Why the Lakehouse?

Craig Porteous

Principal Consultant

https://craigporteous.com

@cporteous

https://github.com/cporteou

databricks PARTNER PROGRAM SOLUTIONS ARCHITECT CHAMPION

MVP Microsoft® Most Valuable Professional

Why the Lakehouse?

Craig Porteous

Principal Consultant

https://craigporteous.com

@cporteous

https://github.com/cporteou

databricks PARTNER PROGRAM SOLUTIONS ARCHITECT CHAMPION

MVP Microsoft® Most Valuable Professional

IS THIS A TIMESHARE?

aws

databricks

Microsoft

ORACLE

snowflake

CLOUDERA

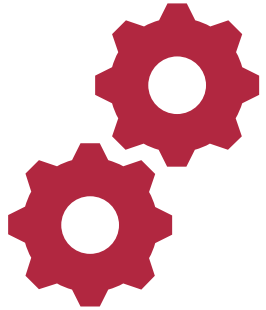
Google Cloud

What problem is it trying to solve?

PATH TO THE LAKEHOUSE

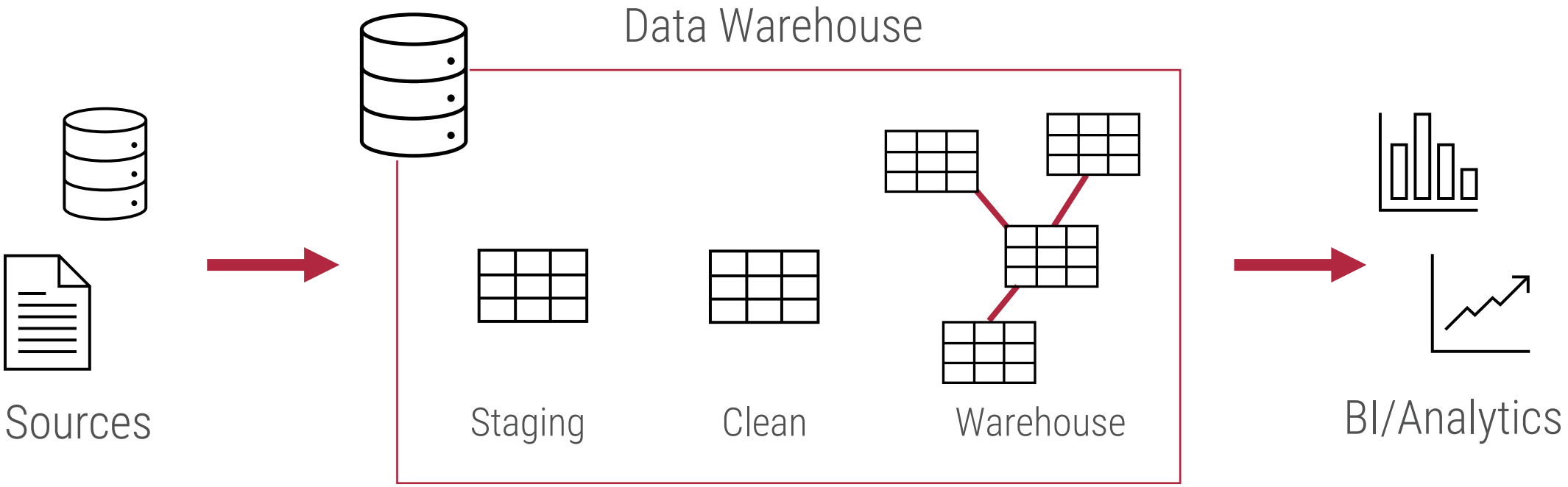
History

Lakehouse



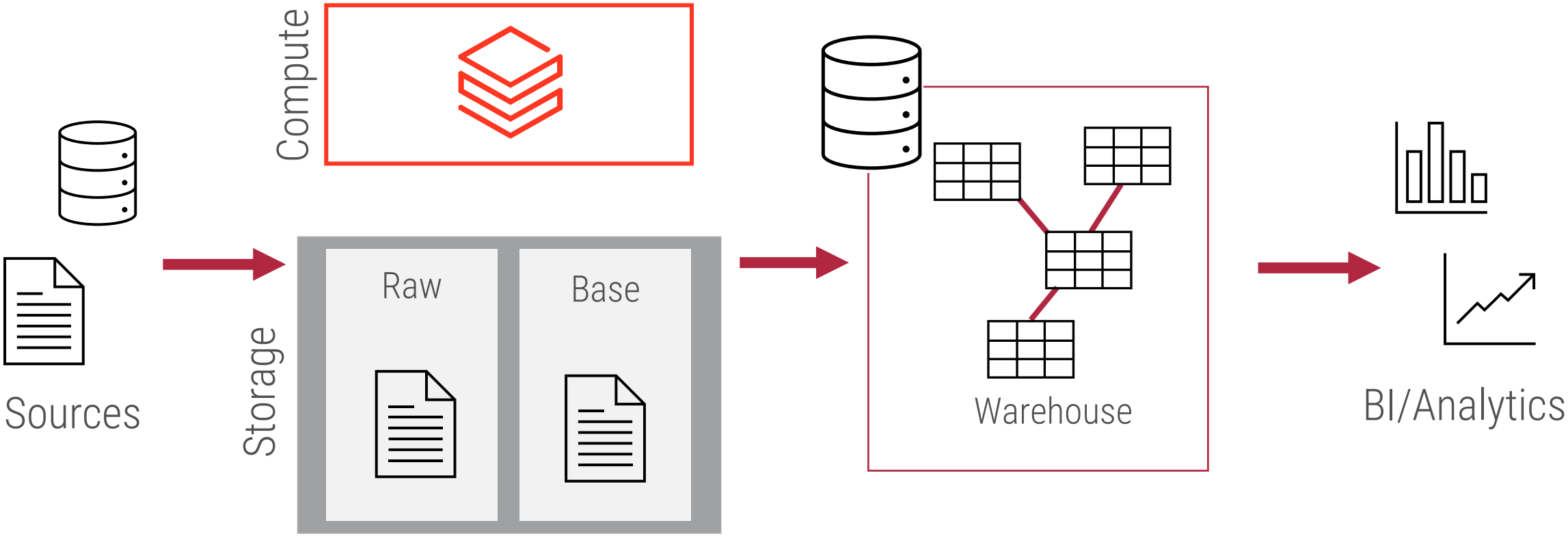
Technologies

THE DATA WAREHOUSE

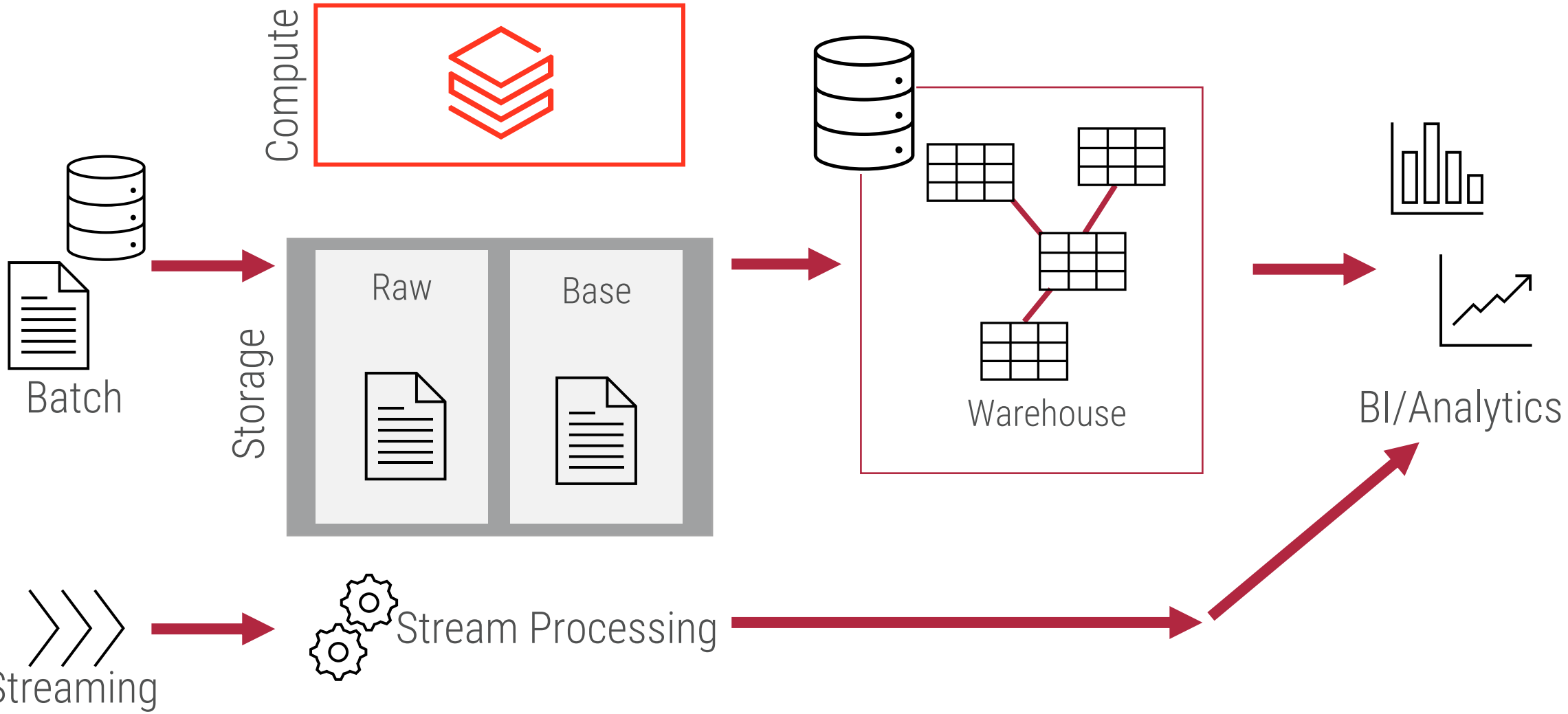


