BUILDING THE DELTA LAKE

1

IMPLEMENTATION

What we will implement?

2

AZURE DATABRICKS

Overview and Setup of Azure Databricks

3

DELTA LAKE IMPLEMENTATION

Implementing the Delta Lake with Azure Databricks

4

DATABRICKS NOTEBOOK

Executing Databricks Notebook Activity with Data Factory

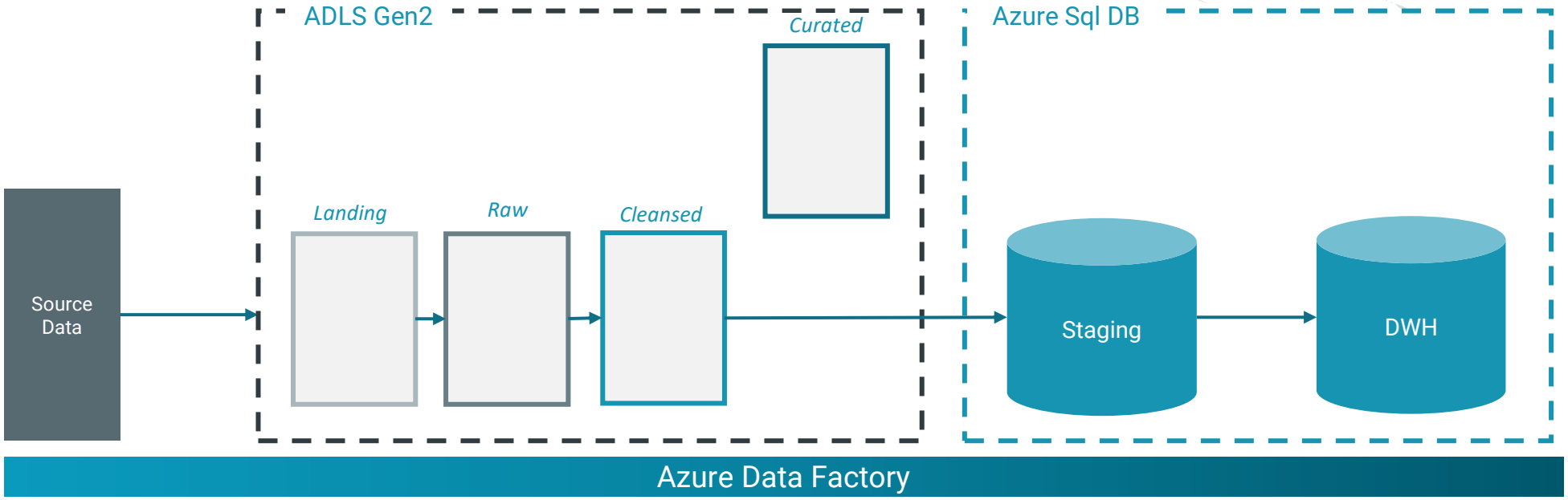
MODULE 12

SECTION 1

IMPLEMENTATION

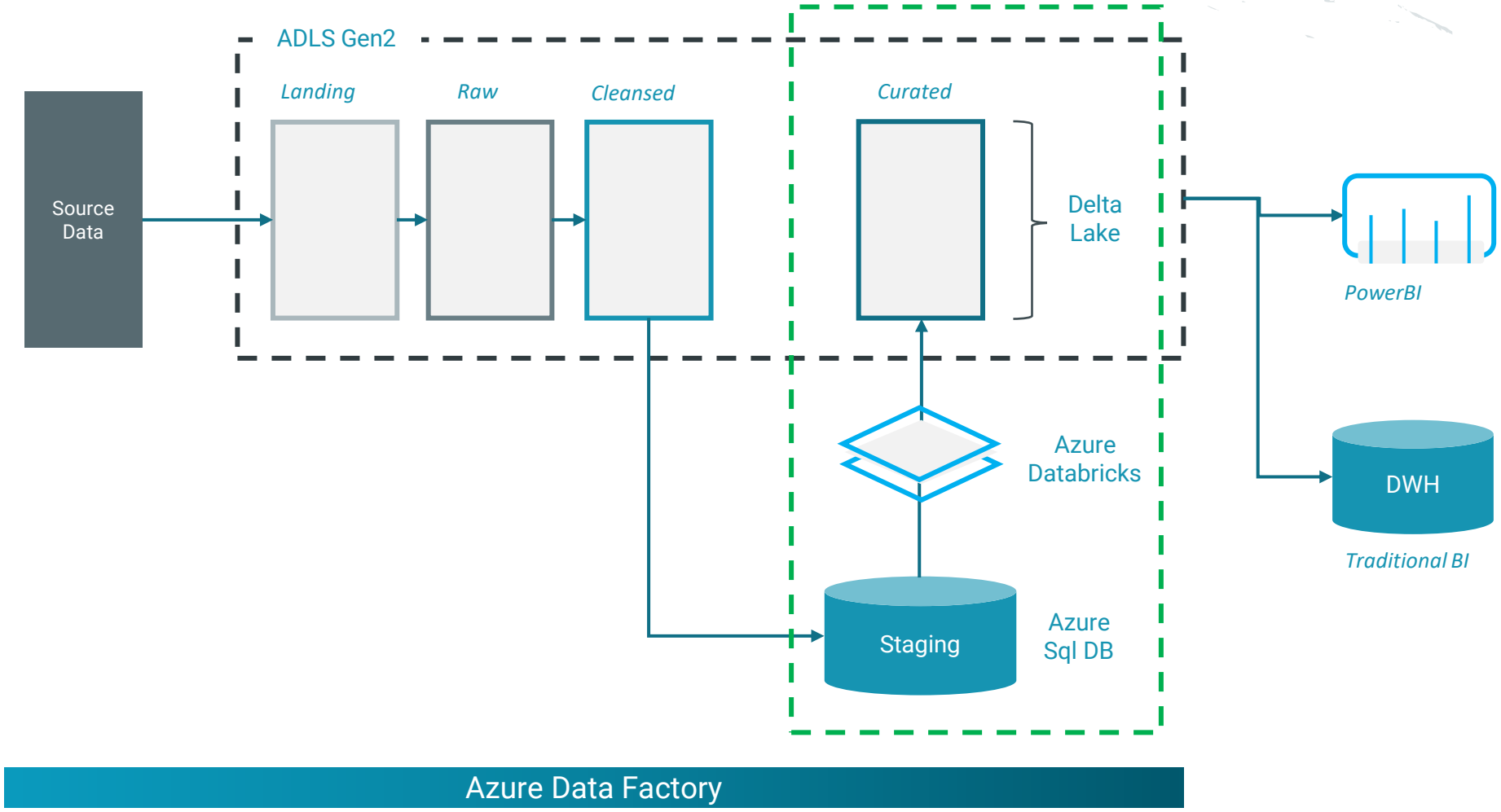
IMPLEMENTATION

What have we implemented?



