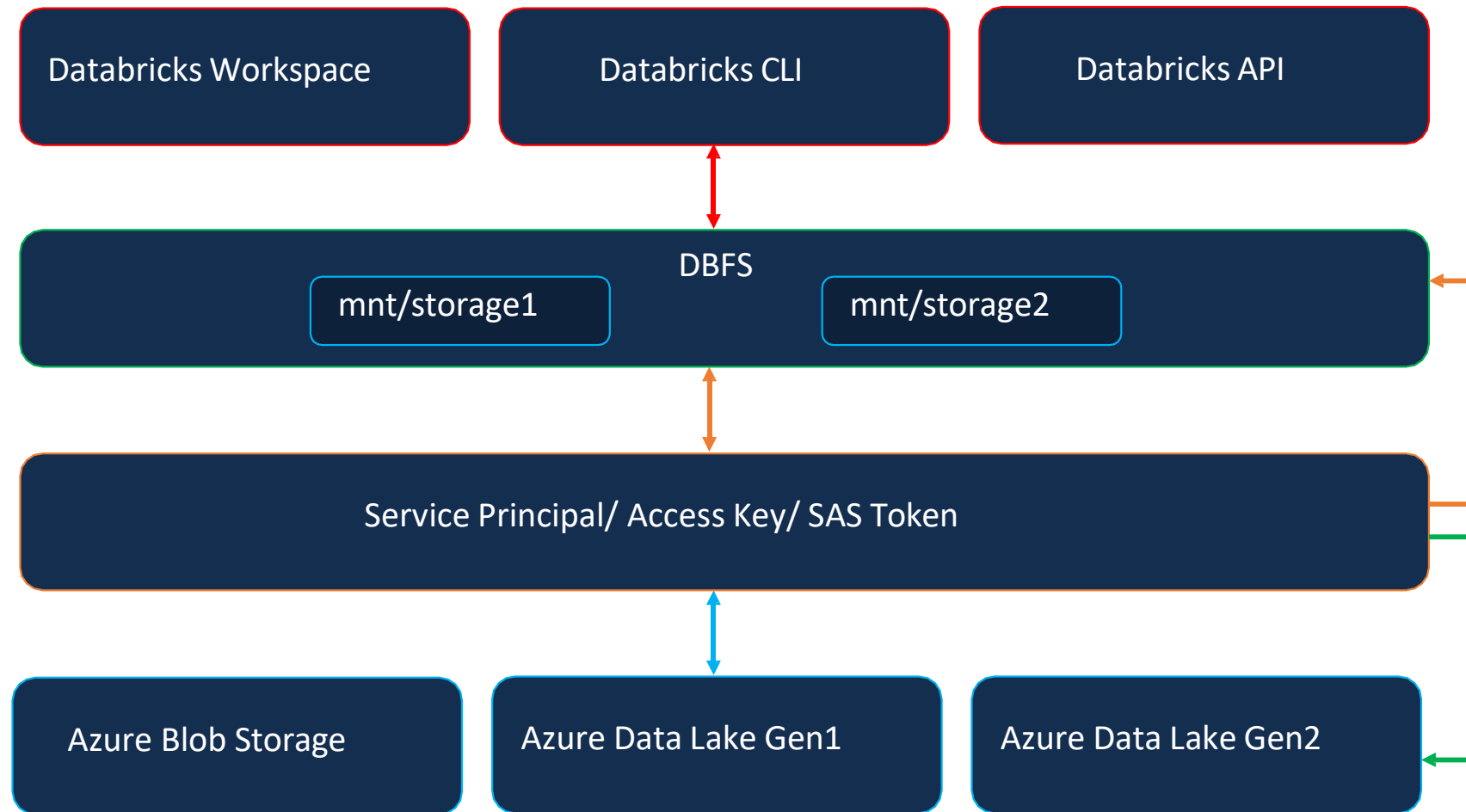
DBFS

Azure Blob Storage

Azure Data Lake Gen1

Azure Data Lake Gen2

Mounting Azure Storage

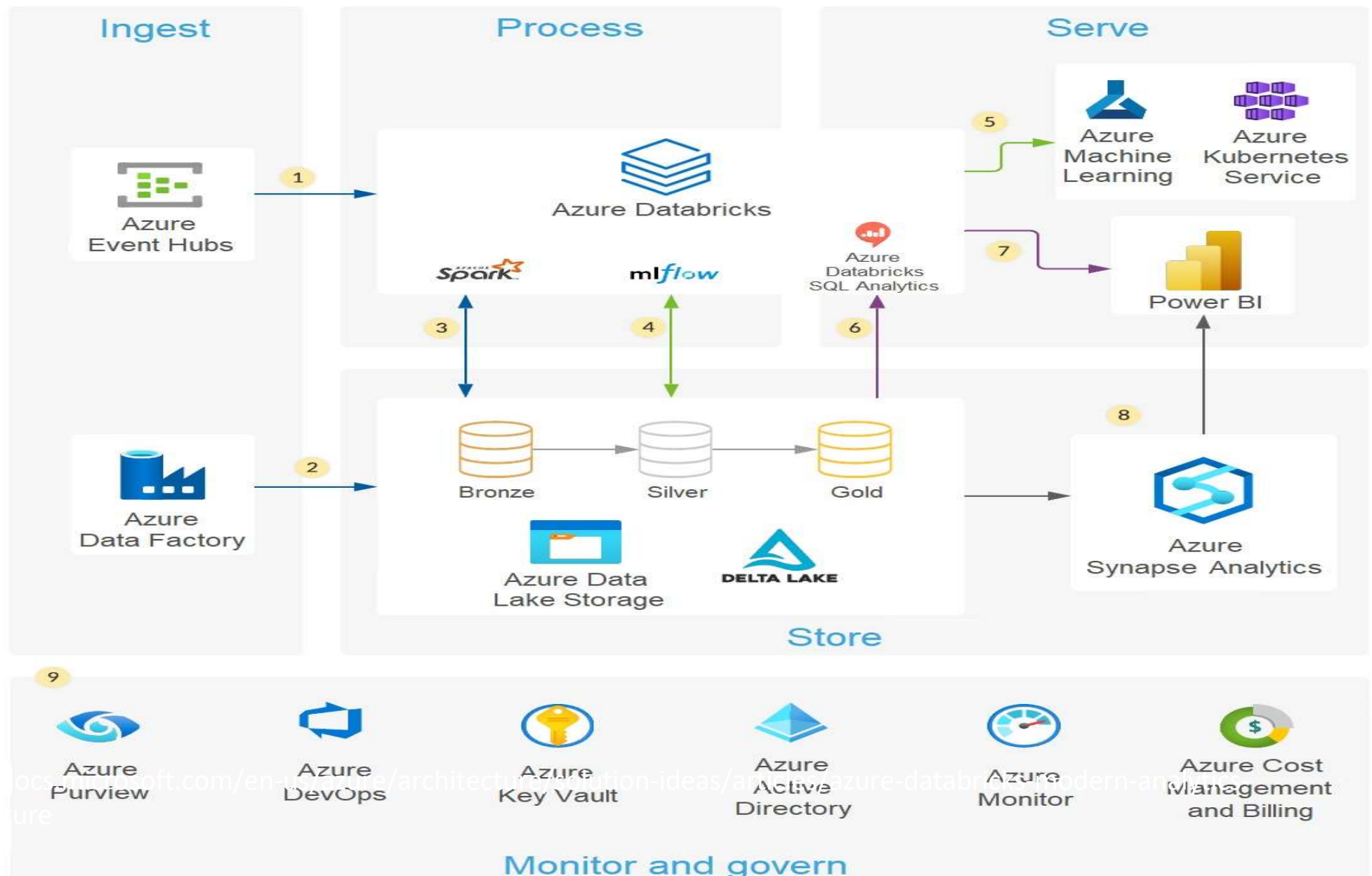


Secret Scope

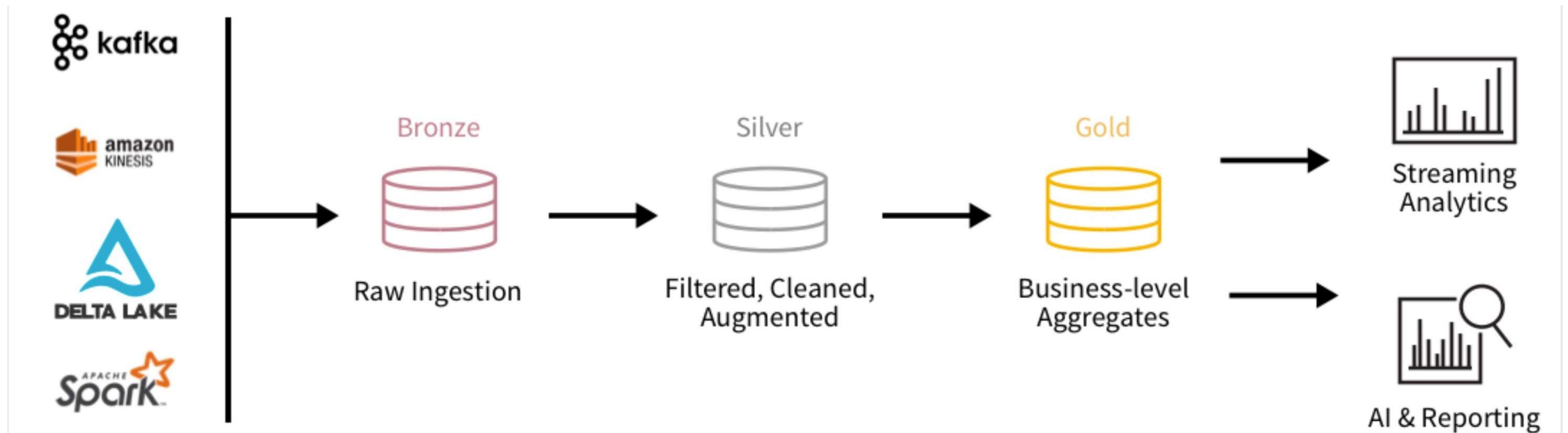
