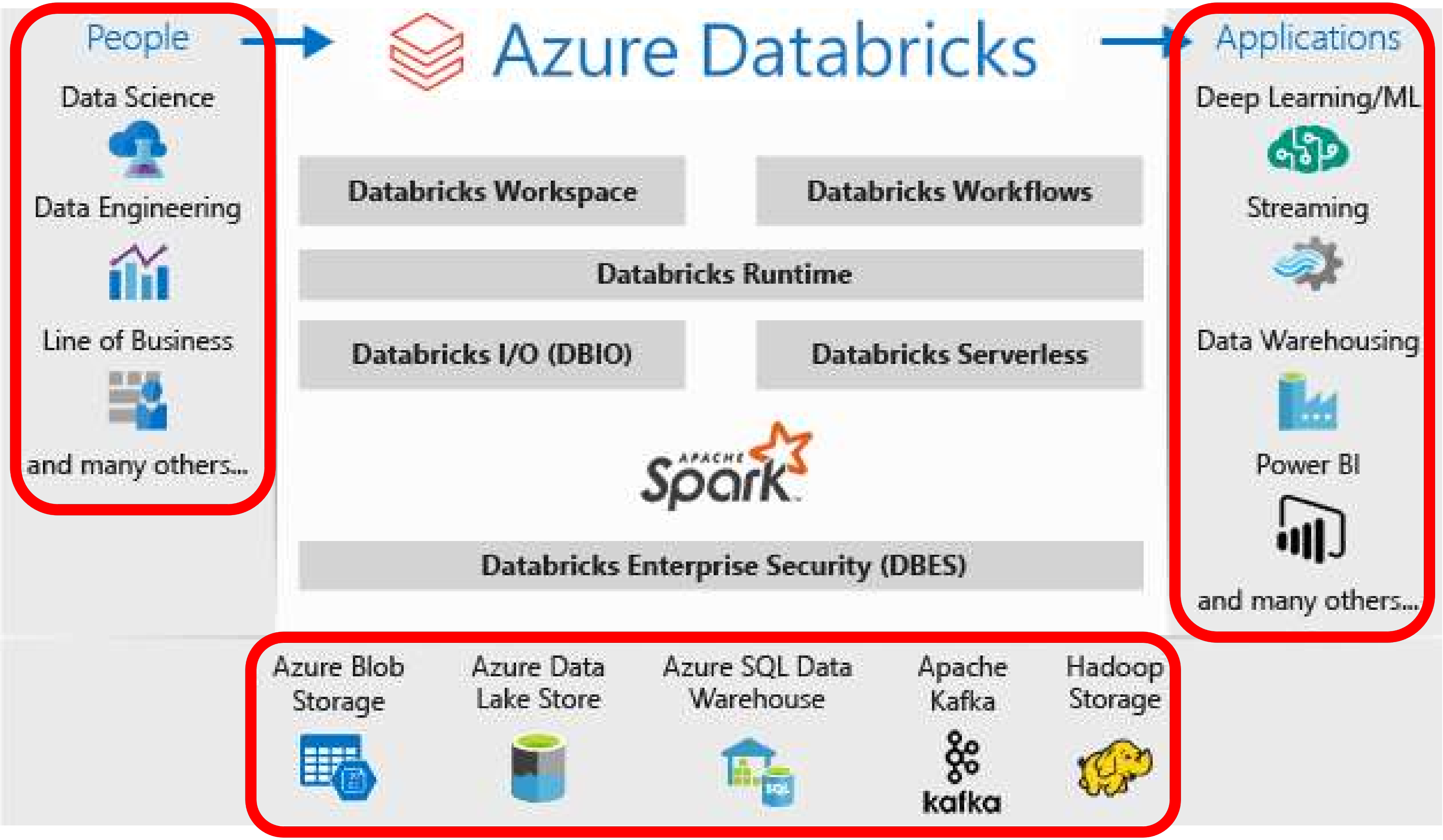
Azure Data Bricks

Duration: 6-8 Hours including labs

Describe Azure Data Bricks

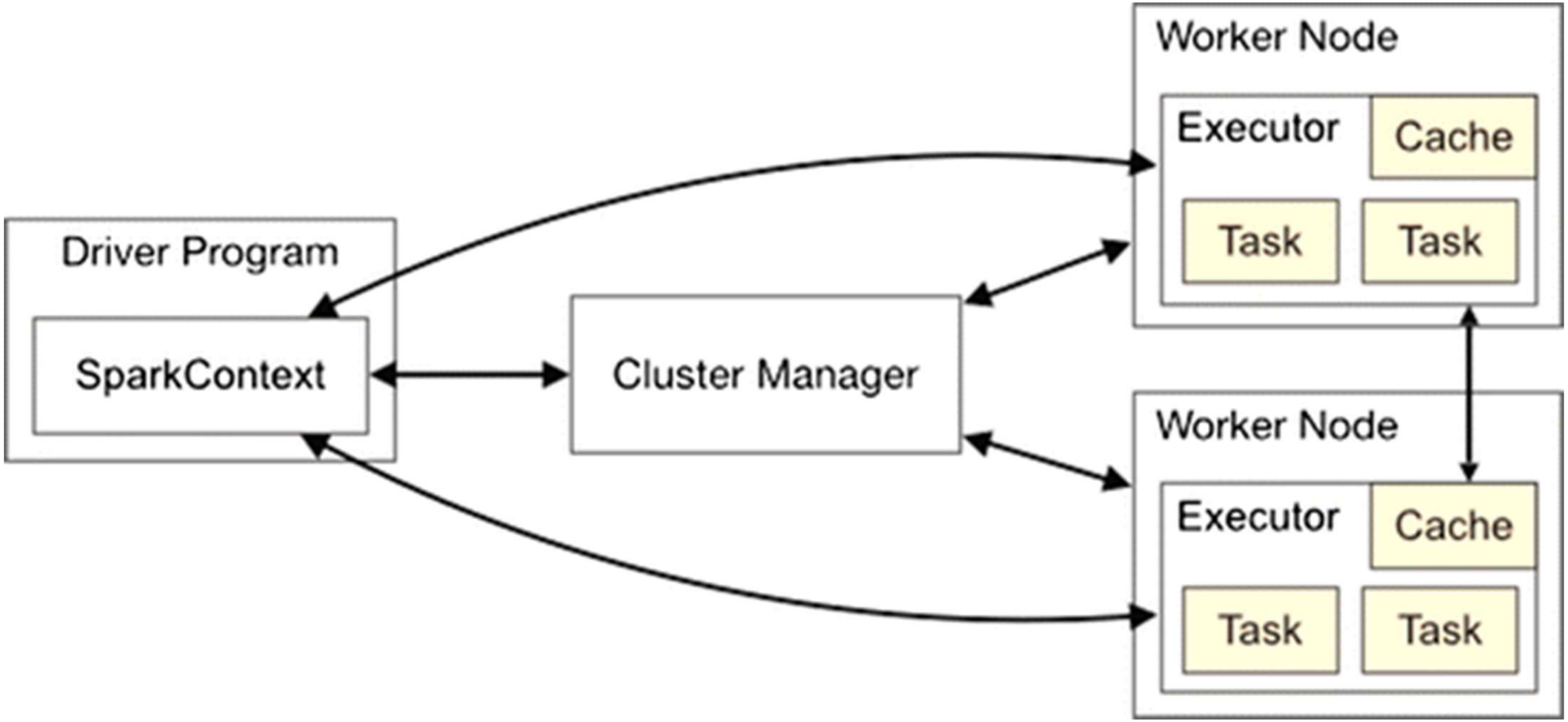
- Introduction
- Explain Azure Data Bricks
- Create an Azure Databricks Workspace and cluster
- Understand Azure Databricks Notebooks
- Exercise: Work with Notebooks


People
Data Science
Data Engineering
Line of Business
and many others...
Azure Databricks
Databricks Workspace
Databricks Workflows
Databricks Runtime
Databricks I/O (DBIO)
Databricks Serverless
APACHE Spark
Databricks Enterprise Security (DBES)
Applications
Deep Learning/ML
Streaming
Data Warehousing
Power BI
and many others...
Azure Blob Storage
Azure Data Lake Store
Azure SQL Data Warehouse
Apache Kafka
kafka
Hadoop Storage

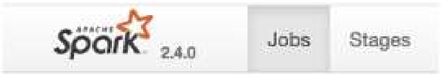
Spark Architecture fundamentals

- Introduction
- Understand the architecture of Azure Databricks spark cluster
- Understand the architecture of spark job