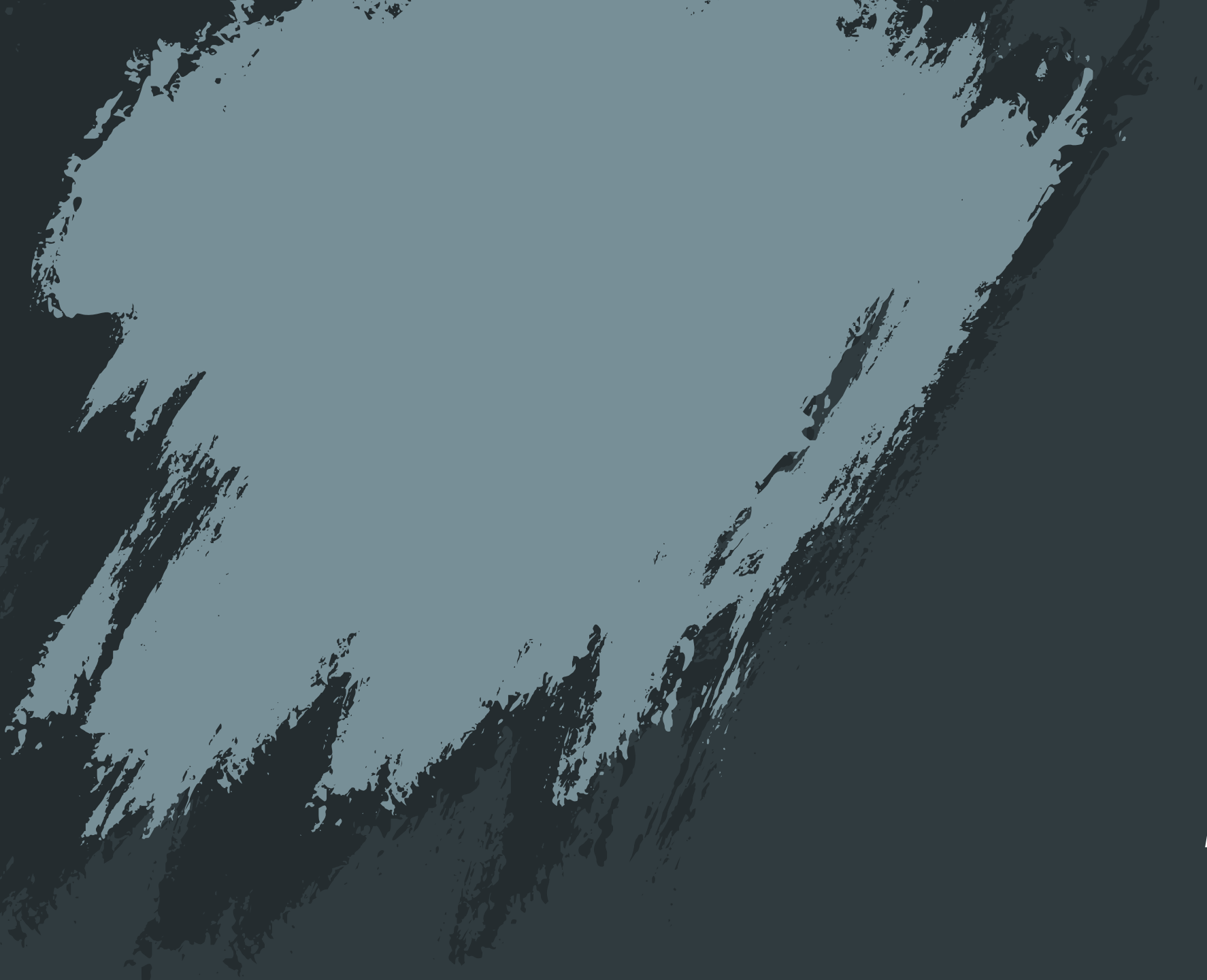
IMPLEMENTATION

What we will implement?



What we will learn?

- *Databricks Setup*
- *How to implement Delta Lake Tables*
- *Delta Lake Concepts*
- *Running Databricks Activity from Data Factory*



