What is Delta Lake?

Learning Objectives

- Understand what Delta Lake is
- Delve into Delta Lake inner engines
- Introduce Delta Lake time travel
- Define Delta Lake advantages

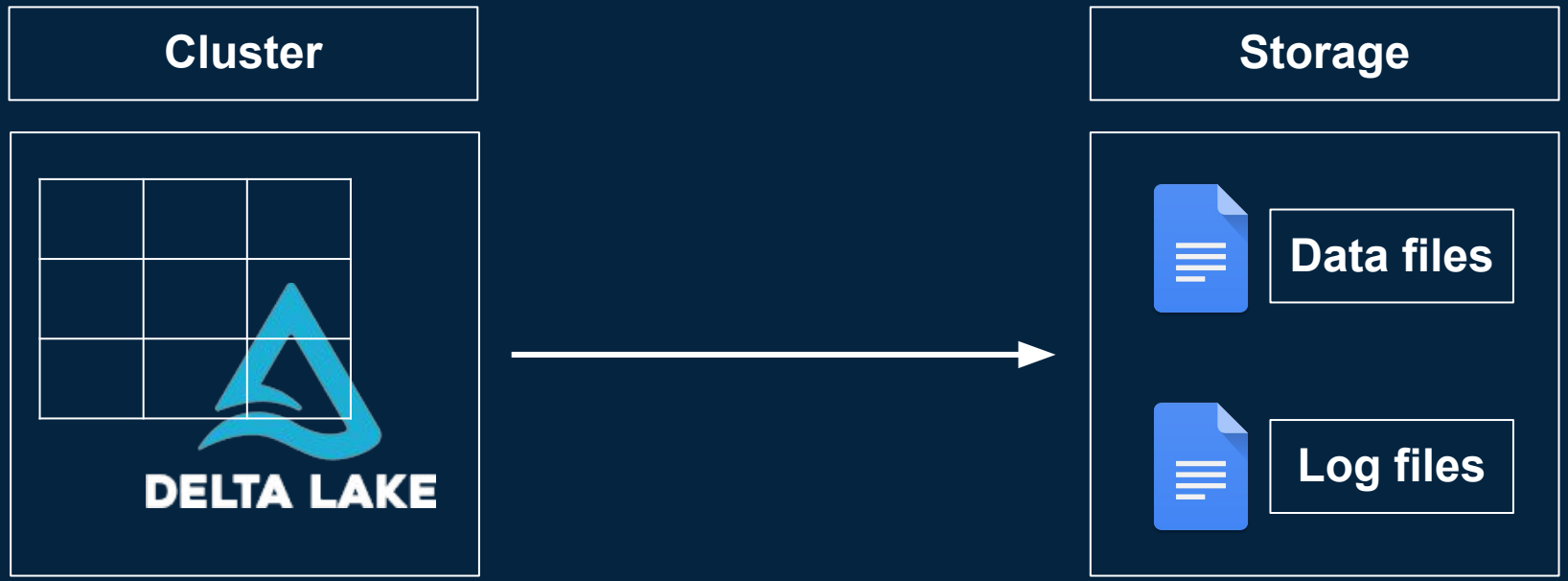
Delta Lake is an open-source storage layer designed to run on top of an existing data lake and improve its reliability, security, and performance.

Delta Lake Is

- Open Source
- Storage Layer
- Optimized for cloud object storage
- Data Lakehouse Foundation

Delta Lake Is Not

- Proprietary
- Storage Format
- Database Service
- Datawarehousing Service


Cluster
DELTA LAKE
Storage
Data files
Log files

Log files

- Record of every transaction performed on a Delta Lake table
- Contains operations names and data files impacted
- Stored in JSON format

1

Write DataFiles

- DataFile1.parquet
- DataFile2.parquet
- LogFile1.json

Read LogFile1
Read DataFiles

2

Update DataFile1

- DataFile2.parquet
- DataFile3.parquet
- LogFile1.json
- LogFile2.json

Read LogFile2
Read DataFiles

3

Write DataFile4

- DataFile2.parquet
- DataFile3.parquet
- DataFile4.parquet
- LogFile1.json
- LogFile2.json
- LogFile3.json

Read LogFile3
Read DataFiles

4

Writing Failure

- DataFile2.parquet
- DataFile3.parquet
- DataFile4.parquet
- LogFile1.json
- LogFile2.json
- LogFile3.json

Read LogFile3
Read DataFiles

Delta Lake Advantages

- Brings ACID transactions to a data lake
- Perfect for scaling metadata
- Keeps track of all changes to a table
- Built on standard parquet and JSON