# Challenges and Opportunities of Big Data

[1]Md Ibrahim Khalil, [2]R. Young Chul Kim, [3]*ChaeYun Seo

***Abstract***

*Big Data is a new concept in the global and local area. This field has gained tremendous momentum in the recent years and has attracted attention of several researchers. Big Data is a data analysis methodology enabled by recent advances in information and communications technology. However, big data analysis requires a huge amount of computing resources making adoption costs of big data technology. Therefore, it is not affordable for many small and medium enterprises. We survey the concepts and characteristics of Big Data along with a number of tools like HADOOP, HPCC for managing Big Data. It also presents an overview of big data like Characteristics of Big data, big data technology, big data management tools etc. We have also highlighted on some challenges and opportunities related to the fields of big data.*

***Keywords:*** *Big data, Hadoop, HDFS, MapReduce, Big data management, Big data analysis*

## I. Introduction

Today, the continuous growth of computational resources generates an enormous amount of data every day. Data is available in abundance, but it is difficult to extract useful information from such Big Data. For example, Twitter processes over 70M tweets per day, thereby generating over 8TB daily [1]. Since 2020, ABI Research estimates that there will be more than 30 billion connected devices [2]. These Big Data possess tremendous potential in terms of business value in a variety of fields such as health care, biology, transportation, online advertising, energy management, and financial services [3, 4]. However, traditional approaches are struggling when faced with these massive data.

The concept of Big Data has been proposed to store, manage, visualize, and analyze such enormous volumes of data generated quickly per day. Cloud Computing provides a platform to the users that is accessible and flexible for storage and processing of such Big Data applications. Most researchers in the database community have then moved on to other problems. “Big Data” is reborn in the 2000’s, with massive, Web-driven challenges of scale driving system developers at companies such as Google, Yahoo!, Amazon, Face- book, and others to develop new architectures for storing, accessing, and analyzing “Big Data”

The Internet of things (IOT) is all set to bring the revolution in the information industry as it provides the interconnectivity of physical objects and allows them to exchange the data with the other connected devices. According to [5], the number of connected devices in 2014 was 3. 7 billion and this number will be estimated to reach 25 billion till 2021. These interconnected devices, growing exponentially in number, are also giving rise to new types of data in large volumes. Big Data here becomes extremely important to convert this data into information.

This paper is organized as follows: Section 2 gives an overview of Big Data Technologies. In Section 3, we discuss some emerging trends in the field of Big Data. Section 4 discusses big data management tools. Section 5 mentions Challenges and Opportunities of big data. Section 6 concludes this paper.

[1] *Graduate student, in Dept. Of SW and Communications Engineering, Hongik University (ibrahim@selab.hongik.ac.kr)*
[2] *Professor in Dept. Of SW and Communications Engineering, Hongik University (bob@hongik.ac.kr)*
[3* Corresponding Author] *Professor in Dept. Of SW and Communications Engineering, Hongik University (chaeyun@hongik.ac.kr)*

Received: June. 10, 2020, Revised: June 22, 2020, Accepted: July. 29, 2020

# II. Overview of Big Data Technologies

## 2.1 Characteristics of Big data.

Volume: organizations [6] collect data from a variety of sources including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden. The name 'Big Data' itself is related to a size which is enormous. Size of data plays very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon volume of data. Hence, 'Volume' is one characteristic which needs to be considered while dealing with 'Big Data'.

Velocity: the term 'velocity' refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data. Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

Variety: data comes in all types of formats – from structured datasets data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions. Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

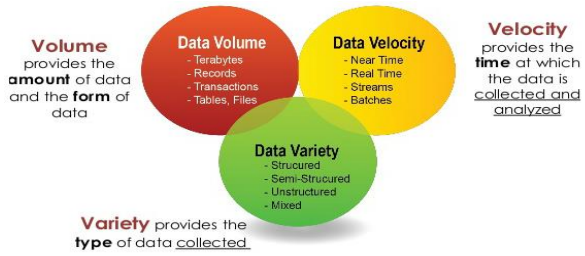


Figure 1. Characteristics of Big data

## 2.2 Big Data Today

Big data can be found both in the public and private sector. From targeted advertising, education, and already mentioned massive industries (healthcare, insurance, manufacturing or banking), to real-life scenarios, in guest service or entertainment. By the year 2020, 1.7 megabytes of data will be generated every second for every person on the planet, the potential for data-driven organizational growth in the hospitality sector is enormous [7]. With the advanced developments in Earth observation systems, various Earth data have been gathered at a high velocity from five major sources**:**