SEPARATION OF COMPUTE AND STORAGE

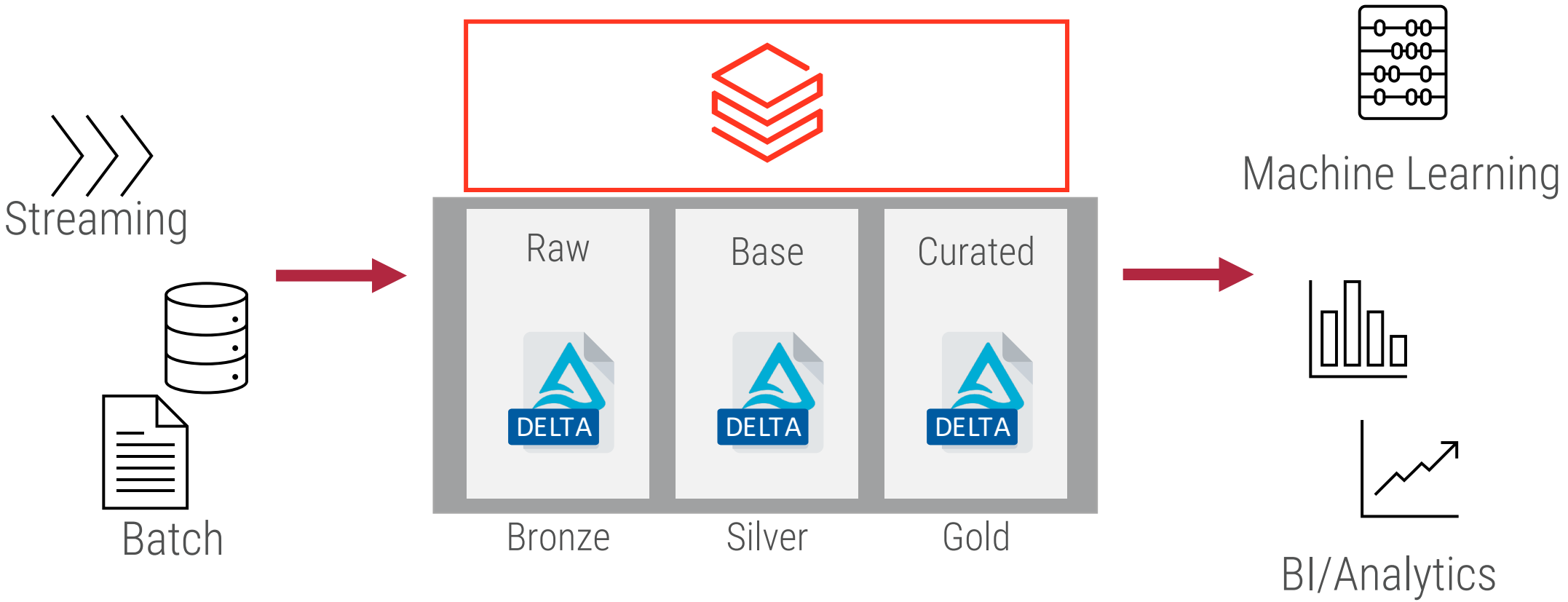
THE MODERN DATA WAREHOUSE



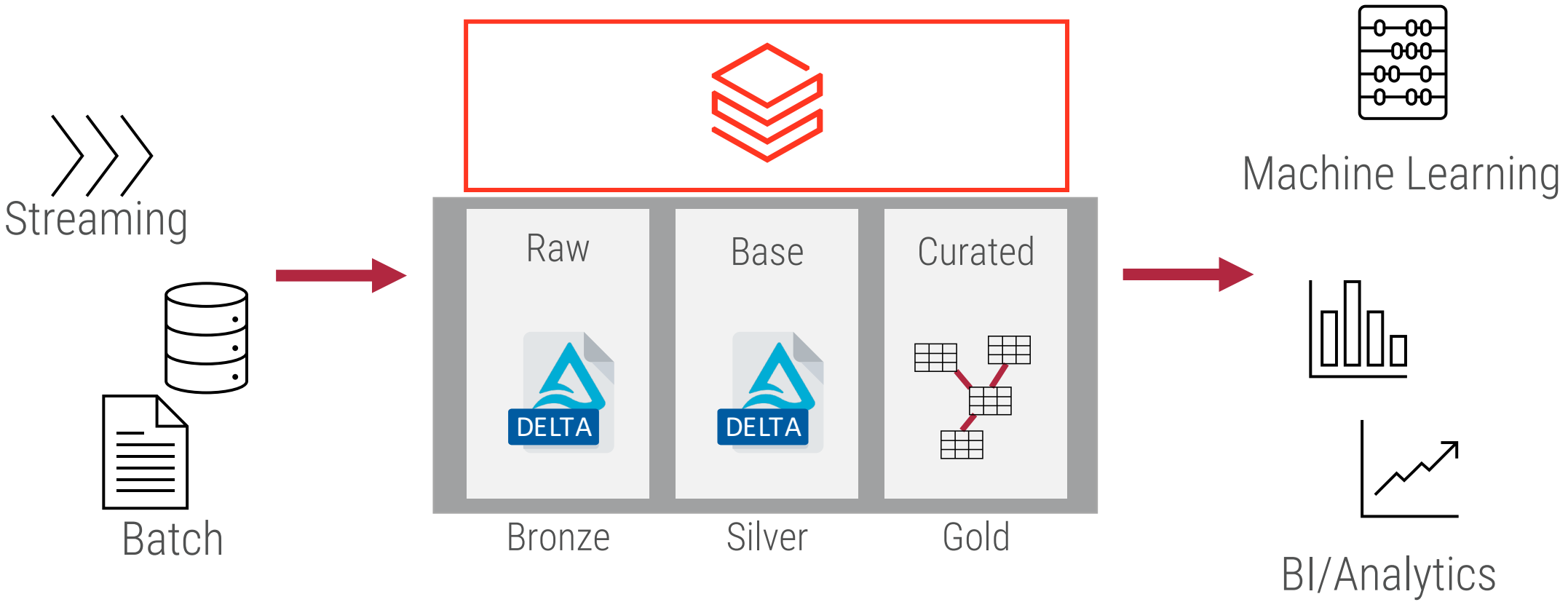
LAMBDA ARCHITECTURE



THE DATA LAKEHOUSE



THE DATA LAKEHOUSE



WHAT IS DELTA?

Delta Lake is an ***optimised, managed format*** *for organising & working with* ***Parquet*** *files*

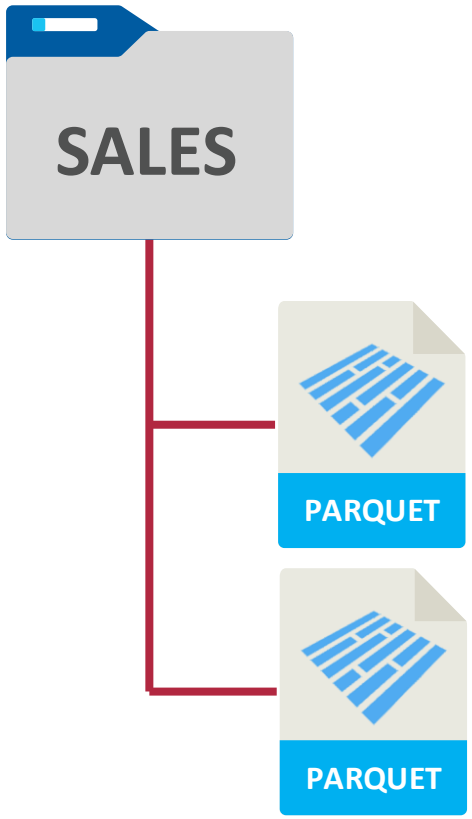
"It's Parquet, but better"

