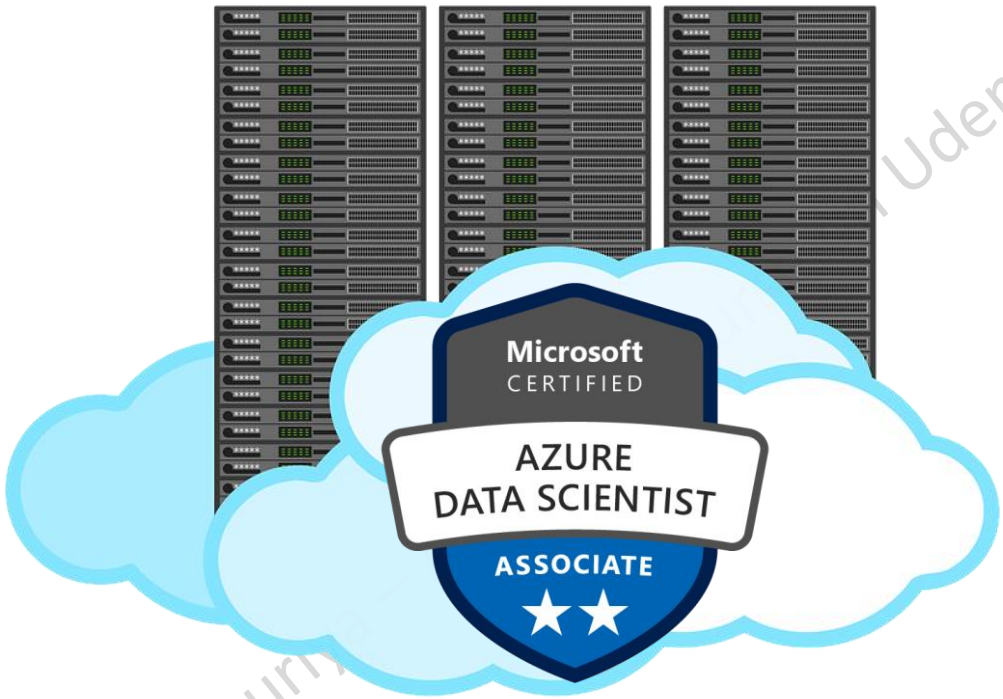


Azure Machine Learning

databricks

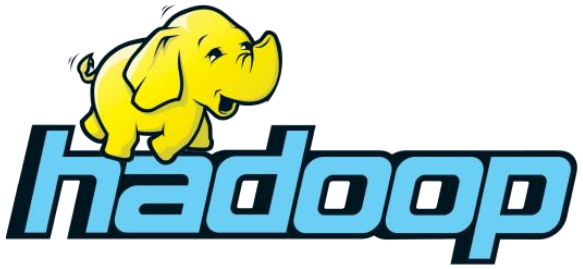
Azure Machine Learning

Databricks Updates to DP-100

- Create an Azure Databricks workspace
- Create an Azure Databricks cluster
- Create and run notebooks in Azure Databricks
- Link and Azure Databricks workspace to an Azure Machine Learning workspace
- Configure Attached Compute resources including Azure Databricks
- Run a training script on Azure Databricks compute
- Use MLflow to track experiments
- Track experiments running in Azure Databricks
- Deploy a model trained in Azure Databricks to an Azure Machine Learning endpoint.

What is Azure Databricks?

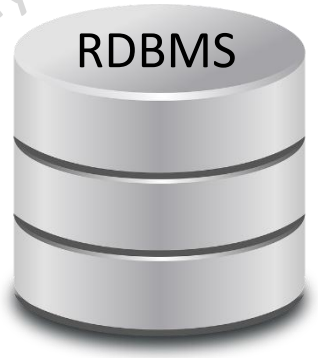
What is Azure Databricks?



What is Big Data?

- **Data Processing** – Voluminous and complex datasets, that traditional database system can not **deal with**

- **Analytics** – Set of techniques and technologies to **reveal insights** from a diverse, complex and large dataset



Big Data Characteristics

Big Data Characteristics

- **Volume – Quantity of data generated and stored**

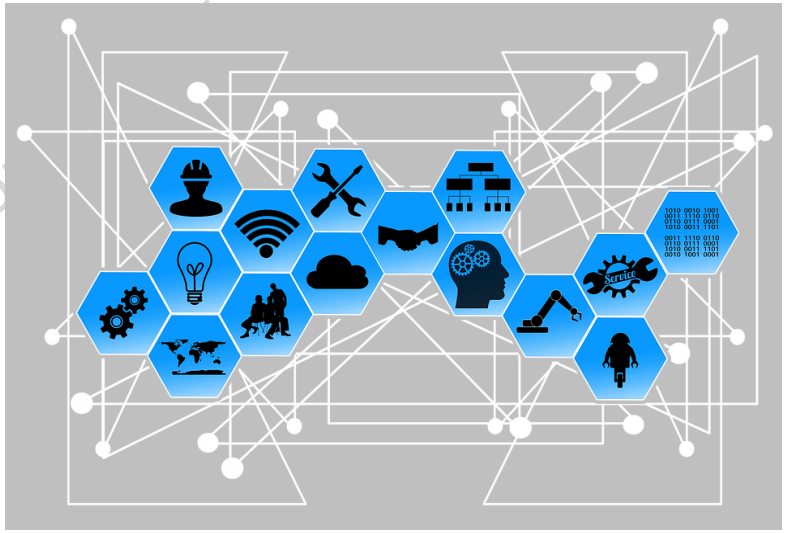
Big Data Characteristics

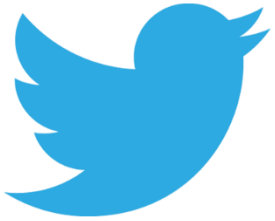
- Volume – Quantity of data generated and stored

- **Variety – Type and nature of data**

Big Data Characteristics

- Volume – Quantity of data generated and stored
- Variety – Type and nature of data
- **Velocity – Speed of data generation and processing**





Big Data Characteristics

- Volume – Quantity of data generated and stored
- Variety – Type and nature of data
- Velocity – Speed of data generation and processing
- **Variability – Inconsistency of dataset**

"Read a book on the flight."

