Homework 5. Spark SQL

Task 1.

Download Covid dataset from Github: https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv

Provide two solutions: SQL and Dataframe/Dataset API for following queries:

1. Find country with the greatest number of cases and the greatest number of cases per million.
2. Find a day with the greatest number of new cases across the world.
3. Find min, max, avg and median of new cases for last month.

Task 2.

To address this task, you need to prepare some data. At first you need to install tool called Presto (use Docker image for Windows). Just follow instructions in doc. Enable - TPC DS connector. Using presto cli export following tables to CSV files: catalog\_sales, warehouse, ship\_mode, call\_center, store\_sales, store, item, customer, customer\_address, customer\_demographics, household\_demographics, income\_band, store\_returns. Using Spark convert files to Avro, Parquet, ORC formats and compare results.

Using Spark SQL execute following queries against all four file formats:

1. For catalog sales, create a report showing the counts of orders shipped within 30 days, from 31 to 60 days, from 61 to 90 days, from 91 to 120 days and over 120 days, grouped by warehouse, call center and shipping mode.
2. List all customers living in Edgewood, with an income between 30000 and 80000.
3. Find top 5 stores for each category based on store sales in 1998 year.

Schema definitions and relationship between tables could be found here: <http://www.tpc.org/tpc_documents_current_versions/pdf/tpc-ds_v2.1.0.pdf>

Task 3.

You were asked to build voter profile for each candidate based on vote data. You have 2 files:

* voter\_demographics.csv with following fields:
  + id
  + age – possible values: 18 – 99
  + gender – possible values: M, F, U
  + education – possible values: “No formal education”, “High school diploma”, “Collage degree”, “Bachelor’s degree”, “Master’s degree”, “Professional degree”, “Doctorate degree”
  + material\_status – possible values: “married”, “single”, “divorced”, “widowed”
* votes.csv with following fields:
  + candidate\_id
  + voter\_id

Provide file with following structure:

|  |  |  |
| --- | --- | --- |
| Attribute | Candidate 1 | Candidate N |
| 18-24 | Count |  |
| 25-34 |  |  |
| 35-44 |  |  |
| 45-54 |  |  |
| 55-64 |  |  |
| 65 and more |  |  |
| M |  |  |
| … |  |  |
| No formal education |  |  |
| … |  |  |
| Married |  |  |
| …. |  |  |

To generate data use generator cli tool: ./generator

Optional task:

Build voter profile using RDD API.