DELTA FEATURES

- OPEN SOURCE
- BASED ON PARQUET
- ACID TRANSACTIONS
- TIME TRAVEL
- SCHEMA EVOLUTION
- BATCH & STREAMING SUPPORT

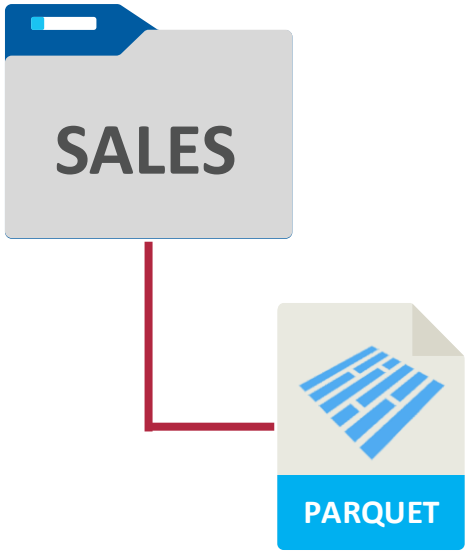
BEFORE DELTA

DELETE * FROM SALES WHERE Segment = 3



PART001.parquet

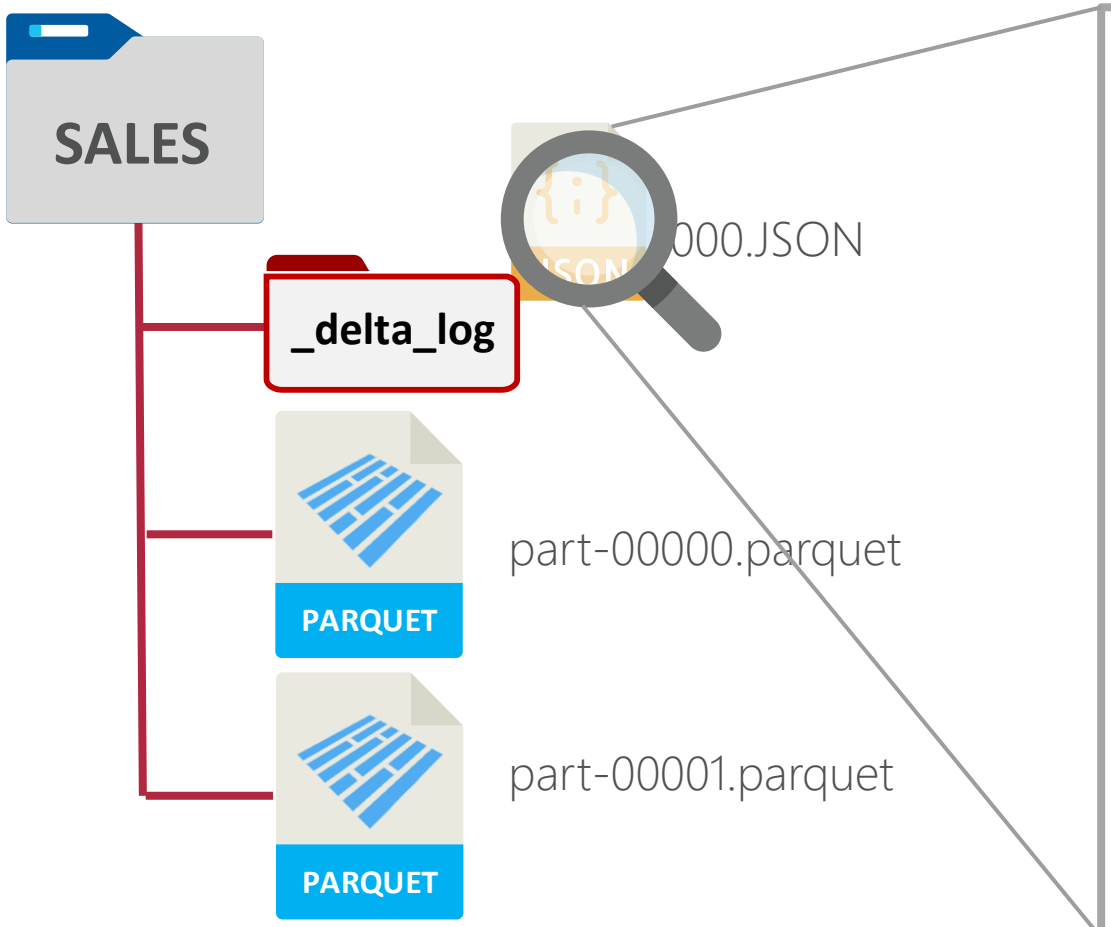
PART002.parquet



PART001.parquet

Only way to delete is to replace the existing files with a new file containing the non-deleted data

