Building a Data Lakehouse with Delta Lake and Spark

Jonathan Neo

Senior Data Engineer, Cuusoo

Who is Cuusoo

Reimagine data without the limits of the status quo, make it happen with databricks

Roadmap, design and strategy

We help you get started linking your overall strategy to design and use cases

Databricks deployment

Help you re-imagine your data environment to be powered by databricks and not constrained by the messy status quo

Databricks tune up

We assess your Databricks platform and usage within your business context and recommend and implement ways to reduce consumption, strengthen security and optimise performance, so you get more from your investment.

Custom use cases

We flesh that out into well-defined use cases and then make them happen on Databricks. It can be any combination of data engineering, analytics, data science and machine learning.

Rapid value accelerators

Helping you to implement the databricks accelerators that already exist across verticals including, but not limited to, financial services, healthcare, retail and consumer goods and manufacturing.

Machine learning at scale

Standing up your scaled machine learning environment and helping you team to implement the best practices and systems to get the most from databricks

Cuusoo

Koo-Soh

Cuusoo, pronounced 'koo-soh', means imagination, vision and clean-slate thinking. Cuusoo will focus on helping businesses imagine data and its applications using a first principles approach.

CUUSOO

We’re are member of the Mantel group

We’re an Australian-owned, technology-led consulting with capabilities from strategy to managed services

Established in November 2017, we’re a dynamic and growing business currently comprised of seven brands. We’ve been recognised in the AFR’s 2020 fastest growing new companies and LinkedIn’s Top Australian Startups. Our plan is to go IPO in 2023. We have hubs in Melbourne, Sydney, Brisbane, Perth, Auckland, Queenstown and Magnetic Island, supporting a team of 400 that will grow to 550 over the next year.

Advisory
CTO Advisory
Security Advisory
Design Advisory

Design
UX and CX Design
Service Design
Customer Research

Data / AI / ML
Data Engineering
AI/Machine Learning
Data Science

Engineering
Software Engineering (Web, Mobile, API)
Test Automation

Cloud
Platform Engineering (AWS, Google, Azure)
Modern Workplace & Devices (G-suite, AWS EUC)

Delivery & Method
Method Coaching
Delivery Leadership
Product Ownership
Business Analysis

Managed Services
Customer Software & Data Security & Access Management Operating System
Network
Cloud Services
Hardware & Global Infrastructure