"Book me a flight."

"This book is good."

"We are fully booked."

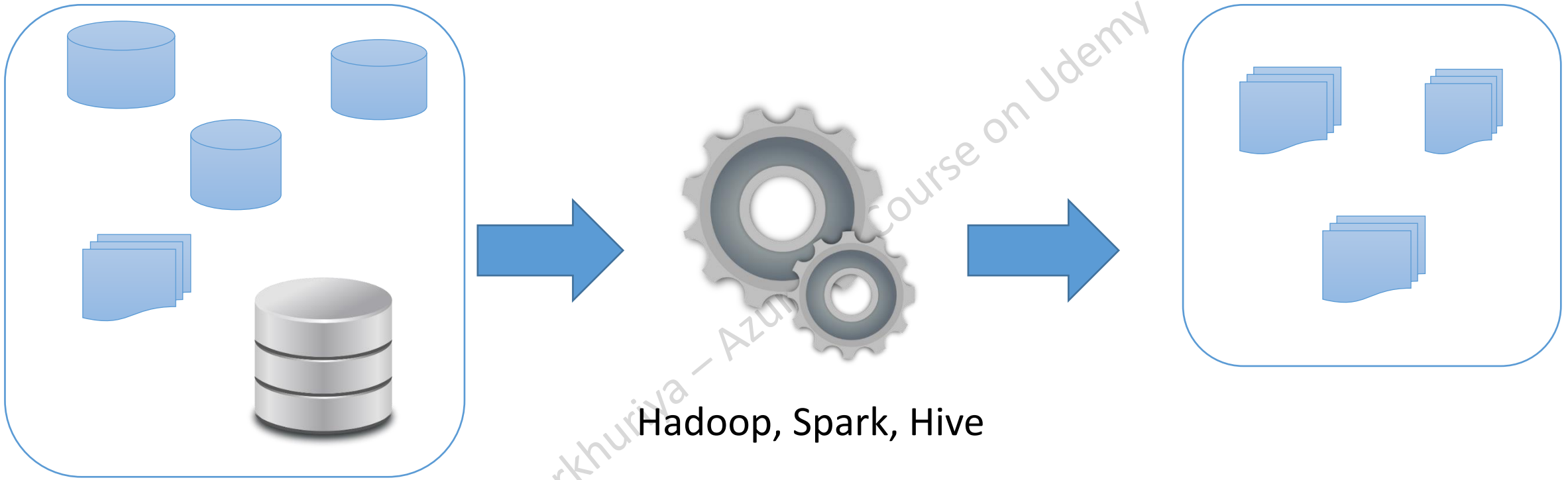
"He was booked for a crime."

Big Data Characteristics

- Volume – Quantity of data generated and stored
- Variety – Type and nature of data
- Velocity – Speed of data generation and processing
- Variability – Inconsistency of dataset
- **Veracity – Quality/Uncertainty of data captured**

Types of Big Data Applications

Batch Processing



Hadoop, Spark, Hive

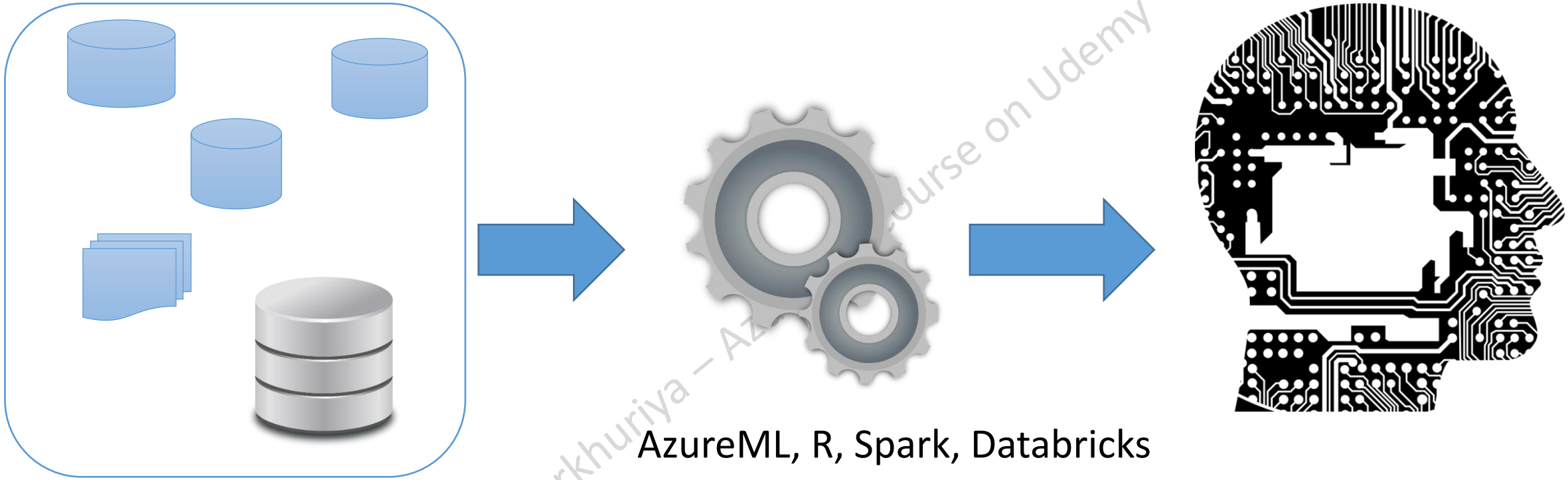
Re-format and clean the data for further processing

