Data lake tables

Understanding and creating

Starburst

Lesson Objectives

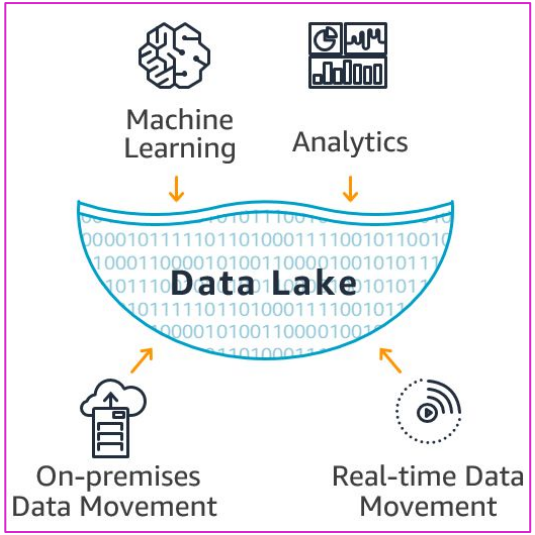
Data lake tables: Understanding and creating

1. Define the data lake.
2. Become aware that multiple file formats can be used in data lakes.
3. Explain the nature of schema on read, and compare it to schema on write in both data lakes and data warehouses.
4. Describe the role that the Hive metastore plays in data lakes.
5. Discuss the difference between external and managed tables. .

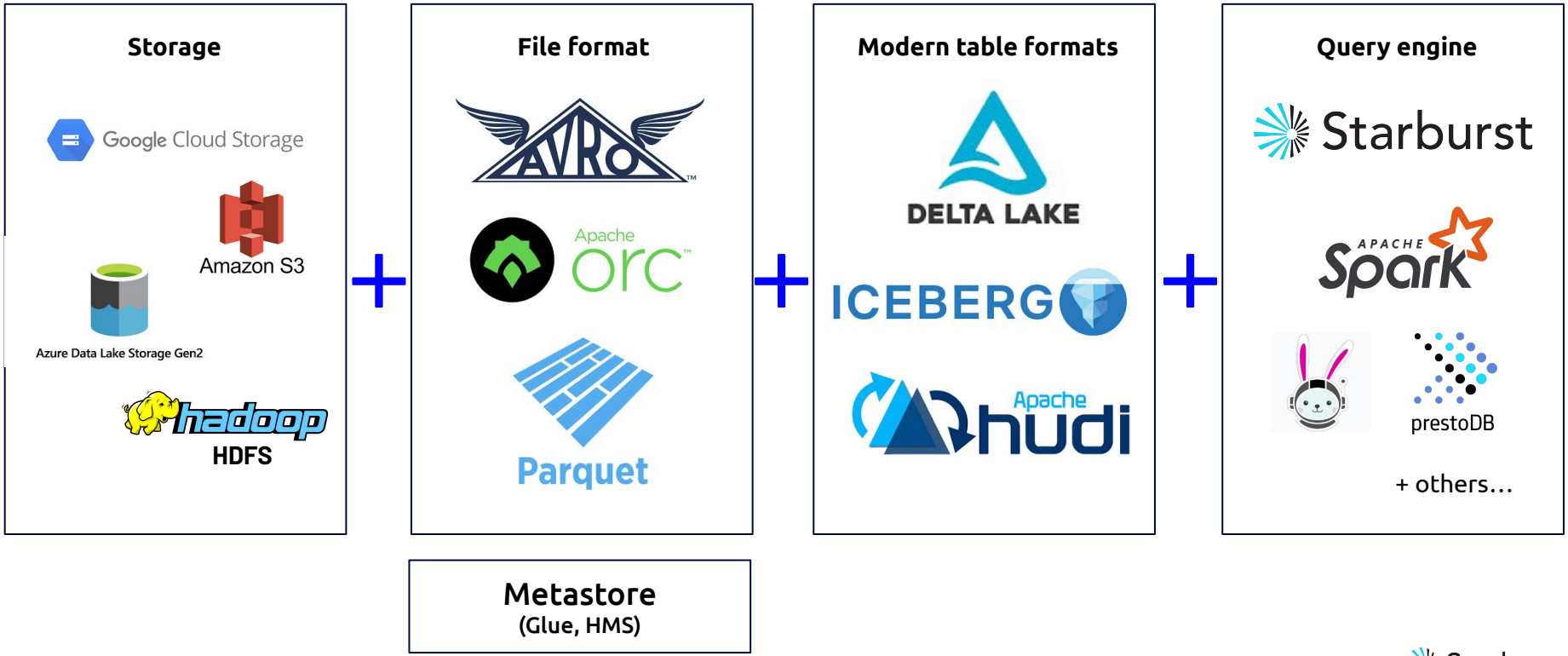
What is a data lake?

