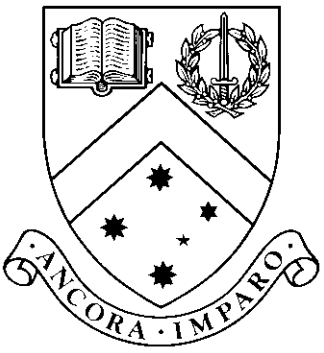
Clayton School of Information Technology
Monash University



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 [11]. Hence, those in society who realised the value of data early immense power over the entire economy and thus society overall [8]. From seemingly inconsequential gains at the macro level, such as the ability to more accurately predict the rise and fall of airline tickets [5], 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 [10] [9]. 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" [12]. These can be thought of as different "dimensions" of big data, and can be summarised as follows [6]:

- *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 to more traditional data warehousing and processing in general, 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.

As will be made clear in the following sections, a lot of the identified classes and characteristics of data directly relate back to these four V's, whether or not it was intentional by the original authors. These can be considered the underlying features of many characteristics of data, both in the sense of big data and traditional data.

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 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.

2.2.1 Characteristics of data, from Mysore et al.

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.

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

2.2.1.1 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).

2.2.1.2 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.

2.2.1.3 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.

2.2.1.4 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).

2.2.1.5 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 Classes of data, from Mysore et al.

The following table highlights the different classes of data put forward by Mysore, et al., in [1]. The table organises each class, along with giving a brief explanation of the class. Furthermore, each class is related back to the previously explained characteristics in an attempt to show the connections between class and underlying characteristics.

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 or even business domains with a different focus on data.

2.2.3 Characteristics of data, from Chen et al.

The second paper sourced is a paper from Chen, Chiang, and Storey, focusing on the impact of big data in the field of business intelligence and analytics [4]. Similarly to the paper looked at in §2.2.1, there is an emphasis on data classes and how they relate to the area of business and organisations. However, this paper has more of an explicit focus on business, being in published in the area of business intelligence and analytics (BI&A). BI&A in itself is a highly data driven field, where data is gathered and analysed to help make informed business decisions [13].

In the paper, Chen et al., discuss the evolution of the field of BI&A, which they categorise into three distinct stages. BI&A 1.0, being the first of the three, focuses on more traditional data processing and analysis. This includes highly structured and relational data. BI&A 2.0 involves more unstructured, web-based content with the rise of "Web 2.0" technologies, including social networks and opinion pieces, such as blogs. BI&A 3.0 looks at more mobile and sensor-based data. This data differentiates itself mostly to do with characteristics such as location-based data and data that is highly context dependent.

Chen et al., elaborate on these different stages of BI&A evolution through showing the major BI&A applications for the previously mentioned evolutionary stages. For each of the BI&A applications presented, they attempt to show the classes of data which are important for the particular application, and subsequently the characteristics associated which each class. The classes and characteristics of data, shown by Chen et al., in relation to BI&A will be presented here. They will be presented in terms of the BI&A application of which they are categorised under.

2.2.3.1 E-Commerce and Market Intelligence:

Types of data include:

- Website logs and analytics data.
- User activity logs for e-commerce websites.
- User transaction records.
- User-generated content, such as reviews, feedback.

Characteristics of the data include:

- Structured web-based data (transactions records, logs, network information).
- Unstructured user-generated content (reviews, feedback).

2.2.3.2 E-Government and Politics 2.0:

Types of data include:

- Government information, such as statistics.
- Rules and regulations.
- Citizen-generated content, such as feedback, comments, and requests.

Characteristics of the data include:

- Fragmented data sources (think high data variety).
- Unstructured data (citizen-generated content).
- Rich textual content.

2.2.3.3 Science & Technology:

Types of data include:

- Machine-generated data from tools and instruments.
- Sensor data.
- Network data.

Types of characteristics include:

- High velocity data collection from instruments and tools and sensors.
- Structured data, often formatted in uncommon structures.

2.2.3.4 Smart Health and Wellbeing:

Types of data include:

- Genomics and sequence data (DNA sequences).
- Electronic health records.
- Health and patient social media.

Types of characteristics include:

- Varying, but interrelated, data.
- Data specific to individual patients.

2.2.3.5 Security and Public Safety:

Types of data include:

- Criminal record data.
- Statistical data (crime maps).
- Media content relating to crime (news articles).
- Cyber-crime data (computer viruses, botnet data).