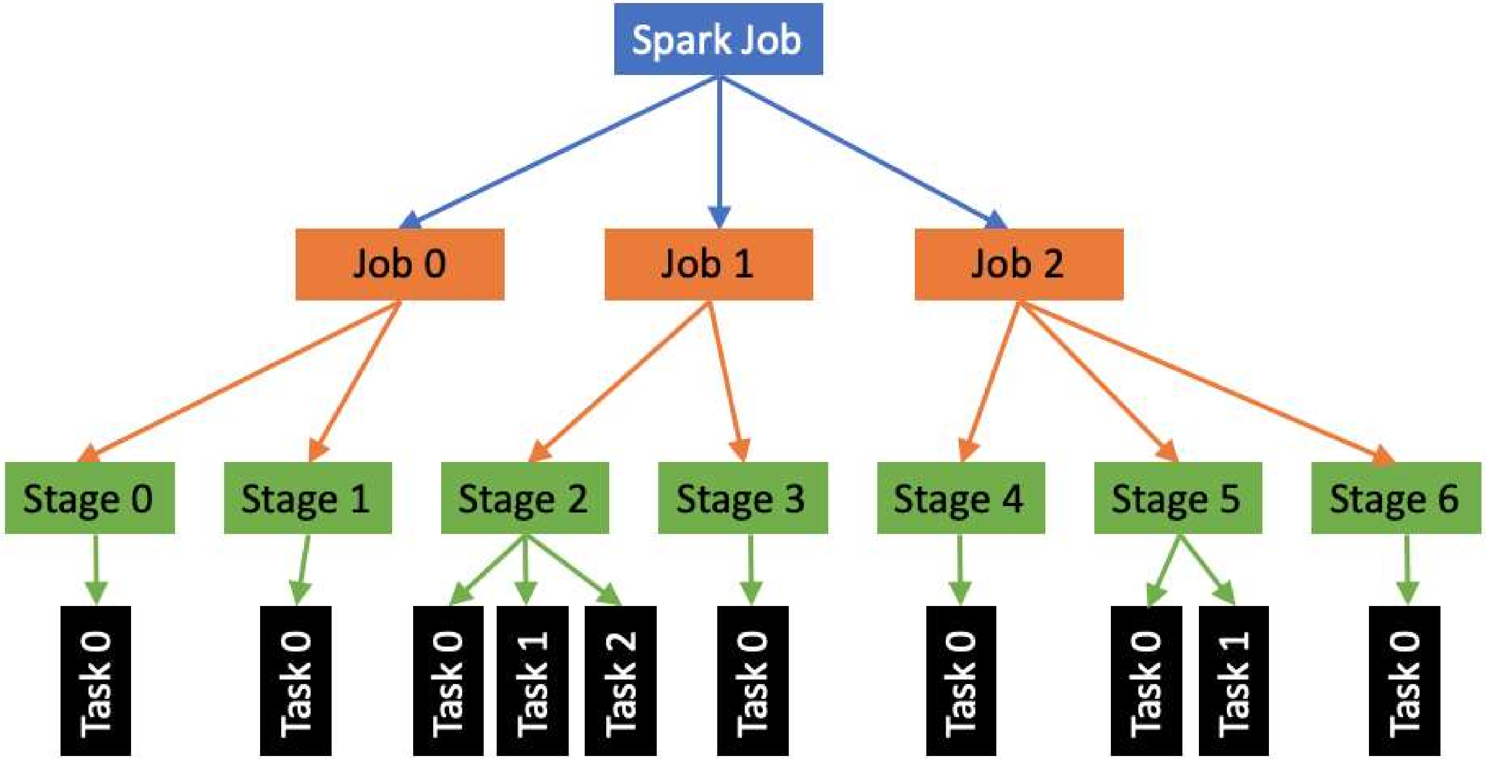
Spark Architecture



Spark Job



Click on a stage for tasks



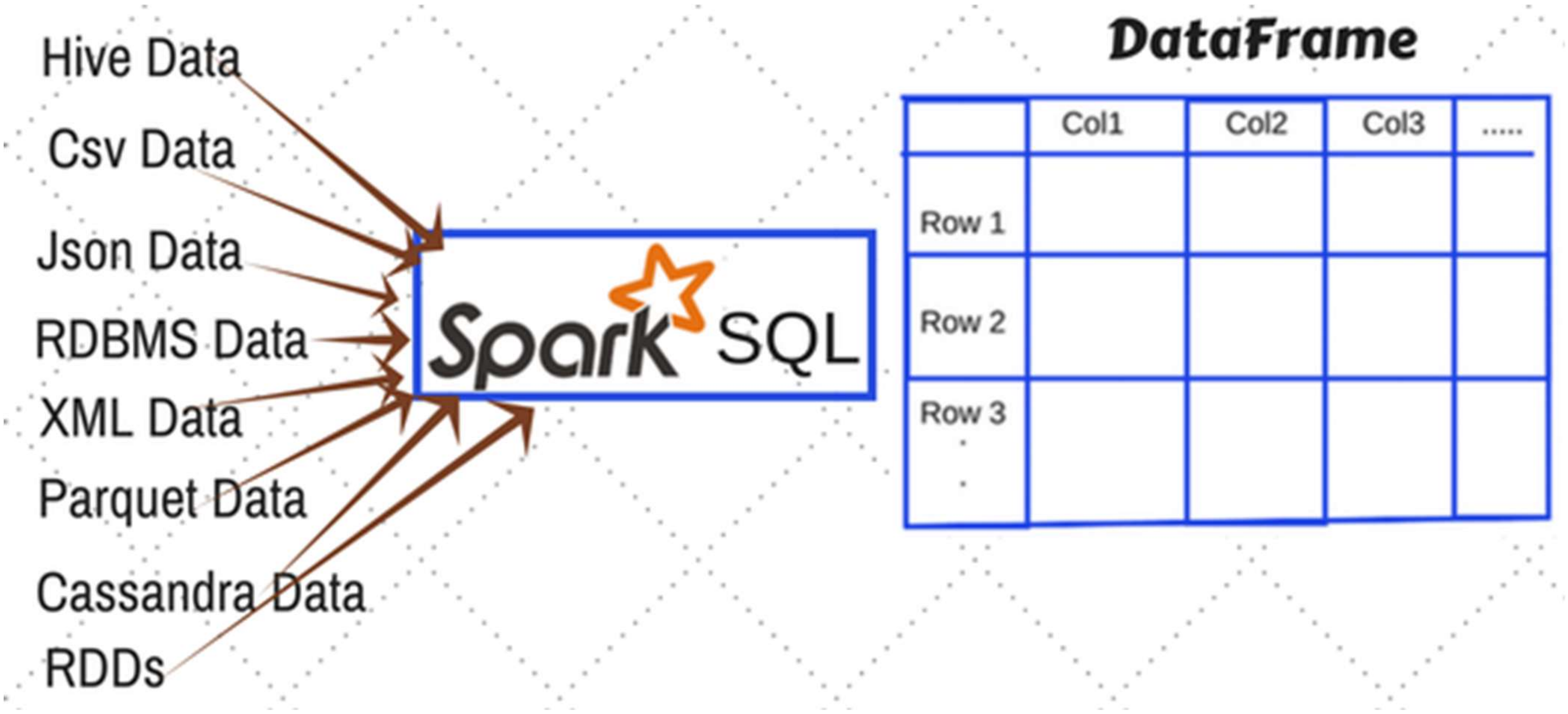
Read and write data in Azure Databricks

- Introduction
- Read data in CSV file
- Read data in JSON file
- Read Data in Parquet file
- Read Data stored in tables and views
- Write data
- Exercise: Read and write data

Work with DataFrames in Azure Databricks

- Introduction
- Describe a DataFrame
- Use Common DataFrame Methods
- Use the display function
- Exercise: Distinct articles