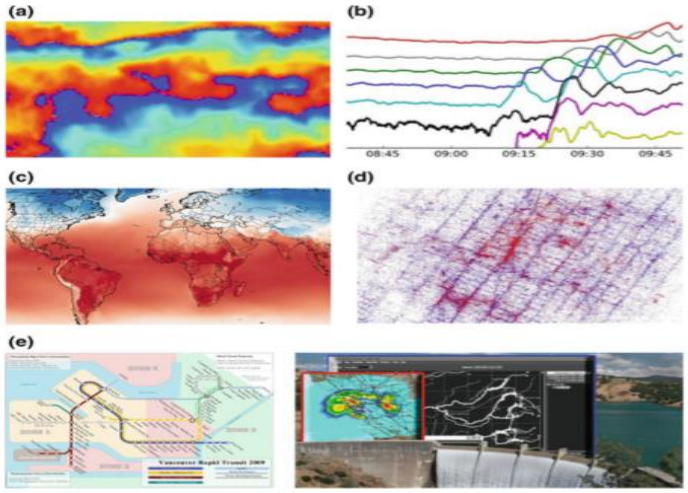


Figure 2. Big Earth data sources: (a) remote sensing data, (b) in situ data (c) simulation data; (d) social media data and (e) infrastructure data

### 2.3 Some Statistics on Big Data

According to Neilson Online currently there are more than 1,733,993,741 internet users. In 2020, the big data market is expected to grow by 14%. Few numbers to understand how much data is generated every year.

- Email
  - ✓ 90 trillion – The number of emails sent on the Internet in 2020
  - ✓ 247 billion – Average number of email messages per day.
  - ✓ 1.4 billion – The number of email users worldwide.
  - ✓ 100 million – New email users since the year before.
- Websites
  - ✓ 234 million – The number of websites as of December 2019.
  - ✓ 47 million – Added websites in 2020
- Domain name
  - ✓ 81.8 million – .COM domain names at the end of 2019.
  - ✓ 12.3 million – .NET domain names at the end of 2019.
  - ✓ 7.8 million – .ORG domain names at the end of 2019
  - ✓ 76.3 million – The number of country code top-level domains (e.g. .CN, .UK, .DE, etc.).
- Social media
  - ✓ 126 million – The number of blogs on the Internet (as tracked by BlogPulse).
  - ✓ 84% – Percent of social network sites with more women than men.
  - ✓ 27.3 million – Number of tweets on Twitter per day (November, 2019)
  - ✓ 57% – Percentage of Twitter's user base located in the United States.
  - ✓ 4.25 million – People following @aplusk (Ashton Kutcher, Twitter's most followed user). 350 million – People on Facebook.
  - ✓ 50% – Percentage of Facebook users that log in every day.
  - ✓ 500,000 – The number of active Face book applications.
- Images
  - ✓ 30 billion – At the current rate, the photos uploaded to Facebook per year.
  - ✓ 2.5 billion – Photos uploaded each month to facebook.

## III. Big Data Technologies

As the big data analytics market rapidly expands to include mainstream customers, which technologies are most in demand and promise the most growth potential [8]. Here are the top technologies used to store and analysis Big Data. We can categorize them into two (storage and Querying/Analysis).

