SOUTHEAST UNIVERSITY-MONASH UNIVERSITY

JOINT GRADUATE SCHOOL (SUZHOU)

# Big Data: A Review

**Bimin Yin**

**9th May 2014**

# Big Data: A review


## Abstract

Big Data are huge volume of rapidly blooming and complex structured data sets coming from multiple heterogeneity sources, which require new architectures and technologies to conduct the analysis process. Big Data can imply useful information and patterns which can be mined effectively and efficiently by data mining. This essay will examine basic concepts as well as characteristics of Big Data, and then examine a dominant platform (Hadoop) and a major technology (Map-Reduce) of Big Data in detail and analyze opportunities and challenges of Big Data by investigating research publications on Big Data. Conclusion will be drawn by proposing some assessments on Big Data and suggestions to corporations.



**Keywords:** Big Data, Map-Reduce, Hadoop, Data Mining.


## Introduction

With the coming of the Information Era, Big Data are expanding in engineering, science and economy domains. Big Data are extremely large and complex data sets which have characteristics such as large emerging velocity and underlying patterns implied. New platforms and technologies are brought up with Big Data,as well as opportunities and challenges, which will be discussed in detail.

## Scope and method

The essay is constructed as follows: First, basic concepts of Big Data will be brought up together with characteristics of Big Data. Second, this essay will examine dominant technologies and platforms of Big Data implementation (Map-Reduce and Hadoop). Third, an analysis on opportunities and challenges of Big Data will also be conducted by investigating relevant research publications. Finally, conclusion will be drawn by proposing assessments on Big Data as well as suggestions to corporations according to opportunities and challenges.

# Body of the review

## Basic concepts of Big Data

Technically, the Big Data technology is defined as a large amount of data implying valuable information which can be deeply analyzed and further utilized by corporations or organization to get an advantage over the competition (Katal, Wazid, & Goudar, 2013). The ‘Big Data’ term which is being used nowadays is a kind of misunderstanding as it indicates only the size of the data but does not put too much of attention to its other properties such as variety and velocity.

The process of research into Big Data to reveal hidden patterns and secret correlations is named as Big Data analytics (Sagiroglu & Sinanc, 2013). Big Data and the analysis are at the center of modern science and business.

## Characteristics of Big Data

Big Data are characterized by five major properties discussed in following paragraphs.

### A. Variety and Heterogeneity of Data

Data types of Big Data are various. Katal et al. (2013) argue that Big Data produced and collected are not of a single category because they not only contain traditional structured data, but also include semi-structured or unstructured data from various resources like web pages, social media and data from sensors. Different structures of data increase the processing complexity of Big Data system and unstructured data are especially difficult to be handled by the existing traditional data mining systems.

### B. Volume and scale

The word ‘big’ in Big Data itself indicates the volume and scale of Big Data. Data volume has been increasing exponentially: up to 2.5 Exabyte of data generated and stored within one day, and this number is expected to double within 40 months by

2015 (Tekiner & Keane, 2013). Handling such large amount of data with rapidly increased volume adds additional complexity to traditional systems which have limited processing speed.

### C. Velocity

Velocity in Big Data is a concept corresponding to the speed of the receiving data from various sources (Katal et al., 2013). This characteristic is being limited not only to the speed of incoming data but also speed of data processing. For example, data from the transactions of a bank would be constantly stored into the database or sent to the analysis system for further processing. Thus traditional database systems are neither capable nor suitable enough to perform analytics on dynamic data.

### D. Complexity

Tekiner and Keane (2013) argue that the complexity of Big Data indicates that data have complex structures, which is also quite a challenge to pre-processing. Discovering and correlating potential relationships and patterns of data, which is necessary for corporations, also increases the complexity of Big Data system.

### E. Value

Big Data are valuable because data collected by different organizations can imply important information and potential patterns by conducting data analytics and data mining (Katal et al., 2013). Corporations, especially business leaders, can make more profit from potential patterns discovered by mining their larger amount of data.

## Dominant technologies and platforms of Big Data

Big Data systems are supported and implemented, dominantly, by Map-Reduce and Hodoop, which will be discussed in following paragraphs.

### Map-Reduce

Map-Reduce is introduced by Google in order to process Big Data. It implements a programming model which allows distributed processing for Big Data.

Katal et al. (2013) present basic concepts and the procedure of Map-Reduce. Map-Reduce basically performs two different tasks: the Map task and the Reduce task. The Map task produces a sequence of key-value pairs from the input and this is done according to the code written for Map function. These values are collected by master controller and sorted by keys and divided among Reduce tasks to assure that the same key values ends with the same Reduce tasks for future processing by Reduce tasks.

Currently, Big Data processing mainly depends on parallel programming models like Map-Reduce, as well as providing a cloud computing platform of Big Data services for the public. The Map-Reduce technology is suitable for distributed tasks and dominant programming languages such as Java and C++ now have Map-Reduce implementations. The concept of Map-Reduce paves the way for development of highly parallel, reliable and distributed applications on large datasets (Wu, X., Zhu, Wu, G.-Q., & Ding, 2014).

**Hadoop**

Hadoop is dominant free implementation of Map-Reduce by the Apache foundation, but it is not a replacement for the database or warehouse (Sagiroglu & Sinanc, 2013). Hadoop mainly consists of: File System (HDFS) and Programming Model (Map Reduce) and thus provides: a reliable shared storage and analysis system.

The popularity of Map-Reduce and Hadoop has made processing on Big Data more effectively and efficiently.