DataFrame in Spark

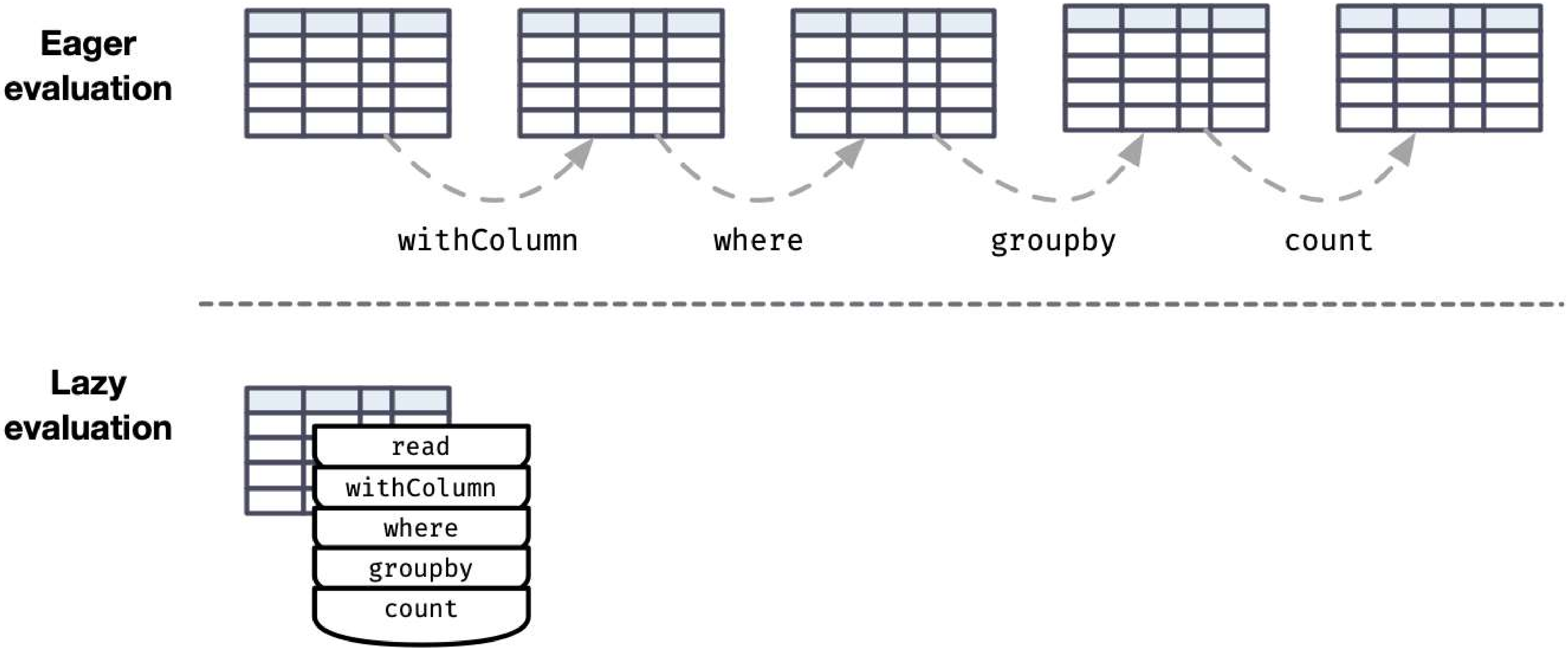


	Col1	Col2	Col3	.....
Row 1				
Row 2				
Row 3				

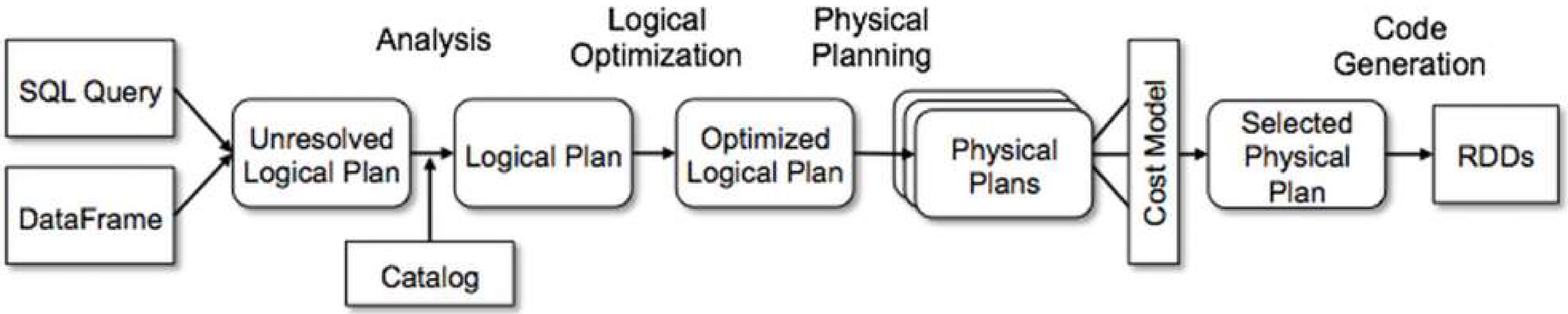
Describe lazy evaluation and other performance features in Azure databricks

- Introduction
- Describe the difference between eager and lazy execution
- Describe the fundamentals of how the Catalyst Optimizer works
- Describe and identify actions and transformations
- Describe performance enhancements by shuffle operations and Tungsten