Real-Time Data processing

Kafka, Storm, Spark

Analyse and Process in real-time

Predictive Analytics



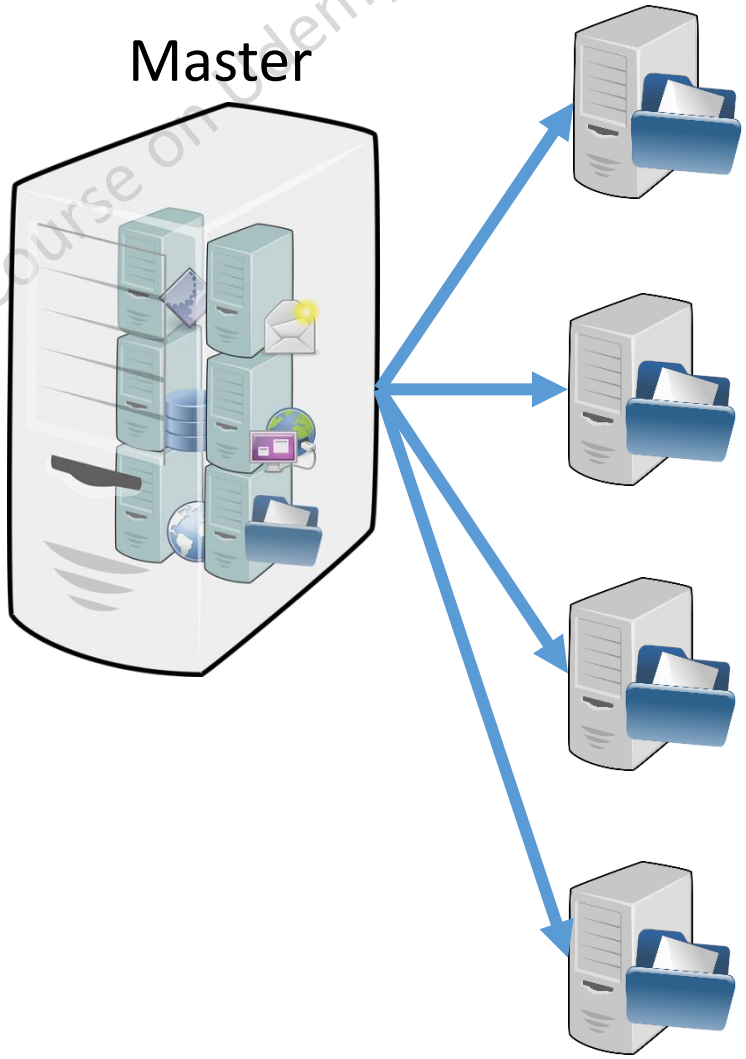
AzureML, R, Spark, Databricks

Machine Learning, AI

What is Hadoop?

Hadoop as a framework

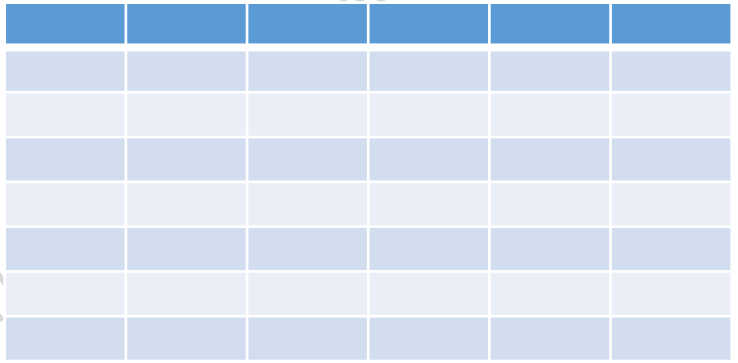
- Open-source framework for distributed processing
- Designed to scale up from single to thousands of servers in parallel
- Uses Hadoop Distributed File System (HDFS)
- MapReduce as data processing function
- YARN – Resource Management

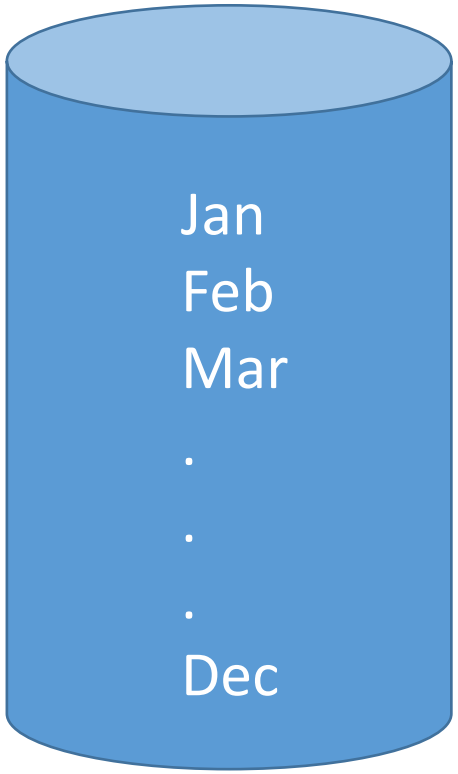


Get a summary report for all the transactions done in the past one year.

