

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

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Big data types background</b>	<b>1</b>
<b>3</b>	<b>Big data processing background</b>	<b>1</b>
<b>4</b>	<b>Relationships between big data classes and big data processing</b>	<b>2</b>
<b>5</b>	<b>Conclusion</b>	<b>2</b>

# 1 Introduction

The realtime processing of big data is of great importance to both academia and industry. The modern economy is said to be built upon such data, with those who control the data controlling the overall course of events in our society. This has 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. For example, 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 often only found in particular applications or use cases of big data processing. This will be elaborated on in later parts of this section.

## 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