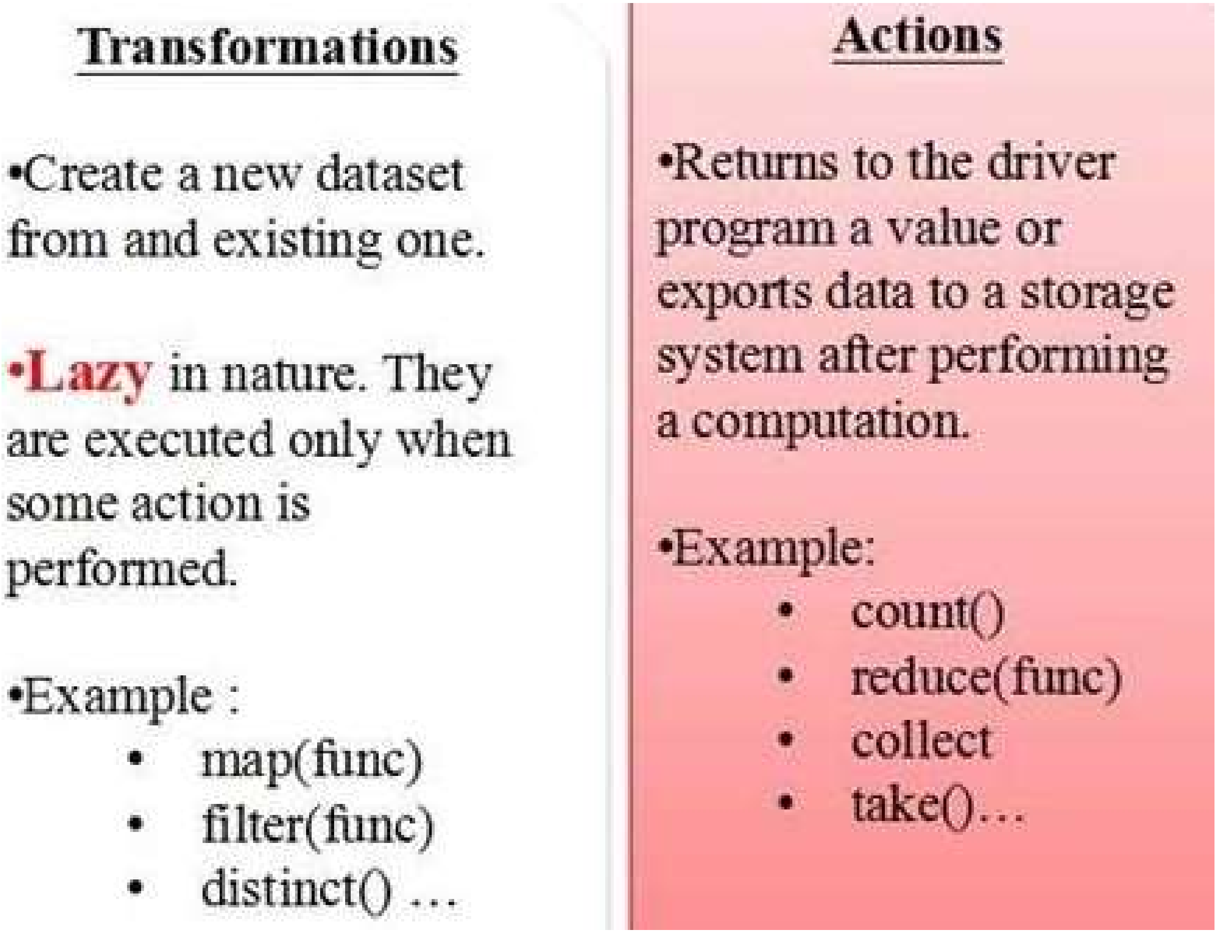
Lazy and Eager Evaluation



Spark Catalyst Optimizer



Identify Transformation and Action



Transformations

- Create a new dataset from and existing one.

- **Lazy** in nature. They are executed only when some action is performed.

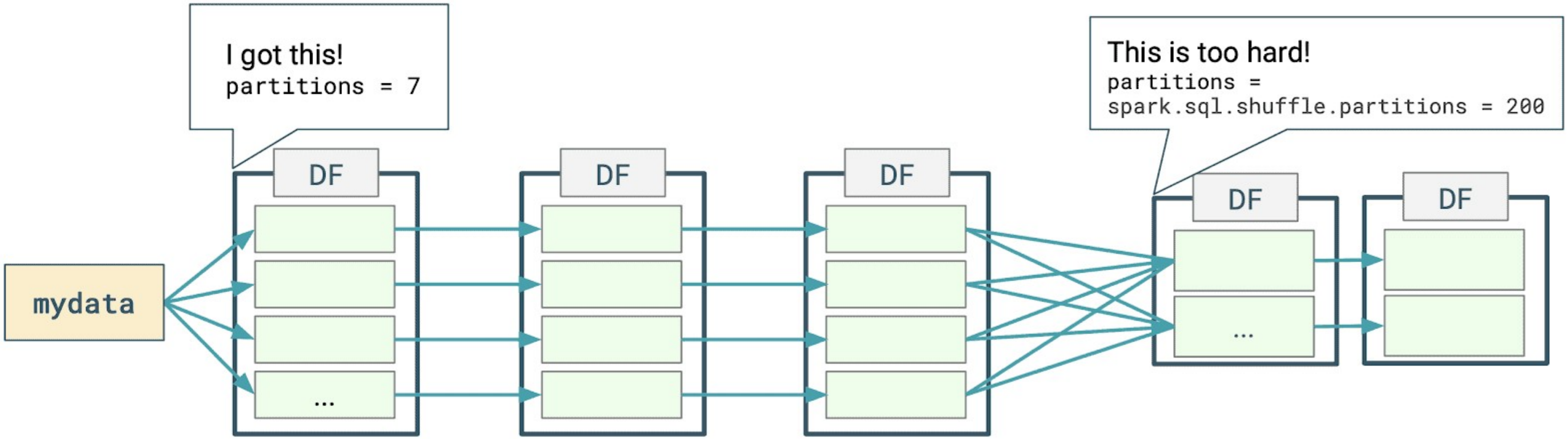
- Example :
  - map(func)
  - filter(func)
  - distinct() …

Actions

- Returns to the driver program a value or exports data to a storage system after performing a computation.

- Example:
  - count()
  - reduce(func)
  - collect
  - take()…

Shuffle Operations



Tungsten

- Codename for the umbrella project to make changes to Apache Spark's execution engine
- It focuses on substantially improving the efficiency of memory and CPU for Spark applications