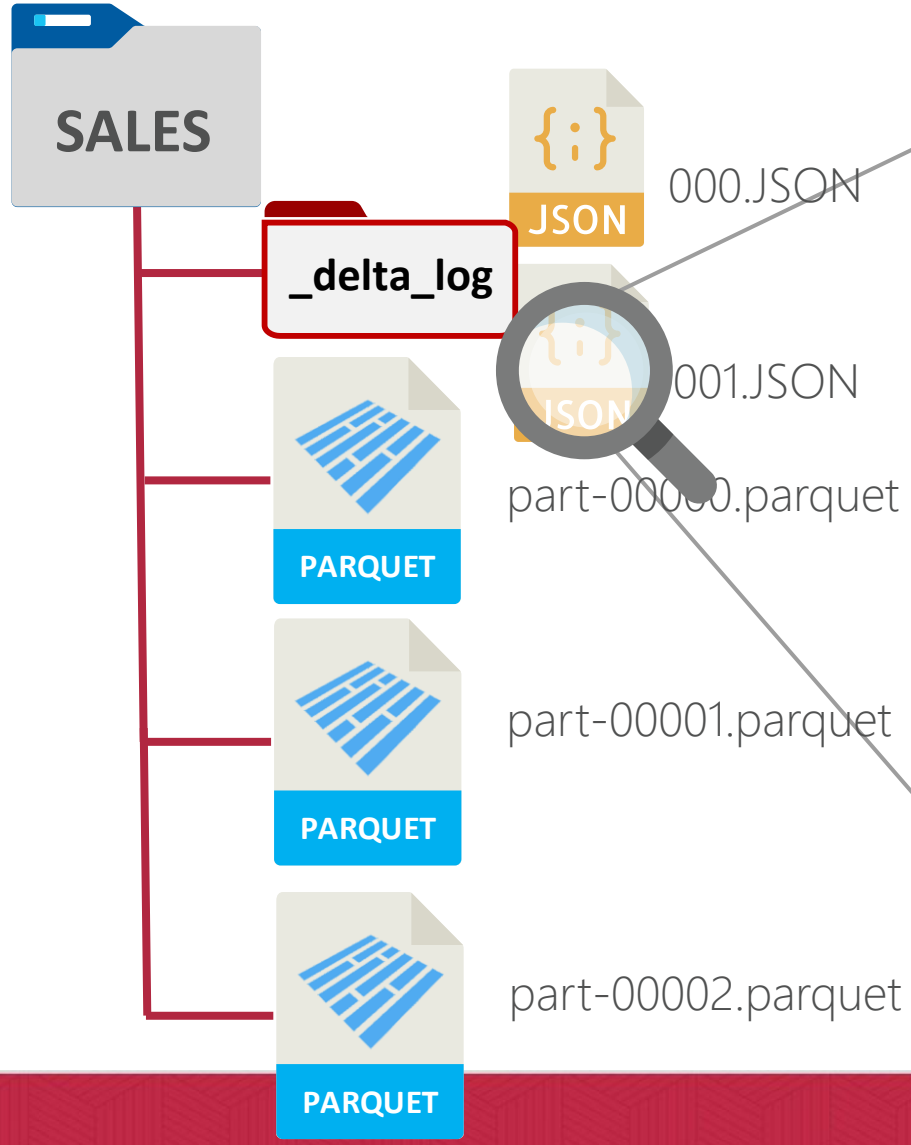
WITH DELTA




```
...
{
   "add": {
      "path": "part-00000.parquet",
      "partitionValues": {},
      "size": 255520,
      "modificationTime": 1572823237000,
      "dataChange": true,
      "stats":[…]},
 "add": {
      "path": "part-00001.parquet",
      "partitionValues": {},
      "size": 242520,
      "modificationTime": 1572823237000,
      "dataChange": true,
      "stats": […]}
}
```