Databricks backed Secret Scope

Azure Key-vault backed Secret Scope

Solution Architecture Overview

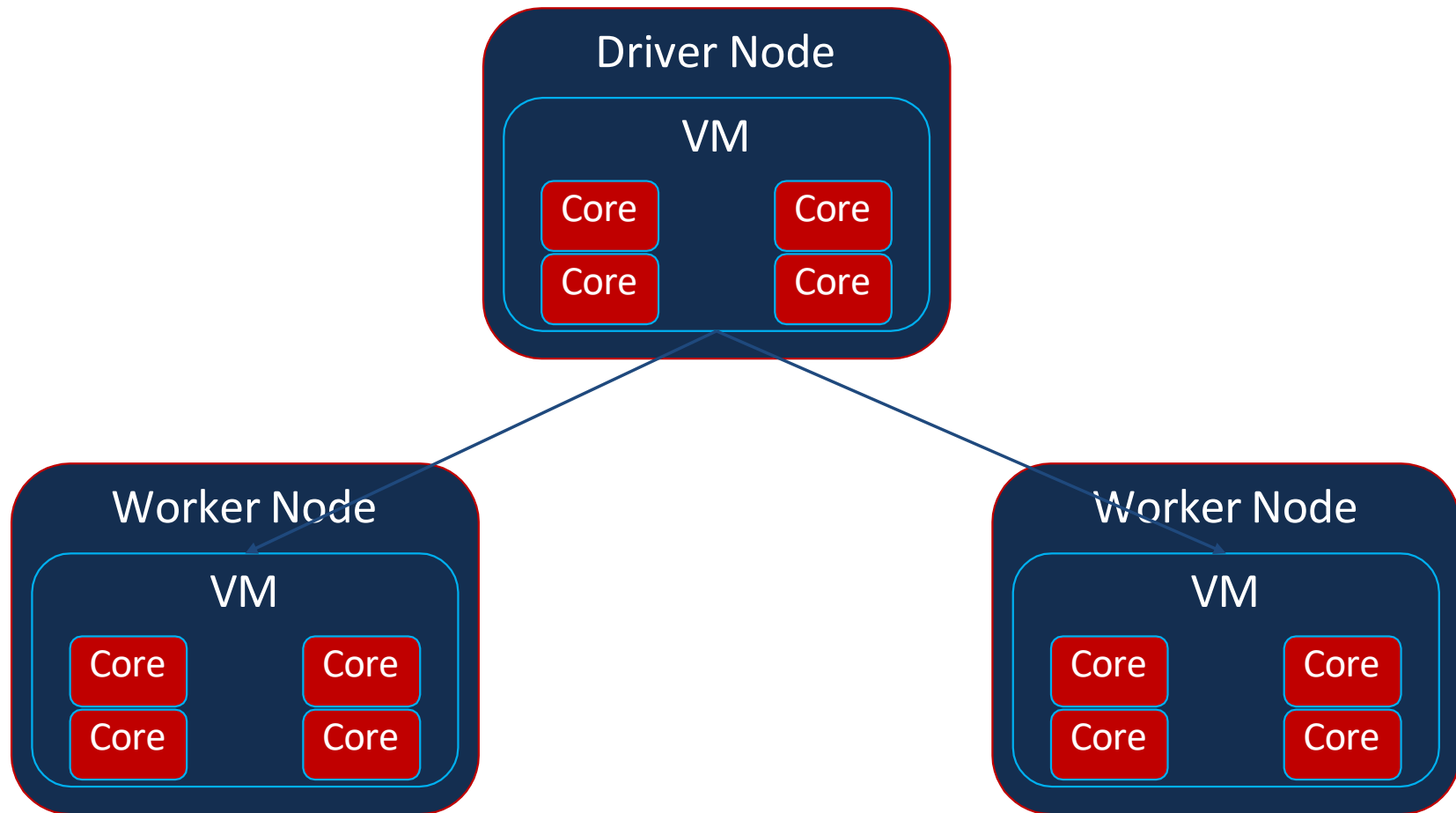


Databricks Architecture

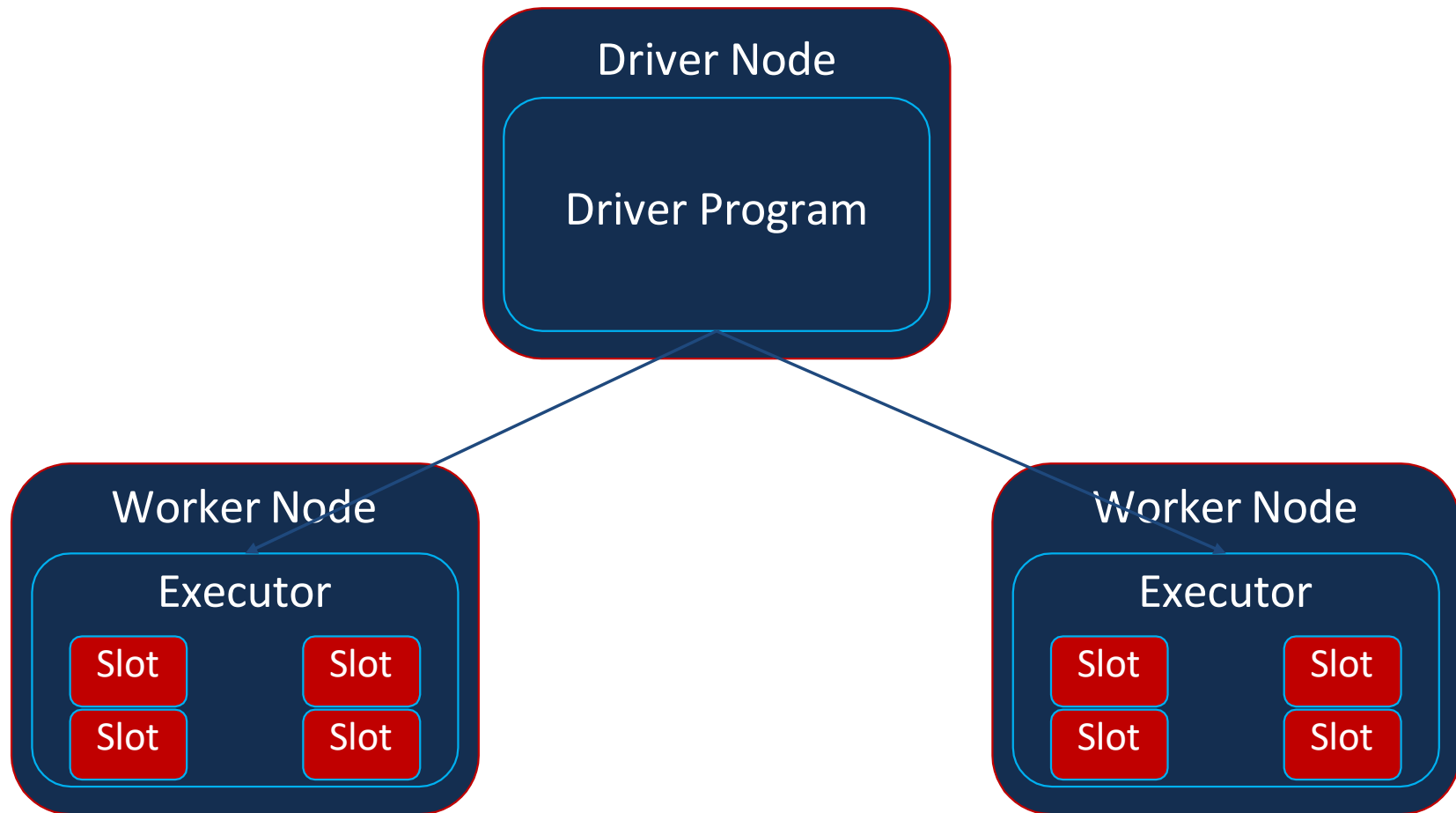


Spark Architecture

