Data Lakes on ACID

JavaZone 2022

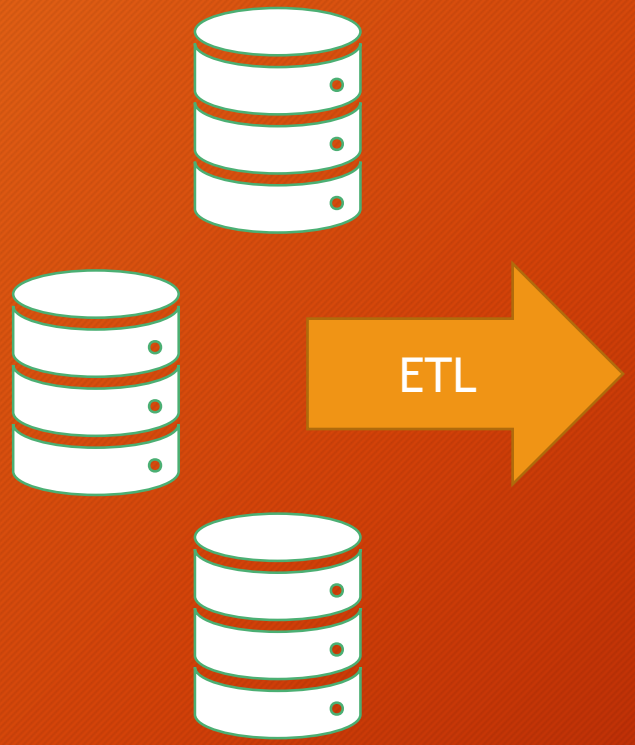
Gareth Western

@gareth
gareth.western@bouvet.no

What do we mean by “data lake”?

Data warehouses handle OLAP queries

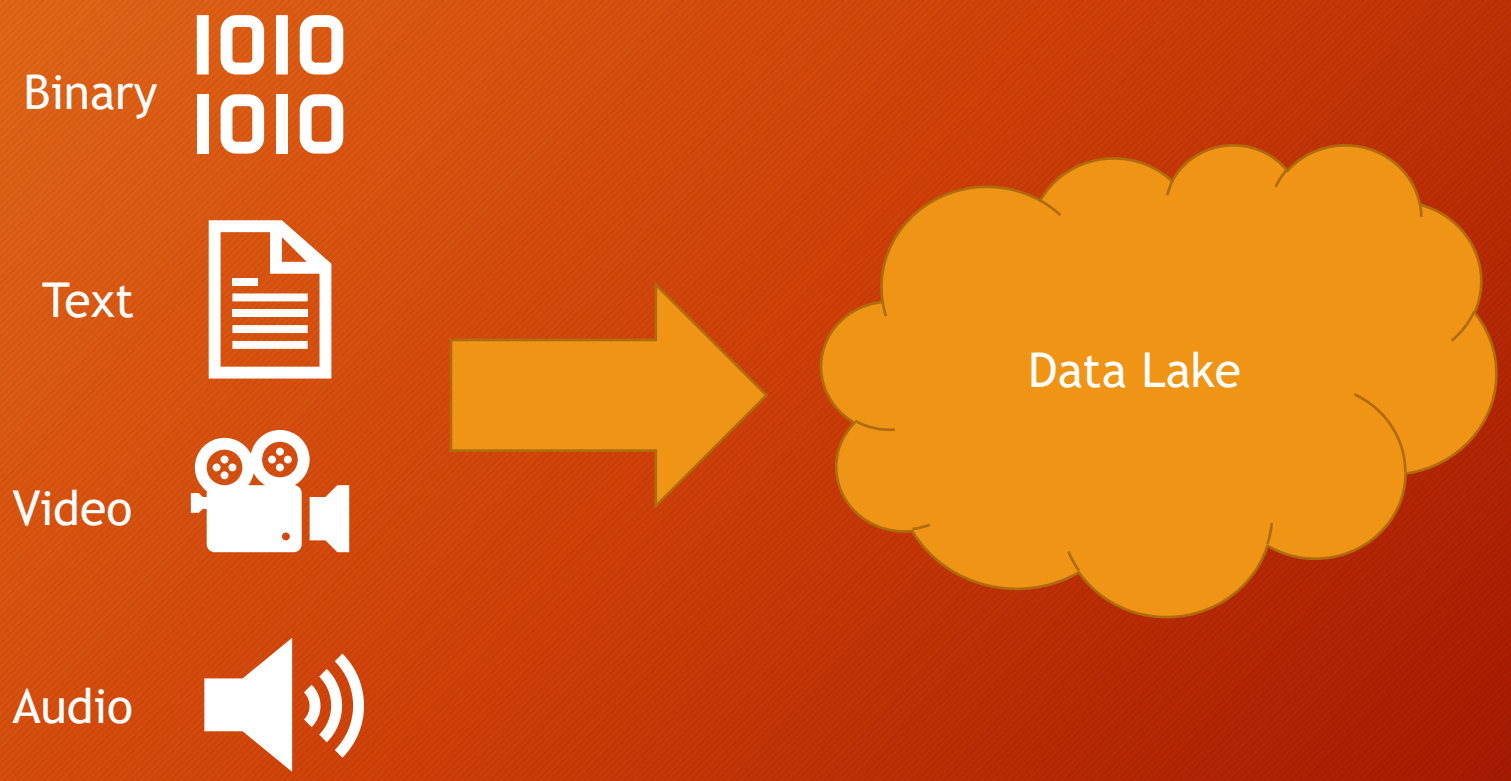
Online Transaction **Processing (OLTP)**



Data Warehouse

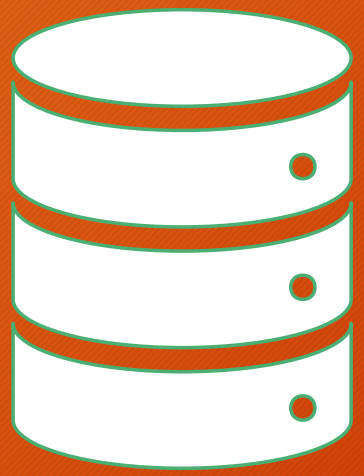
Online Analytical Processing (OLAP)

Data lake stores semi/un-structured data



(This data is usually text-based)