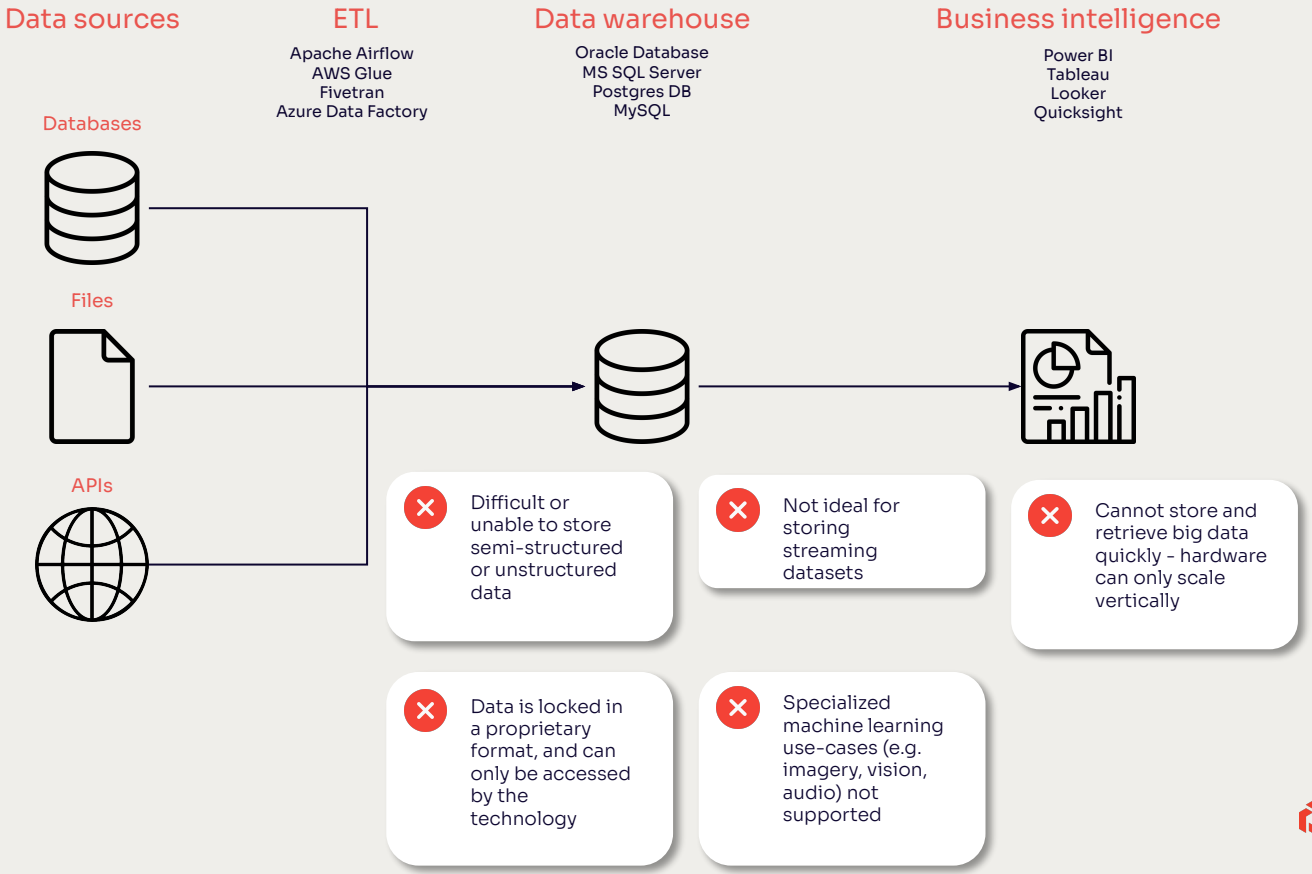
Our Customers

Diverse experience across a range of industries

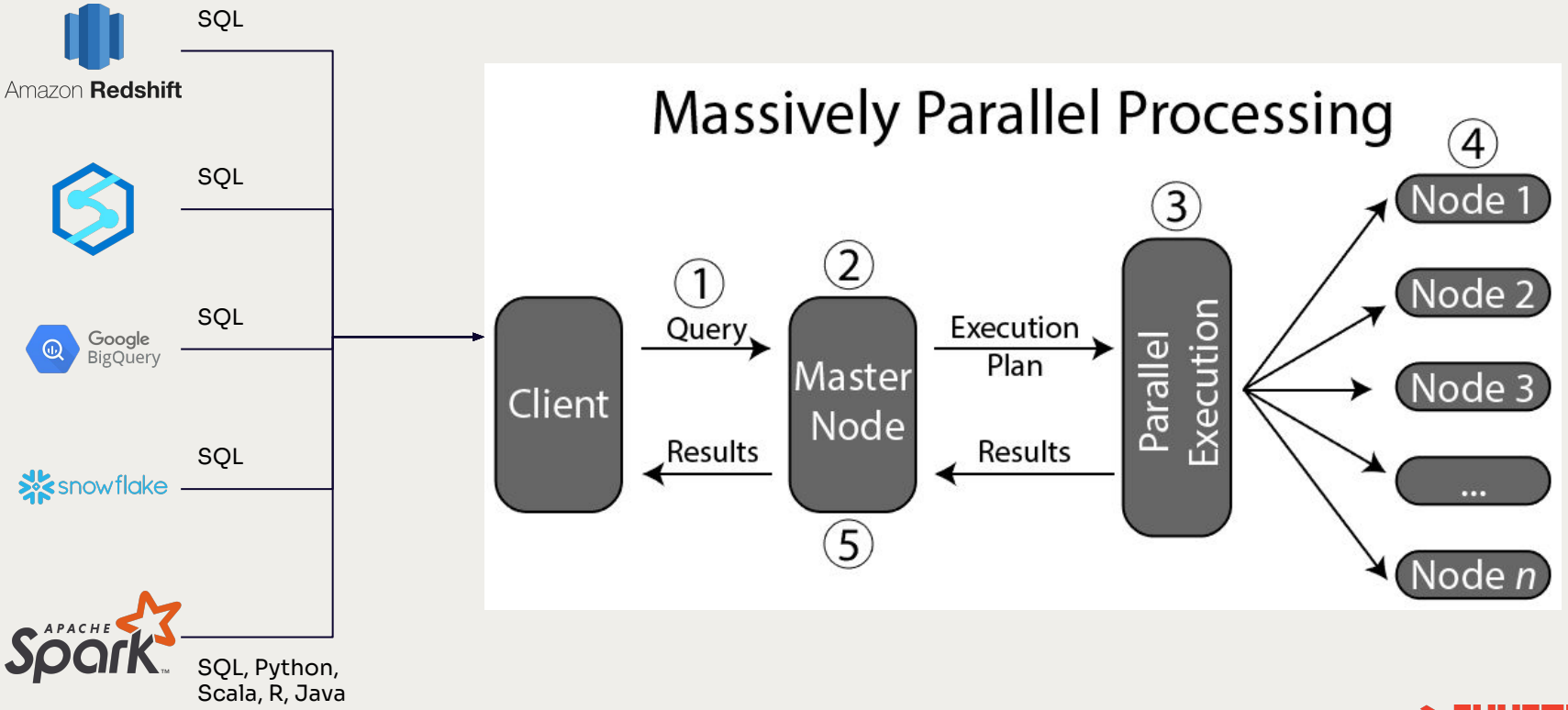
Agenda

1. Introduction to the Lakehouse architecture (10 mins)
