Data Formats

In this lab exercise you will analyze the size of CSV and Parquet files for various data sets. Moreover, you analyze the performance of queries against CSV and Parquet files.

Exercise 1: Business Data

In this exercise, you use the following data set from a database benchmark:

- customer.tbl

Step 1:

- Read the data set into a Spark DataFrame and store it as a **CSV-file**.
- What is the size of the file?
- Perform a query against the attribute “nationkey” and measure the performance.

Step 2:

- Store the DataFrame as a **Parquet file**.
- What is the size of the file now?
- How is the file stored?
- Perform the same query as in step 1 and compare the performance.

Step 3:

- Run all the experiments at least five times and measure the average query performance along with the standard deviation.
- Plot the results of the sizes and query performance to show a direct comparison between CSV-file and Parquet file.

Exercise 2: Scientific Data

In this exercise, you create your own data set to simulate scientific data.

Step 1:

- Write a program to generate a file with 10 columns (c1 to c10) and 1 million rows.
- Each column should be populated with random floating point values in the range of 0 and 1.
- Store the resulting DataFrame as a CSV file.
- What is the size of the file?
- Perform a range query on one of the attributes and measure the performance.

Step 2:

- Write the DataFrame as a Parquet file.
- What is the size of the file?
- What is the query performance now?

Exercise 3: Comparative Analysis

After you have performed experiments with business data and scientific data, you should analyze and compare the results. Answer the following questions:

- What is the difference in size of CSV vs. Parquet?
- How does the difference change, if you increase the file size by a factor of 10?
- What is the difference in the outcome (data size and query performance) between the business data and the scientific data and why?
- What is the expected behavior and what did you measure?

Exercise 4: Spark Delta Tables

Perform the same experiments as in Exercises 1 and 2 on Spark Delta tables and compare the results. What data format is performs best?