Data lakehouse is a combination of the two



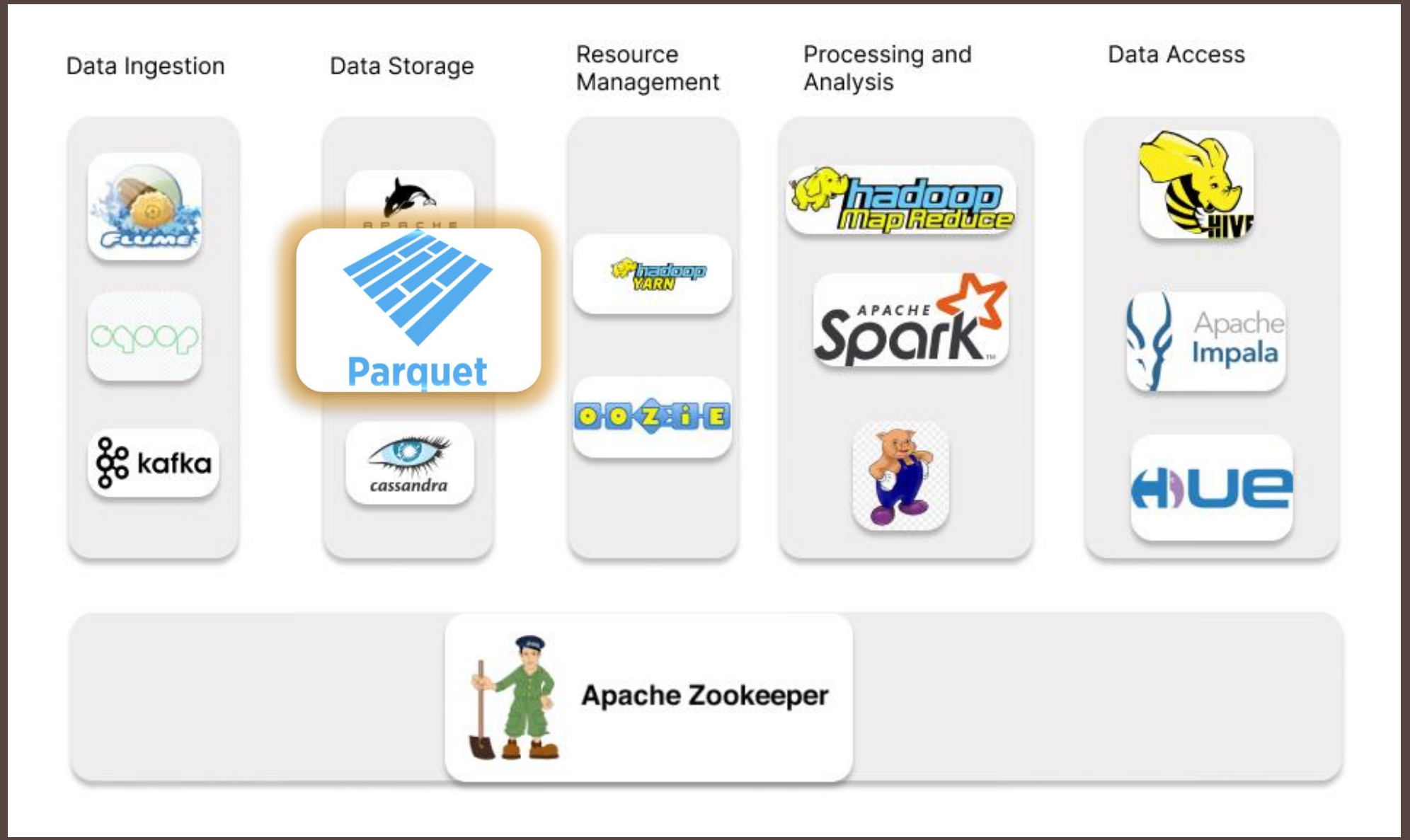
Data Warehouse

+

The “lake” poses some challenges

1. Inconsistent formats
2. Difficult to query
3. Inefficient storage
4. Updating existing data
5. Recovering from bad writes
6. Keeping track of changes

The rise of Big Data led to new technologies



Source: https://www.analyticsvidhya.com/blog/2022/05/an-introduction-to-hadoop-ecosystem-for-big-data/

A bit of Parquet

- Inspired by the Dremel paper
- Binary, column-oriented format
- Well-suited for analytical queries
- Includes metadata
- Supports multiple encoding and compression schemes

Dremel: Interactive Analysis of Web-Scale Datasets

Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer,
Shiva Shivakumar, Matt Tolton, Theo Vassilakis
Google, Inc.
{melnik,andrey,jlong,gromer,shiva,mtolton,theov}@google.com

ABSTRACT

Dremel is a scalable, interactive ad-hoc query system for analysis of read-only nested data. By combining multi-level execution trees and columnar data layout, it is capable of running aggregation queries over trillion-row tables in seconds. The system scales to thousands of CPUs and petabytes of data, and has thousands of users at Google. In this paper, we describe the architecture and implementation of Dremel, and explain how it complements MapReduce-based computing. We present a novel columnar storage representation for nested records and discuss experiments on few-thousand node instances of the system.

1. INTRODUCTION

Large-scale analytical data processing has become widespread in web companies and across industries, not least due to low-cost storage that enabled collecting vast amounts of business-critical data. Putting this data at the fingertips of analysts and engineers has grown increasingly important; interactive response times often make a qualitative difference in data exploration, monitoring, online customer support, rapid prototyping, debugging of data pipelines, and other tasks.

Performing interactive data analysis at scale demands a high degree of parallelism. For example, reading one terabyte of compressed data in one second using today's commodity disks would require tens of thousands of disks. Similarly, CPU-intensive exchanged by distributed systems, structured documents, etc. lend themselves naturally to a *nested* representation. Normalizing and recombining such data at web scale is usually prohibitive. A nested data model underlies most of structured data processing at Google [21] and reportedly at other major web companies.

This paper describes a system called Dremel[1] that supports interactive analysis of very large datasets over shared clusters of commodity machines. Unlike traditional databases, it is capable of operating on *in situ* nested data. *In situ* refers to the ability to access data 'in place', e.g., in a distributed file system (like GFS [14]) or another storage layer (e.g., Bigtable [8]). Dremel can execute many queries over such data that would ordinarily require a sequence of MapReduce (MR [12]) jobs, but at a fraction of the execution time. Dremel is not intended as a replacement for MR and is often used in conjunction with it to analyze outputs of MR pipelines or rapidly prototype larger computations.

Dremel has been in production since 2006 and has thousands of users within Google. Multiple instances of Dremel are deployed in the company, ranging from tens to thousands of nodes. Examples of using the system include:

- Analysis of crawled web documents.
- Tracking install data for applications on Android Market.
- Crash reporting for Google products.
- OCR results from Google Books.

