UNIVERSITY OF LIVERPOOL

COMP529/336: COURSEWORK ASSIGNMENT #1 (BATCH ANALYTICS)

Dr. Bakhtiar Amen

Coursework Data: **06 Nov 2020**

Due Data: **27 Nov 2020**

INTRODUCTION

This assessed coursework assignment is worth 20% of your mark for COMP529/336. Failure on this assignment can be compensated by higher marks on other assessment on the module. The assignment aims to test your understanding of batch data analytics, with a focus on your ability to use Spark to solve Big Data Analytic problems. More specifically, it aims to partially assess the following learning outcome for COMP529/336: ***"understanding of the middleware that can be used to enable algorithms to scale up to analysis of large datasets".***

ASSESSMENT

The report will be assessed according to the following criteria:

Criterion	Percentage
Clarity of presentation (including succinctness) of main report	20%
Quality of pyspark code and middleware configuration description (including assessment of how easy it is to understand)	40%
Quality of analysis performed	40%

SUBMISSION

Please submit your coursework online using the COMP529/336 page on **CANVAS by 16:00 on Friday 27th November 2020.** Standard lateness penalties will apply to any work handed in after this time. The report and the python program must be written by ***yourself*** using your own words (**see the University guidance on academic integrity for additional information**).

PROJECT BACKGROUND

COVID-19 is a new virus that can affect human's lungs and chest. This virus is rapidly spreading across the world, there are 47 million confirmed cases in 190 countries and about 1.2 million deaths. For this coursework, your task is to process and analyse a sample of COVID-19's dataset which is collected by EU open data portal and available at;
https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data

This assignment enables you to gain experience in implementing Big Data Batch Analytic framework to learn from large-scale of dataset by using apache Spark's different Resilient Distributed Dataset -RDD operations.

DATASET

The dataset is in a csv file format called (Covid19.csv) and available on Canvas > Files > Courseworks > Coursework-1. This dataset contains the latest available public data on COVID-19 including a daily situation update, the epidemiological curve and the global geographical distribution (EU/EEA and the UK, worldwide). For this coursework, we are only using part of the data as described in below table;

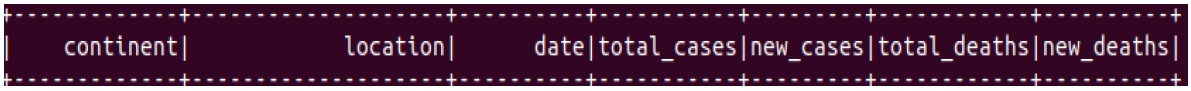

Table 1: Data Filed Description

Field	Description
Continent	COVID-19 cases across of each continent (e.g., Asia, EU, Africa, USA)
Location	COVID-19 cases per country
Date	Data of COVID-19 cases confirmed
Total cases	Total COVID-19 Confirmed cases
New cases	New COVID-19 Confirmed cases
Total deaths	Total COVID-19 death cases per each country
New deaths	New COVID-19 death cases

YOUR TASKS

1. Implement Big Data Batch Analytic **Spark framework** (Standalone Mode is preferred).
2. Write a pyspark program to do the following tasks;
   a. Load the csv dataset file into spark and create **DataFrame** with header=True.
   b. To see if you have created the DataFrame successfully, run a command to show your dataframe table as well as run another command to print your schema.
   c. Use Spark RDD transformation **filter function** to filter NULL values and show your output evidences before and after the filter function.
3. Use Spark RDD aggregate and groupby functions to see the highest **total death cases** in each country (e.g., Sweden 986… and so on). Your output result should consist of the highest/max total death cases per each country (*shown list of at least 20 countries are recommend*).
4. Use Spark RDD **max and min** functions to see which country has a highest and lowest cases based upon using total_cases column (*shown list of at least 20 countries are recommend*).
5. In your report, comment on how this analysis could be extended to consider larger datasets (e.g., 4 Terabyte of COVID-19 for 2 years). Briefly Describe how to use your Spark skills to solve other problem (Chose your own case study)/ draw data flow diagram for such data analytic case.

YOUR OUTPUT REPORT

The output from this coursework is a brief report suggested to have the following sections:

1. **Middleware Configuration**: How you configured Spark middleware (including a description of your Spark and your rationale for this choice).
2. **Data Analytic Design**: How did you design your DataFrame (what command did use to load data and read the data), what RDD operation functions did you use (including your rationale for your design, briefly state/draw data flow model for your work).
3. **Results and Discussion**: The results obtained from your data analytic tasks.

4. **Conclusions and Recommendations:** This will be including discussion of how you would perform the task if it were to be undertaken at larger scale.

FORMAT OF YOUR REPORT

1. The output from this coursework is a brief report to be less than or equal to **two**[1] **A4 pages** excluding any appendices, text size is 12-point, justify text, and in only **pdf/docx formats**.
2. Make sure to save your file under your **surname + module code** (e.g., Abcd_COMP336).
3. List of your **Pyspark program** should be in the appendix (no longer than 2 pages).

End of your Coursework

[1] While the requirement is to produce no more than 4 pages, it is anticipated that the challenge will be to fit everything into those 4 pages: it is unlikely that a report of much less than 2 pages will result in a high mark.