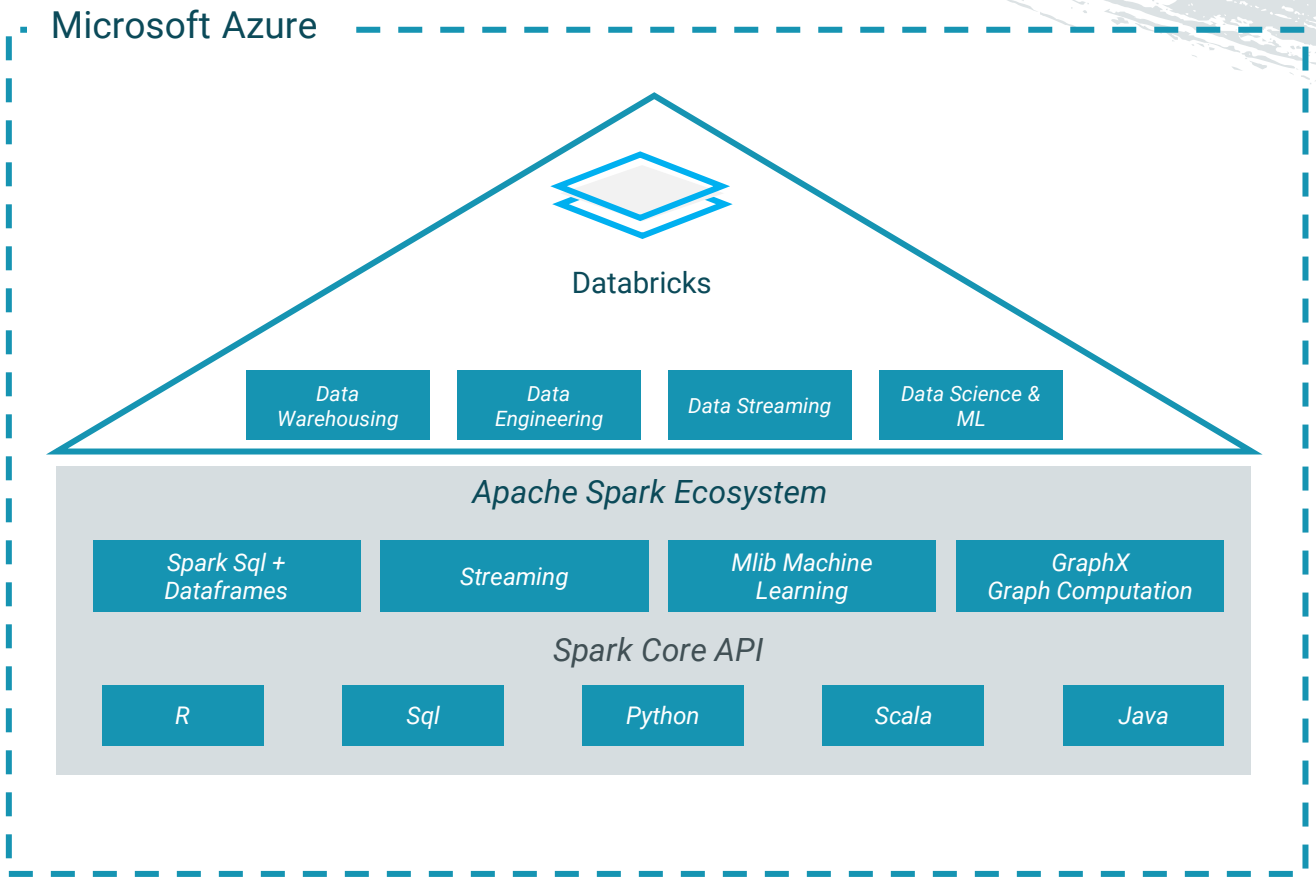
SECTION 2

AZURE DATABRICKS

AZURE DATABRICKS

What is Azure Databricks?

Databricks + Microsoft Azure



Power of Spark in Microsoft Azure

Native Integration with Azure Services

Compute with preferred language

Delta Lake

Machine Learning Environments

Unifies Data, Analytics and ML in a single platform within Azure

AZURE DATABRICKS

What are the main artifacts of Azure Databricks?