Row-based works well for reading entire rows

Name	Age	Favourite Colour
Anders	40	Blue
Bjørn	40	Red
Oda	31	Red
Ada	29	Green

- Horizontal partitioning
- OLTP ☑
- OLAP ✕

Anders	40	Blue	Bjørn	40	Red	Oda	31	Red	…

Column-based is better for analytics

Name	Age	Favourite Colour
Anders	40	Blue
Bjørn	40	Red
Oda	31	Red
Ada	29	Green

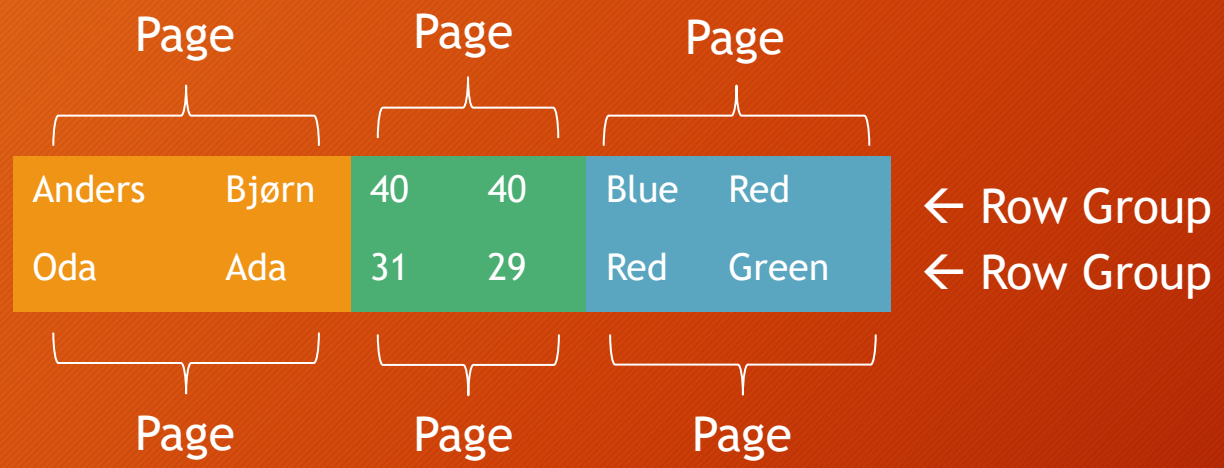
- Vertical partitioning
- OLTP ✗
- OLAP ☑

Anders	Bjørn	Oda	Ada	40	40	31	29	Blue	…

Hybrid approach brings the best of both

Name	Age	Favourite Colour
Anders	40	Blue
Bjørn	40	Red
Oda	31	Red
Ada	29	Green

- Vertical & Horizontal partitioning
- Divide the data into *Row Groups* and *Pages*



Metadata and statistics optimizes queries

Name	Age	Favourite Colour
Anders	40	Blue
Bjørn	40	Red
Oda	31	Red
Ada	29	Green

Schema
Name: String
Age: Integer
Fav. Colour: String

Stats
Min age: 29
Max age: 40

Dictionary encoding helps to reduce filesize

- Build a dictionary of column values
  - Anders → 0
  - Bjørn → 1
  - Oda → 2
  - Ada → 3

