Join Performance Analysis

In this lab exercise you will analyze the performance of joins for CSV and Parquet files based on a TPC database benchmark data set.

https://drive.switch.ch/index.php/s/LoshPoCXH8U6guv

In particular, you will work with the following three files:

- customer.tbl
- orders.tbl
- lineitem.tbl

Exercise 1: Reading and Writing Files

Read the three CSV-files. Print the schema of each file and perform a SQL statement to return the number of records per file.

Store the files as Parquet files and check the file sizes. How much is the difference between the CSV and the Parquet files?

Exercise 2: Joins

Perform a join on the three tables. Measure the performance for both CSV and Parquet formats. Analyze the query plans in the Spark UI.

Exercise 3: Self-Joins

Rather than performing joins on two tables, perform a self-join on one of the large tables. Compare the performance of CSV vs. Parquet.

Exercise 4: Join Ordering

Perform a join on the three tables where you modify the order of the tables. What is the performance difference between the following two joins?

- A join B join C (where A is the smallest and C is the largest table)
- C join B join A

What is the impact for CSV and Parquet files?