# COMP 1702
# Big Data Coursework Part 1.

Student ID - 001002629

15th March 2021

# Contents

# 1 Introduction

Big Data plays an important role in every aspect of modern life, but how this is achieved and why requires investigation. This document contains a brief literature review on the subject of Big Data and answers to four investigations. Section One answers what the main features of Big Data are, Section Two compares Hadoop to Relational Databases and describes an application that would suit features that Hadoop has. Section Three describes the steps that MapReduce takes to solve a problem using key-value pairs while Section Four compares NoSQL Databases to Relational Databases and has a use case for NoSQL.

# 2 Literature Review

The article Big Data: The Management Revolution published by McAfee et al. (2012) only had 3 Vs, Volume, Velocity and Variety where as Kamilaris, Kartakoullis, and Prenafeta-Boldú (2017) document five V's which include the previously mentioned and expand these features to, *"Veracity (V4): The quality, reliability and potential of the data, as well as its accuracy, reliability and overall confidence. Valorization (V5): The ability to propagate knowledge, appreciation and innovation."* (Kamilaris, Kartakoullis, and Prenafeta-Boldú, 2017) This alludes to no definition that is defined and agreed too by the community or that some features are seen as more important then others depending on their point of view. But what is agreed on is often that growth of this sector will continue as *"an increasing number of people are directly or indirectly storing significant volumes of their unstructured information online in India"* (Garg, Popli, and Sarao, 2021) but even within this paper they have referred to six V's.

With the figurative tsunami of data flowing consistently and growing exponentially, Big Data, even if experts cannot agree on a concrete definition, agree that the the increase of data being produced presents new challenges to be tackled.

# 3 Section One

To describe Big Data and its main features, there can be no looking past what is commonly assigned the four V's, Volume, Velocity, Variety, and Veracity. Volume refers to how much data there is or the scale of data. *"From 2005 to 2020, the digital universe will grow by a factor of 300, from 130 exabytes to 40,000 exabytes"* (Gantz and Reinsel, 2012) and as it stands, according to Desjardins (2019), 44 zettabytes of data would comprise the digital universe by 2020.

Velocity describes the speed at which data is generated, as real time or near real time data from a multitude of sources is captured it much also be processed for this data to hold any value.

Variety indicates how many sources data can be captured from, it is not a single source with a single form of data, but in forms such as social media post, data captured from wearable smart devices, vehicle sensor data or smart phone GPS data in the form of unstructured data to name a few.

The last V is Veracity of the data, or is how accurate, trustworthy a source of data is and the quality of the data being used. Poor quality and inaccurate data is costly and will lead to poorly informed business decisions.

## 4 Section Two

Relational Databases (RDBMS) store data in highly structured forms, with tables, rows, columns and keys that determine relationships between them where as Hadoop uses key value pairs. As RDBMS use structured forms, it is used mainly for structured data unlike Hadoop which can be used for structured or unstructured data using Hadoop Distributed File System (HDFS) to segment data into

Schemas for both systems take different approaches also with RDBMS using a Schema on Write approach, this requires the tables, columns and keys to be prepared before consuming data, this slows the ingestion of data and makes it a ridged format but gives fast reads of the data. Hadoop uses Schema on Read meaning that the data does not have a structure to follow before being consumed.

Hadoop is an open source project that makes use of commodity hardware which is cost effective when scaling Horizontally to gain more performance. RDBMS requires proprietary licences to use and scales Vertically which adds more performance by adding more resources like CPU and memory but is limited by the limit of which the servers can handle making it more expensive to scale.

Hadoop can be applied effectively when there are extremely large, varied, unstructured datasets, upwards of terabytes of data, that needs processed to gain insights to the data. By taking advantage of the Hadoop eco-system incorporating multiple nodes, parallel processing of data can be achieved much faster then traditional RDBMS. For instance log files from websites that track user clicks and shopping cart data can be used to gain insights on how users interact with the site, allowing business decisions based on the data to improve users experiences therefore increasing users purchasing from the business.

## 5 Section Three

By implementing the framework for MapReduce within Hadoop, processing large datasets quickly can be achieved. MapReduce has two parts to the algorithm, the mapping stage takes the input data and distributes data across multiple nodes while the Reduce stage takes the output from the mapper and reduces data

Solving for an average using MapReduce would first require data to be loaded into the Hadoop Distributed File System (HDFS) using the put command, let us assume that the file containing the information is already loaded and split across the HDFS and and the code that will be used to compute the averages compiled.

