Data Lakes

Kristo Raun

Data Engineering 2024 Fall

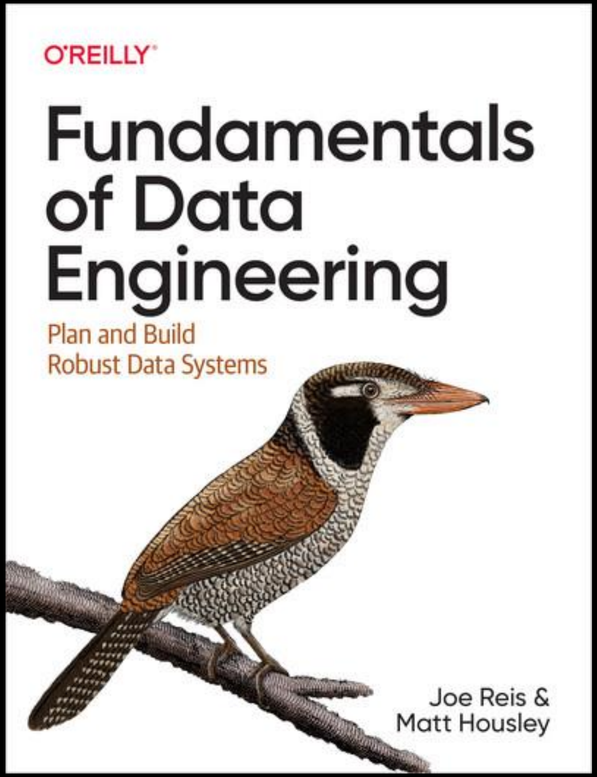
UNIVERSITY OF TARTU

Agenda

- History of Data Lakes
- Data Lake characteristics
- Quiz session

Reading

- Chapter VI


O'REILLY
Fundamentals of Data Engineering
Plan and Build Robust Data Systems
Joe Reis & Matt Housley

History of Data Lakes

- What’s wrong with a database?
  - Data diversity
    - Mostly only handles structured (table-format) data
  - Flexibility
    - Pre-defined schemas (schema-on-write)
    - (potentially) complex ETL
  - Scalability
    - Complex to handle rapidly growing data volumes
  - Data silos
    - Isolated use, difficult to integrate data

History of Data Lakes

- 2000s
  - Internet boom
    - Much more data generated
  - Social media
    - All sorts of data generated
  - Big data
    - Hadoop, HDFS (Java)

Flexibility

- 2010s
  - Cloud storage
    - Object storage – S3, etc
  - SQL support
    - Spark SQL
    - Presto (Trino)
  - Issues
    - Data governance (+security +quality +…)

Scalability

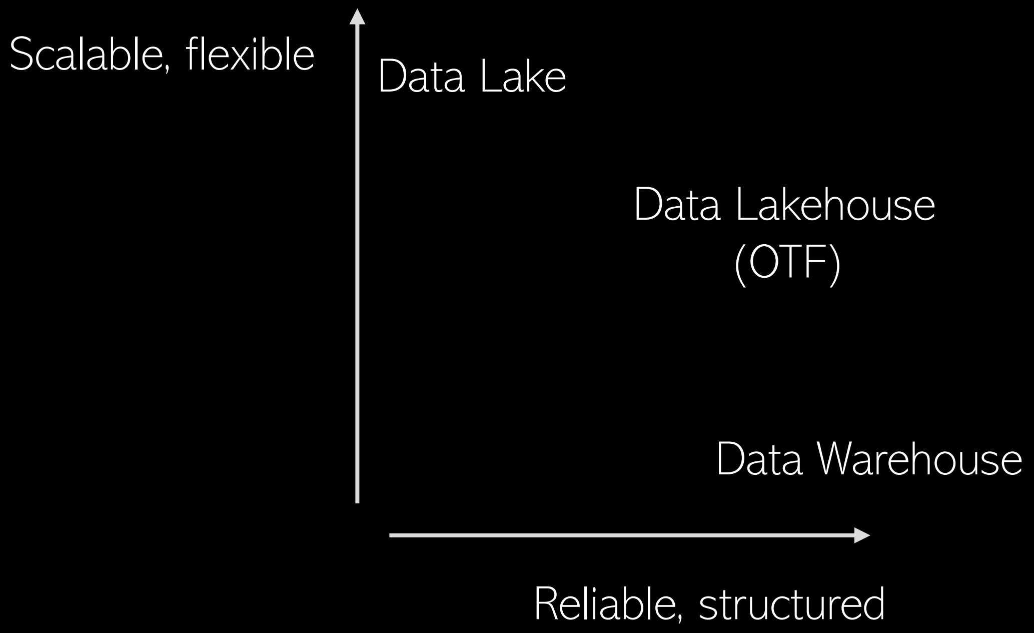
0001
1000
39563
10017
DOGUMAN
DOOOOAR
0000
BOUI6 GOD
8001
D007000 7
VARNONI
1957
0897

