CSCI 620/Section 02/Mior, Introduction to Big Data, Spring 2235

Assignment 9 – Distributed data processing

Description

In this assignment, you are going to work with a JSON version of the original Stack Exchange dataset files you used in assignment 1.

You can download Apache Spark here:
https://www.apache.org/dyn/closer.lua/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz

Extensive documentation is available for both Java and Python.

Your tasks

Provide a program to query the original data files using Apache Spark DataFrames. You should use unmodified files downloaded directly from the link above. The queries below are some of the same queries you wrote in assignment 2 although. All queries should use the Spark SQL API, but *not* to simply define tables and write SQL queries.

You should instead compose operations on the DataFrames using `select`, `where`, and other functions in the DataFrame API. (This is similar to directly writing relational algebra expressions for each query.) If you are unable to write the complete query, you should write as close an approximation as you are able.

Include in your submission the time taken to return *all* results for each query and the actual values for the first 10 results.

(25 points per query)

1. Users whose display name starts with “J” and did not make any posts in 2014.

2. The most popular tags on posts created in 2017. (Hint: You can use the `split` function to turn tags into an array and `array_contains` to check for a particular tag.)

3. The display names of users with the most number of comments on posts with the tag “networking”.

4. The most popular badges earned by users with less than 100 reputation.