For this instance the key value pair would consist of a zip code and price with the former being the key and the latter being the value. The first step after execution of the program would be for the mapper to work sequentially and assign records to single instances of the mapper and in the case a single record would be assigned to a single mapper. By implementing combiners during this stage, the output of the mapper can be reduced to save on the volume of data being transferred, but this intermediate key value pair will contain zip codes and values from each record.

The reducer has three tasks to perform, these are shuffling, sorting and reducing data. The shuffle phase of the process takes the output from the mapper and feeds the data to be sorted into groups by their keys. Both tasks are performed in parallel and do not wait for the mapper task to finish completely before starting. The sorting function will output data that is grouped by the key to the reducer.

The reducer can then compute the average price for each zip code in the final phase by keeping a running total of the value and a count as variables to then divide the totals to produce a final key value pair of zip code and average price as the output.

# 6 Section Four

Relational databases store data in tables, columns and rows, relationships are defined by keys within tables and need to be defined before the database is used. The structure of the database is ridged and is hard to change if the needs change also requiring full compliance to ACID transactions or the transaction will fail completely.

NoSQL databases come in four flavours, Document Store which uses JSON files to store data, Key-Value stores where data is stored using keys and values, Wide-Column Stores which uses tables, row and dynamic columns to store data and finally graph databases which use nodes and edges to form relationships between data. All these types of databases can handle data that is decentralised, has a large volume and is not ridged in types of data that need to be stored.

An application that could take advantage of NoSQL strengths could be social gaming, where user bases can explode overnight and data that is collected and stored starts to scale massively. In this case a RDBMS would not cope with a large influx of data and could cause failures to occur corrupting data that has already been collected, resulting in a user base not able to access the social game. Basic Availability Soft-state Eventual consistency (BASE) would be a prime candidate for this as basic availability means that there will be a response from a request, soft-state indicates that the system can change over time and eventual consistency states that they system will find its own consistent, NoSQLs flexibility with how it stores data, the volume of data collected and the velocity of the data coming from multiple sources would make it a prime candidate for this type of application.

## 7 Summary

The importance of Big Data cannot be understated, NoSQL databases flexibility in handling semi and unstructured data puts them at the forefront of this digital revolution that is happening. However, as relational databases excel in ACID transactions they cannot be forgotten or seen as obsolete. Both NoSQL and relational databases have use cases within the Big Data universe, but their use cases are very different and require forward thinking and planning to fully capture the strengths of both technologies.

# Reference

Desjardins, Jeff (Apr. 2019). *How much data is generated each day?* URL: https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/.

Gantz, John and David Reinsel (2012). "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east". In: *IDC iView: IDC Analyze the future* 2007.2012. Accesses 16 Feb 2021, pp. 1–16.

Garg, A, R Popli, and B S Sarao (Jan. 2021). "Growth of Digitization and its Impact on Big Data Analytics". In: *IOP Conference Series: Materials Science and Engineering* 1022. Accesses 16 Feb 2021, p. 012083. DOI: 10.1088/1757-899x/1022/1/012083. URL: https://doi.org/10.1088/1757-899x/1022/1/012083.

Kamilaris, Andreas, Andreas Kartakoullis, and Francesc X. Prenafeta-Boldú (2017). "A review on the practice of big data analysis in agriculture". In: *Computers and Electronics in Agriculture* 143. Accesses 16 Feb 2021, pp. 23–37. ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2017.09.037. URL: https://www.sciencedirect.com/science/article/pii/S0168169917301230.

McAfee, Andrew et al. (2012). "Big data: the management revolution". In: *Harvard business review* 90.10. Accesses 16 Feb 2021, pp. 60–68.

# Bibliography

Davenport, Thomas H, Paul Barth, and Randy Bean (2012). "How'big data'is different". In: Accesses 15 Feb 2021.

Ianni, Michele et al. (2020). "Fast and effective Big Data exploration by clustering". In: *Future Generation Computer Systems* 102. Accessed: 15 Feb 2021, pp. 84–94. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2019.07.077. URL: https://www.sciencedirect.com/science/article/pii/S0167739X19303838.

Oussous, Ahmed et al. (2018). "Big Data technologies: A survey". In: *Journal of King Saud University - Computer and Information Sciences* 30.4. Accesses 15 Feb 2021, pp. 431–448. ISSN: 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2017.06.001. URL: https://www.sciencedirect.com/science/article/pii/S1319157817300034.

Sagiroglu, S. and D. Sinanc (2013). "Big data: A review". In: *2013 International Conference on Collaboration Technologies and Systems (CTS)*. Accesses 15 Feb 2021, pp. 42–47. DOI: 10.1109/CTS.2013.6567202.