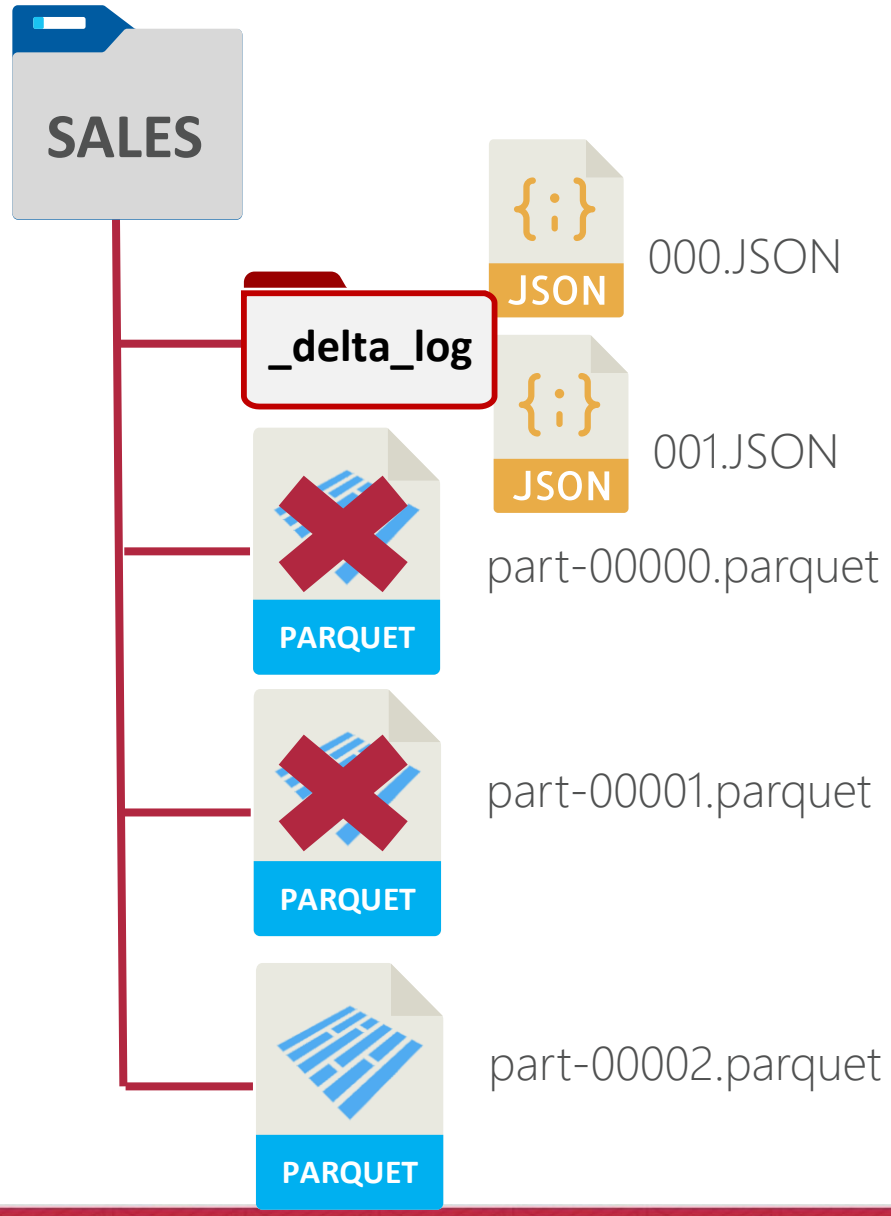
WITH DELTA

DELETE * FROM SALES WHERE Segment = 3

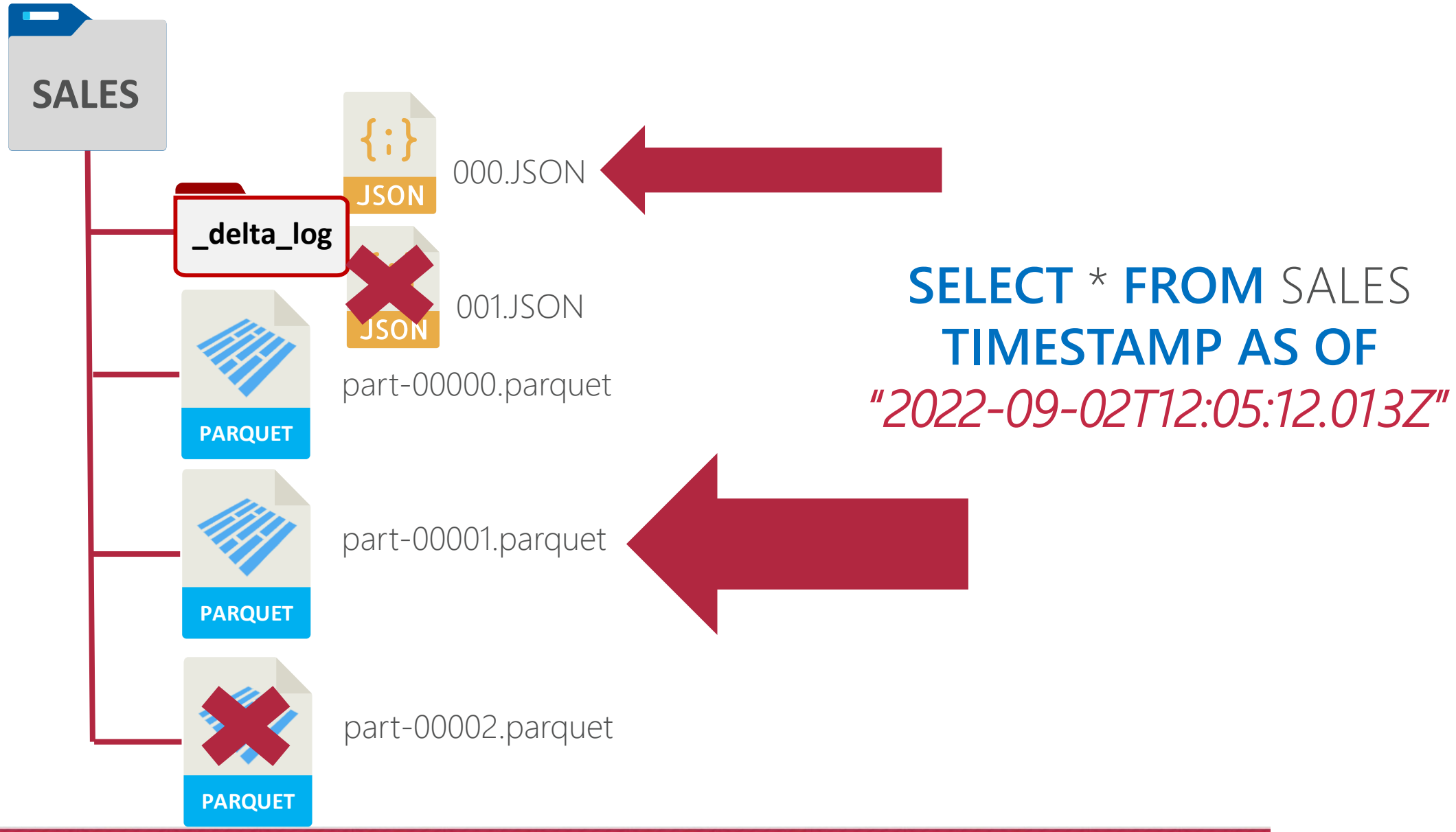



```
...
{
   "add": {
      "path": "part-00002.parquet",
      "partitionValues": {},
      "size": 255520,
      "modificationTime": 1572823237000,
      "dataChange": true,
      "stats":[…]},
 “remove": {
      "path": "part-00000.parquet",
            "modificationTime": 1572823237000,
      "dataChange": true},
 “remove": {
      "path": "part-00001.parquet",
            "modificationTime": 1572823237000,
      "dataChange": true}
}
```