"A data lake is a system or repository of data stored in its natural/raw format"

- Supports business units having TB/PB of data in different formats
- Cheap decoupled storage that can be reused for multiple purposes by multiple engines & frameworks
- Strategically aligned with Cloud Vendors (compute on-demand)
- Everyone is trying to do ML / AI projects
- New generation of Developers/Architects/Data People with skills acquired after 2006+ (Hadoop Era)



Components of a modern data lake

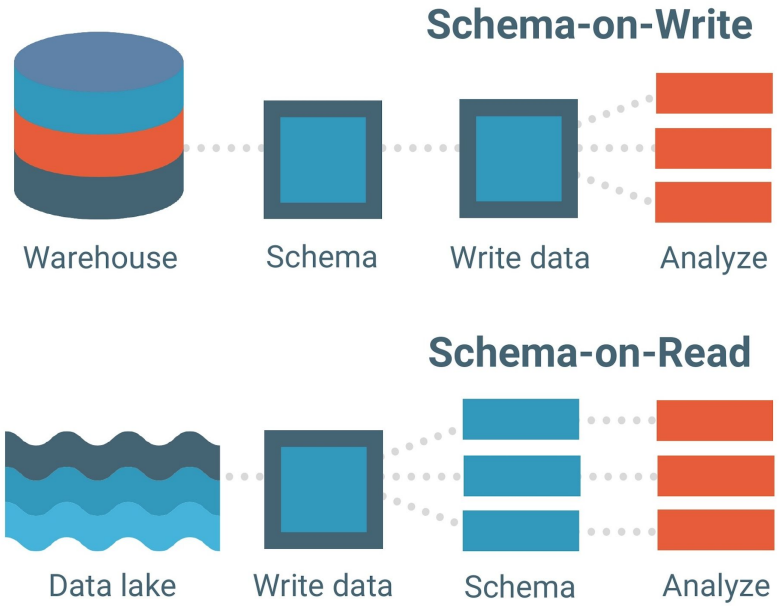


Schema considerations

Data lake vs data warehouse

- Data warehouse: *Schema on Write*
  - Schema must be pre-defined
    - Full relational data model required
  - Data must be “transformed” to match schema & data model before it can be loaded into DW

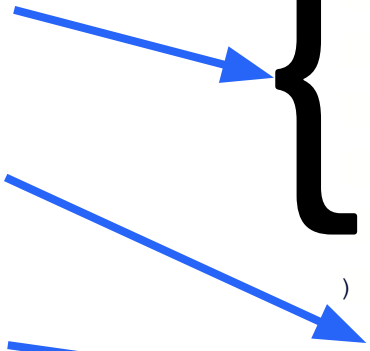
- Data lake: *Schema on Read*
  - Load data in its raw form (historically immutable)
  - Schema can be defined or inferred from raw data as it’s read
  - Schema can evolve easily as data changes
  - Lacks PK/FK integrity checks



Schema on read requirements

3 pieces of information are required

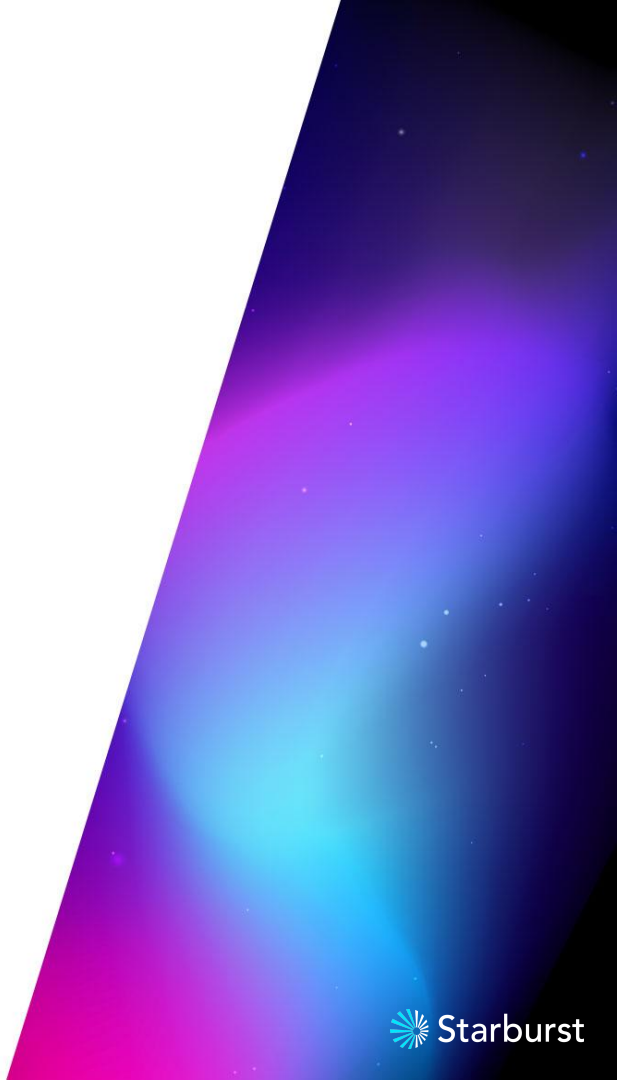
- What does it look like (the schema)
- Where is it (the directory location)
- What will I find there (the file type)



```
CREATE TABLE mycat.myschema.mytable (
    event_time       TIMESTAMP,
    ip_address       VARCHAR( 15),
    app_name         VARCHAR( 25),
    process_id       SMALLINT,
    log_level        VARCHAR( 15),
    message_details  VARCHAR( 555)
) WITH (
    external_location = 'objStore/mytable/',
    format = 'ORC'
);
```

