Query execution[*]

Distributed Data Processing Environments

Lab Guide 5

This session aims at using a SQL tool to run data processing tasks. It uses the Free Tier[1] and should not consume credits.

Steps

1. Setup:
   - Alternative 1) install DuckDB in your laptop.
   - Alternative 2) use the Vagrant recipe from Lab 4 to start a host with DuckDB in the cloud. Connect with `ssh -L4213:localhost:4213 ...` to use the UI remotely.
2. Generate synthetic data:
   - Install TPC-H extension: `INSTALL tpch;`
   - Load TPC-H extension: `LOAD tpch;`
   - Generate data: `CALL dbgen(sf = 1);`[2]
3. Explore the schema and data in the UI or using `SHOW TABLES` / `SHOW TABLE ...` commands.
4. Explore how queries are converted to operators:
   - Run queries and use `EXPLAIN` / `EXPLAIN ANALYZE` to observe execution plans. Create simple queries or use standard TPC-H queries.[3]
   - Explain the purpose of each operator.
5. Observe the impact of data size and format:
   - Repeat step 4 with *lineitem* table in a Parquet file.
   - Repeat step 4 with data created with diffrent scale factors (*sf*=...).

Learning Outcomes Describe the purpose of each relational operator in a query plan. Identify factors that impact the choice of query plan.

[*] Use of AI tools is encouraged in steps 4 and 5.

[1] `https://cloud.google.com/free/docs/free-cloud-features#free-tier`

[2] Generates 256MB of data. Use other values for different sizes.

[3] `https://github.com/ibis-project/tpc-queries/tree/master/sqlite_tpc`.