2. Deep dive into the backbone of Lakehouses – the Delta Lake format (5 mins)
3. Demo - Building a data lakehouse with Delta Lake and Spark (30 mins)
4. Q&A (5 mins or less)

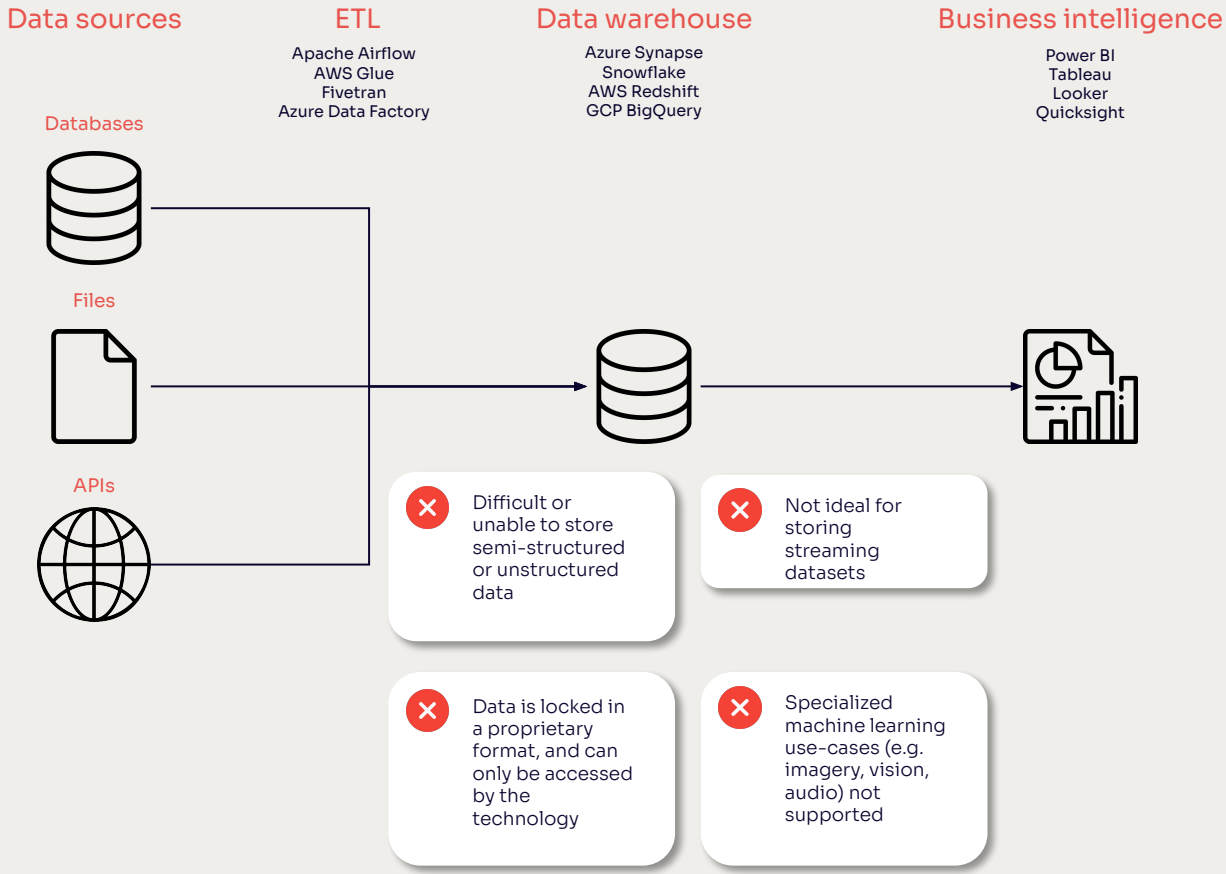
Traditional data warehouse architecture (1980s - 2010s)



