Course: Computing clusters, grids, and clouds
Lecturer: Andrey Shevel
ITMO University – St. Petersburg

Problems to store, transfer and process the Big Data

Giang Tran – tttgiang2510@gmail.com

1. Introduction

In the last few years, big data has emerged due to the fact that we are living in a society which makes increasing use of data intensive technologies. Big data is being generated by multiple sources such as social media, systems, sensors and mobile devices at an alarming velocity, volume and variety. Big data is defined as data sets that are so large and complex that traditional database management concepts and tools are inadequate. It can bring various benefits to business organization. Insights from big data can enable all employees to make better decisions, optimizing operations, preventing threads and fraud, deepening customer engagement and capitalizing on new sources of revenue.

However, big data, due to its various properties, has many issues such as storage, transfer, processing, real time analytics, search, sharing, visualization, and information privacy. In this report, we are going to present big data characteristics in section 2. In the next section, 3 main problems of big data including storage, transfer and processing will be discussed. Finally, section 4 concludes the report.

2. Big data characteristics

Big data is often described using five Vs: Volume, Velocity, Variety, Veracity and Value.

- **Data Volume**: Data volume is the amount of data available to an organization. Currently, the data is in petabytes (10^{15}), exabytes (10^{18}) and is supposed to increase to zettabytes (10^{21}). For example, Facebook currently processes more than 500 terabytes of new data each and every day. This increase in data volume makes data sets too large to store and analyze using traditional database technology. With big data technology we can now store and use these data sets with the help of distributed systems, where parts of the data is stored in different locations and brought together by software.
- **Data Velocity**: Data velocity is the concept dealing with the speed of data coming from various different sources, the speed of data flows and data processing. Different types of big data may need to be processed at different speeds.

- **Data Variety**: Data variety measures the richness of data representation such as text, image, video, voice. In the past, we only focused on structured data that neatly fitted into tables or relational database. Nowadays, big data can include traditional data (structured data), semi-structured data and unstructured data from various sources such as social networks, logs, emails, and document.
- **Data Veracity**: Data veracity refers to the messiness or trustworthiness of the data. With many forms of big data, quality and accuracy are less controllable. The volumes often make up for the lack of quality or accuracy. Take Twitter for example, posts are with hashtags, abbreviations, typos and colloquial speech as well as the lack of reliability and accuracy of content. However, big data now allows us to work with these types of data.
- **Data Value**: Data value measures the usefulness of data in decision making processes. User can run certain queries in the database, deduct critical results and find the business trend, which can help them to adjust their business strategies.

Based on the characteristics, big data is also defined as “high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization” [1].

3. Problems

The characteristics of big data put forward many issues. In this section, we discuss the problems to storage, transfer and process the big data.

3.1. Storage

There are three problems in terms of big data storage. Firstly, current technologies of data management systems are not able to satisfy the needs of big data, and the increasing speed of storage capacity is much less than that of data. Table 1 below demonstrates some examples of big data.

Table 1 – Some examples of big data [2]

Data set/domain	Description
Large Hadron Collider/Particle Physics (CERN)	13-15 petabytes in 2010
Internet Communications (Cisco)	667 exabytes in 2013
Social Media	12+ terabytes of tweets every day and growing. Average retweets are 144 per tweet.
Human Digital Universe	1.7 Zbytes (2011) -> 7.9

	Zbytes in 2015
Others	RFIDS, smart electric meters, 4.6 billion camera phones w/ GPS

Current disk technology limit is 10 terabytes (10^{12}) per disk. Therefore, an Exabyte (10^{18}) would require 10^5 disks. Even if an Exabyte of data could be processed on a single computer system, it would be unable to directly attach the requisite number of disks. Secondly, previous computer algorithms are not able to effectively store big data because of the heterogeneity of the big data. Last but not least, how to re-organize data is another problem in big data management.

The crucial requirements of big data storage are (1) it can handle very large amount of data and keep scaling to keep up with data growth, and (2) it can provide the input/output operations per second (IOPS) necessary to deliver data to analytics tools. Currently, Google, Facebook and Apple – the largest big data users are using hyperscale computing environments for big data storage [3]. Hyperscale storage provides rapid, efficient expansion to handle big data use cases such as Web serving, database applications, data analysis, high-performance computing [4]. The hyperscale computing environments run Hadoop and Cassandra as analytics engines.