Core Artifacts of Azure Databricks

CLUSTERS

Two types of Clusters – Interactive and Job cluster

WORKSPACES

Enables users to organize/share – Notebooks, Libraries, and Dashboards

NOTEBOOKS

Web-based interface containing runnable spark code, visualizations, and text

LIBRARIES

Containers residing within workspaces, holding Python, R, Java/Scala Libraries

JOBS

Mechanism to run a notebook or JAR on the Databricks cluster

AZURE DATABRICKS

Setting up the Databricks Environment

Setup Databricks Service in Azure Portal

DATABRICKS SERVICE

01

DATABRRICKS CLUSTER

02

Create the Azure Databricks Cluster

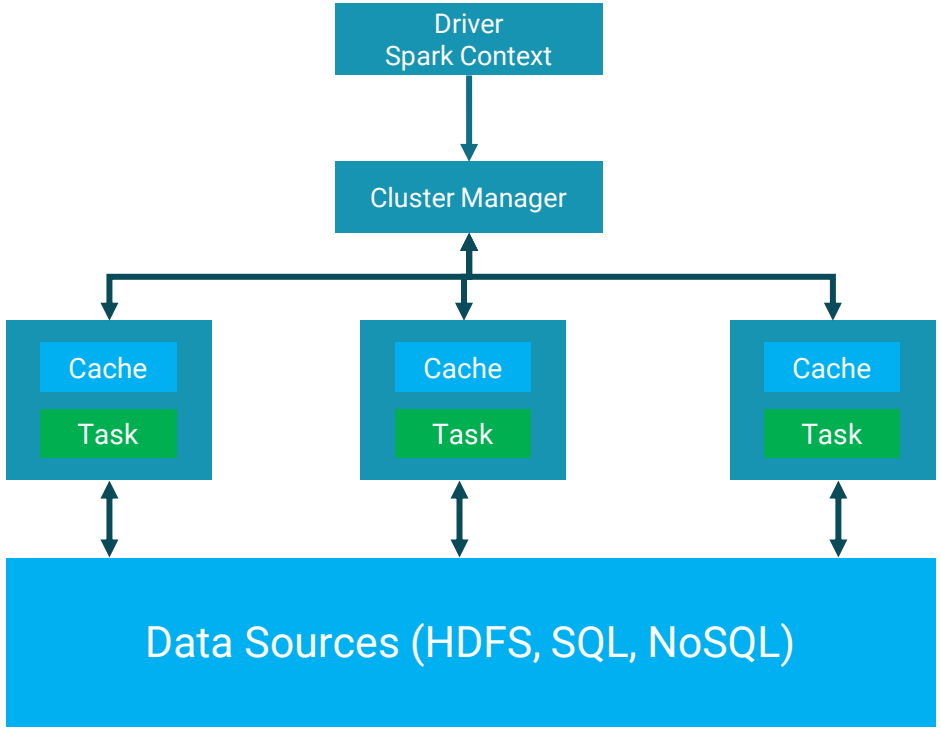
03

MOUNT STORAGE

Mount Azure Data Lake Storage

AZURE DATABRICKS

Spark Cluster Architecture



Driver runs the user's main function and executes operations in parallel on worker nodes

Worker nodes and driver nodes execute as VMs in Azure

Worker nodes read and write data from/to data sources

Worker nodes cache transformed data in memory as RDDs (Resilient Data Sets)

Results of the operation are collected by the driver

The spark cluster architecture powers Azure Databricks bringing distributed computing

AZURE DATABRICKS

Types of Clusters

There are two basic types of Clusters in Databricks

ALL PURPOSE/INTERACTIVE CLUSTERS

- Can be created using the Databricks UI, CLI, or API
- Used for analyzing data using interactive notebooks
- Can share them with multiple users to collaborate
- Can manually terminate and restart them

JOB CLUSTERS

- Used for running automated jobs using the UI or API
- Job scheduler creates a job cluster when a job is run
- Job scheduler terminates the job cluster when the job completes
- Cannot restart a job cluster

