AY 2020/2021

Politecnico di Milano

Middleware Technologies Analysis of COVID-19 Data

Federico Armellini Luca Pirovano Nicolò Sonnino

Professor
Alessandro Margara

Version 1.0
June 23, 2021

Contents

1 Introduction and assignment

1.1 Description of the project

In this project, you have to implement a program that analyzes open datasets to study the evolution of the COVID-19 situation worldwide. The program starts from the dataset of new reported cases for each country daily and computes the following queries:

1. Seven days moving average of new reported cases, for each county and for each day
2. Percentage increase (with respect to the day before) of the seven days moving average, for each country and for each day
3. Top 10 countries with the highest percentage increase of the seven days moving average, for each day

You can either use real open datasets 1 or synthetic data generated with the simulator developed for Project 4. A performance analysis of the proposed solution is appreciated (but not mandatory). In particular, we are interested in studies that evaluate (1) how the execution time changes when increasing the size of the dataset and/or number of countries; (2) how the execution time decreases when adding more processing cores/hosts.

1.2 Assumptions and guidelines

- When using a real dataset, for countries that provide weekly reports, you can assume that the weekly increment is evenly spread across the day of the week.

2 Solution Overview

2.1 Architecture chosen

Apache Spark

2.2 Assumptions and Definitions

- The dataset is in the csv format. Each line contains: day, rank, number of infected people, number of sane people, number of new infected people (than the day before), the number of new sane people (than the day before).

2.3 General solution

2.3.1 Main and Calculate

- Doing a query that gets how many days and how many countries are in the data.
- Doing a for loop, we calculate for each country the necessary values for the 3 queries.
- Printing the results

2.3.2 calculate (single country)

We handle differently the first 7 days to the other ones (because they are special, the first doesn't have, for example, a previous moving average to compare to).

So, for each day:

- Query the data to get the number of newly reported cases
- Calculating the needed values for the 3 queries and storing them.

2.4 Data structures

We use 2 main data structures: DayCountryInfo and Top10Countries.

Then we have highscore (defined as HashMap¡Integer, Top10Countries¿ indexed by day) and query1and2Result (defined as HashMap¡Integer, HashMap¡Integer, DayCountryInfo¿¿, indexed by day and country rank) to store the overall result of the 3 queries.

2.4.1 DayCountryInfo

DayCountryInfo stores information about a certain country in a certain day (its movingAverageValue, its movingAverageIncrease);

2.4.2 Top10Countries

Top10Countries stores in an arraylist, for each day, the information about the top 10 countries, as for query 3.

2.5 Performance evaluation

2.5.1 Variables and parameters

- $queryWhereOnRank = maxContries$
- $arrayCopyDays = maxDays/2$
- $queryWhereOnDay = maxDays$
- $query3UpdateResults = log2(10) + 10$
- $query1and2UpdateResults = 1$

2.5.2 Equations

- $maxCountries * (queryWhereOnRank + maxDays * (arrayCopyDays + queryWhereOnDay + query3UpdateResults + query1and2UpdateResults))$
- $maxCountries * (maxContries + maxDays * (maxDays/2 + maxDays + (log2(10) + 10) + 1))$
- $maxCountries * (maxContries + maxDays * (maxDays * 3/2 + 14.32))$
- $maxCountries * (maxContries + maxDays^2 * 3/2 + 14.32 * maxDays))$
- $maxContries^2 + maxCountries * maxDays^2 * 1.5 + 14.32 * maxDays * maxCountries)$

2.5.3 Conclusion

Overall performance is (simplified):
$maxContries^2 + maxCountries * maxDays^2 * 1.5$