UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science
University of Pretoria

Big Data
MIT 805

Assignment 2

4 August 2017

1 Introduction

Due to an increase in the world's population, advances in technology, and the expansion of big corporations, many organisations face challenges when collecting and analysing large amounts of data on a daily basis. This assignment provides a broad overview and a hands-on approach to the main phases of dealing with big data, namely, collecting, processing, reducing, and visualising the data.

This assignment consists of three parts that you will be completed during the course of the semester. These parts can be broadly classified as follows:

1. **Collection and process:** Collect and analyse a big dataset.
2. **Process and map-reduction:** Process the dataset and reduce it to a smaller sensible output.
3. **Visualisation:** Visualise the dataset and the produced output to find useful relationships.

You should keep all three parts in mind while completing each part of the assignment. Since the parts follow on from one another, it is important to thoroughly think about the decisions you will be making for each part.

2 Instruction

Part 1 - Collection

The first part in any big data project is to collect the necessary data which is later processed and analysed in order to make decisions. Since you do not have the time, budget, and resources to collect your own data, you will use an existing dataset. There are a wide variety of free datasets available online and your task is to choose one that you will use for the other parts of this assignment. You can choose any dataset you like, but if you do not know where to start you search, check here:

`http://hadoopilluminated.com/hadoop_illuminated/Public_Bigdata_Sets.html`

https://github.com/caesar0301/awesome-public-datasets

When choosing a dataset you should consider the following:

- **Size:** Do not choose a dataset that is too large, some of these sets contain terabytes of data. You will have to process the dataset on your own computer and if the dataset is too large, it will take a lot of time to processes the entire set. Also do not choose a dataset that is too small, since you will not find useful information and relationships in the data which is essential for the other parts of the assignment. We recommend choosing a dataset between 1GB and 10GB, depending on the machine you will use for the processing.
- **Diversity:** Make sure there is enough diversity in the dataset so that you can nicely reduce, summarise, and visualise the data later on.
- **Format:** You will find that the datasets have a wide variety of formats. We recommend using raw text, CSV, or SQL-based datasets. Technically any format will work, as long as you understand the format and can manipulate it through some code. Try to avoid datasets with binary entries (e.g.: images, audio, and video), since it will increase your processing time and will not be useful for the level of data analysis you will be doing in this module.

Once you have chosen a dataset, write a report about the data. The report should contain some technical aspects of the dataset, such as the size, format, and age. Provide a brief overview of the data contained in the set and the reasons/procedures followed by the organisation who collected the data. Also predict which relationships and correlations you expect to find in the data (the actual processing of the data will only be done in the second part of the assignment). The majority of your report should focus on describing your dataset according to the V's of the collection and processing phases, variety, veracity, volume, and velocity. Note that since you are not working on a live database, the velocity might not be obvious or even determinable. Some datasets contain dates or timestamps that can guide you with the velocity. You may also incorporate other V's if applicable.

This part of the assignment has to be completed individually. Hence, everyone has to find their own dataset and write their own report.

Date issued: 4 August 2017.

Deliverable: Written report completed individually.

Due date: Upload your report on the 31st of August 2017 before 21:00 to the module website.

Part 2 - Process and Map-Reduction

For this part of the assignment you will work in groups of between 4 and 6. As a group decide which of the datasets from your members you want to use.

You should install Hadoop on a computer and familiarise yourself with the way framework works. Based on your dataset, decide which interesting information could be extracted from it. If you are working on a set collected by a commercial company, you may ask yourself which information the company might want to extract in order to give them a competitive advantage. Then write a small Java program for Hadoop that will extract and reduce the information of your dataset. Remember to start early with this part, since coding and processing might take some time, depending on the size of your dataset.

Hadoop is a free and open-source framework developed by Apache for processing big datasets. Hadoop is written in Java and you will therefore be able to run it under any operating system, as long as you have Java installed. If you are not familiar with Hadoop, we suggest setting up a Ubuntu virtual machine, since it will make the installation, compilation, and execution a lot easier. You can use any version of Hadoop, but we suggest using the latest stable version (2.8.1). Below are the download links for Java and Hadoop:

Java (JRE): `http://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html`
Hadoop: `http://apache.is.co.za/hadoop/common/hadoop-2.8.1/hadoop-2.8.1.tar.gz`

Extract the Hadoop archive with any supporting tool, such as 7-Zip. There are many tutorials on the web explaining how to setup Hadoop. Here are some easy ones to get started:

`https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html`
`http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster`
`https://www.digitalocean.com/community/tutorials/how-to-install-hadoop-in-stand-alone-mode-on-ubuntu-16-04`

Note that many of those tutorials are written for enterprise systems and you do not have to follow most of those installation steps (ssh, IP routing, security, etc). Your Hadoop node only has to interact locally (aka localhost) and does not have to interact with other nodes on the network/server. The basics for your setup should be to install the Java runtime environment, download and extract Hadoop, and then link Hadoop to Java. That should be enough for you to get started.

Once Hadoop is installed, familiarize yourself with the software and their MapReduce framework. MapReduce will allow you to write a small Java code snippet to extract and summarize your dataset (eg: calculate and average of one column, similar to how SQL does it). Hadoop does most of the work for you. You only have to write the code, compile it and then execute it in Hadoop.

Since some of you might be unfamiliar with MapReduce, we created a small example for you. The example includes a small dataset, the MapReduce Java code, and small scripts to compile and run your example. The scripts are for Ubuntu, but you can edit them to also work under Windows and Mac. The example uses a dataset of Titanic passengers to determine how many female and male passengers were saved. The input dataset is located in the `input` directory. You can use any data format (CSV, TXT, SQL database, etc), but you have to write code to access/interpret the data. The MapReduce code is located in `Titanic.java`. The `compile.sh` script (execute with `sh compile.sh`) can be used to compile your Java class into a JAR executable. Run the JAR file in Hadoop using the `run.sh` script (execute with `sh run.sh`). The results will be written to the `output` directory. Note that the given MapReduce example is very simple. You should create a more interesting reduction that incorporates multiple attributes from your dataset.

As a group, decide which of your datasets is most interesting. Determine which useful information can be extracted from the set. Code the MapReduce algorithm to extract and summarize your data. You may use separate algorithms to extract different narrow information, or create a complex algorithm that combines a number of attributes to find previously unseen information. The latter approach is advised, since it applies to Big Data in industry. For instance, a grocery store might want to determine which other product a consumer might purchase if he/she already has product X in the basket (eg: macaroni and cheese). This stands in contrast to traditional data reduction, such as simply calculating what the average consumer spends in the store.

For the deliverable you should submit your group's MapReduce code. Do not upload your datasets, since they might be very large. Also submit a short report briefly explaining your dataset, the approach you followed for the MapReduce algorithm, the reasons why you chose your specific algorithm/attributes, and a short discussion on the results you retrieved from the MapReduce algorithm. The format should be similar to part 1.

Date issued: 1 September 2017.

Deliverable: Written report and Hadoop code completed as a group.

Due date: Upload your report on the 28th of September 2017 before 21:00 to the module website.

Part 3 - Visualisation

For this part of the assignment you will continue working in groups.

Choose a visualisation tool of your choice to view your data. We suggest using Hue, since it provides direct support for Hadoop, is free, has SQL and file functionality, and supports R and Java. You may however decide to use a different tool more applicable to your expertise and/or dataset. For instance, if you have a geographical dataset, you might want to use a tools that can display your data over a world map.

Further information on this part will be published at a later stage.

Date issued: TBA

Deliverable: TBA

Due date: TBA