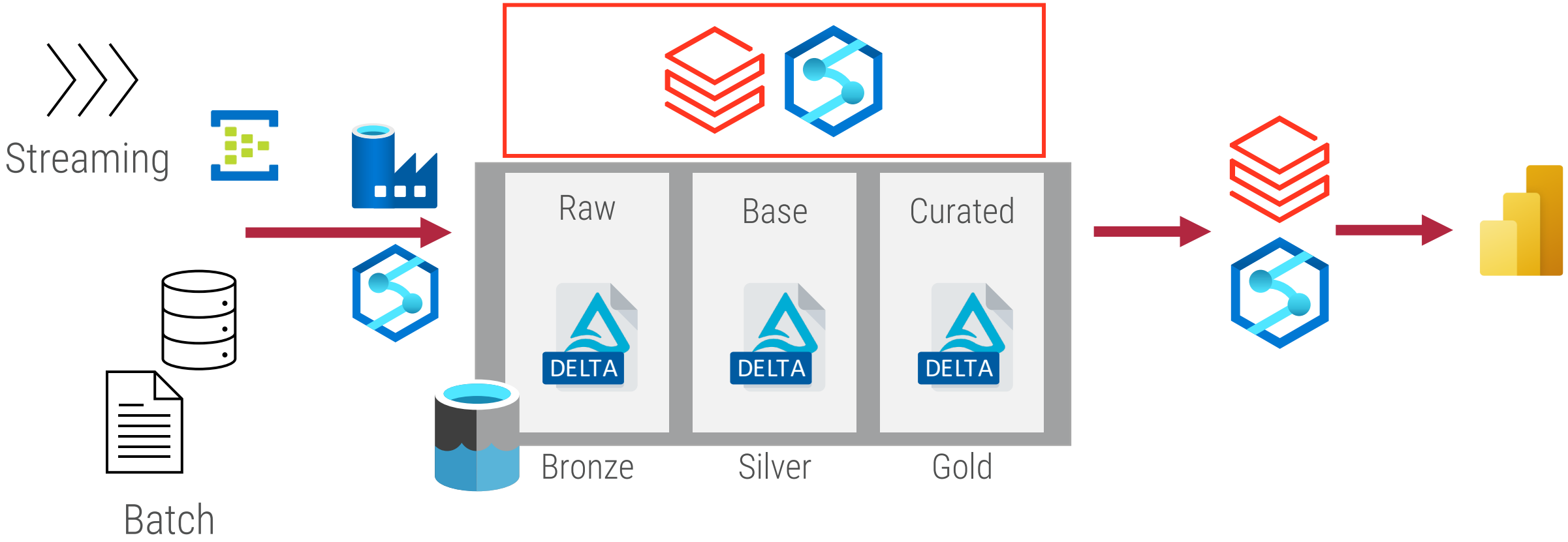
ADVANCING ANALYTICS



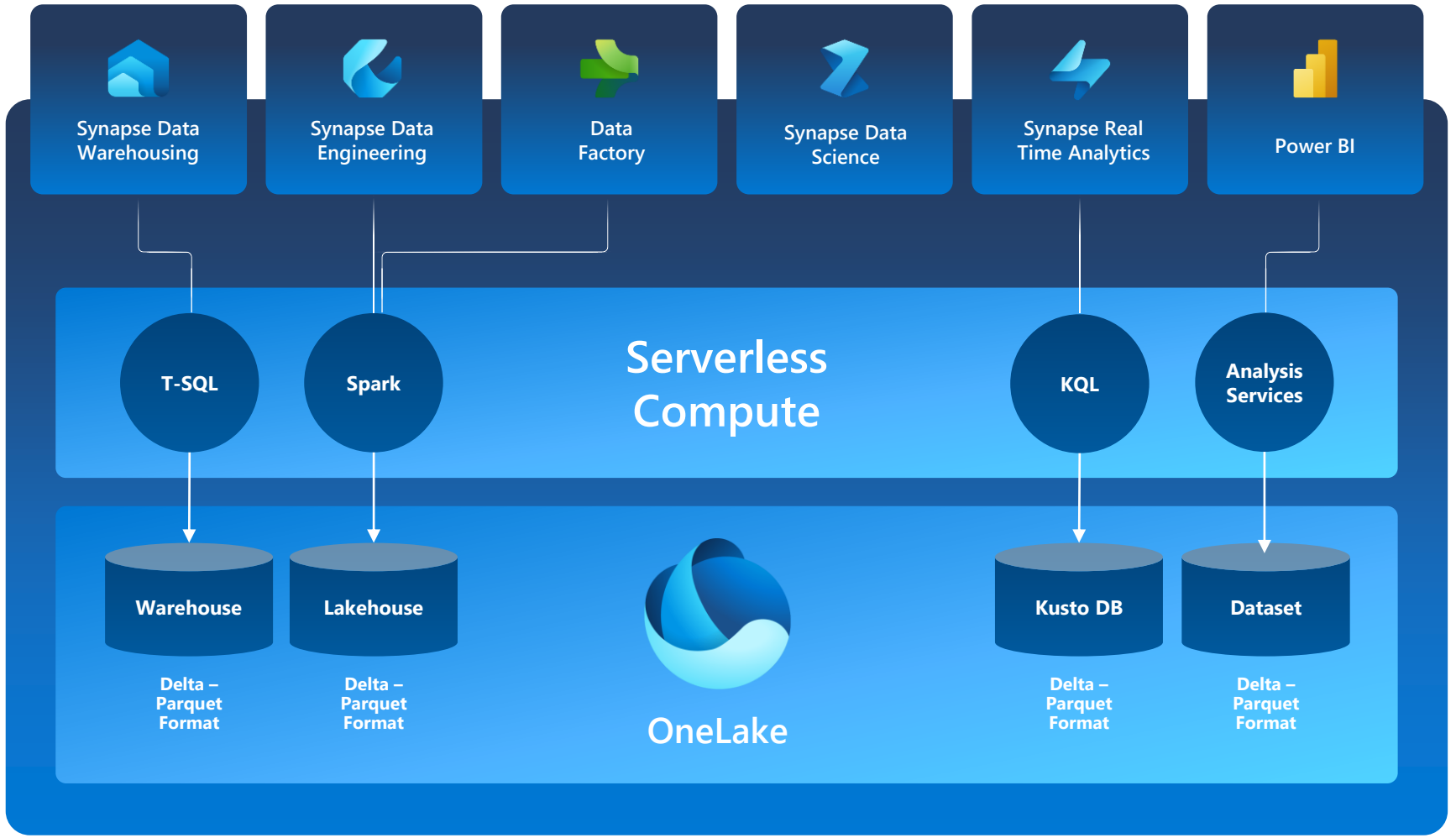
SELECT * **FROM** SALES


SALES
_delta_log
JSON
000.JSON
JSON
001.JSON
PARQUET
part-00000.parquet
PARQUET
part-00001.parquet
PARQUET
part-00002.parquet
SELECT * FROM SALES
TIMESTAMP AS OF
"2022-09-02T12:05:12.013Z"