All hail – Massively Parallel Processing (MPP) technologies

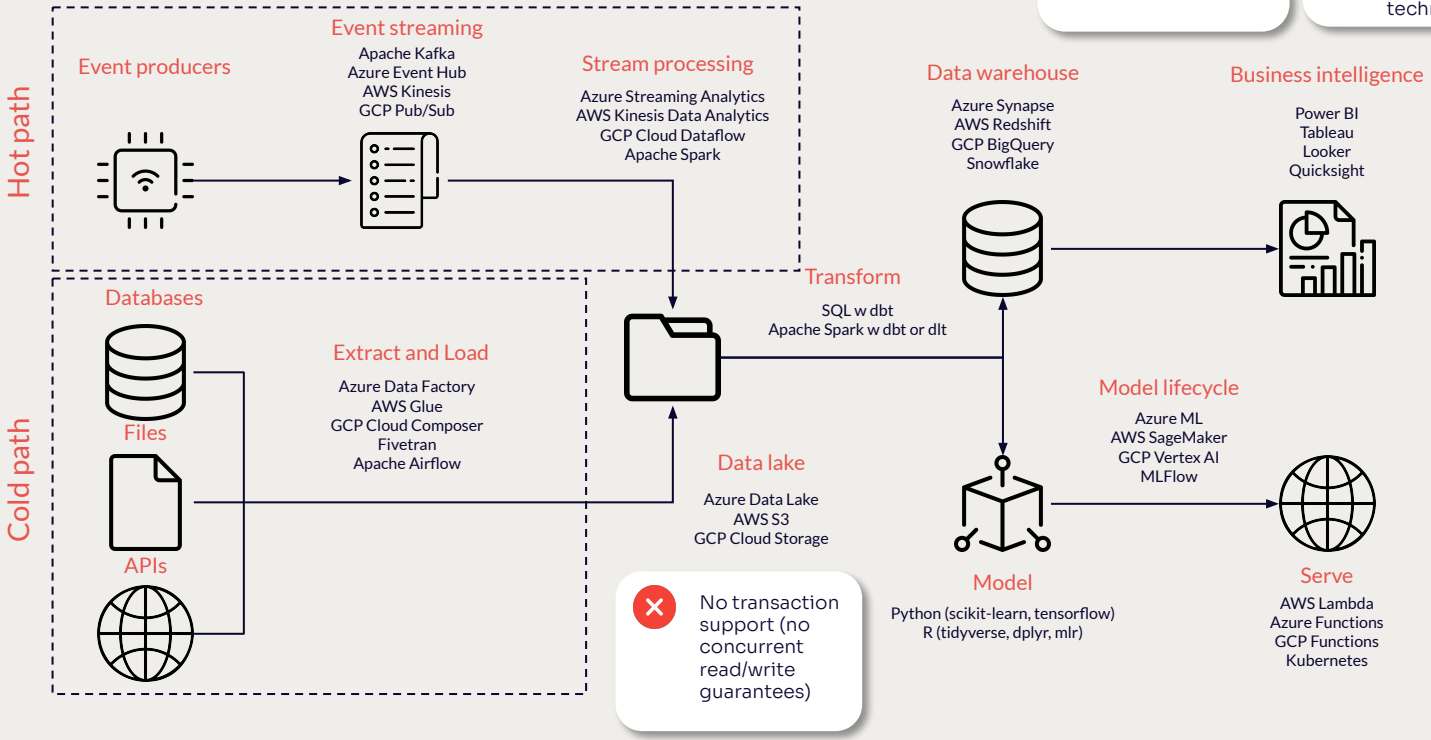


Big data warehouse architecture (2010s)