AZURE DATABRICKS

Mounting Azure Data Lake Storage

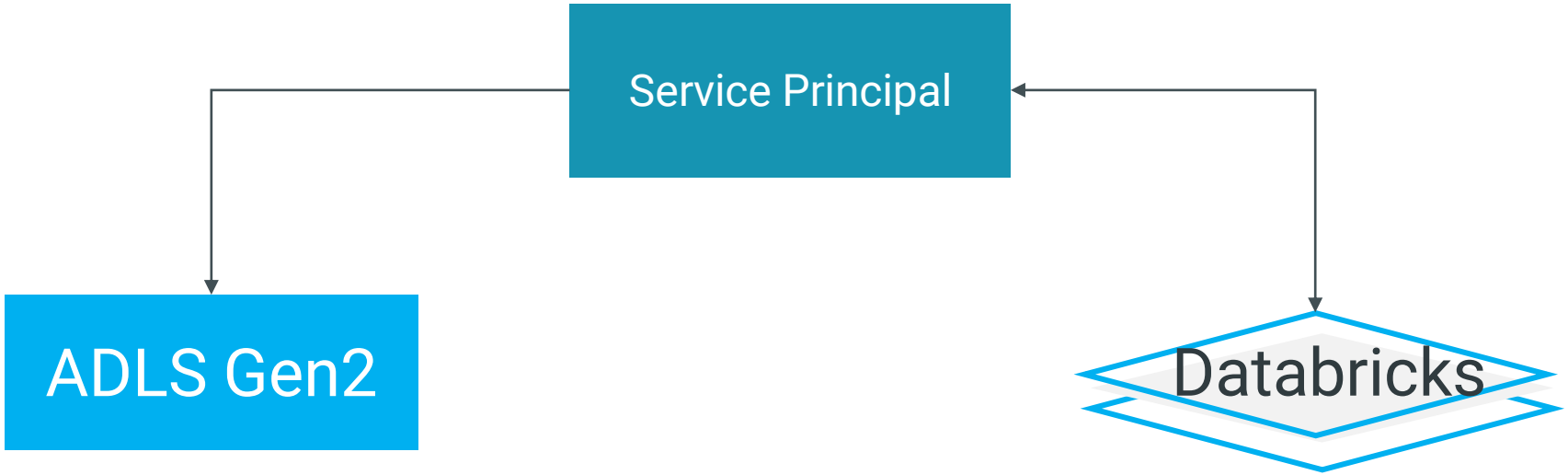
1

- *Create Service Principal*
- *Grant Service Principal access to the Storage Account*
- *Provide the role*

Storage Blob Data Contributor

2

- *Mount storage account in Databricks via Service Principal*
- *Use* dbutils.fs.mount *command with the* abfss *scheme*
- *Provide the tenant ID, directory ID, and secret name*

