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 back to data. The value of data has become more apparent, and data has become a sort of currency for the information economy [8]. Hence, those in society who realised the value of data early immense power over the entire economy and thus society overall [5]. From seemingly inconsequential gains at the macro level, such as the ability to more accurately predict the rise and fall of airline tickets [3], 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 [7] [6]. 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.

2 Data types and characteristics background

2.1 Velocity, variety, volume, and veracity

Data, and more specifically, big data, are often characterised into what is known as the "four V's" [9]. These can be thought of as different "dimensions" of big data, and can be summarised as follows [4]:

- *Velocity:* The rate at which data is being collected and made available to the data consumers.
- *Variety:* The heterogeneity of data. Big data often exhibits substantial variations in both the structural level and the instance level (representations of real-world entities). This is often highlighted by data systems that depend on acquiring of data from a number of non-conforming, and sometimes unrelated, data sources.

- *Volume:* The amount of data that is obtained by the data consumer from the data source/s.
- *Veracity:* The quality, in terms of accuracy, coverage, and timeliness, of data that is consumed from the data source/s. Veracity of data can widely differ between sources.

While the four V's are often described in terms of big data, they can also apply in general to more traditional data warehousing and processing, albeit on a far smaller scale. In the domain of big data processing, data will exhibit signs of high velocity, variety, and volume [2], and hence the veracity of the data may also fluctuate. Meanwhile, in more traditional data processing, the scope may be limited, especially in terms of factors such as variety and, as a consequence, there is less need of an emphasis on veracity due to limited variety in data sources.

2.2 Classification of data

Data, in general, can be categorised into a number of different classes or types. In this paper, we will define the concept of a data class to mean the same as the terms of "data type", "data category", or "data format", as all terms were often used interchangeably in other literature.

Each class of data can be further defined and categorised via the characteristics they exhibit. Furthermore, these characteristics exhibited by data classes can be exploited and it is often possible to optimise the processing of each class of data by processing it using a specific method 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 optimisations that could be applied to the overall process from correctly identifying the data class beforehand, and processing it accordingly.

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

There is no concrete, universally accepted standard for the classification of data. While the study of big data processing could arguably be considered still in its infancy (or at least temperamental toddler stage), data handling and processing in general is relatively mature. From preliminary research on looking at past work and literature in this area, it must be noted that in the there is a significant lack of research on the classification of data. The literature that will be reviewed in this section is often not wholly focused on the idea of data classification, hence data classification is presented relative to whatever the overall topic of the literature is on. This is important to note, as one attempt at data classification may not be appropriate under a different context. This also explains the large variation in different classification attempts, although we will also highlight the recurring similarities between different data classification literature.

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 that we intend to produce from this research project. The paper looks at identifying the different data classes, or "formats" as they were 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 the type of processing needed would be dependent on those characteristics.

2.2.1 Data characteristics, from Mysore et al.

The characteristics of data, as put forward by Mysore et al., in [1], include the following:

Analysis type:

- Whether or not the data would be processed/analysed in realtime, or batched for later processing.
- Often this data class characteristic is dependent on the application of the data (*e.g.* The processing of social media data for the analysis of currently occurring events would want to be processed in realtime, regardless of the type of data that is involved).

Processing methodology:

- This characteristic involves the approach used when processing the data.
- Some examples of different processing methodologies include: predictive processing, analytical, ad-hoc queries, and reporting.
- Often the processing methodology for a particular class is determined by the business requirements or application of the data.
- Depending on the processing methodology used, many different combinations of big data technologies can be used.

Data frequency and size:

- The amount of data expected to arrive to the processing system, along with the speed and regularity of the incoming data.
- Knowing this characteristic beforehand can determine the methods for data storage and preprocessing, if needed.
- Examples of data frequency includes: on-demand data (social media), continuous/realtime (weather data, transactions), time-series (email).
- Considering the four V's, the characteristic of data frequency and size directly relates back to velocity and volume.

Content format:

- This characteristic relates back to the structure of the underlying data.
- Examples of data content format include: structured (JSON, XML), unstructured (human-readable literature), semi-structured (email).

Data source:

- This characteristic relates back to where the data originated from.
- As discussed previously in §2.1, the origin of data can have a great effect on whether or not that data is usable, as data often varies greatly, especially when many different sources are used which may or may not conform to a specific content format.
- Another thing that is dependent on the data source is whether or not the data can be trusted.
- Considering the four V's, the characteristic of data source directly relates back to veracity and variety.

2.2.2 Data classifications, from Mysore et al.

The following table will highlight the different classes of data put forward by Mysore, et al., in [1]. The identified characteristics for each class of data will also be given.

Data class	Explanation	Characteristics
Machine generated data	• Data that is automatically generated as a by-product of some interaction with a machine. • While Mysore et al. present this as being a distinct class in itself, it could be argued that this class is an umbrella class which many other data classes presented in their paper fall under. This will be touched upon further in later sections.	• Structured data (JSON, XML). • Frequency of data varies depending on application.
Web and social data	• Data that is automatically generated through use of the Internet or social media, such as Facebook or Twitter.	• Unstructured text (long: blogs, short: microblogs, Facebook). • Miscellaneous multimedia (video, image, audio). • On-demand frequency. • Can be continuous feed of data in cases such as Twitter.
Transaction data	• Data that is automatically generated as a by-product of transactions, such as money transactions or otherwise.	• Structured text (JSON, XML, logs). • Continuous feed.
Human generated data	• Data that is solely produced by humans. • Examples of human generated data, as it is defined here, include such things as music, literature, recordings, and emails.	• Unstructured text (mail, literature). • Miscellaneous multimedia (audio, video, images). • Semi-structured text (email, online messaging services). • On-demand frequency.
Biometrics data	• Data that relates to human bioinformatics.	• Structured data. • On-demand frequency. • Continuous feeds of data in cases such as persistent health monitoring sensors (*i.e.* hospital patients).

The classes and characteristics of data presented by Mysore et al., in [1], are highly oriented towards industry and business users, coming from an IBM-published paper. While this is not an issue as such, as noted earlier in this section, these characteristics and data classes are defined within the domain relevant to this paper. As such, they may not be as relevant or appropriate for usage in other, non-business domains.

2.2.3 Characteristics of data, from Géczy

Géczy attempts to characterise

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] Beyer, M. Gartner says solving 'big data'challenge involves more than just managing volumes of data. *Gartner http://web. archive. org/web/20110710043533/http://www. gartner. com/it/page. jsp* (2011).

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

[4] Dong, X. L., and Srivastava, D. Big data integration. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on* (2013), IEEE, pp. 1245–1248.

[5] Lievesley, D. Increasing the value of data. *BLRD REPORTS 6122* (1993), 205–205.

[6] 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.

[7] 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.

[8] St Amant, K., and Ulijn, J. M. Examining the information economy: Exploring the overlap between professional communication activities and information-management practices. *Professional Communication, IEEE Transactions on 52*, 3 (2009), 225–228.

[9] Wang, L., Zhan, J., Luo, C., Zhu, Y., Yang, Q., He, Y., Gao, W., Jia, Z., Shi, Y., Zhang, S., et al. Bigdatabench: A big data benchmark suite from internet services. *arXiv preprint arXiv:1401.1406* (2014).