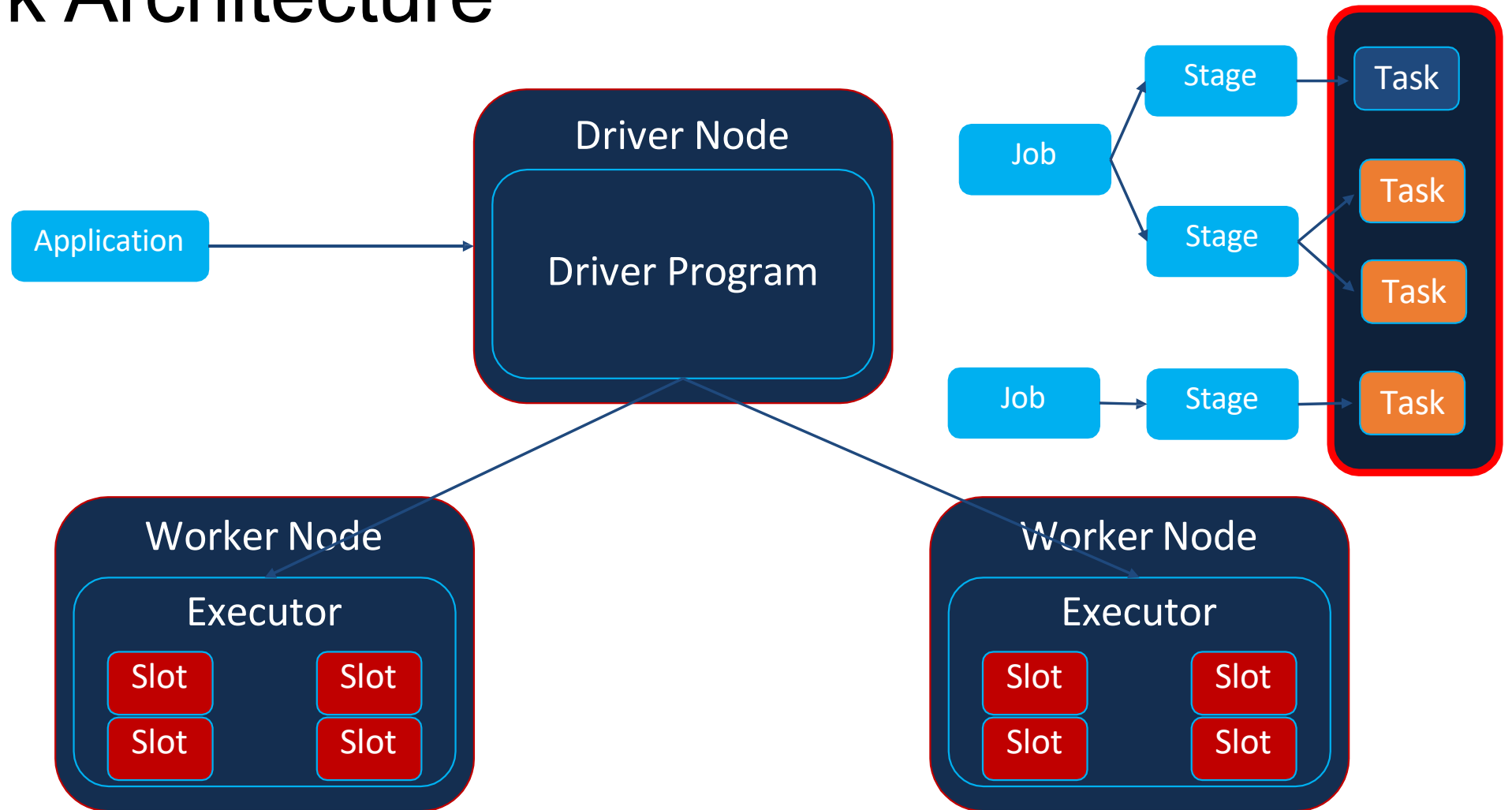
Spark Architecture



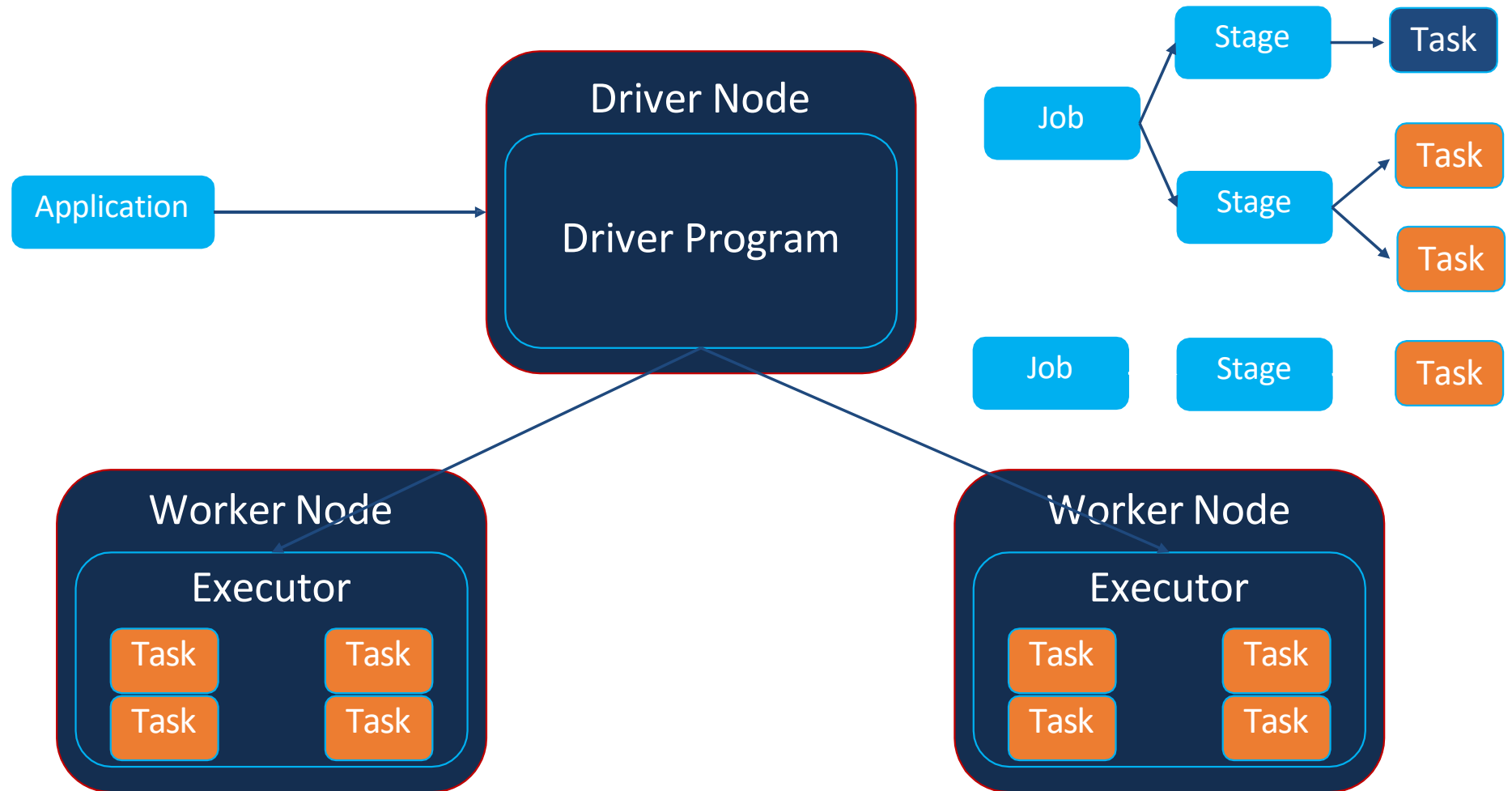
Spark Architecture



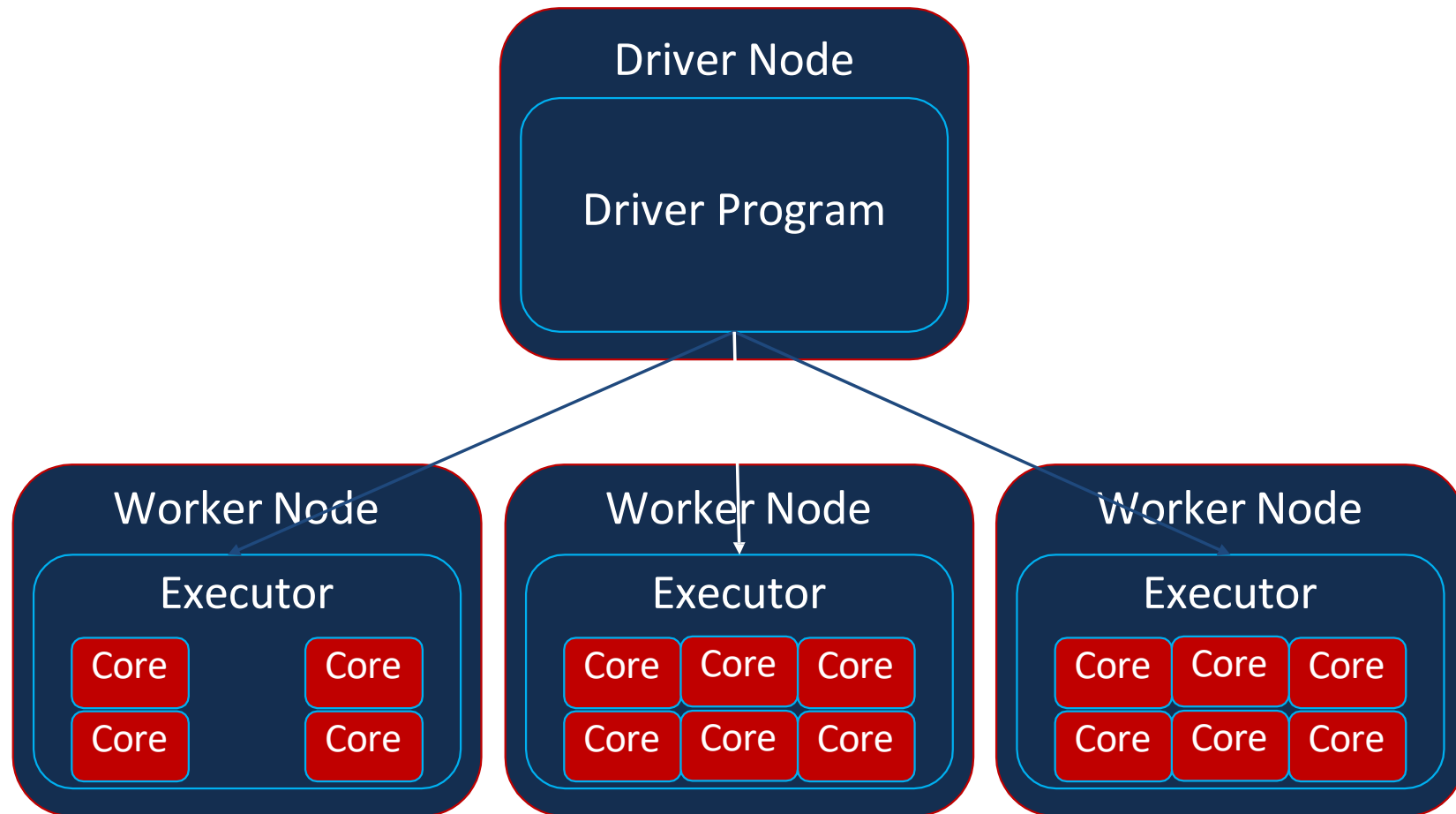
Spark Architecture



Spark Architecture

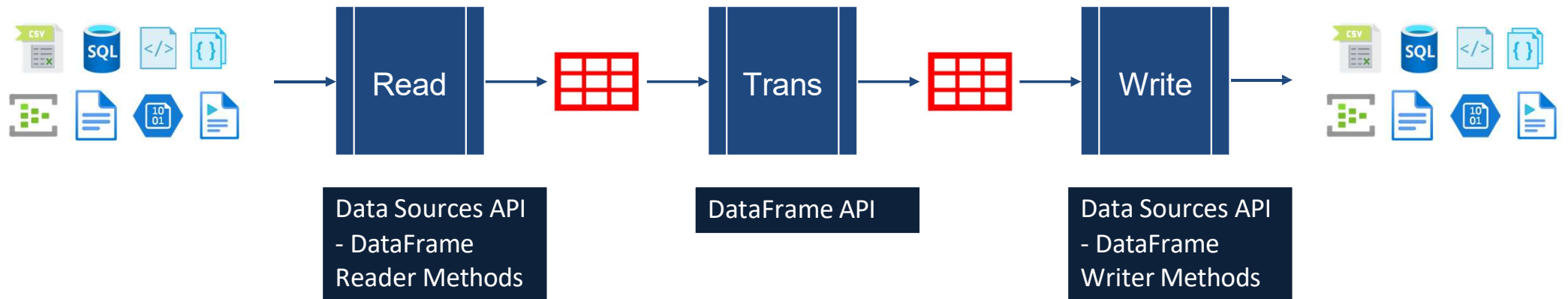