- property:
  - spark.sql.tungsten.enabled to true

Work with DataFrame Columns

- Introduction
- Describe the columns class
- Work with Columns expressions

Work with DataFrames advanced methods

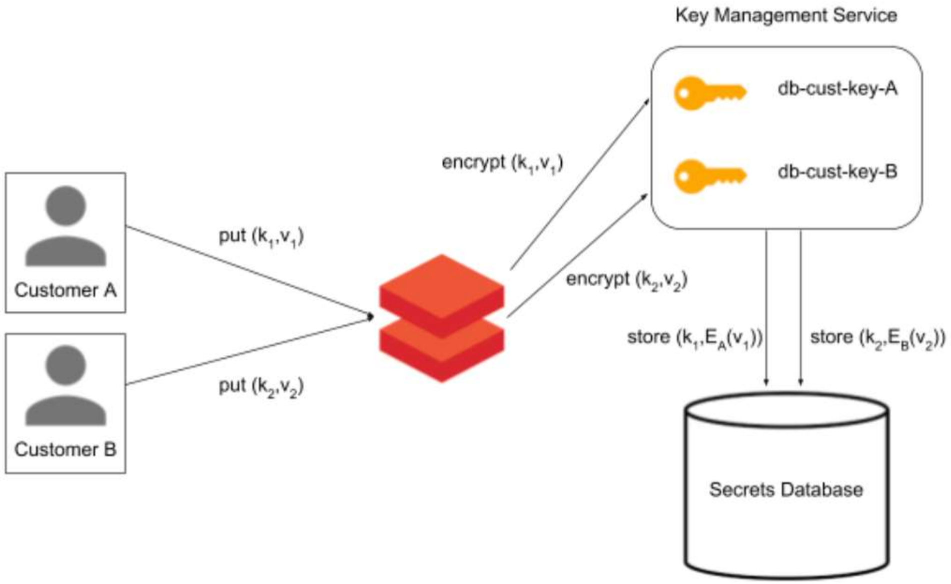
- Introduction
- Perform date and time manipulations
- Use aggregate functions
- Exercise: Deduplication of data

Platform architecture, security and data protection

- Describe Azure key vault and Databricks security scopes
- Secure access with Azure IAM and authentication
- Describe security
- Exercise: Access Azure storage with key vault backed secrets

Azure key vault and Databricks security scopes

- Azure Key Vault
  - A cloud service for securely storing and accessing secrets
- Databricks security scopes
  - Collection of secrets identified by a nam
  - Stored in an encrypted database owned



Secure access with Azure IAM and authentication

- Lab

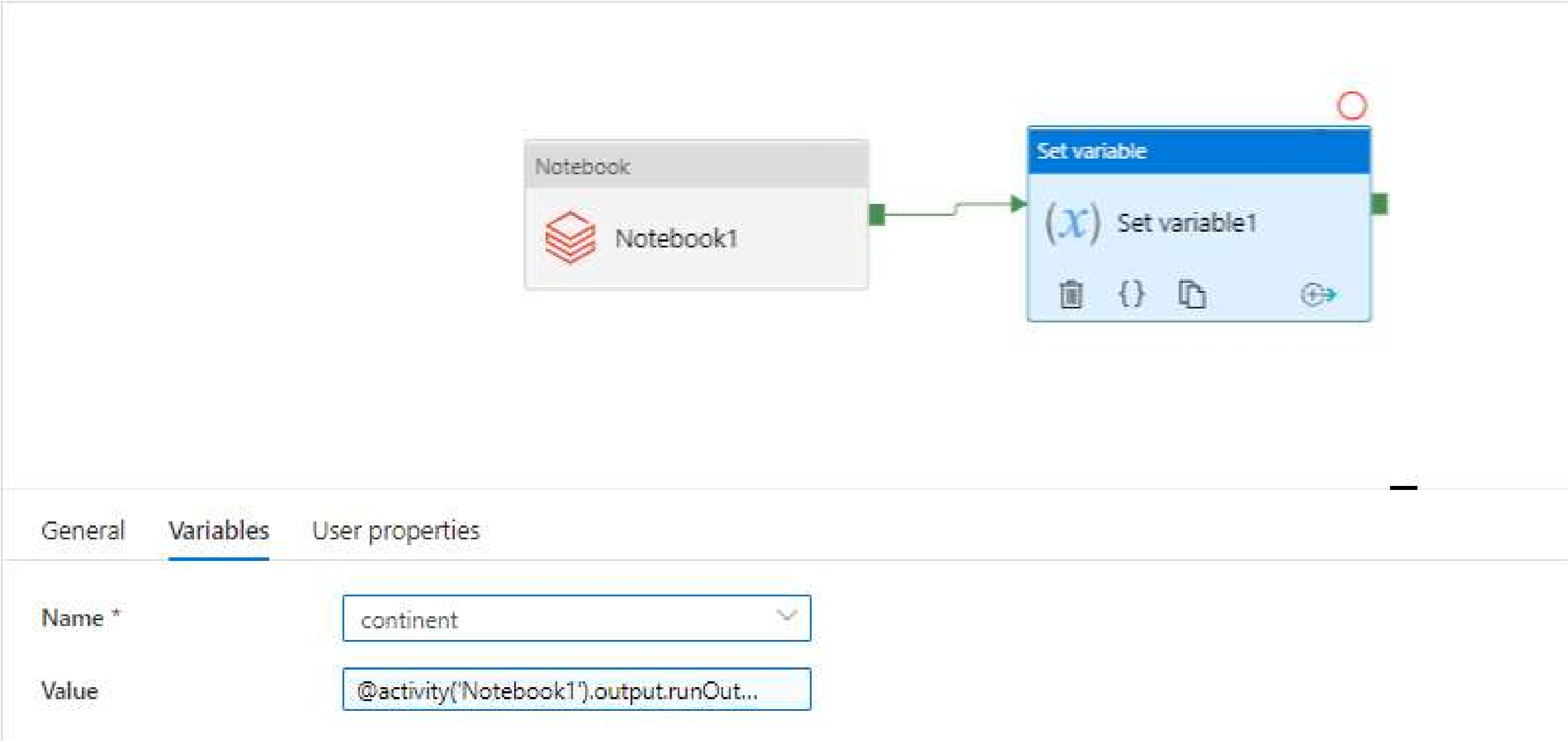
Access Azure storage with key vault backed secrets