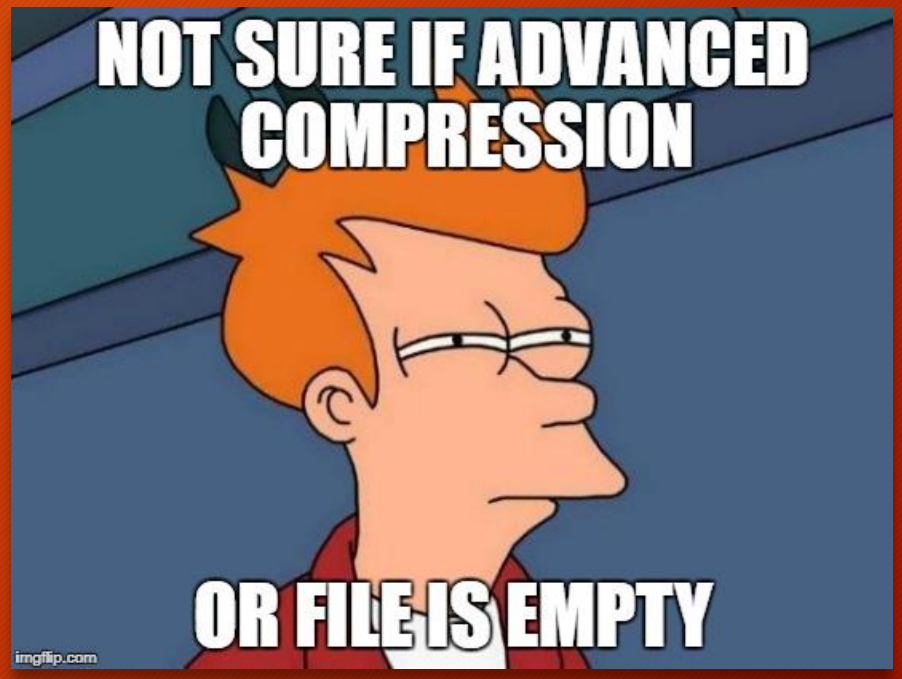
  - Blue → 0
  - Red → 1
  - Green → 2

Name	Age	Favourite Colour
Anders	40	Blue
Bjørn	40	Red
Oda	31	Red
Ada	29	Green

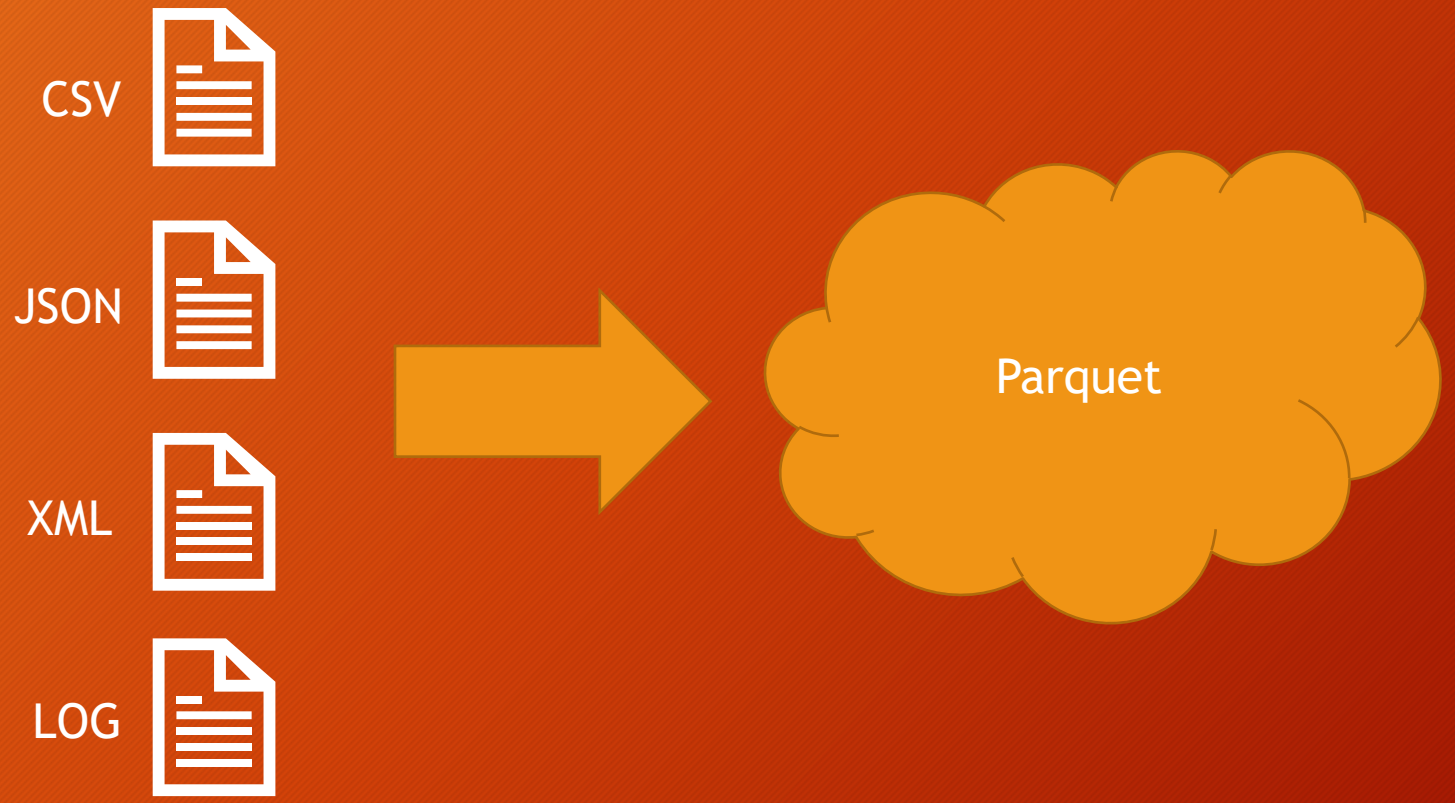
Name	Age	Favourite Colour
0	40	0
1	40	1
2	31	1
3	29	2

Compression reduces filesize even more

- Snappy – well-balanced compression / speed
- LZO – similar to snappy
- GZIP – better compression, but slower



Standardising on parquet improves the lake



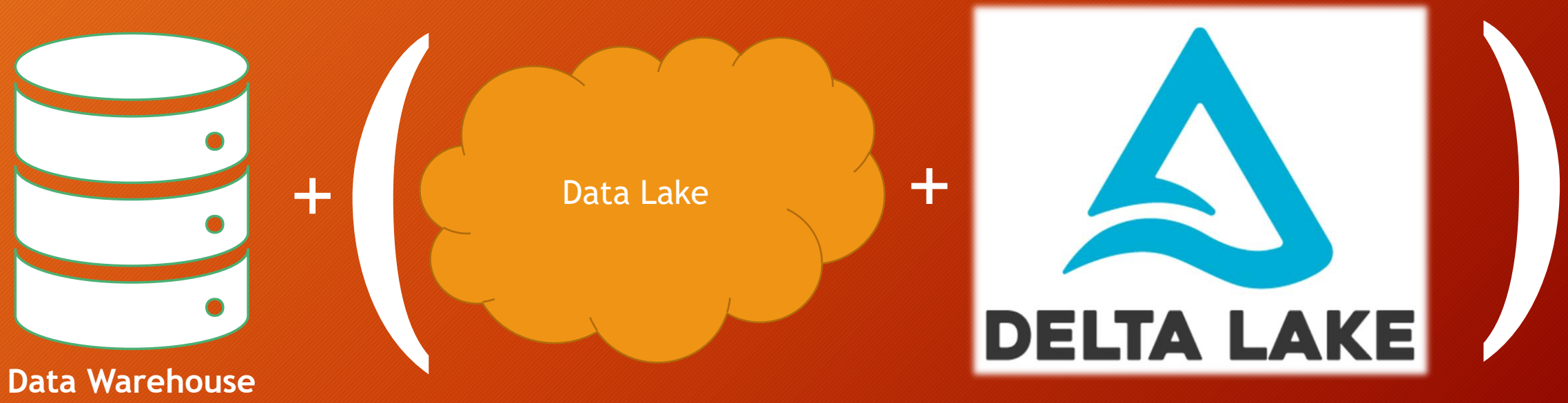
Using parquet has solved *some* problems

1. ~~Inconsistent formats~~
2. ~~Difficult to query~~
3. ~~Inefficient storage~~
4. Updating existing data
5. Recovering from bad writes
6. Keeping track of changes

Adding ACID to the data lake

- **Atomicity**
  - Transactions must be treated as a single “unit” which either fail or succeed completely.
- **Consistency**
  - Transactions must ensure that the database goes from one valid state to another.
- **Isolation**
  - Concurrent transactions must be handled as if the transactions were executed sequentially.
- **Durability**
  - One committed, a transaction must stay committed.