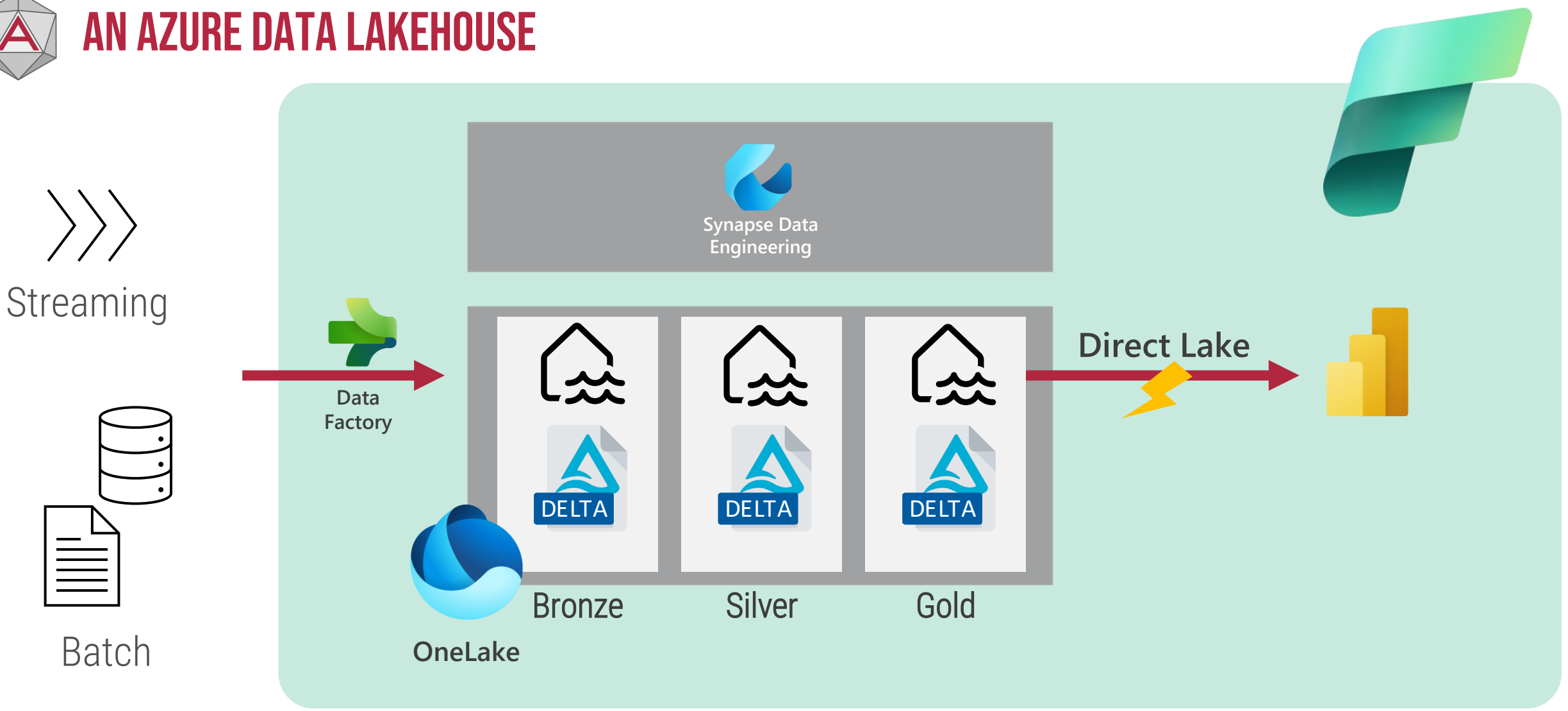
AN AZURE DATA LAKEHOUSE



A FABRIC DATA LAKEHOUSE



AN AZURE DATA LAKEHOUSE



COMMON LAKEHOUSE TECHNOLOGIES

COMPUTE

STORAGE

SERVE

SKILLS GAP

- SQL
- PYTHON
- SCALA

WHAT IS A DATA LAKEHOUSE

DATA WAREHOUSE

- ACID TRANSACTIONS
- GOVERNANCE

+

DATA LAKE STORAGE

- CHEAP
- FLEXIBLE

=

DATA LAKEHOUSE

- ANALYTICS
- BUSINESS INTELLIGENCE
- MACHINE LEARNING
- BIG DATA

THANK YOU

https://craigporteous.com

@cporteous

https://github.com/cporteou

Let's thank our..

GOLD SPONSORS

quorum

SIOS

solarwinds

TIMEXTENDER

SILVER SPONSORS

ADVANCING ANALYTICS

redgate

POWER BI SENTINEL
Governance, Disaster Recovery and Auditing for Power BI

BRONZE SPONSOR

GE
GETHYNELLIS.COM

Community Partner Support

DATA SATURDAYS

ORGANISERS

Satya Jayanty

Enterprise Architect specialised in Data Platforms and Cloud Computing, Speaker, Mentor, and Community Organiser.

Tom Sykes

Senior Cloud Data Architect with many years of experience in the Cloud and Database industry, and Certified Trainer.

Volunteers:

- Syama Rudra
- Aditya Yembuluru
- Venkatesh Gaddam
- Samrat Ravuri
- Saikrishna Prathipati
- Saimohan Inala
- Mark Adamson

@SQLScot