- Lab

Create production workloads on Azure Databricks with Azure Data Factory

- Introduction
- Schedule Databricks jobs in a data factory pipeline
- Pass parameters into and out of Databricks jobs in data factory

Introduction



Name * continent

Value @activity('Notebook1').output.runOut...

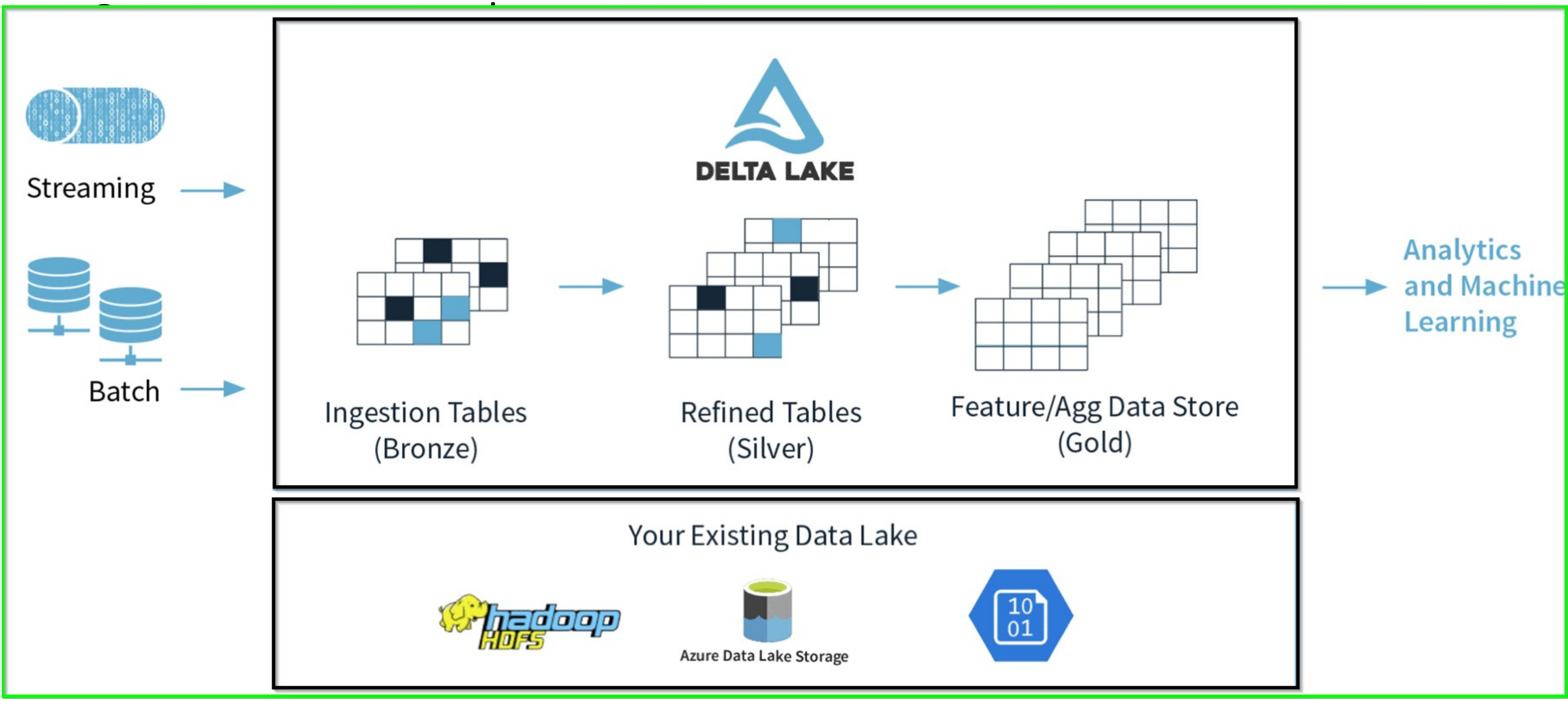
Schedule Databricks jobs in a data factory pipeline

- Duration: 10-20 minutes

Pass parameters into and out of Databricks jobs in data factory

- Duration: 10-20 minutes

What Is Delta Lake?



Lab: ETL using Batch and Streaming

- Ingest data in batch.
- Do basic transformations to move the data from Bronze -> Silver -> Gold
- Do basic transformation for Streaming data

Thanks