Introducing the *Delta* Lakehouse



Three Delta Features to Win Friends and Influence People

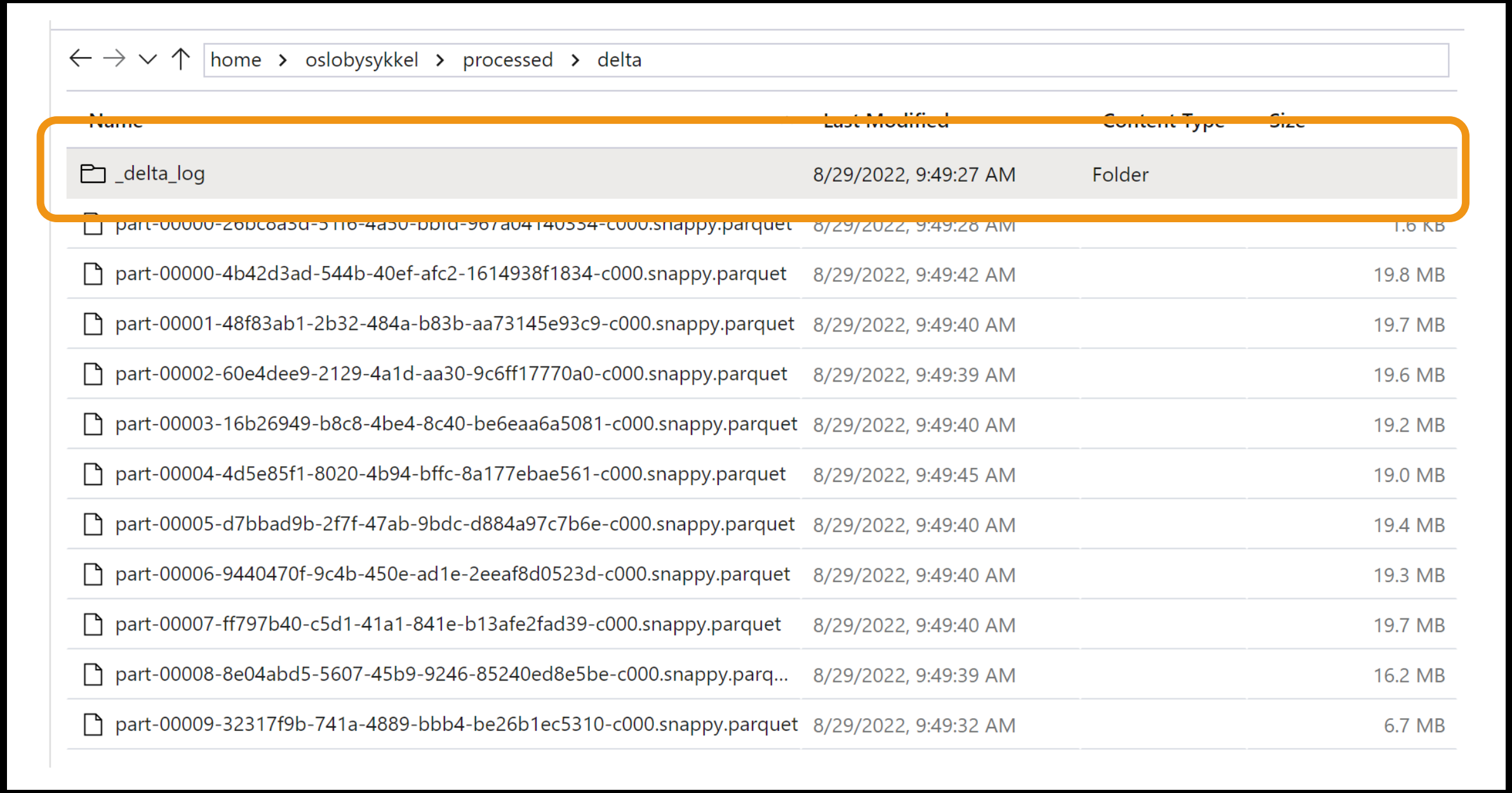
Transaction Log

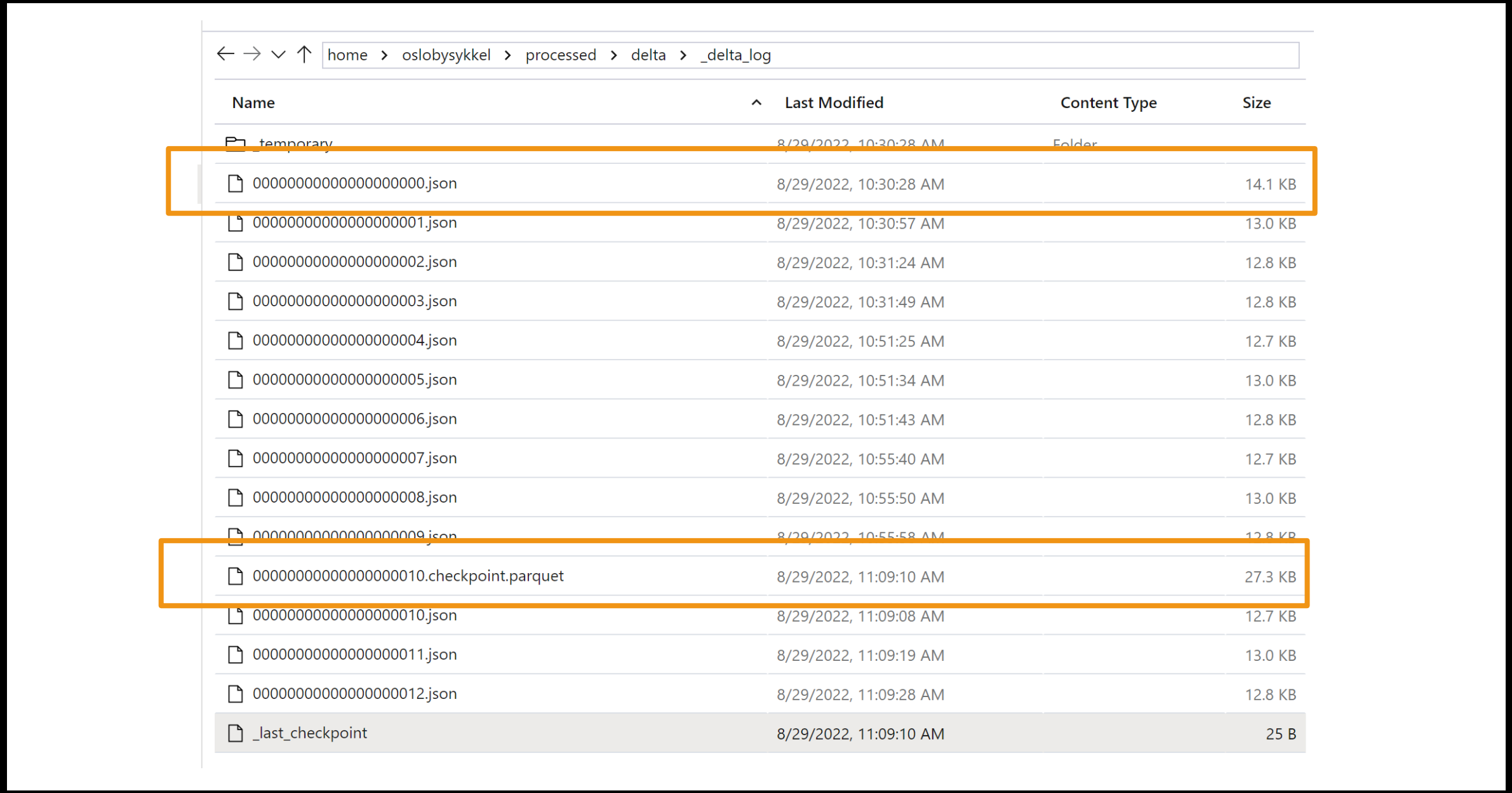
Optimize / Z-Order

Time Travel


https://www.pinterest.com/pin/giant-fir-log-oregon--564849978240801202/

The transaction (Delta) log is the key to ACID


home › oslobysykkel › processed › delta
Name
Last Modified
Content Type
Size
_delta_log
8/29/2022, 9:49:27 AM
Folder
part-00000-4b42d3ad-544b-40ef-afc2-1614938f1834-c000.snappy.parquet
8/29/2022, 9:49:42 AM
19.8 MB
part-00001-48f83ab1-2b32-484a-b83b-aa73145e93c9-c000.snappy.parquet
8/29/2022, 9:49:40 AM
19.7 MB
part-00002-60e4dee9-2129-4a1d-aa30-9c6ff17770a0-c000.snappy.parquet
8/29/2022, 9:49:39 AM
19.6 MB
part-00003-16b26949-b8c8-4be4-8c40-be6eaa6a5081-c000.snappy.parquet
8/29/2022, 9:49:40 AM
19.2 MB
part-00004-4d5e85f1-8020-4b94-bffc-8a177ebae561-c000.snappy.parquet
8/29/2022, 9:49:45 AM
19.0 MB
part-00005-d7bbad9b-2f7f-47ab-9bdc-d884a97c7b6e-c000.snappy.parquet
8/29/2022, 9:49:40 AM
19.4 MB
part-00006-9440470f-9c4b-450e-ad1e-2eeaf8d0523d-c000.snappy.parquet
8/29/2022, 9:49:40 AM
19.3 MB
part-00007-ff797b40-c5d1-41a1-841e-b13afe2fad39-c000.snappy.parquet
8/29/2022, 9:49:40 AM
19.7 MB
part-00008-8e04abd5-5607-45b9-9246-85240ed8e5be-c000.snappy.parq...
8/29/2022, 9:49:39 AM
16.2 MB
part-00009-32317f9b-741a-4889-bbb4-be26b1ec5310-c000.snappy.parquet
8/29/2022, 9:49:32 AM
6.7 MB


home › oslobysykkel › processed › delta › _delta_log
Name
Last Modified
Content Type
Size
_temporary
8/29/2022, 10:30:28 AM
Folder
00000000000000000000.json
8/29/2022, 10:30:28 AM
14.1 KB
00000000000000000001.json
8/29/2022, 10:30:57 AM
13.0 KB
00000000000000000002.json
8/29/2022, 10:31:24 AM
12.8 KB
00000000000000000003.json
8/29/2022, 10:31:49 AM
12.8 KB
00000000000000000004.json
8/29/2022, 10:51:25 AM
12.7 KB
00000000000000000005.json