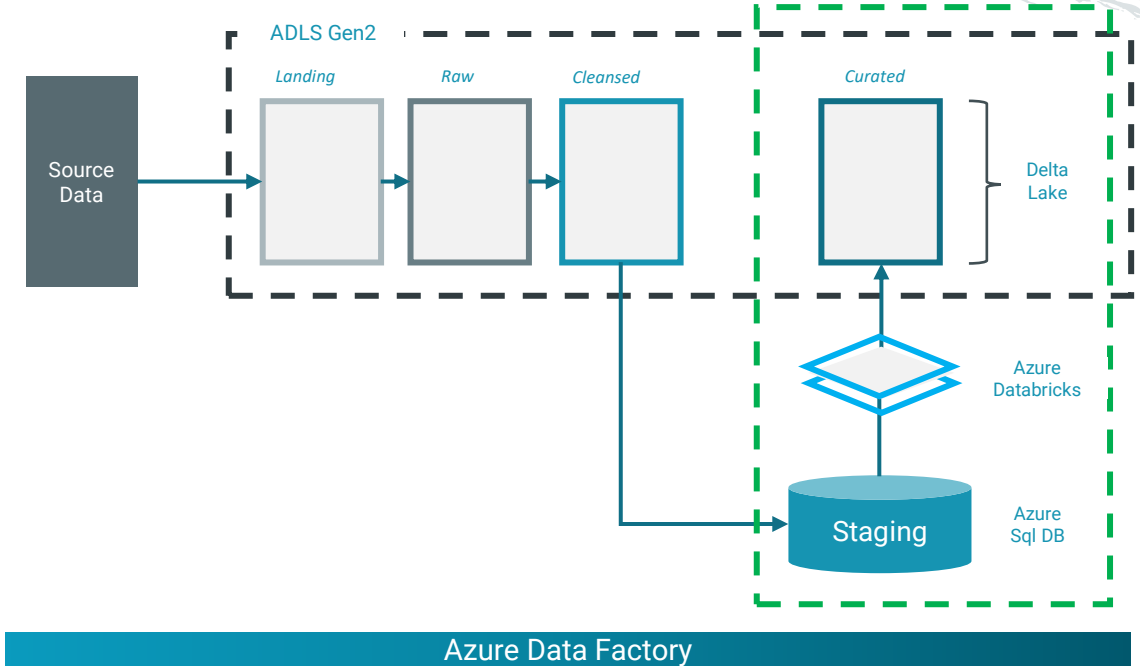


SECTION 3

DELTA LAKE IMPLEMENTATION

DELTA LAKE IMPLEMENTATION

Overview of Implementation

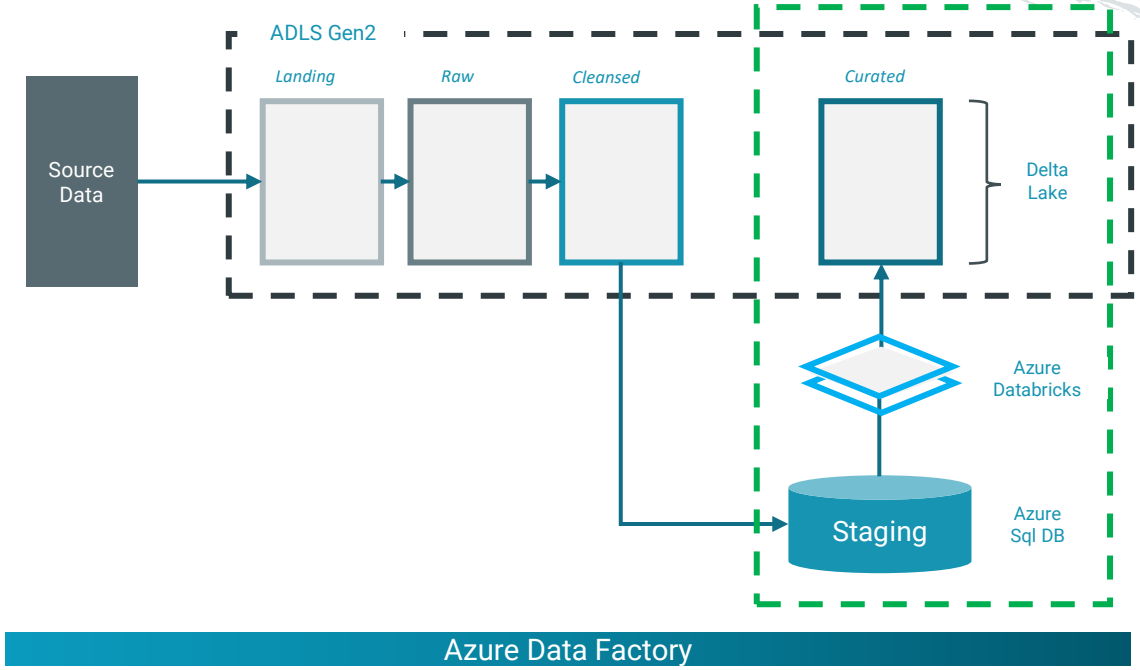


What we will Implement?

- *Connect to Azure Sql DB from Databricks*
- *Create dimStore as a delta table in Azure Databricks*
- *Load dimStore from staging using the Type 1 dimension load stored procedure*
- *Review the structures created in Databricks and ADLS Gen2*
- *Review some of the concepts of a Delta Lake Table*

DELTA LAKE IMPLEMENTATION

What is a Delta Lake?



