MSc in CSTE, CIDA option
Machine learning & Big Data Assignment

Analysis of the COVID-19 global dataset using Hadoop/Spark

Dr Jun Li
Cranfield University

September 13th, 2023

Hand-in date: TBC (FT), TBC (PT)

1. Introduction

The goal of the assignment is to address some queries regarding the Covid-19 pandemic at a global level, using either the Apache Hadoop or Apache Spark data processing framework. As dataset, the following file is provided in CSV format:

`time_series_covid19_confirmed_global` available at URL https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv

Such file contains data on the worldwide trend of confirmed cases of Covid-19. In particular, each line of the file indicates: the *province/state*, the *country/region*, the *latitude* and *longitude*. Then, for each day starting from January 22nd, 2020 to today, each column with the date *mm/dd/yy* reports the total number of confirmed cases up to that day (cumulative data).

The dataset is updated every day, with a new daily row added at 23:59 UTC and corrections are possible if inaccuracies are found.

2. Tasks

The queries to address are:

1. For each country, calculate the mean number of confirmed cases daily for each month in the dataset.

2. For each continent, calculate the mean, standard deviation, minimum and maximum of the number of confirmed cases daily for each week. When calculating the statistics, consider only the 100 states most affected by the pandemic. If the state is not indicated, consider the country. To determine the most affected states in the entire dataset, consider the trend of the daily increases of confirmed cases through the *trendline coefficient*. To estimate it, calculate the slope of the regression line that approximates the trend of the daily increments.
   Observe that the continent to which each state/country belongs is not explicitly indicated in the dataset, but has to be identified. To this end, consider 6 continents: Africa, America, Antarctica, Asia, Europe, Oceania.

3. Considering the 50 most affected states calculated on a monthly basis according to the trendline coefficient, for each month in the dataset, apply the clustering algorithm *K*-means [1, 2] with *K*= 4. Determine the states (or nations) that are part of each cluster. Each cluster should group the states that have a similar pattern of daily increases in confirmed cases.
   Optional: For the clustering algorithm K-means, compare the performance of a naive implementation of the algorithm and the implementation provided by the Spark MLlib or Apache Mahout [3] library.

In your report (see below), you should include the results of each query in CSV format, specifying the date of the dataset that has been used for the analysis. You should also experimentally evaluate the processing times of the 3 queries on the reference platform that you have used for the assignment and discuss the results in the report. The input dataset can be read either from a local file or from HDFS. The resulting files can be saved to a local file or to HDFS.

In addition, at the end of the report you need a paragraph to discuss the potential ethical issues and challenges facing related to Machine Learning and Big Data using scenarios such as the COVID-19 case used in this assignment.

3. Source Code and Report requirements

Write a report to present and discuss your findings. The report should be no less than 1500 words and must not exceed 3000 words. The report can contain any number of figures/tables, however all figures/tables should be numbered and discussed. All code used in the analysis should be included in an Appendix along with appropriate documentation.

The is an individual assignment.

4. Assignment Submission

The source code and documentation should be submitted electronically via the technical work submission points. The report should be submitted electronically via the TurnItIn submission point by the prescribed deadline, for the assignment submission to be considered complete.

5. Marking

The assignment will be assessed based on the following marking scheme:

- 20% Introduction, methodology, conclusions
- 40% Software: system architecture, code organization and efficiency
- 30% Discussion and analysis of the results
- 10% Report structure, presentation, clarity, references

References

[1] K-means clustering, https://en.wikipedia.org/wiki/K-means_clustering

[2] J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of Massive Datasets*, chapter Clustering. Cam- bridge University Press, USA, 3rd edition, 2020. http://infolab.stanford.edu/~ullman/mmds/ch7.pdf

[3] Apache Mahout, https://mahout.apache.org/