8/29/2022, 10:51:34 AM
13.0 KB
00000000000000000006.json
8/29/2022, 10:51:43 AM
12.8 KB
00000000000000000007.json
8/29/2022, 10:55:40 AM
12.7 KB
00000000000000000008.json
8/29/2022, 10:55:50 AM
13.0 KB
00000000000000000009.json
00000000000000000010.checkpoint.parquet
8/29/2022, 11:09:10 AM
27.3 KB
00000000000000000010.json
8/29/2022, 11:09:08 AM
12.7 KB
00000000000000000011.json
8/29/2022, 11:09:19 AM
13.0 KB
00000000000000000012.json
8/29/2022, 11:09:28 AM
12.8 KB
_last_checkpoint
8/29/2022, 11:09:10 AM
25 B

Single transaction

```
{
  "protocol": {…
  }
} {
  "metaData": {
    "id": "c2c87577-83a0-4db5-8281-906301a4a5b4",
    "format": {…
    },
    "schemaString": "{\"type\":\"struct\",\"fields\":[{\"name\":\"started_at\",\"type\":\"string\",\"nullab
    "partitionColumns": [],
    "configuration": {},
    "createdTime": 1661761823592
  }
} {
  "add": {
    "path": "part-00000-4c88d73b-452f-4c01-b50f-7f506a7f90a3-c000.snappy.parquet",
    "partitionValues": {},
    "size": 8461535,
    "modificationTime": 1661761828700,
    "dataChange": true,
    "stats": "{\"numRecords\":285149,\"minValues\":{\"started_at\":\"2019-04-02 22:18:47.926000+00:00\",\"e
  }
} {
  "remove": {…
  }
} {
  "commitInfo": {
    "timestamp": 1661761828758,
    "operation": "WRITE",
    "operationParameters": {…
    },
    "isolationLevel": "Serializable",
    "isBlindAppend": true,
    "operationMetrics": {
      "numFiles": "8",
      "numOutputRows": "2245247",
      "numOutputBytes": "66540451"
    },
  }
}
```

Parquet metadata


```
{
    "protocol": {…
    }
} {
    "metaData": {
        "id": "c2c87577-83a0-4db5-8281-906301a4a5b4",
        "format": {…
        },
        "schemaString": "{\"type\":\"struct\",\"fields\":[{\"name\":\"started_at\",\"type\":\"string\",\"nullab
        "partitionColumns": [],
        "configuration": {},
        "createdTime": 1661761823592
    }
} {
```


