

Delta Lake

Delta Lake

- ▶ Open-source storage framework that brings reliability to data lakes

Delta Lake is/is not

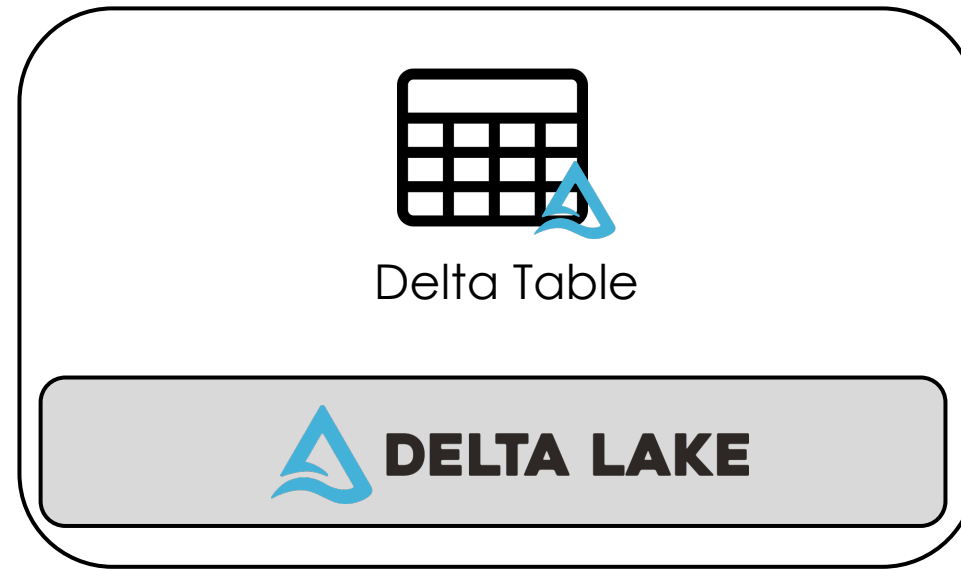
Is

- ▶ Open-source technology
- ▶ Storage framework/layer
- ▶ Enabling building Lakehouse

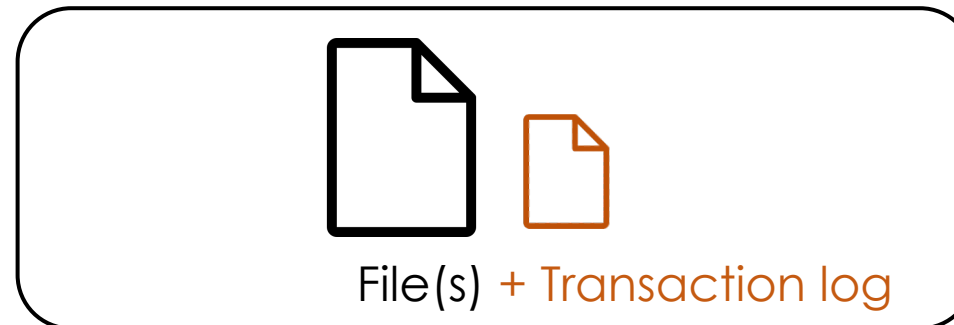
Is Not

- ▶ Proprietary technology
- ▶ Storage format/medium
- ▶ Data warehouse/Database service