Types of characteristics include:

- Highly sensitive information (identity data).
- Incomplete and deceptive content (speculative media content).
- Multilingual content.

As can be seen from the data classes and characteristics identified by Chen et al., in [4], there is a far greater variation to those previously presented by Mysore et al., in [3]. As explained earlier, this is mainly because of the more domain specific content of this piece of literature, while the paper from Mysore et al., while still having underlying tones of business and industry, had a less explicit focus on their particular domain.

2.2.4 Characteristics of data, from Géczy

Coming away from the business point-of-view, Géczy attempts to characterise data, and more specifically big data, in a more generic way [7]. He uses what he labels as "aspects" to determine what he believes to be the deciding characteristics of data, in terms of the way they should be processed and also simply their intrinsic traits.

Géczy uses the following aspects to determine the different intrinsic characteristics of data:

2.2.4.1 Sensitivity:

- Relates to whether or not given data contains sensitive information, *i.e.* personally identifiable information, confidential information, etc.
- The sentivity of the data determines the requirements relating to how it should be handled.
- Often it is either a legal requirement, or in the owners' interest, to keep protected the handled data deemed sensitive.

2.2.4.2 Diversity:

- Relates to the range of different data elements present within the data.
- The example given explains the ability of smart phones to produce highly diverse data; *e.g.* audio, video, location data, gyroscopic data, etc.
- Having high diversity in data can both be beneficial and detrimental; diversity can add factors of complexity, although also makes for a more rich dataset.
- Note that this data characteristic relates directly back to the *Variety* dimension, of the four V's.

2.2.4.3 Quality:

- Quality characteristics of data are defined to be features that affect data quality; *e.g.* completeness, accuracy, timeliness.
- Often the quality of data may be subject to the qualitative metrics of an organisation, or predefined standards.
- The quality of data relates back to the *Veracity* dimension, of the four V's.

2.2.4.4 Volume:

- Volume refers to the size of data in terms of its basic forms of measurement, bits and bytes.
- Volume is an important characteristic to take into consideration when it comes to determining the type of processing needed.
- Volume, as the name suggests, directly relates back to the Volume dimension of the four V's.

2.2.4.5 Speed:

- Data speed refers to the inflow and outflow speeds; inflow being the data that is being acquired, while outflow being the data leaving the system (often results of computations).
- Different classes of data often require different data speeds. *e.g.* audio is often streamed at a far lesser speed than video, due to the relatively low amount of data in audio when compared with video.

2.2.4.6 Structure:

- Structure relates to whether data are in structured or unstructured formats.
- Generally unstructured data is more suitable for human consumption, such as literature or music.
- Structured data is usually structured in such a way that it is easily able to be parsed by an algorithm, often automated by computers.
- The structure of data directly relates to the difficulty of processing that data, as unstructured data usually will need some pre-processing or artificial intelligence to process.

Géczy later goes on to talk about the aspects of data that relate to data processing. This will be looked at further in §**??**

Overall, Géczy looks at data characteristics, not from any particular perspective, but from one that attempts to capture the interests and be relevant to a number of disciplines. This impartiality is a nice refreshment from most other literature available on the topic, which have been shown to have been looking at data classification from a certain point-of-view. However, this should not be misinterpreted as a criticism of the previous literature. It is simply that the classes of data identified in other author's literature was more appropriate for the topic on which the rest of their research was focussed on. Hence, the way they treated data changed accordingly. The paper presented by Géczy, simply titled *BIG DATA CHARACTERISTICS*, was focussed on nothing other than characteristics of data, hence there was no reason to attempt to classify those characteristics based on any other domain-related biases. Additionally, Géczy's paper was published in a notable interdisciplinary journal, rather than one aimed at a particular discipline. This difference in terms of impartiality is the important difference to note with this paper, especially as we revisit this paper in §**??**

2.2.5 Criticisms of previously presented models

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] Bifet, A. Mining big data in real time. *Informatica 37*, 1 (Mar. 2013), 15+.

[4] Chen, H., Chiang, R. H., and Storey, V. C. Business intelligence and analytics: From big data to big impact. *MIS quarterly 36*, 4 (2012), 1165–1188.

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

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

[7] Géczy, P. BIG DATA CHARACTERISTICS.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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).

[13] Watson, H. J. Tutorial: Business intelligence-past, present, and future. *Communications of the Association for Information Systems 25*, 1 (2009), 39.