Instructor demonstration

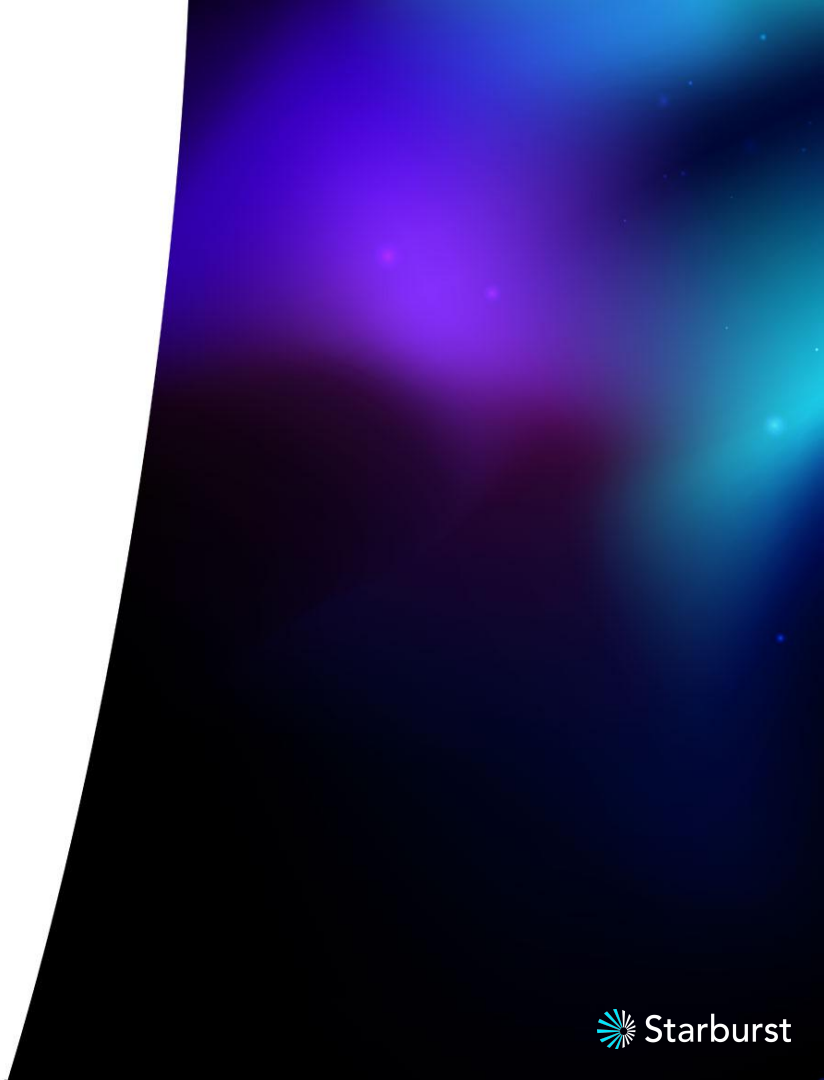
Create tables (15 mins)



Hands-on exercises

Lab 1: Create a schema and tables (20 mins)

Lab 2: Investigate Hive's special columns (15 mins)



Starburst

Lesson summary

Data lake tables: Understanding and creating

1. Data lakes store structured, semi-structured, and unstructured data and make use of traditional table formats like Hive, often using cloud object storage or other storage means.

2. Data lakes make use of multiple file formats like ORC, Parquet, and Avro.

3. Schema on read applies schemas only when data is read, whereas schema on write applies schemas when data is added to a data lake.

4. Data lakes use metastores to manage metadata, and the Hive Metastore is one of the most common implementations.

5. External tables often are populated, and accessed, by other systems, whereas underlying files for managed tables are created directly by the query engine.