Spark Architecture – Cluster Scaling



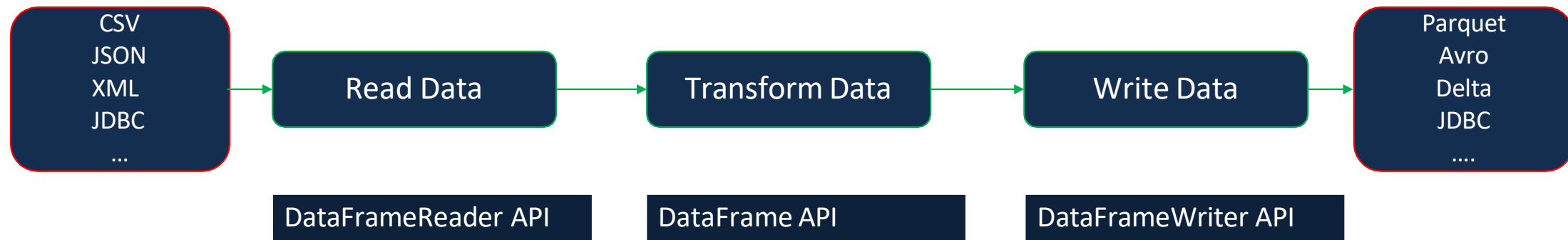
Spark DataFrame

Spark DataFrame

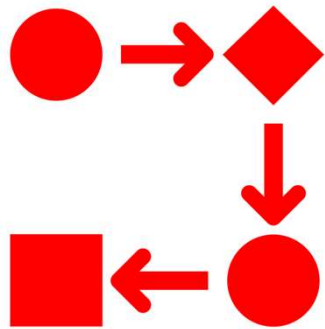


Data Ingestion Overview

Data Ingestion Overview



Databricks Workflows