History of Data Lakes

- Data swamps
  - Lack of data governance
    - Security
    - Data lineage
    - Data discoverability
  - Data quality problems
    - Raw data, no validations
  - Corrupted data states
    - No ACID compliance

History of Data Lakes

- 2017-2019
  - Open Table Formats (OTF)
    - Apache Iceberg, Delta Lake, Apache Hudi



History of Data Lakes

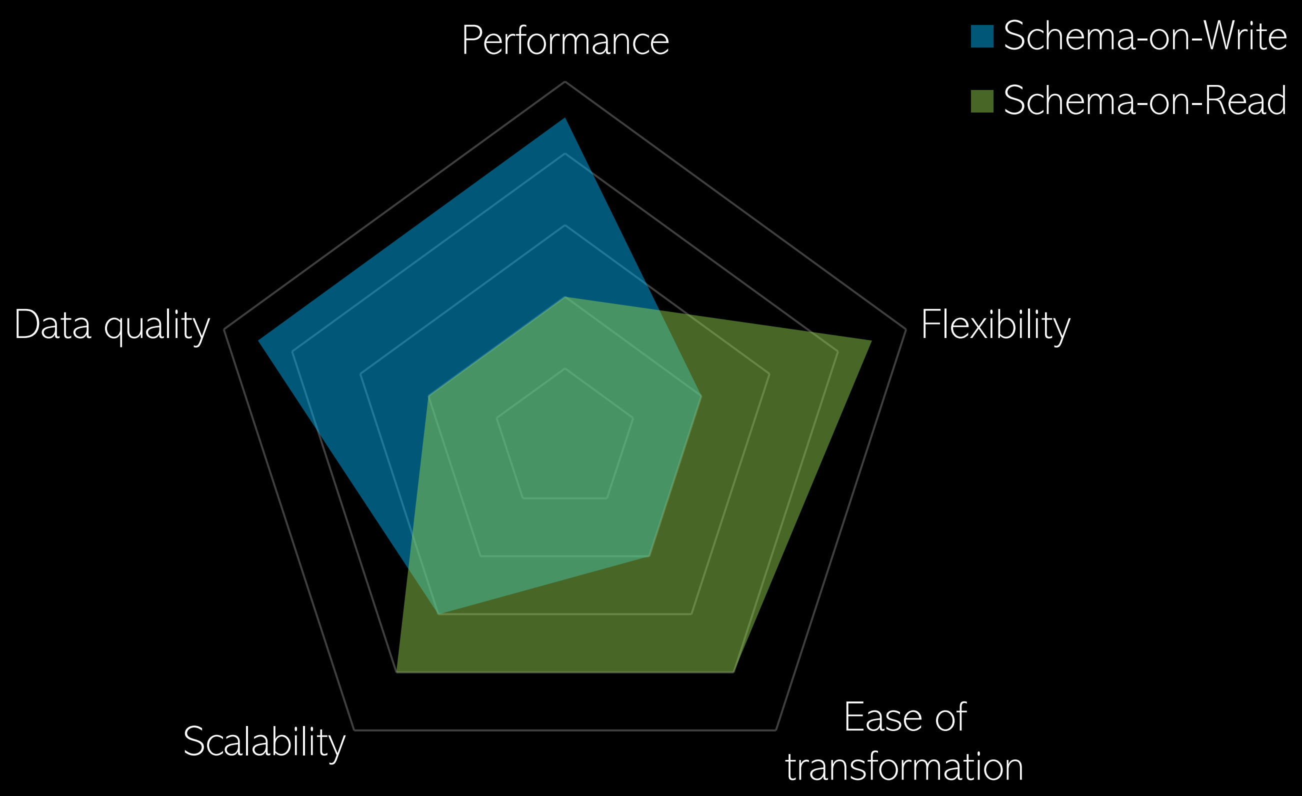
“You should know that the popularity of separating storage from compute means the lines between OLAP databases and data lakes are increasingly blurring. Major cloud data warehouses and data lakes are on a collision course. In the future, the differences between these two may be in name only since they might functionally and technically be very similar under the hood.”