Traditional Database System

Select * from table1
Where Year = “2017”
group by month



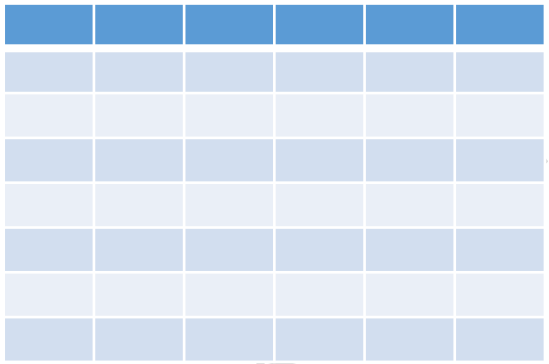


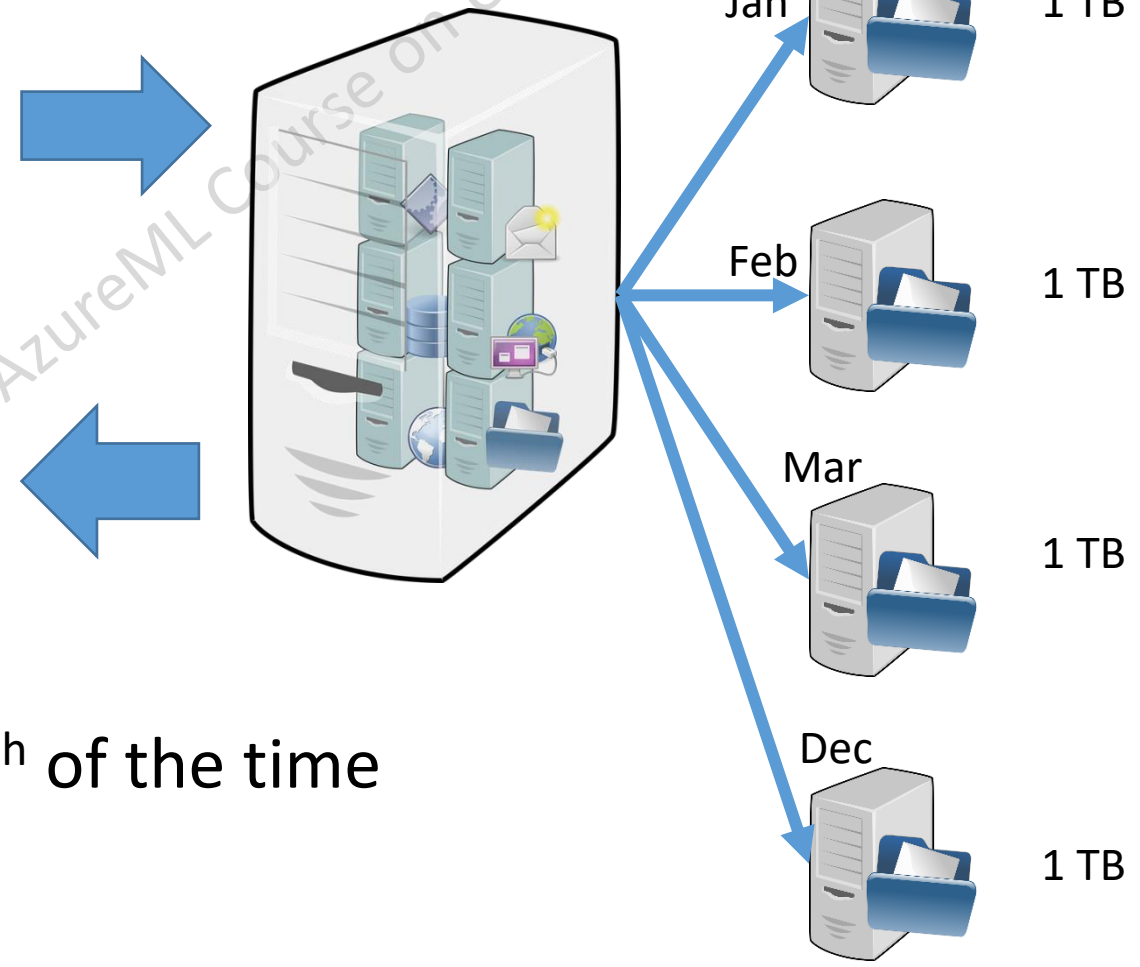
12TB

Hadoop

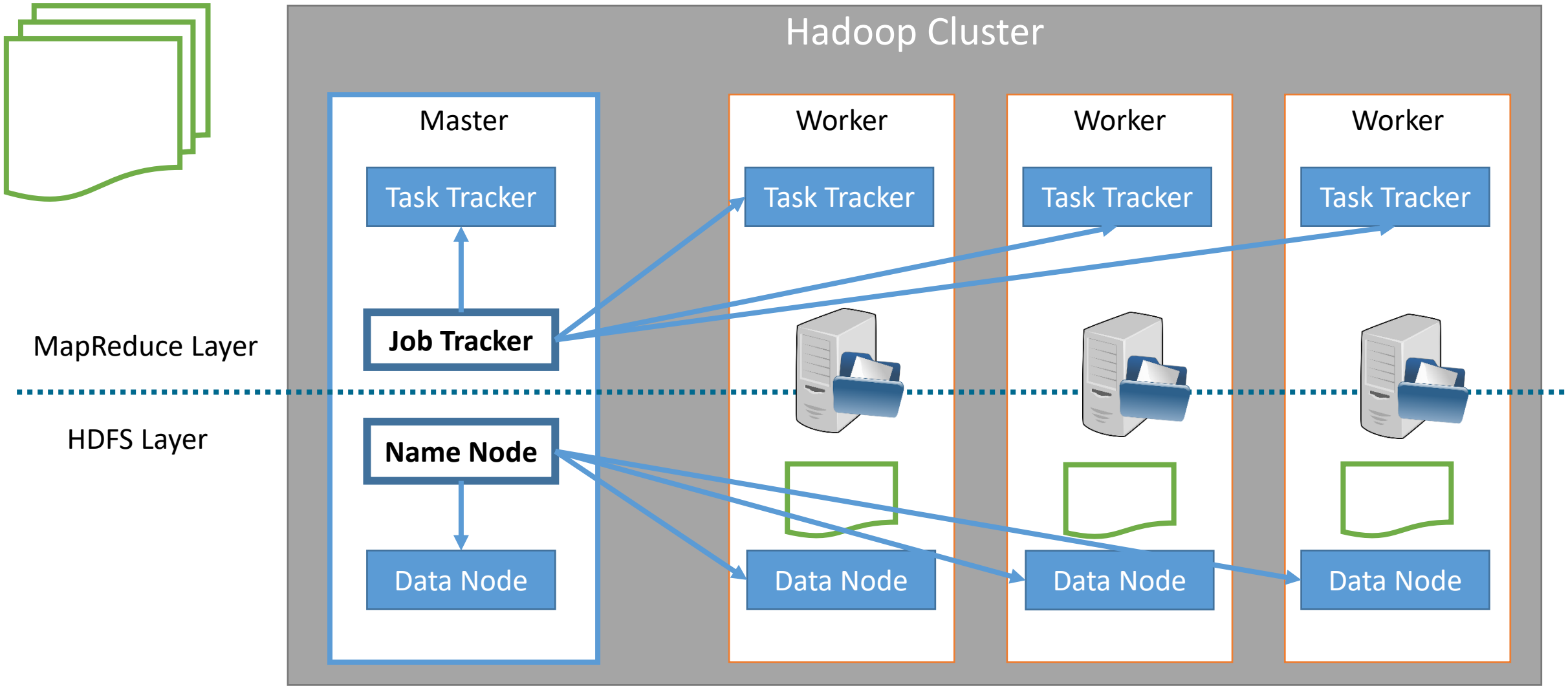
Hive query

```
Select * from table1
Where Year = "2017"
group by month
```