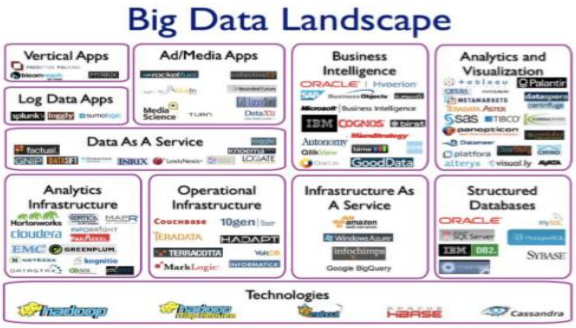


Figure 3. Big Data technology

**Apache Hadoop:** Apache Hadoop is a java based free software framework that can effectively store large amount of data in a cluster. This framework runs in parallel on a cluster and has an ability to allow us to process data across all nodes.

**Microsoft HDInsight:** It is a Big Data solution from Microsoft powered by Apache Hadoop which is available as a service in the cloud.

**NoSQL:** While the traditional SQL can be effectively used to handle large amount of structured data.

Artificial Intelligence: The big data trend has driven advances in AI, particularly in two subsets of the discipline: machine learning and deep learning. Deep learning is a type of machine learning technology that relies on artificial neural networks and uses multiple layers of algorithms to analyse data.

**Data virtualization**: a technology that delivers information from various data sources, including big data sources such as Hadoop and distributed data stores in real-time and near-real time.

Data preparation: software that eases the burden of sourcing, shaping, cleansing, and sharing diverse and messy data sets to accelerate data's usefulness for analytics

## IV. Big Data Management Tools

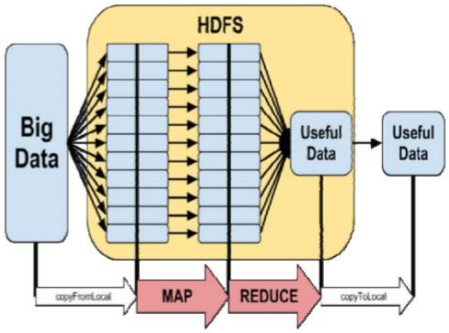


Figure 4. HADOOP Environment

Hadoop [9] is written in Java and is a top-level Apache project that started in 2006. It emphasizes discovery from the perspective of scalability and analysis to realize near-impossible feats. Doug Cutting developed Hadoop as a collection of open-source projects on which the Google MapReduce programming environment could be applied in a distributed system. Presently, it is used on large

amounts of data. With Hadoop, enterprises can harness data that was previously difficult to manage and analyze. Hadoop is used by approximately 63% of organizations to manage huge number of unstructured logs and events (Sys.con Media, 2011). Hadoop is composed of HBase, HCatalog, Pig, Hive, Oozie, Zookeeper, and Kafka; however, the most common components and well-known paradigms are Hadoop Distributed File System (HDFS) and MapReduce for Big Data. Hadoop ecosystem consists as follows-

HDFS: Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop applications. HDFS creates multiple replicas of data blocks and distributes them on compute nodes throughout a cluster to enable reliable, extremely rapid computations.

Map Reduce: Map Reduce is a software framework introduced by Google to support distributed computing on large data sets on clusters of computers.

Pig: Pig is a platform for analysing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs.

Hive: Hive is a data warehouse infrastructure built on top of Hadoop that provides tools to enable easy data summarization, querying and analysis of large datasets data stored in Hadoop files.

## V. Big Data Opportunities and Challenges

Big data opportunities: Cost reduction - Big data technologies like Hadoop and cloud-based analytics can provide substantial cost advantages. While comparisons between big data technology and traditional architectures (data warehouses and marts) are difficult because of differences in functionality, a price comparison alone can suggest order-of-magnitude improvements.

Faster, better decision making - Analytics has always involved attempts to improve decision making, and big data doesn't change that. Following the Big data analytics really makes the business managers good decision makers. Large organizations are seeking both faster and better decisions with big data, and they're finding them.

New products and services - Perhaps the most interesting use of big data analytics is to create new products and services for customers. Online companies have done this for a decade or so, but now predominantly offline firms are doing it too

Product recommendation - It is obviously very clear that the adoption of big data and analytics have proved to be a very powerful strategy for online businesses. Storing and working on huge data has been always a challenge for any trade. Big data has constructed the road for managing such huge data making business much simpler and profitable

Fraud Detection - Insurance frauds are a common incidence. Big data use case for reducing fraud is highly effective.