3.2. Transfer

Conventionally, we have 2 methods of transfer data: (1) transfer via the network, using TCP-based transfer methods such as FTP or HTTP and (2) use storage medium such as hard disk drives. However, in terms of big data, both options are impractical and impose costs on the business because of delayed access to data, risks of loss or damage, and the need to invest resources.

Current communication network are unsuitable for such massive volume of big data, introducing unacceptable delays in moving data. Assuming that a 1 gigabit per second network has an effective sustainable transfer rate of 80%, the sustainable bandwidth is about 100 megabytes per second. Thus, transferring an Exabyte would take about 2800 hours, if we assume that a sustained transfer could be maintained. It would take longer to transfer the data from a storage point to a processing point than it would to actually process it [2].

There are two approaches to tackle problem to transfer big data: (1) process the data “in place” and transmit only the resulting information and (2) perform triage on the data and transmit only that data which is critical to downstream analysis. In either case, integrity and provenance metadata should be transmitted along with the actual data. In addition, there are currently different techniques to transfer a large amount of data across the globe such as parallel transmission techniques used on the Internet and NICE model for big data transfer. The parallel transmission

techniques are designed to grab as much bandwidth as possible by establishing concurrent data transfer streams. Whereas, the NICE model ensures that big data transfers only occur during low demand periods when there is idle bandwidth that can be utilized for these large transfers [5].

3.3. Process

Processing big data is a major challenge, even more than storage and transfer problems. Currently, being able to extract real-time information from a large stream of data remains difficult. Speed is a crucial demand when processing a query in big data. However, it is impossible to traverse all the related data in the whole database in a short time. For instance, an Exabyte of data needs to be entirely processed. Assuming that a processor expends 100 instructions on one block (8 words) at 5 gigahertz, end-to-end processing time would be 20 nanoseconds. The time required to process an Exabyte of data is around 635 years [2]. Besides, the traditional serial algorithm is inefficient for the big data.

Processing big data requires extensive parallel processing and new analytics algorithms to provide timely and actionable information. Application parallelization and divide-and-conquer are natural computational paradigms for tackling the problem. Currently, distributed computing framework such as Apache Hadoop and Google MapReduce are being used to process large amounts of data in parallel.

4. Conclusion

Big data is not a new concept but very challenging. This report describes the problems in storing, transferring and processing the big data, caused by its characteristics. It is a fundamental problem that data is too massive to store, transfer and process conventionally. Additionally, some current approaches and solutions to address the problems are also discussed. The problems seem to be solvable in the near-term, but present a long-term challenges that require a lot of research from industry and academia.

5. References

[1] MA Beyer and D. Laney and, "The importance of "big data": a definition," *Stamford, CT: Gartner,* pp. 2014-2018, 2012.

[2] S. Kaisler, F. Armour, J. a Espinosa, and W. Money, “Big Data: Issues and Challenges Moving Forward,” 46th Hawaii Int. Conf. Syst. Sci., pp. 995–1004, 2013.

[3] A. Adshead, "Big data storage: Defining big data and the type of storage it needs," [Online]. Available: http://www.computerweekly.com/podcast/Big-data-storage-Defining-big-data-and-the-type-of-storage-it-needs. [Accessed 1 June 2016].

[4] C. Sliwa, "Understanding stripped-down hyperscale storage for big data use cases," [Online]. Available: http://searchstorage.techtarget.com/podcast/Understanding-stripped-down-hyperscale-storage-for-big-data-use-cases. [Accessed 1 June 2016].

[5] K. K. Reddi and D. Indira, "Different Techniques to Transfer Big data: a Survey," *Int. Journal of Engineering Research and Applications,* vol. 3, no. 6, pp. 708-711, 2013.

[6] "Big data", Wikipedia, 2016. [Online]. Available: https://en.wikipedia.org/wiki/Big_data#cite_note-10. [Accessed: 04- Jun- 2016].