Agenda

- ~~History of Data Lakes~~
- Data Lake characteristics
- Quiz session

Data Lake characteristics

- Schema-on-Read vs Schema-on-Write
    - When is schema applied in the data pipeline?
- Data Warehouses
    - Schema-on-Write
- Data Lakes
    - Schema-on-Read
- Data Lakehouses
    - Support both

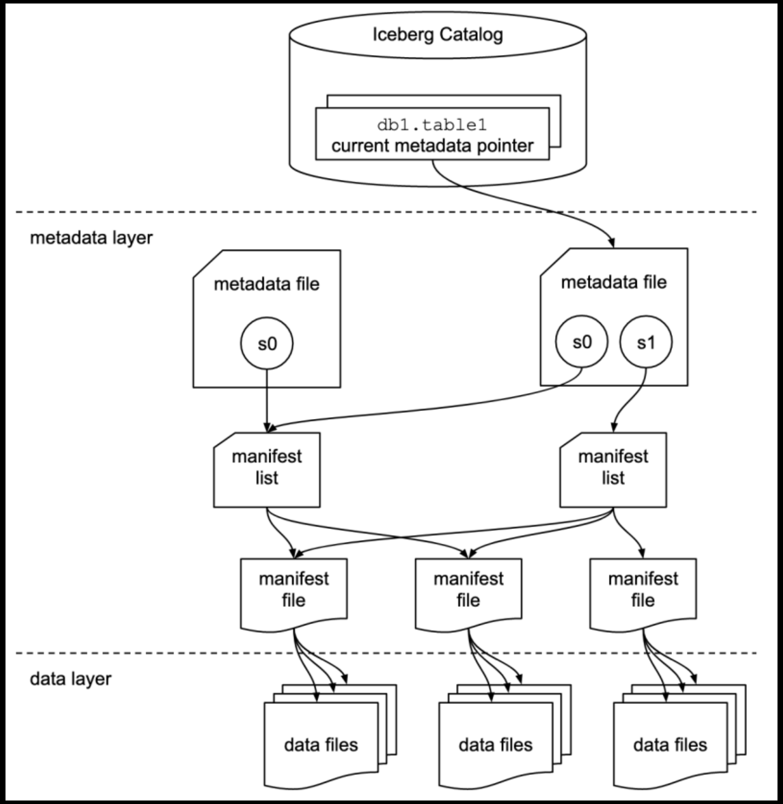


Data Lake characteristics

- DLs usually run on object storage
  - What is a characteristic of object storage?
  - Data immutability
    - → versioned files
- Time travel
  - Return to a previous version of the file(s)