Modern data lake architecture (2010s)

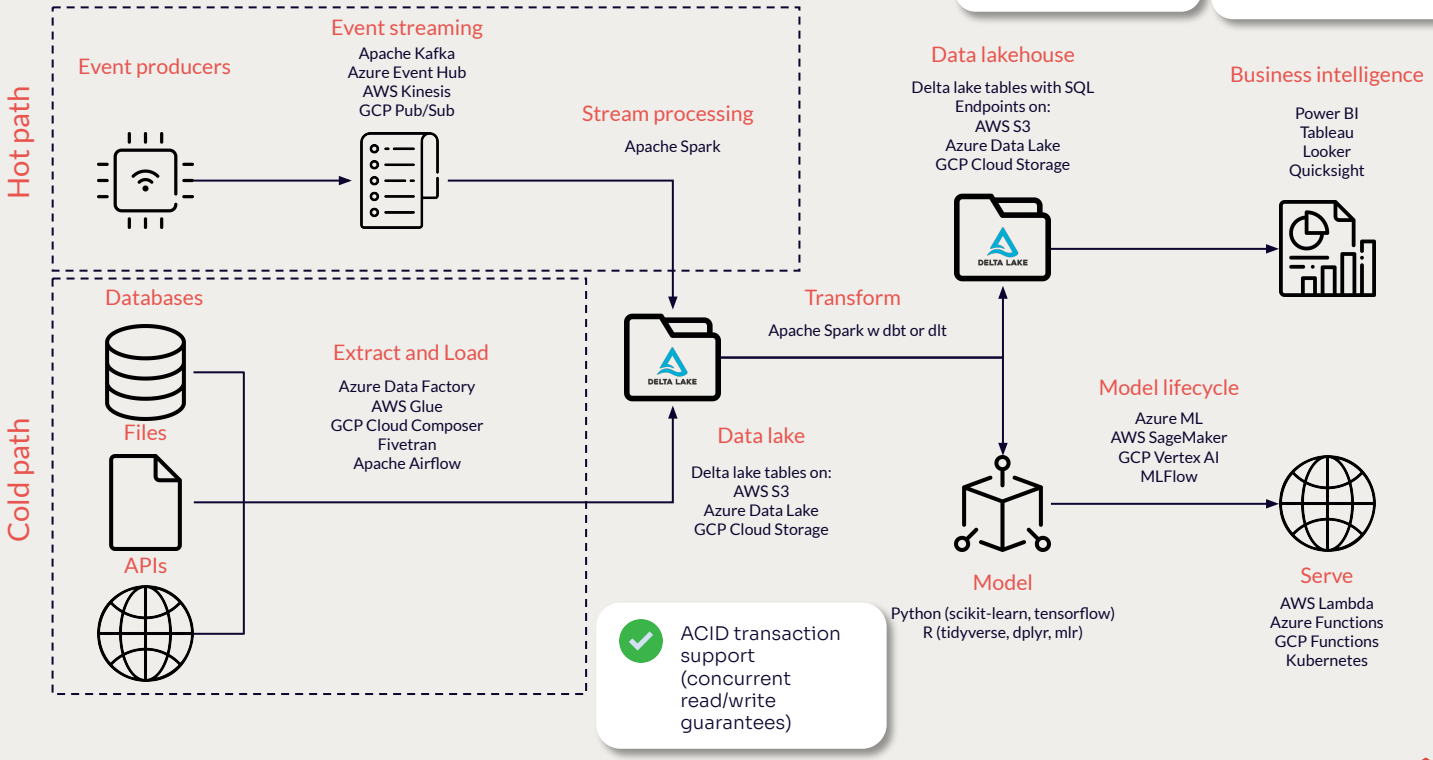
- Duplication of data – one copy in the lake, one copy in the warehouse
- Data is locked in a proprietary format, and can only be accessed by the technology



- No transaction support (no concurrent read/write guarantees)

Data lakehouse architecture (2020s)

- No duplication of data – single source of truth
- Open format – no lock-in to technology provider
- Storage and compute de-coupled



- ACID transaction support (concurrent read/write guarantees)

EUUSOO

Delta lake – key features

ACID guarantees

Delta Lake ensures that all data changes written to storage are committed for durability and made visible to readers atomically. In other words, no more partial or corrupted files.

