BDA2 - Spark SQL - Exercises

Assignments

Redo the exercises BDA1 using Spark SQL whenever possible. The initial processing of csv files (such as splitting on ;) can be done using Spark's map.

There are two ways to write queries in SparkSQL - using built-in API functions or running SQL-like queries. **To pass this lab, you need to use built-in API functions for all the 6 exercises and, in addition, SQL-like queries for the second exercise**. **The slides of this link (https://www.ida.liu.se/~732A54/lab/SparkSQLQuickIntro.pdf) show some examples of Spark SQL.**

For each exercise include the following data in the report and sort it as shown:

1.
year, station with the max, maxValue ORDER BY maxValue DESC
year, station with the min, minValue ORDER BY minValue DESC

2.
year, month, value ORDER BY value DESC
year, month, value ORDER BY value DESC

3.
year, month, station, avgMonthlyTemperature ORDER BY avgMonthlyTemperature DESC

4.
station, maxTemp, maxDailyPrecipitation ORDER BY station DESC

5.
year, month, avgMonthlyPrecipitation ORDER BY year DESC, month DESC

6.
year, month, difference ORDER BY year DESC, month DESC