List of actions

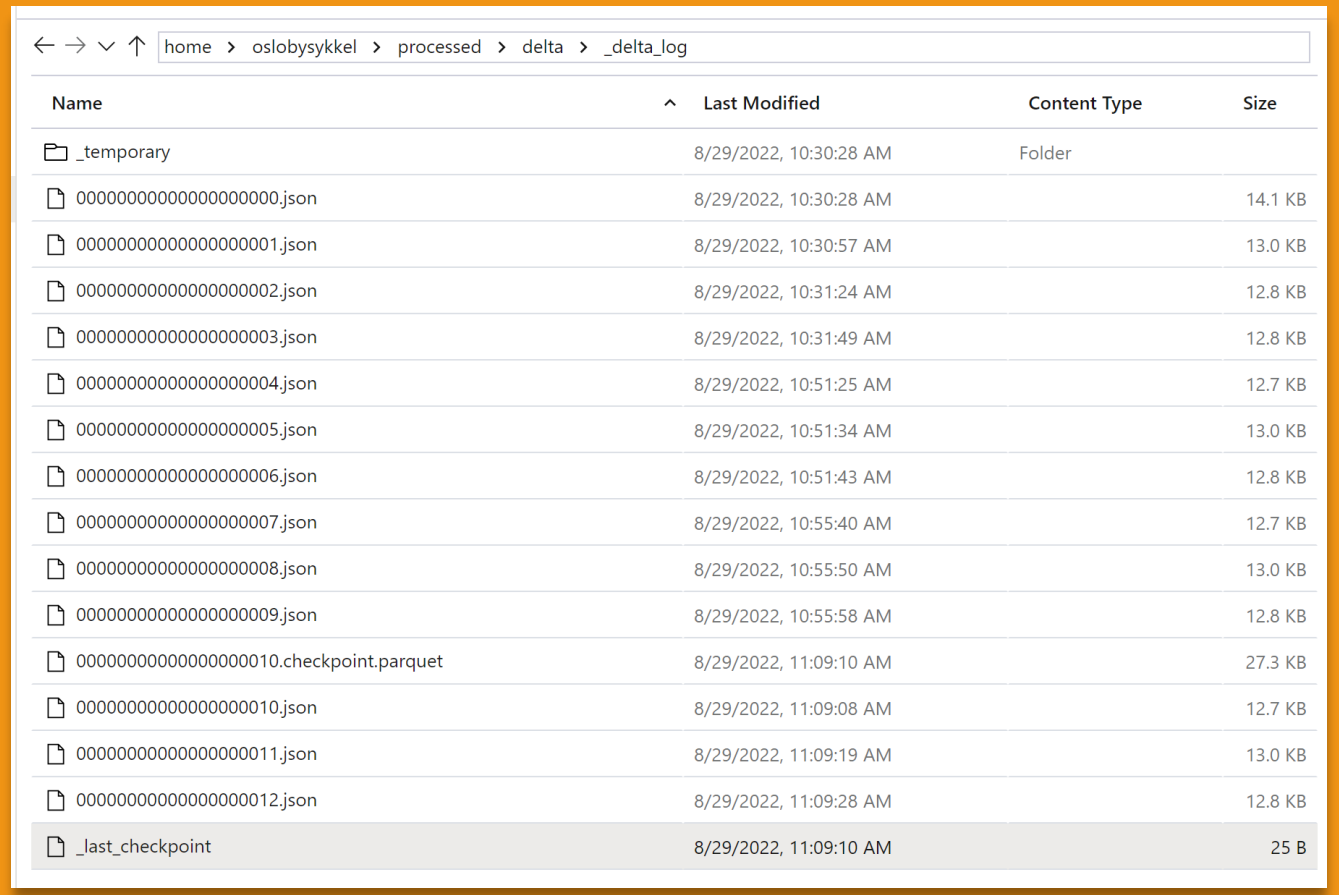
```
18  } {
19      "add": {
20          "path": "part-00000-4c88d73b-452f-4c01-b50f-7f506a7f90a3-c000.snappy.parquet",
21          "partitionValues": {},
22          "size": 8461535,
23          "modificationTime": 1661761828700,
24          "dataChange": true,
25          "stats": "{\"numRecords\":285149,\"minValues\":{\"started_at\":\"2019-04-02 22:18:47.926000+00:00\",\"e
26      }
27  } {
28      "remove": {…
30      }
31  } {
```

Commit metadata


```
} {
    "commitInfo": {
        "timestamp": 1661761828758,
        "operation": "WRITE",
        "operationParameters": {…
        },
        "isolationLevel": "Serializable",
        "isBlindAppend": true,
        "operationMetrics": {
            "numFiles": "8",
            "numOutputRows": "2245247",
            "numOutputBytes": "66540451"
        },
    }
}
```

Mutual Exclusion and Optimistic Concurrency Control ensures “Isolation”

Recording all transactions to disk ensures “Durability”



Optimize / Z-Order to optimize data files

Name	Age
Anders	..
Bjørn	…

Name	Age
Oda	..
Ada	..

Name	Age
Gareth	..
Hanne	..

Name	Age
Dag	..
Claudio	..

OPTIMIZE
Z-ORDER BY NAME

Name	Age
Ada	..
Anders	..
Bjørn	..
Claudio	..

Name	Age
Dag	..
Gareth	..
Hanne	..
Oda	..

SELECT COUNT(*)
WHERE name = “Gareth”

Time Travel

https://slate.com/human-interest/2015/10/great-scott-who-was-scott-on-back-to-the-future-day-the-origin-of-doc-browns-favorite-phrase.html

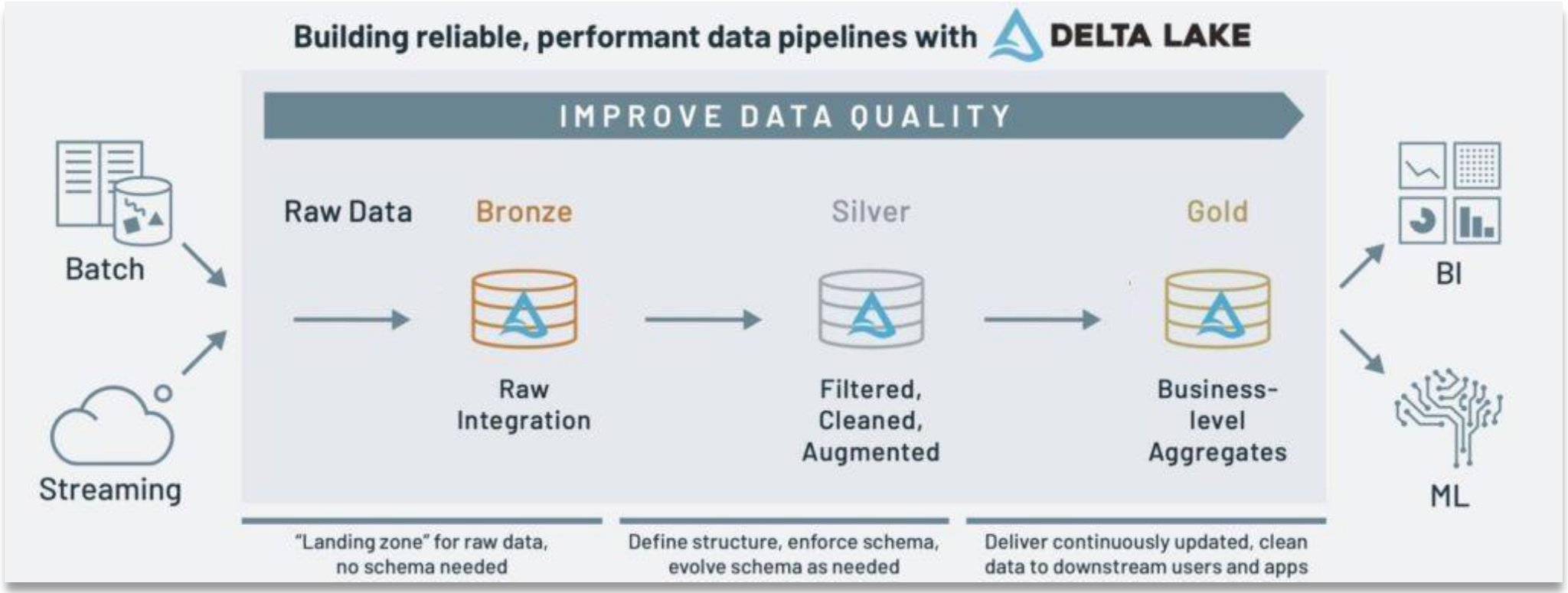
Azure Synapse Analytics

About Oslo City Bike

The most efficient way to get around Oslo is on a bike. City bikes are primarily used for short rides and as a supplement to public transport. About 100,000 people in Oslo use city bikes, and in 2018 over 2.7 million bike trips were made.