Include notebook

Defining notebook parameters

Notebook workflow

Databricks Jobs

Filter/ Join Transformations

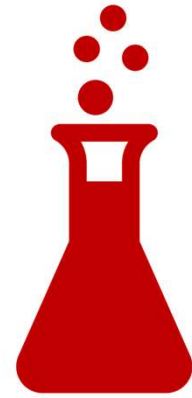


Filter Transformation

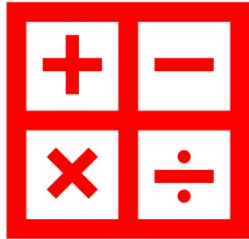
Join Transformations

Apply Transformations to F1 Project

Set-up Environment Presentation Layer



Aggregations



Simple Aggregations

Grouped Aggregations

Spark SQL Introduction



SQL Basics

Simple Functions

Aggregate Functions

Joins

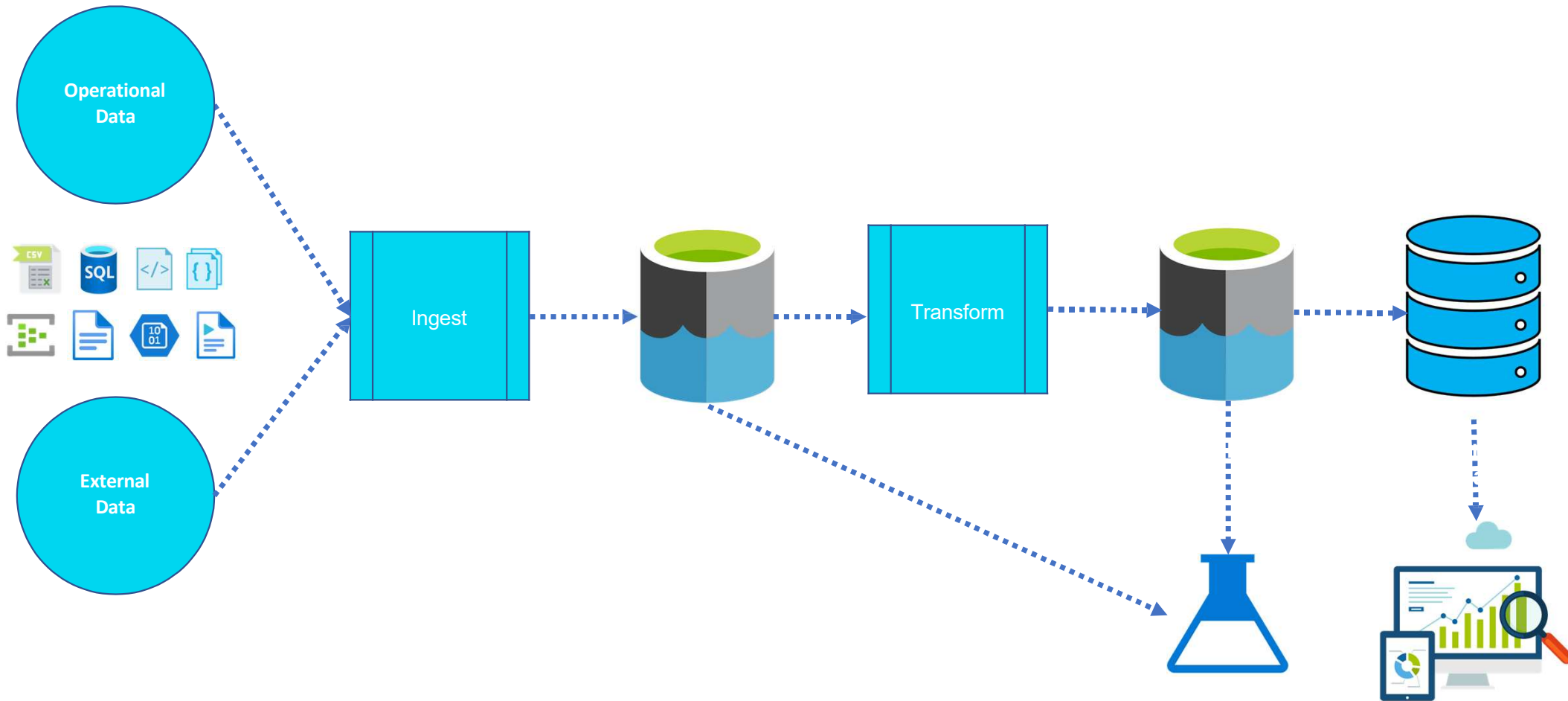
Delta Lake

Data Warehouse



- Lack of support for unstructured data
- Longer to ingest new data
- Proprietary data formats Scalability
- Expensive to store data
- Lack of support for ML/ AI workloads