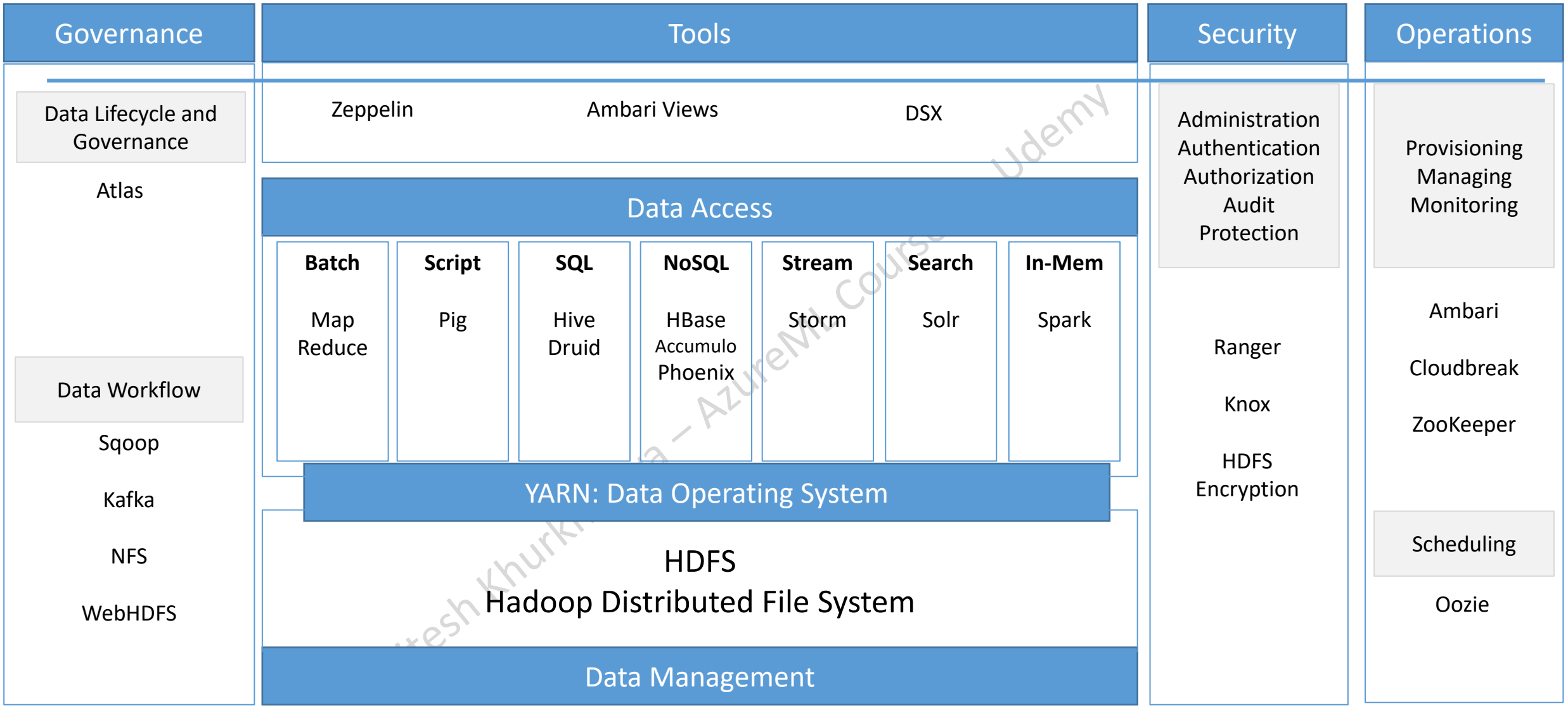




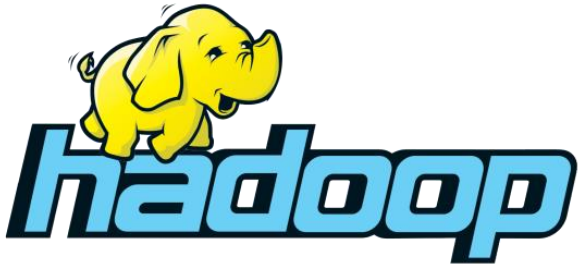
1/12th of the time



Big Data Ecosystem



From Hadoop to Databricks

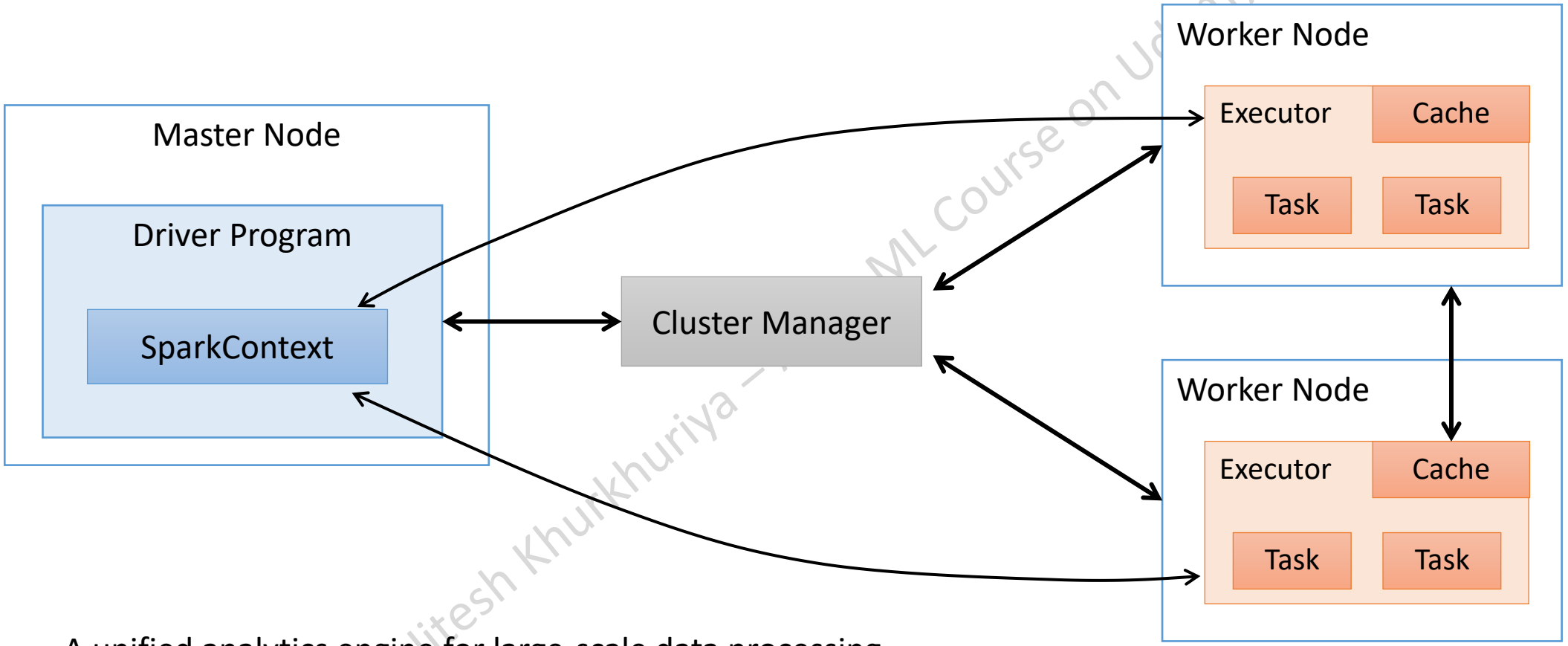


- Inefficient processing
- Batch Only
- Huge ecosystem
- No in-memory processing.

What is Apache Spark?

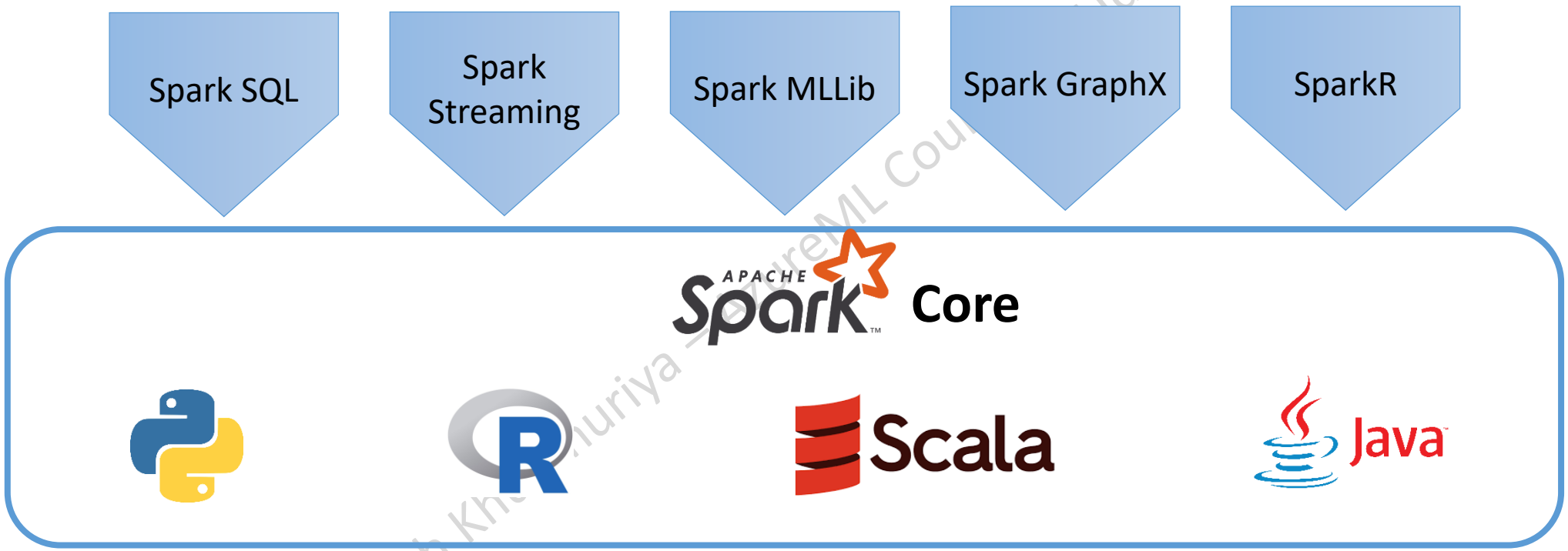
- A unified analytics engine for large-scale data processing.
- In-Memory Distributed cluster computing
- Provides APIs for development in Java, Python, Scala and R
- Supports batch and real-time processing
- Very high speed of execution.

Spark Architecture

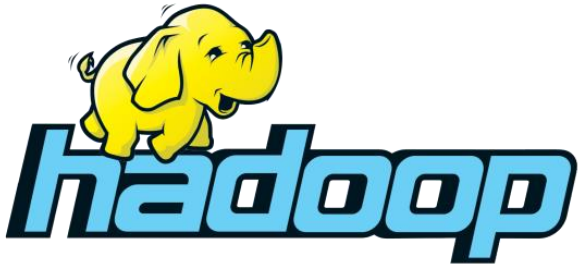


A unified analytics engine for large-scale data processing.