Schema enforcement and schema evolution

Delta Lake automatically prevents the insertion of data with an incorrect schema, and when needed, it allows the table schema to be explicitly and safely evolved to accommodate ever-change data.

Scalable data and metadata handling

Since Delta Lake is built on data lakes, all reads and writes using Spark or other distributed processing engines are inherently scalable to petabyte-scale.

Support for deletes updates, and merge

Most distributed processing frameworks do not support atomic data modification operations on data lakes. Delta Lake supports merge, update, and delete operations to enable complex use cases such as CDC and SCD operations.

Audit History and Time travel

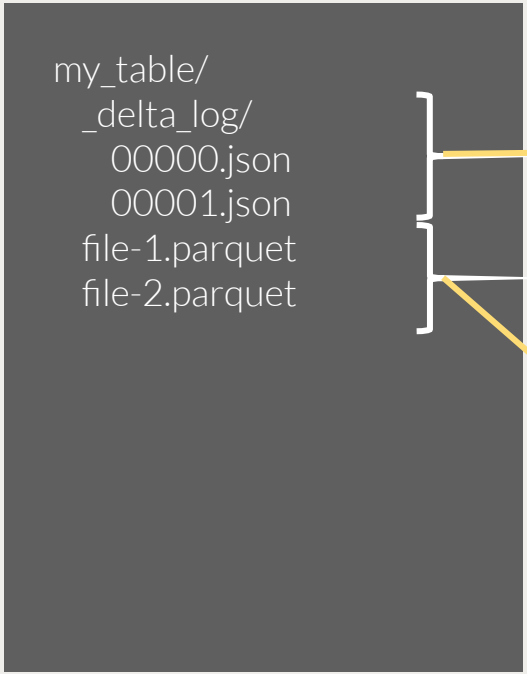
The Delta Lake transaction log records details about every change made to data providing a full audit trail of the changes. These data snapshots enable developers to access and revert to earlier versions of data for audits, rollbacks, or to reproduce experiments.

Streaming and batch unification

A Delta Lake table has the ability to work both in batch and as a streaming source and sink.

Taken from: Delta Lake: The Definitive Guide Modern Data Lakehouse Architectures with Delta Lake

Delta lake – key components



Transaction log

The transaction log makes it possible for multiple readers and writers on a given Delta table to work at the same time. It serves as a **single source of truth** - a central repository that tracks all changes that users make to a table.

Delta files

Delta Lake uses Parquet files stored in object storage. Apache Parquet is designed for efficient and performant flat columnar storage of data, compared to row based files like CSV or TSV files.

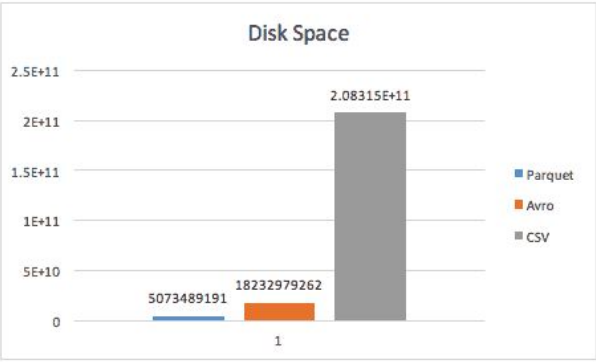
Delta lake – key components

```
my_table/
  _delta_log/
    00000.json
    00001.json
  file-1.parquet
  file-2.parquet
```

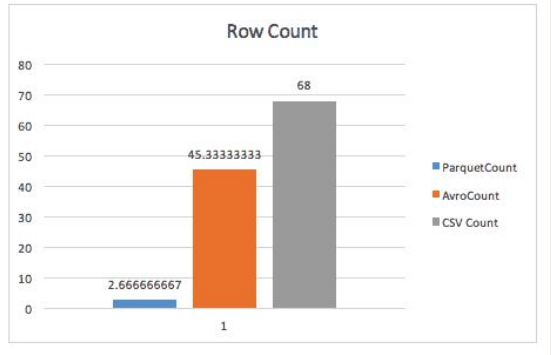
Delta files

Delta Lake uses Parquet files stored in object storage. Apache Parquet is designed for efficient and performant flat columnar storage of data, compared to row based files like CSV or TSV files.