Key Features of a Delta Lake

- *ACID Transactions on a Data Lake*
- *Time Travel*
- *Schema Enforcement*
- *Metadata handling*

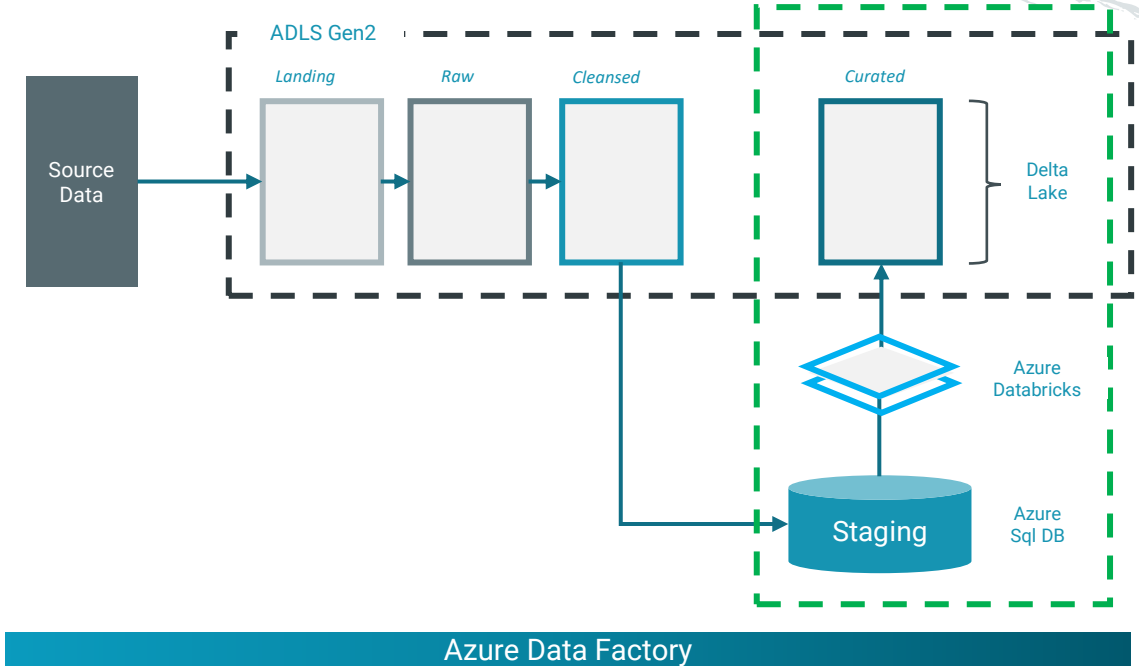
Delta Lake is an open-source storage layer that provides ACID transactions and metadata handling for data lakes.

SECTION 4

DATABRICKS NOTEBOOK

DATABRICKS NOTEBOOK

Executing Databricks Notebook Activity through Data Factory



What we will Implement?

- *Create a Databricks Linked Service*
- *Build a data factory pipeline with a Databricks Activity*
- *Execute the Databricks Activity from Data Factory*
- *Review the results*

MODULE SUMMARY

In this module we learnt

OVERVIEW

We got an overview of Azure Databricks and Delta Lake Concepts

We learnt about the benefits of a Delta Lake

INTEGRATION

We learnt how Azure Databricks and Azure Data Lake Gen 2 integrate

We learnt how Azure Databricks Azure Data Factory integrate

HANDS-ON

We learnt how to build a delta lake table from Azure Databricks

We learnt how to build the delta lake table by executing the Databricks Notebook from Data Factory

REFERENCES

Delta Lake Tutorial

Tutorial: Delta Lake - Azure Databricks | Microsoft Learn

Connect to Azure Data Lake Storage Gen2

Tutorial: Connect to Azure Data Lake Storage Gen2 - Azure Databricks | Microsoft Learn

Running an ETL Workload on Azure Databricks

Run your first ETL workload on Azure Databricks - Azure Databricks | Microsoft Learn