Cluster



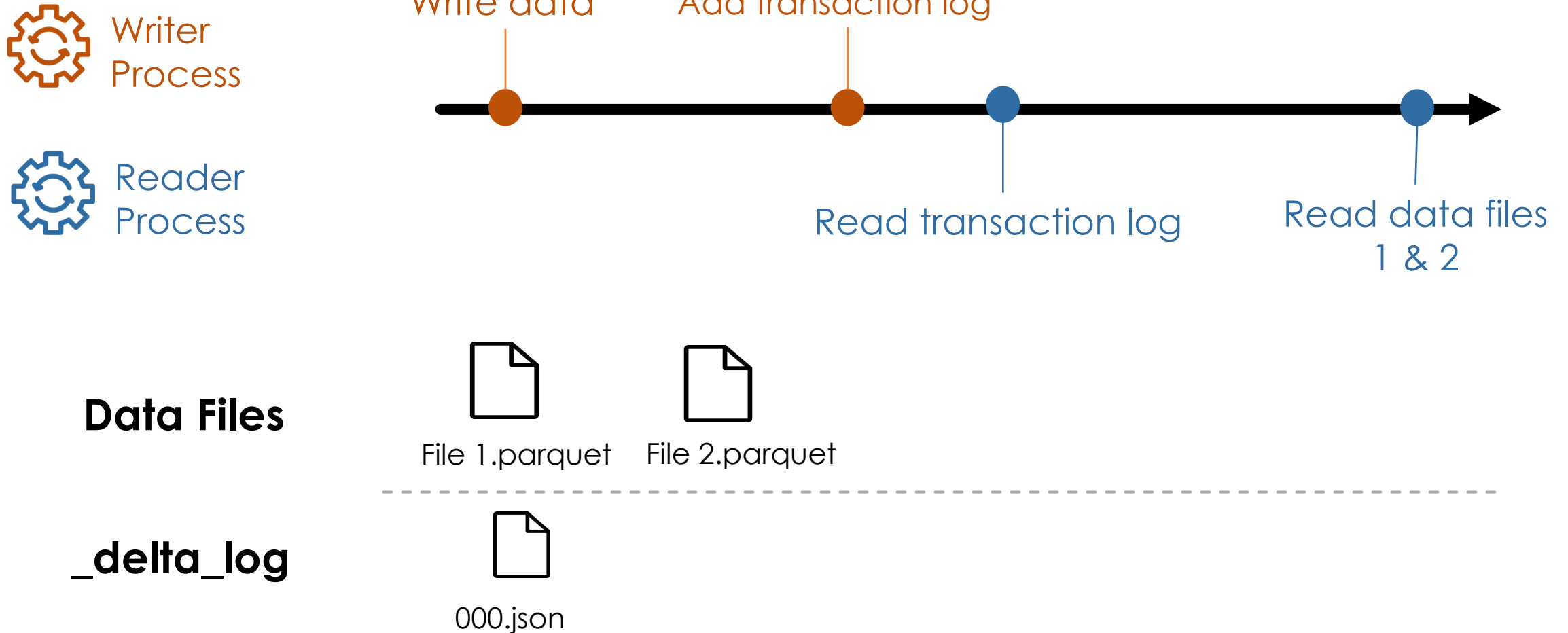
Storage



Transaction log (Delta log)

- ▶ Ordered records of every transaction performed on the table
- ▶ Single Source of Truth
- ▶ JSON file contains commit information:
 - ▶ Operation performed + Predicates used
 - ▶ data files affected (added/removed)

