ICS5114 Project
Big Data Processing

Charlie Abela, Joseph Bonello, Jean-Paul Ebejer

March 12, 2019

This document contains the details for the Big Data Processing project. Individuals are encouraged to **work in teams of at most two**. While discussions between individual groups are considered healthy, the final deliverables need to be your own and not plagiarised in any way. This project is worth 100% of the total, final mark for ICS5114 (Big Data Processing). Questions related to the problem should ideally be discussed with us. Other questions related to the general aspects of the project and/or questions related to the different technologies etc should be posted in the project's VLE forum.

The deadline for this project is **12:00pm Monday 10th June 2019**. Deliverables, with attached and signed plagiarism form, must be submitted on the VLE. Furthermore, you will need to submit a pendrive (will be returned) with a docker version of your solution as well as any complimentary datasets that were used, to the AI department's secretary, Ms Francelle Scicluna (Level 1, Block A, Room 4). Late projects will be penalised.

1 Introduction

This year's project is based around the theme of scientific publications. Each group is expected to develop a **Sci**entific **P**ublication Analyt**i**cs prototype *SciPi* through which it is possible to analyse and discover interesting collaborations among scientists and researchers. The implemented artefact needs to allow for the analyses and visualisation of "strong" and successful collaborations among authors as well as collaborations which do not exhibit such success.

Some existing **Datasets** include:

i. Open Academic Graph[1] which has a total size of about 140GB (split across a number of files). The OAG links two large academic graphs:

[1] `https://aminer.org/open-academic-graph`

the Microsoft Acadmic Graph (MAG) and ArnetMiner (also called AMiner);

ii. DBLP[2] with a size of about 350M;

iii. PubMed[3] which includes over 29 Million citations for biomedical literature, life science journals and online books;

iv. Any other dataset that is freely available and/or can be accessed via some dedicated API.

SciPi is intended to showcase the potential of, and the challenges behind working with Big Data. You are required to use this project to demonstrate your depth of understanding of the tools and techniques that were discussed in class and how these can be used to process very large datasets with the goal of providing increased value.

2 Requirements

SciPi needs to deal effectively with main the three main Vs: **Variety**, **Volume** and **Velocity**.

Data Variety: *SciPi* must be able to handle data that comes from a variety of sources. This data could be structured and/or unstructured and preferably having different formats. The goal is to integrate this variety of data types quickly into useful information that can be exploited by *SciPi*. Which model is better suited to represent the data? How will the variety issue be resolved?

Data Volume: *SciPi* must clearly make use of a variety of technologies to handle the large volume of data. Which big data technologies are more suitable to address the defined problem? What resources are required and how can these be used efficiently? Which storage structure is ideal?

Data Velocity: Data streams vary quickly, so big data streams need to be understood, prioritized and integrated into *SciPi* in real-time. What are your strategies in dealing with Velocity? The application should either function in real-time or if pre-computation is needed, have a real-time realization (but we will take a wide view of real time depending on the scale of what is done).

2.1 Problems

SciPi needs to address the following problems:

[2] `http://dblp.uni-trier.de/xml/`

[3] `https://www.nlm.nih.gov/databases/download/pubmed_medline.html`

i. Find dense communities of interest. Given a set of labels denoting research domains/topics of papers etc. from an end user, the goal is to present dense communities in the publications network that closely connect all the entities satisfying the labels, and summarize it as a collaboration pattern. A collaboration pattern connects a set of authors, papers and the venues/conferences/journals the papers are published in.

ii. Track the dynamics in the publication networks. Given a collaboration pattern, how does the information change over time? Is it possible to discover interesting events in the evolving network for some particular pattern?

iii. Association and correlation analysis. Given a set of keywords describing research area/topics:

   a. define and discover associations between authors and keywords (taken from title and/or the abstract);
   b. allow Information Retrieval of authors and associated sub-graphs based on topic keywords;
   c. perform clustering of authors based on their association to keywords so as to recommend potential collaborators.

3 Deliverables

The following is information about the deliverables that you will need to present.

3.1 D1: Implemention of *SciPi*

Each team is expected to implement a solution that solves the problems described in Section 2.1. The solution must make use of the techniques/technologies that you have learned/used during the course or some other technique/technology that you think is suitable for the solution. Your solutions should make use of some cloud platform such as **Microsoft Azure**, **AWS**, **Google's Cloud** etc.

Marks allocated for this deliverable: **(60 marks)**

3.2 D2: Scientific Paper

The documentation should be formatted as a scientific paper using the IEEEtran LaTeX class template[4]. Please make sure to include page numbers. The

[4] `http://ctan.mirror.garr.it/mirrors/CTAN/macros/latex/contrib/IEEEtran/IEEEtran_HOWTO.pdf`

maximum page limit for this report is 8 pages, excluding figures, tables and references. A two-column layout should be used and font size should be set to 10pt with default line spacing. You are not allowed to adjust margins.

Your documentation should, at least, include the following sections:

a. *Introduction*: a brief explanation of the problem that is being addressed. This should be concise and highlights the main aim and objectives; **(10 marks)**;

b. *Related research*: discusses existing research related to problems addressed in *SciPi* **(20 marks)**;

c. *Methodology*: this section should deal with the design and implementation of the various features of *SciPi*. Which challenges were encountered and how were these resolved? **(60 marks)**

d. *Conclusion and Future work*: what were the strong and the weak points of your approach? What worked well and what did not give the desired results? A brief description of how this work can be extended should also be included **(10 marks)**.

Scientific paper writing is a very important aspect of this project and research in general, you should allow time to compile it with due diligence. References to existing research and correctly referencing such research will be rewarded.

Marks allocated for this deliverable: **(25 marks)**

3.3 D3: Presentation and Demo

Every team will be expected to deliver a 10 minute presentation and a 5 minute demo. Questions will also be asked. A video of the 5 minute demo should also be delivered.

Marks allocated for this deliverable: **(15 marks)**

3.4 Summary of deliverables

D1: Implementing the solution	60 marks
D2: Scientific paper	25 marks
D3: Presentation	15 marks

Submissions:

i. On the VLE you need to upload all the deliverables D1-D3 (by the team representative), unless size is an issue, in which case include on the pendrive;

ii. D2 needs to include a duly signed plagiarism form;

iii. Furthermore, you will need to submit a pendrive with a docker version of your solution as well as the dataset/s that were used and any other document/resource related to your solution.

Final suggestion: if you have difficulties do not hesitate to contact us and/or post questions on the VLE 'Project' forum.

Good Luck!