Postgraduate coursework
DATA7201 Data Analytics at Scale (2021)

Project Report – Report on Dataset Analytics (Coursework)

1. Introduction

This assessment for "DATA7201 Data Analytics at Scale" consists of a piece of individual coursework. Given a dataset (see Section 2), you should use big data analytics techniques to explore the data and to draw some conclusions that inform decision makers. You will also need to select the most appropriate techniques and justify your choices using supporting evidence from academic literature.

You should **write a 1,500 word structured report** (see Section 3) that describes the approach you have taken to analyse the chosen dataset using big data analytics techniques. The report should focus on summarising your approach on the chosen dataset and presenting your main findings. You should pay particular attention on communicating clearly the results of your analysis and on helping the reader interpret your findings. Charts, tables, and appendices are not included in the word count.

This assessment is worth 50% of the overall course mark for DATA7201. **Submission deadline: 4pm Monday 24th May 2021 (Week 13) via Turnitin.**

2. Given datasets: Twitter (1% API) or Facebook (Ad Library API)

The dataset to be used in this assessment is to be chosen among two possible options: A Twitter and a Facebook dataset. You need to choose only one among these two options and perform the analysis on it.
The **first option** is a collection of tweets available through the free Twitter API which covers a 1% random sample of the entire Twitter stream over 6 months (07/2014-12/2014). A description of the data structure is available starting from: https://developer.twitter.com/en/docs and https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview. The dataset you should use covers 6-month worth of data collected from this API. The format in which the data is provided by Twitter is compressed JSON files. There is no need to analyse the entire dataset and it is instead acceptable to focus on a specific subset (e.g., based on a period of time around an event) for your analysis. You can find the data on the data7201 cluster HDFS under /data/ProjectDatasetTwitter.
The **second option** is a collection of sponsored political posts on Facebook targeted at US users during 8 months (03/2020-11/2020) preceding the US Presidential election in 2020. A description of the data structure is available starting from: https://www.facebook.com/ads/library/api/. The dataset covers 8-month worth of data collected from this API. The format in which the data is provided by Facebook is JSON files. Each file is the result of a request for active ad campaigns performed every 12 hours during the 8 months period, thus a lot of ad campaigns are duplicated across files (i.e., if they run for more than 12 hours) and should be properly managed during pre-processing. Given the limited size of this dataset, it is expected that projects would analyse most of the available data. You can find the data on the data7201 cluster HDFS under /data/ProjectDatasetFacebook.

You can integrate the chosen dataset with external data if you want (e.g., with weather data via time information and mentioned locations), although this is not mandatory. The emphasis of this coursework assignment is on how you engage with big data analytics techniques, select appropriate big data analytics technologies, and on how well you communicate your analysis and findings. You are allowed to use any other data analytics tool (e.g., for producing visualisations or data summaries) as long as you also use, in some steps of your analysis (e.g., to pre-process the entire dataset to select a relevant sample of the data), the cluster where the data lies (e.g., Pig, Python, SQL, etc.).

Examples of possible analysis include, but are not restricted to, the following:

- Look at tweet volume over time for a certain hashtag.
- Focus on certain users (e.g., edu.au users and see which academics from which university are most active).
- Look at URLs included in tweets to understand which internet domains are most popular on Twitter.
- Look at a specific event or hashtag and look at who is tweeting about it.
- Look at sponsored content spend per demographic group during the US Presidential election in 2020.
- Look at the duration of ad campaigns over topics and political alignment during the US Presidential election in 2020.

You should investigate the chosen dataset using tools on the DATA7201 cluster and write up your findings into a report also providing the code/scripts/queries (if any) you used as an appendix. You will be evaluated according to the learning objectives of the module as specified in the report structure (Section 3).

3. Report structure

You are required to produce a structured report that includes all the sections detailed in Table 1. You can structure sub-sections as you prefer. Overall, 90 marks will be awarded based on the content of your report. In addition, 10 marks will be awarded based on the presentation of the report and how well you communicate your findings. You must state the word count somewhere in the report. As there is a word count limit you should aim to make your writing as concise and informative as possible. Note also that your work will be assessed taking into account the word limit; therefore, we are not expecting multiple detailed analyses in the report; rather the emphasis should be on the clarity, accuracy and quality in communicating your findings.

Table 1: Required content of the structured report.

Section	Description	Maximum allocated marks	Learning Objective
Structured abstract	This should provide a summary of your report in a structured manner. This is not included in the word count.	Required, but 0 marks	
Table of contents	This should include section titles and page numbers. This is not included in the word count.	Required, but 0 marks	
Introduction	This section should briefly describe the general area of big data analytics and motivate the need for distributed system solutions with practical examples on why these solutions are needed.	15 marks	1. Solve challenges and leverage opportunities in dealing with Big Data
Dataset Analytics	This section should provide a brief description of the dataset used in your report and the pre-processing steps you took (e.g., focus on tweets containing a certain hashtag). You should also list any additional datasets you used (e.g., weather data), if any. Describe all steps performed to analyse the data and present the results of your analysis. You can select in which way to analyse your data (e.g., Pig, Python, SQL, etc.) using the DATA7201 cluster, what specific dimensions to look at, and what questions to investigate. You should use at least one of the tools available on the cluster and you can use additional external tools, If desired.	50 marks	3. Apply data analytics infrastructures to best support data science practices for non-technical stakeholders (e.g., executives). 5. Judge in which situations Big Data analytics solutions are more or less appropriate. 6. Design the most appropriate Big Data infrastructure solution given a use case where to deploy Big Data solutions.
Discussion and conclusions of the analysis	In this section, you should summarise and discuss the main findings of your analysis and lessons learned. You should state the main message the reader should come away with from your data analysis.	25 marks	3. Apply data analytics infrastructures to best support data science practices for non-technical stakeholders (e.g., executives).
Appendix	Include the code/scripts/queries you used as an appendix. The code quality will not be assessed.	Optional, and 0 marks	