Better compression (97.56% smaller than CSV)



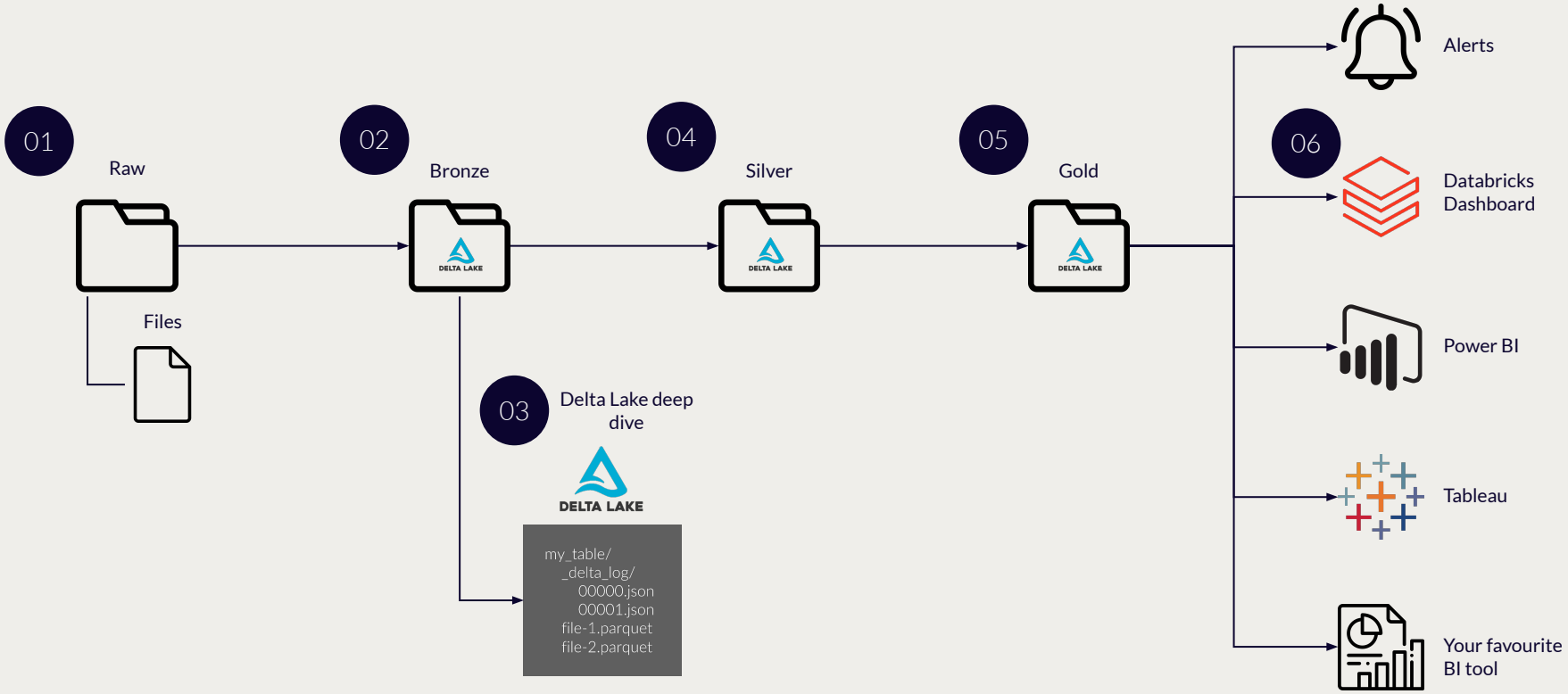
Better performance (95.59% faster than CSV)



https://blog.cloudera.com/benchmarking-apache-parquet-the-allstate-experience

EUVSOO

End-to-end demo



EUUSOO

Thank you