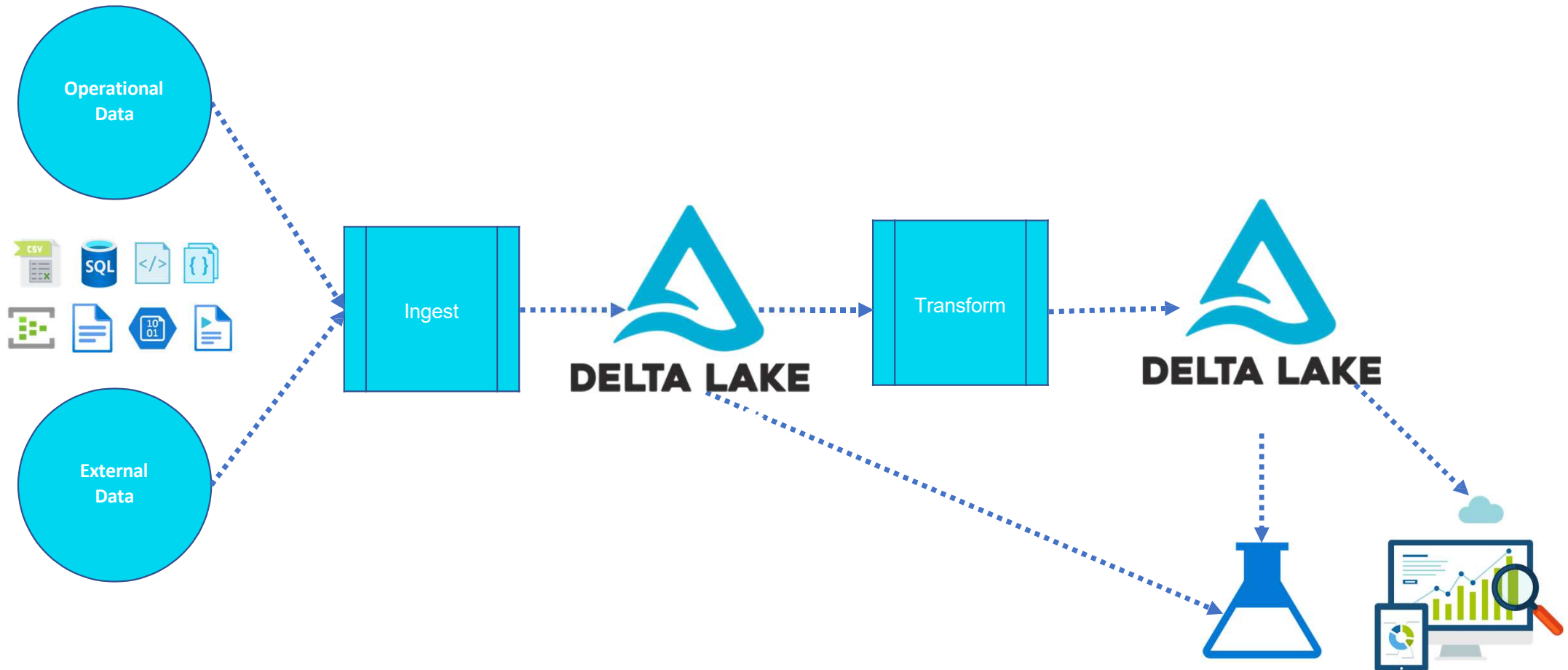
Data Lake



Data Lake

- No support for ACID transactions
- Failed jobs leave partial files
Inconsistent reads
- Unable to handle corrections to data
- Unable to roll back any data.
- No history or versioning
- Poor performance

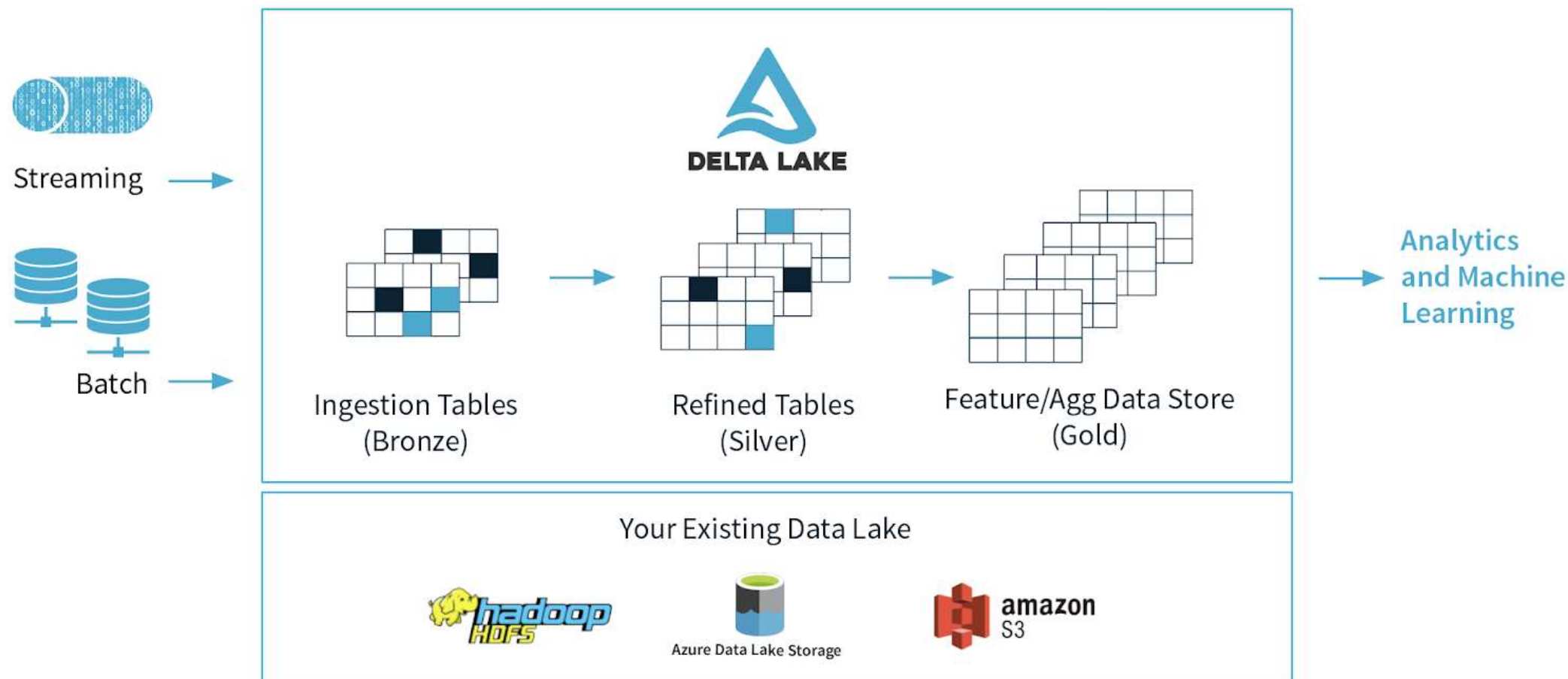
Data Lakehouse



Data Lakehouse

- Handles all types of data
- Low cost cloud object storage
- Uses open source parquet format
- ACID support
- History & Versioning
- Better performance Simple architecture

Delta Lake Architecture



Thank you