- Big data challenges: the fact that the valuable enterprise data will reside outside the corporate firewall raises serious concerns. Some of the most common challenges are discussed below [10, 11]:

**Data Storage -** Storing and analysing large volumes of data that is crucial for a company to work requires a vast and complex hardware infrastructure. With the continuous growth of data, data storage device is becoming increasingly more important, and many cloud companies pursue big capacity of storage to be competitive.

**Data Quality** - Accuracy and timely availability of data is crucial for decision-making. Big data is only helpful when an information management process is implemented to guarantee data quality.

**Security and Privacy** - Security is one of the major concerns with big data. To make more sense from the big data, organizations would need to start integrating parts of their sensitive data into the bigger data. To do this, companies would need to start establishing security policies which are self-configurable: these policies must leverage existing trust relationships, and promote data and resource sharing within the organizations, while ensuring that data analytics are optimized and not limited because of such policies.

## VI. Conclusions and Future work

Big data is a potential research area receiving considerable attention from academia and IT communities. In the digital world, the amounts of data generated and stored have expanded within a short period of time. This paper presents the fundamental concepts of Big Data. These concepts include the overview of big data, big data management tools. The challenges and opportunities were identified highlighting the pros and cons of the Big Data. However, little work has been done in the field of big data. We need further details research about big data in the future.

## VII. Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2020R1I1A1A01072928)

## VIII. References

[1] R.Krikorian,“TwitterbytheNumbers,”Twitter,2010.[Online].Available:http://www.slidesshaer.net/raffikrikorian/twitter-by-thenumbers?ref=http://techcrunch.com/2010/09/17/twitter-seeing-6-billionapi-calls-per-day-70k-per-second/.
[2] ABI, “Billion Devices Will Wirelessly Connect to the Internet of Everything in 2020,” ABI Research 2013, [Online]Available:https://www.abiresearch.com/press/more-than-30-billion-devices-willwirelessly-conne/.
[3] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: promise and potential,” HealthInformation Science and Systems, vol. 2, no. 1, pp. 1–10, 2014.
[4] Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, “Efficient Machine Learning for Big Data: A Review,” Big Data Research, vol. 2, no. 3, pp. 87–93, Apr. 201
[5] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li, Big Data Processing in Cloud Computing Environments, 2012 IEEE
[6] Characteristics of Big data. [online]Available:https://www.rd-alliance.org/group/big-data-ig-data-development-ig/wiki/big-data-definition-importance-examples-tools
[7] Today big data [online]Available:rd-alliance.org/group/big-data-ig-data-development-ig/wiki/big-data-definition-importance-examples-tools
[8] Big data technology [online] Available: https://www.datamation.com/big-data/big-data-technologies.html
[9] Hadoop, “Hadoop,” 2009, http://hadoop.apache.org/.View at: Google Scholar
[10] Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Research on Hadoop Cloud Computing Model and its Applications.", Hangzhou, China: 2012, pp. 59 – 63, 21-24 Oct. 2012
[11] Puneet Singh Duggal, Sanchita Paul, “Big Data Analysis: Challenges and Solutions”, International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGP

## Author(s)

***MD IBRAHIM KHALIL***

2020 : Graduate Student in Dept. Of Software and Communications Engineering, Hongik University

Research Interests : Big Data, Low Power, Programming Language, Software Visualization

***Robert Young Chul Kim***

2000 : Ph. D., Dept. of CS, Illinois Institute of Technology, Chicago, USA
2020 : Professor in Dept. Of Software and Communications Engineering, Hongik University

Research Interests : Software Visualization, Design driven Testing, Scenario based Testing, Maturity Model (CMM, DMM, TMM, ETMM), Process Model, Big Data Analysis, Big Data Visualization

***ChaeYun Seo***

2014 : Ph.D in Dept. Of Electronics and Computer Engineering, Hongik University

Research Interests : Business Process, Business Process Modeling, Big Data Visualization, Software Coding Education, Software Visualization, Software Engingeering, Computational Thinking, Software process