Writes/Reads



Updates



Writer
Process



Reader
Process

Update data

Add transaction log

Read transaction log

Read data files
2 & 3

copy & update



File 1.parquet



File 2.parquet



File 3.parquet

Data Files

_delta_log

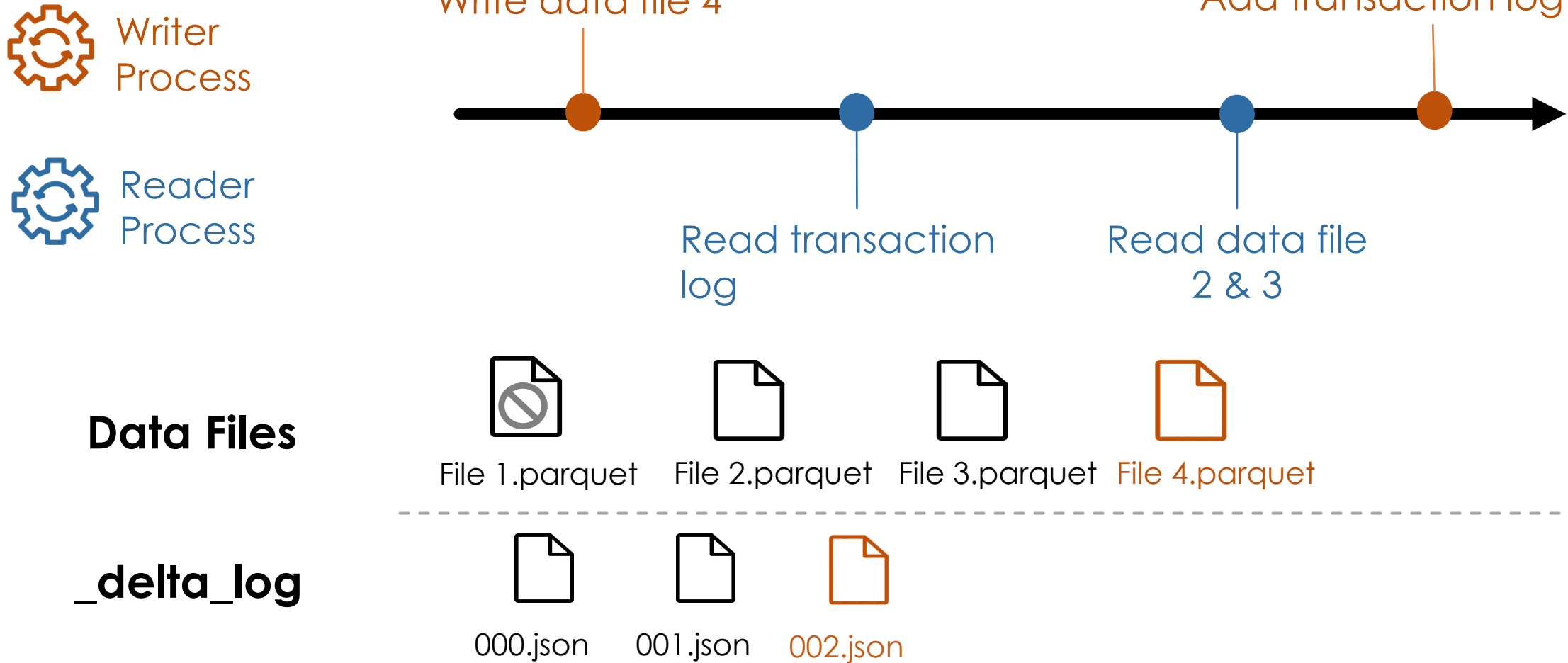


000.json

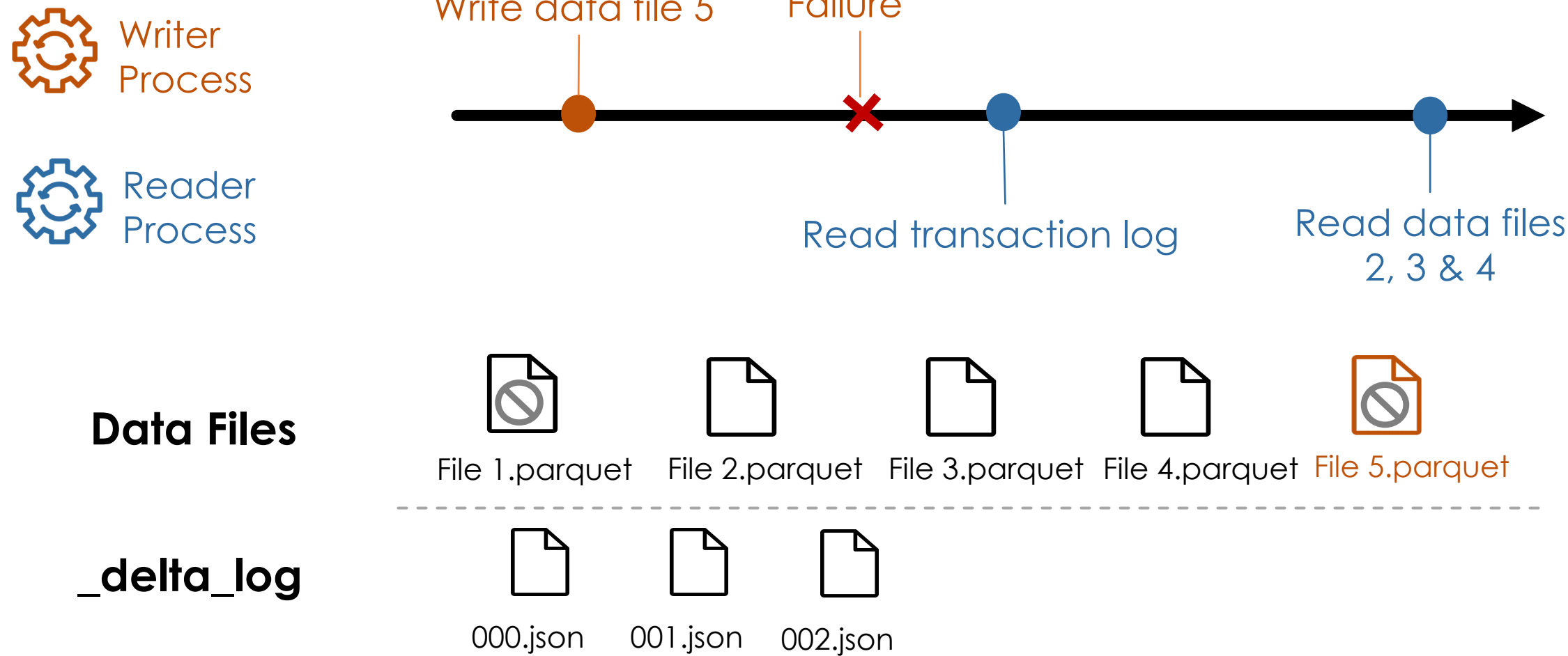


001.json

Simultaneous Writes/Reads



Failed Writes



Delta Lake Advantages

- ▶ Brings ACID transactions to object storage
- ▶ Handle scalable metadata
- ▶ Full audit trail of all changes
- ▶ Builds upon standard data formats: Parquet + Json