Apache Spark Ecosystem



From Hadoop to Databricks



- Inefficient processing
- Batch Only
- Huge ecosystem
- No in-memory processing

- Not Easy to use
- Develop environment on your own
- Collaboration of work
- Not cloud-first

databricks

- Platform optimized for efficient working with Spark
- Provides workspace to manage Spark and its infrastructure
- Ease of collaboration and integration

Databricks Workspace

Collaborative Notebooks, libraries, experiments

Databricks Run Time

DELTA LAKE

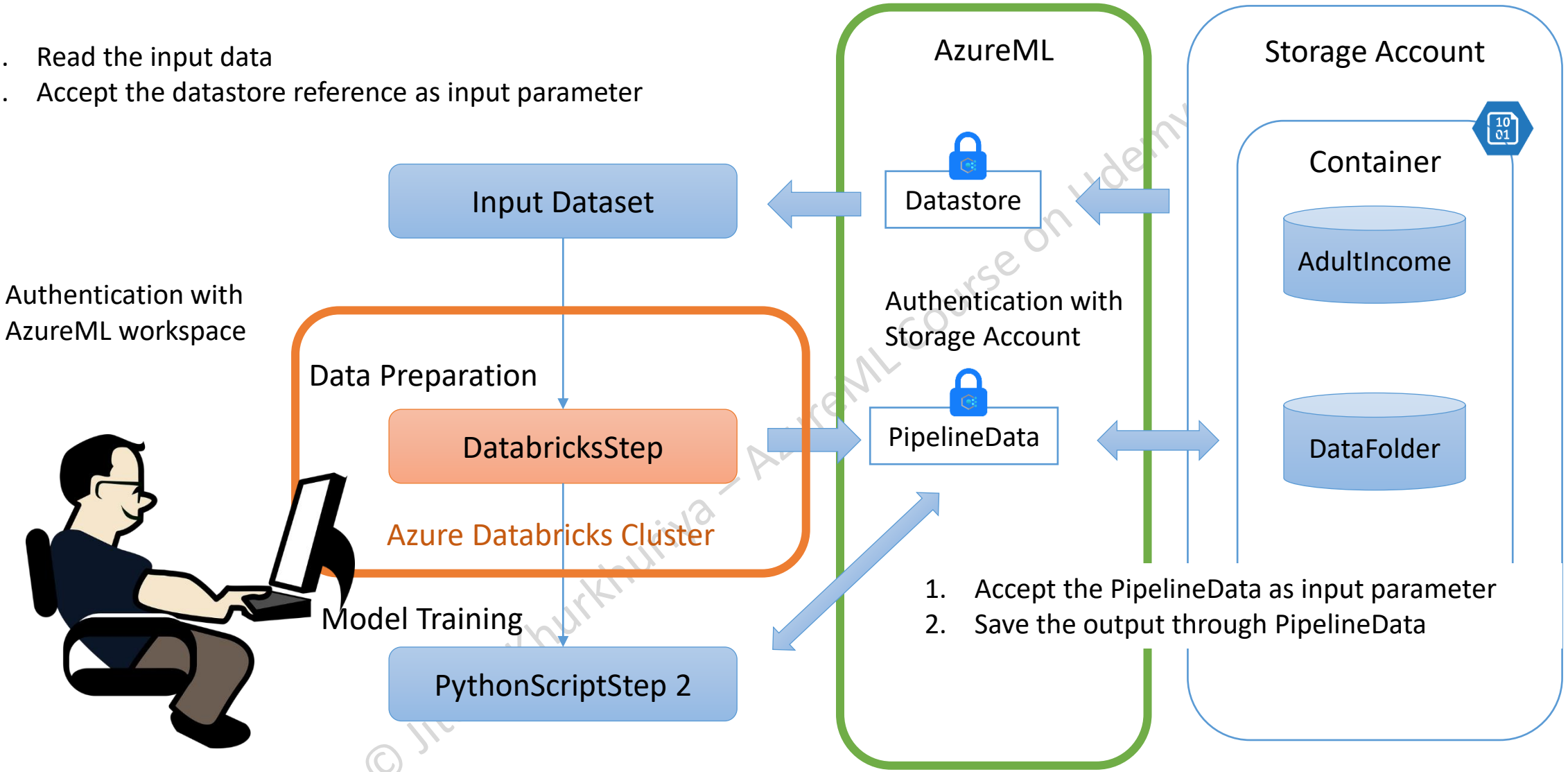
APACHE Spark

mlflow

Cloud Services

aws

1. Read the input data
2. Accept the datastore reference as input parameter



Steps to run a Pipeline with DatabricksStep

Set-up Steps

- Create Azure Storage Account
- Create Blob Container
- Copy the access key for storage account
- Upload the data/csv to container

- Create AzureML Workspace
- Create AzureML Datastore
- Create AzureML Dataset

- Create Databricks Workspace
- Create Databricks Cluster
- Create and Copy Databricks workspace access key

Python Job Steps

- Create workspace object from the config file
- Create custom environment and cluster
- Create run_config for python script step
- Create data reference for input dataset
- Create PipelineData objects for Input/Output

