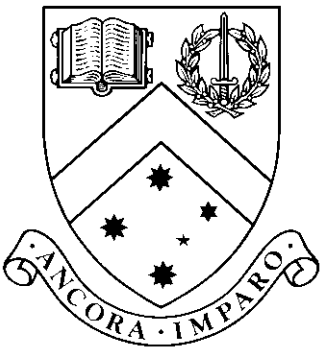
Clayton School of Information Technology
Monash University



Honours Literature Review — Semester 2, 2014

A study of the Hadoop ecosystem for pipelined realtime data stream processing

Jonathan Poltak Samosir
[2271 3603]

Supervisors: Dr Maria Indrawan-Santiago
Dr Pari Delir Haghighi

Contents

1 Introduction

The realtime processing of big data is of great importance to both academia and industry. Advancements and progress in modern society can be directly attributed to the rise of data, with those in society who control the data controlling the overall course of events. From seemingly inconsequential gains, at the macro level, such as the ability to more accurately predict the rise and fall of airline tickets [2], to those of utmost importance for society as a whole, such as predicting and tracking the spread of the Swine Flu Pandemic in 2009 more accurately that the United States Centers for Disease Control and Prevention could [4] [3]. It is example applications of big data processing like these that have been recognised by academics and organisations in industry alike, with the last decade seeing a major shift in research and development into new methods for the handling and processing of big data.

This paper will give a background on the types and classes of big data, as well as the various methods employed to process those given classes of data. We will more specifically focussing on the methods that are involved with the analysis and processing of realtime data streams, as opposed to the batch processing of big data. This paper will look into detail at previous work that has been done in the field of big data, specifically those works that have had a greater influence on the field as a whole. This includes both works looking specifically at the processing of streaming data, and works involving processed big data in batch mode, given that batch mode processing arguably led onto the current hot-topic of realtime stream processing.

This paper will be structured in two main sections. In §2, an overview of the different classes and types of big data will be presented. This includes an overview of the big data classes presented through others' findings as well as our own proposed classes for big data, based on the criticisms of those prior findings. In §3, an overview will be given of the major open-source big data processing systems. A special emphasis will be given on data stream processing systems (DSPSs), given that the main area of this research is focusing on realtime data processing, or data stream processing.

§4 will then give a discussion relating to future work we have planned to form data processing recommendations based on the classification of specific data classes. All of the sections will then be summarised in the conclusion in §5.

As a an outcome of this paper, we will identify a gap in previous research and development in the big data processing field, upon which our future work will attempt to work towards filling.

Note that for the duration of this paper, the terms "data" and "big data" will be used interchangeably, and can be assumed to refer to the same idea.

2 Big data types background

The data underlying big data can be categorised into a number of different classes or types. Each class of data comes with their own characteristics, and it is often possible to optimise the processing of each class of data by processing it in a specific way depending on those characteristics.To give an example of this, data that is expected to have highly iterative processing applied to it would benefit from a data processor that does not have to unnecessarily write to disk after every single iteration. The elimination of this I/O overhead is an example of the savings that could be gotten from correctly identifying the data class beforehand, and processing it accordingly.

Furthermore, particular types of data are generally only found in particular applications or use cases of big data processing.As this is the case, it narrows amount of classification needed, depending on the application that is being looked at. This will be elaborated

on in later parts of this section.

From preliminary research on looking at past work and literature in this area, it must be noted that in the subtopic of data classes and types for big data processing, there is a significant lack of research. Because of this, it was difficult to find conclusive information to base our research on, although it also highlights a further related area of research that may prove to be of benefit to look into.

The main piece of literature that this section sources is a white paper, from IBM Architects Mysore, Khupat, and Jain, published by IBM in 2013 [1]. The white paper is targeted towards beginners in the area of big data processing; much like the set of recommendations we intend to produce from this research project. It looks at identifying the different classes, or "formats" as it was labelled in the paper, that are commonly encountered in big data. For each of these formats what was identified was the underlying characteristics of the data, and it was noted that depending on those characteristics, the type of processing needed would vary.

The following subsections will highlight the different types or classes of big data, and give examples of their use in applications of big data processing.

2.1 Machine generated data

2.2 Web and social data

2.3 Transaction data

2.4 Human generated data

2.5 Biometric data

3 Big data processing background

Much work has been done in the area of big data processing. As discussed in

4 Relationships between big data classes and big data processing

5 Conclusion

References

[1] Big data architecture and patterns, part 1: Introduction to big data classification and architecture, Sept. 2013.

[2] Darlin, D. Airfares made easy (or easier). *The New York Times 1* (2006), B1.

[3] Mayer-Schönberger, V., and Cukier, K. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. An Eamon Dolan book. Houghton Mifflin Harcourt, 2013.

[4] Ritterman, J., Osborne, M., and Klein, E. Using prediction markets and twitter to predict a swine flu pandemic. In *1st international workshop on mining social media* (2009), vol. 9.