## Opportunities of Big Data

Wu et al. (2014) argue that Big Data offers opportunities to process less structured data such as social media, e-mails and photographs which cannot be done by traditional relational databases. Major business intelligence companies, such as IBM and Oracle, have produced their own products to help corporations to acquire and process customers' data to discover potential patterns and hidden relationships.

The Web technology provides opportunities for Big Data as well (Sagiroglu & Sinanc, 2013). For example: potential patterns analysis of Big Data such as mining user shopping patterns for more targeted advertising relies on the cookie technology and the semantic web.

In addition, predictive analytics on traffic flows or crime prediction by mining historical data indicate opportunities of Big Data in fields besides computer science.

## Challenges of Big Data

Despite considerable opportunities, the Big Data technology still have to face challenges because of characteristics of Big Data. Five major challenges will be discussed in following paragraphs.

### A. Privacy and Security

Katal et al. (2013) argue that the issue of privacy and security is the most challenging one with Big Data. Big Data used by the government will increase the possibility of certain tagged people to suffer from bias or even having no knowledge that they are being discriminated. Personal information such as email, phone number and address, which should be remaining private, can be leaked if data collected are managed carelessly.

### B. Data access and sharing of information

If data are to be used for making accurate decisions, it becomes necessary that data should be available in an accurate, complete and timely manner. This makes the Data management a more complex to make data access in a standardized and timely manner with standardized APIs. Sharing of data between corporations is awkward because of the need for competition. Sharing data about their clients and operations threatens the culture of secrecy and competitiveness.

### C. Storage and processing issues

The storage available is not large enough for storing the large amount of data

which are being produced by almost everything (Tekiner & Keane, 2013). Despite of technologies of Map-Reduce and Hadoop, how to pre-process and mine Big Data remains to be challenging.

### D. Skill Requirement

New skills are required because the Big Data technology is an emerging and developing technology. Skills should not be limited to technical ones but also should include research, analytical and creative ones. These skills need to be well acquired by employees thus training programs held by corporations or organizations are required. Moreover, universities need to introduce not only concepts of Big Data, but also implementation by conducting experiments.

### E. Technical Challenges

Wu et al. (2014) argue that technical challenges, including heterogeneous data, the quality of data and complex algorithms, are more significant.

#### a) Heterogeneous Data:

How to pre-process heterogeneous data including traditional structured data, semi-structured data and unstructured data is a major technical challenge. Structured data can be stored and processed effectively and efficiently, while unstructured data cannot, which are completely raw and unorganized. The process of converting all unstructured data into structured data is costly.

#### b) Quality of Data

Better results of predictive analysis or decision making in business require more data, which need costly storage and processing capability. The Big Data technology basically needs to focus on qualified data storage rather than having very large irrelevant data to ensure that better analysis and predictions can be conducted. This further leads to various questions like how it can be ensured that what kind of data is relevant, how much data would be enough for decision making and how to judge whether data stored are qualified or not.

**c) Complex algorithms**

Data mining algorithms usually need to scan through the training data to obtain the statistics for future analysis and prediction, which needs intensive computing capability to access the large-scale data frequently. Mining algorithms of Big Data can be very complex, such as parallelization with multiplicative algorithms proposed by Luo, Ding, and Huang (2012) and a granular computing (GrC) approach of a mathematical framework proposed by Xie, J., Chen, Xie, G., and Lin (2013).

# Interpretation and conclusion

With the coming of the Information Era, Big Data are expanding in engineering, science and economy domains. Big Data have characteristics of large-volume, outstanding growing velocity, heterogeneous data types, requiring complex computing and implying valuable information.

Information implied within Big Data together with the widely use of web technology provides opportunities for corporations and organizations, especially business leaders such as Google, IBM and Microsoft. Potential patterns and further prediction can be conducted by analyzing and mining Big Data, which can make more profit for large corporations and organizations. Characteristics of rapid growing velocity, various data types and complex computing capability pose challenges for Big Data. The major challenge of Big Data is the issue of privacy and security which is increasingly essential and disputable in the Information Era. Other challenges involving the issue of storage capability and complexity of processing can be overcome in the near future with the development of relevant platforms and technologies such as Map-Reduce and Hadoop.

Business leaders need to take their advantages of larger amount of customers, larger volume of data and more powerful computing capability supporting by better infrastructures. The era of Big Data has arrived.

# References

Katal, A., Wazid, M., & Goudar, R. (2013). Big data: Issues, challenges, tools and Good practices. *Proceedings of the 2013 Sixth International Conference on Contemporary Computing* (IC3), 20(1), pp. 404-409.

Luo, D., Ding, C., & Huang, H. (2012). Parallelization with Multiplicative Algorithms for Big Data Mining. *Proceedings of the 2012 IEEE 12th International Conference on Data Mining* (DM), 11(1), pp. 489-498.

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *Proceedings of the 2013 International Conference on the Collaboration Technologies and Systems* (CTS), 15(3), pp. 42-47.

Tekiner, F., & Keane, J. A. (2013). Big Data Framework. *Proceedings of the 2013 IEEE International Conference on the Systems, Man, and Cybernetics* (SMC), 23(2), pp. 1494-1499.

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *Proceedings of the IEEE Transactions on Knowledge and Data Engineering*, 26(1), pp. 97-107.

Xie, J., Chen, Z., Xie, G., & Lin, T. Y. T. (2013). Knowledge Mining in Big Data—A Lesson From Algebraic Geometry. *Proceedings of the 2013 IEEE International Conference on the Granular Computing* (GrC), 21(1), pp. 362-367.