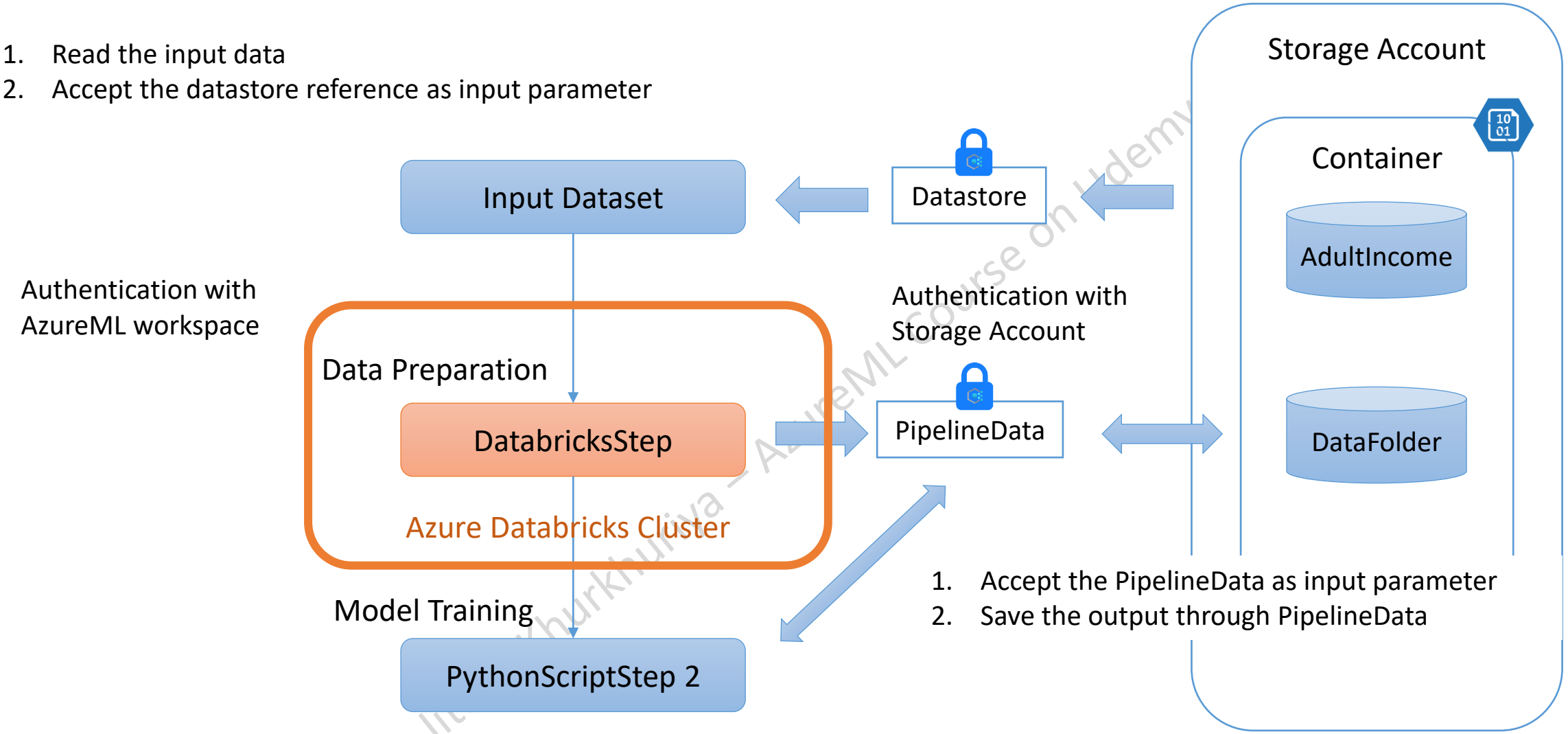
- Create Databricks compute configuration parameters with key
- Attach the databricks cluster as attached compute

- Create DatabricksStep step
- Create PythonScriptStep as the second step
- Create Pipeline using DatabricksStep and PythonScriptStep

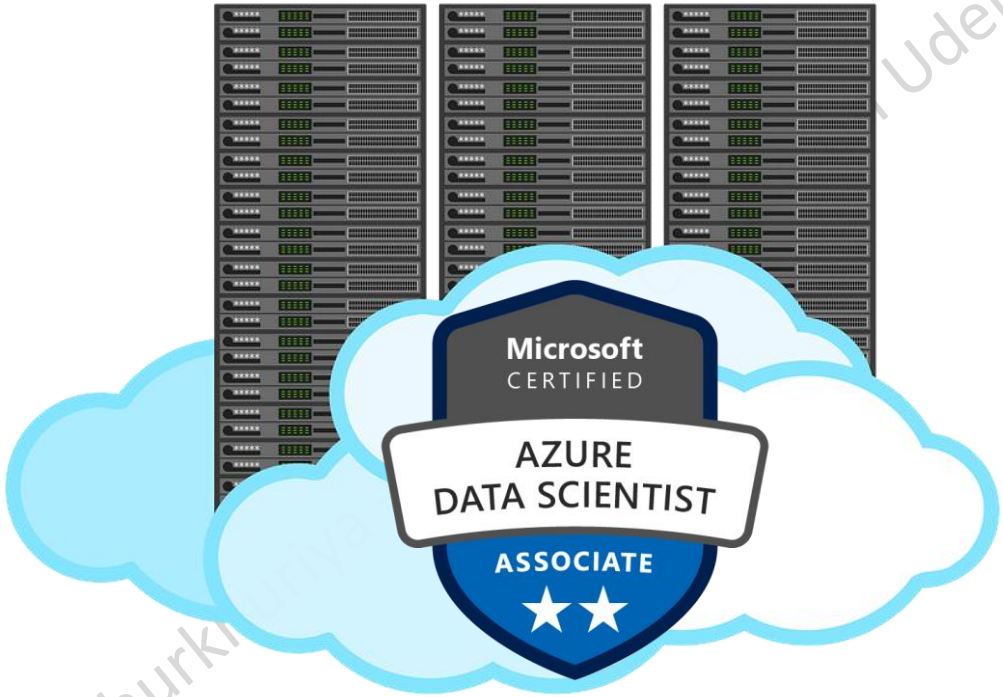
Databricks notebook steps

- Unmount the input and output data mounts
- Get the Inputs and Outputs parameters using dbutils.widgets.get
- Create conf_key and key_value for the storage account
- Mount input and output blob storage folders as dbfs directory
- Read data from the mounts
- Perform data processing or functions as desired
- Make output directories on blob storage using dummy blob
- Save output files using the dbfs mount to blob storage

1. Read the input data
2. Accept the datastore reference as input parameter



Azure Machine Learning



Thank You..!!