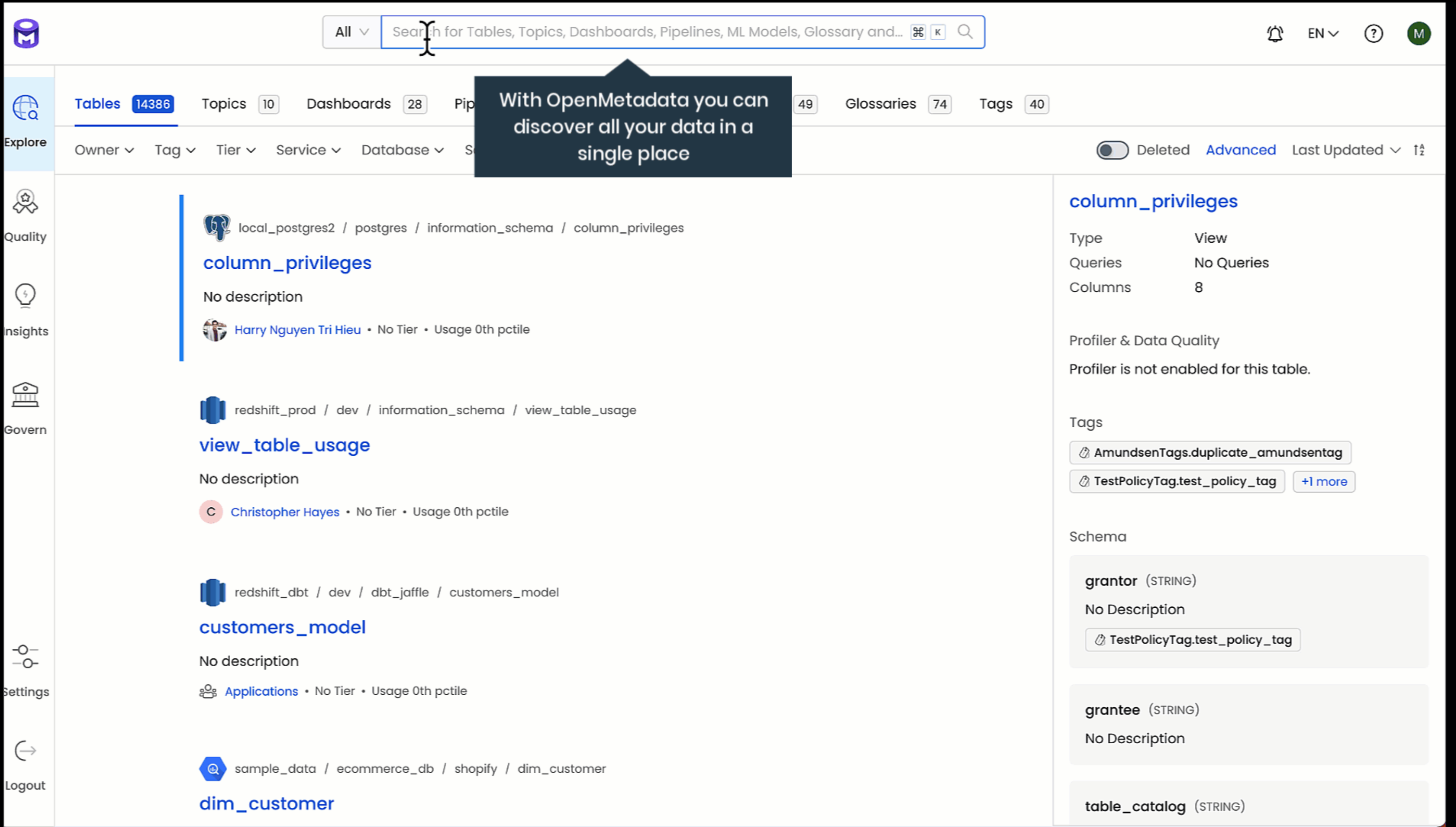
Data Lake characteristics

- DML of data lakes
  - Most typical DML statements?
    - *SELECT*
    - INSERT
    - DELETE
    - UPDATE
  - Update & Delete → won't work on immutable data
  - Track updates & deletes as part of separate files



Data Lake characteristics

- Data catalogs
  - Where's the data?



All ⌄ Search for Tables, Topics, Dashboards, Pipelines, ML Models, Glossary and... ⌘ K

EN ⌄ M

Explore

Quality

Insights

Govern

Settings

Logout

Tables 14386 | Topics 10 | Dashboards 28 | Pip 49 | Glossaries 74 | Tags 40

Owner ⌄ Tag ⌄ Tier ⌄ Service ⌄ Database ⌄ S

Deleted Advanced Last Updated ⌄

With OpenMetadata you can discover all your data in a single place

local_postgres2 / postgres / information_schema / column_privileges

column_privileges

No description

Harry Nguyen Tri Hieu • No Tier • Usage 0th pctile

redshift_prod / dev / information_schema / view_table_usage

view_table_usage

No description

C Christopher Hayes • No Tier • Usage 0th pctile

redshift_dbt / dev / dbt_jaffle / customers_model

customers_model

No description

Applications • No Tier • Usage 0th pctile

sample_data / ecommerce_db / shopify / dim_customer

dim_customer

column_privileges

Type	View
Queries	No Queries
Columns	8

Profiler & Data Quality

Profiler is not enabled for this table.

Tags

AmundsenTags.duplicate_amundsentag

TestPolicyTag.test_policy_tag +1 more

Schema

grantor (STRING)

No Description

TestPolicyTag.test_policy_tag

grantee (STRING)

No Description

table_catalog (STRING)

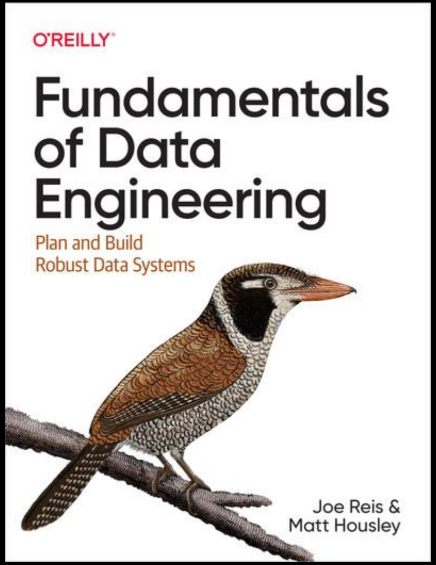
Data Lake characteristics

- Separation of storage and compute
  - Colocation = high performance
  - Why separate?
- Scalability
  - Don't need to migrate data
- Cost efficiency
  - VMs not running 24/7
- Fault tolerance
  - Cloud storage (object storage) typically highly available

Agenda

- ~~History of Data Lakes~~
- ~~Data Lake characteristics~~
- Quiz session

Further reading


O'REILLY
Fundamentals of Data Engineering
Plan and Build Robust Data Systems
Joe Reis & Matt Housley

- Chapter VI
  - Trends in data storage
    - Data catalogs
    - Data sharing
    - …
- Databricks perspective:
  - https://www.databricks.com/discover/data-lakes
- Microsoft perspective (and comparisons):
  - https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-a-data-lake/