The city bike scheme in Oslo is a collaboration between the city of Oslo and Clear Channel Norway AS. The municipality makes public advertising space available and gets a city bike space in return. The scheme is financed by subscriptions, advertisements on the stations, and sponsorship.

Demo – Oslo City Bike Trips



Medallion Architecture

Apache Spark™

docs | source code

Spark

This connector allows Apache Spark™ to read from and write to Delta Lake.

Apache Flink

docs | source code

Flink **standalone**

This connector allows Apache Flink to write to Delta Lake.

PrestoDB

docs | source code

PrestoDB **standalone**

This connector allows PrestoDB to read from Delta Lake.

Trino

docs | source code

Trino

This connector allows Trino to read from and write to Delta Lake.

Delta Standalone

docs | source code

Scala **Java** **standalone**

This library allows Scala and Java-based projects (including Apache Flink, Apache Hive, Apache Beam, and PrestoDB) to read from and write to Delta Lake.

Apache Hive

docs | source code

Hive **standalone**

This connector allows Apache Hive to read from Delta Lake.

Delta Rust API

docs | source code

Rust **Python** **Ruby**

This library allows Rust (with Python and Ruby bindings) low level access to Delta tables and is intended to be used with data processing frameworks like datafusion, ballista, rust-dataframe, vega, etc.

SQL Delta Import

docs | source code

SQL **JDBC**

This utility is for importing data from a JDBC source into a Delta Lake table.

Kafka Delta Ingest

docs | source code

Kafka **Rust**

This project builds a highly efficient daemon for streaming data through Apache Kafka into Delta Lake.

Azure Synapse

docs

Azure **Synapse**

The current version of Delta Lake included with Azure Synapse has language support for Scala, PySpark, and .NET.

MinIO

docs | source code

MinIO **community**

This connector allows you to read and write from Delta tables on MinIO storage.

Athena

docs | source code

Athena **AWS** **manifest**

This utility allows Athena to read from Delta Lake using a manifest file.

Snowflake (Beta)

docs

Snowflake

This preview allows Snowflake to read from Delta Lake via an external table.

DataHub

source code

DataHub **community**

This connector allows DataHub to extract Delta Lake metadata.

Redshift

docs | source code

Redshift **AWS** **manifest**

This utility allows AWS Redshift to read from Delta Lake using a manifest file.

Ceph

source code

Ceph **community**

This connector allows you to read and write from Delta tables on Ceph storage.

Datastream Connector

source code

GCS **Datastream** **badal.io** **community**

As Datastream streams changes to files to Google Cloud Storage, this connector streams these files and writes the changes to Delta Lake.

dlt | SparkR

docs | source code

SparkR **community**

This package allows SparkR to read from and write to Delta Lake.

Power BI

docs | source code

PowerBI **community**

This connector allows Power BI to read from Delta Lake.

Java | Delta Sharing (Beta)

source code

Java **Delta Sharing** **community**

This connector allows as Java client to read from Delta Sharing endpoint.

Apache Beam (Beta)

docs | source code

Beam **standalone** **community**

This connector allows Apache Beam to read from Delta Lake.

Power BI | Delta Sharing (Beta)

docs

PowerBI **Delta Sharing**

This connector allows Power BI to read from Delta Sharing endpoint.

Databricks

docs

Databricks **Azure** **GCP** **AWS**

Delta Lake is included within Databricks allowing it to read from and write to Delta Lake.

Athena Query Federation (Beta)

docs | source code

AWS **Athena** **standalone** **community**

This connector allows AWS Athena to read from Delta Lake.

node.js | Delta Sharing (Beta)

source code

node.js **Delta Sharing** **community**

This connector allows node.js to read from Delta Sharing endpoint.

Starburst

docs

Starburst **Azure** **GCP** **AWS**

The Starburst Delta Lake connector is an extended version of the Trino/Delta Lake connector with configuration and usage identical.

Delta Integrations

Delta Sharing – an open protocol for secure data sharing

DuckDB – SQLite for OLAP

read_delta.py > ...

```python
import duckdb
from deltalake import DeltaTable

trips = DeltaTable("./delta").to_pyarrow_dataset()

conn = duckdb.connect()

df = conn.execute('''
SELECT start_station_id, end_station_id, start_year, start_month, COUNT(*) AS cnt
FROM trips
GROUP BY start_station_id, end_station_id, start_year, start_month
ORDER BY cnt DESC
''').df()

print(df)
```

PROBLEMS 2 OUTPUT DEBUG CONSOLE TERMINAL AZURE JUPYTER

	start_station_id	end_station_id	start_year	start_month	cnt
0	606	504	2017	6	1235
1	504	606	2018	5	1234
2	606	504	2017	8	1124
3	504	606	2017	8	1115
4	606	504	2017	7	1107
...	...	...	...	...	...
1481070	381	575	2016	4	1
1481071	782	532	2016	4	1
1481072	553	506	2016	4	1
1481073	545	560	2016	4	1
1481074	534	545	2016	4	1

[1481075 rows x 5 columns]

Mission accomplished!

1. ~~Inconsistent formats~~
2. ~~Difficult to query~~
3. ~~Inefficient storage~~
4. ~~Updating existing data~~
5. ~~Recovering from bad writes~~
6. ~~Keeping track of changes~~

- Delta Lake = Parquet++

Learn More

- Notebook:
  - https://github.com/gdubya/data-lakes-on-acid
- ACID for your data lake
  - Delta: https://delta.io
  - Iceberg: https://iceberg.apache.org
  - Hudi: https://hudi.apache.org
- Hadoop
  - Home: https://hadoop.apache.org
  - Parquet: https://parquet.apache.org
  - Spark: https://spark.apache.org
- DuckDB: https://duckdb.org