Big Data: The next frontier for advance, competition, and efficiency

Author, Rakesh Saroj

A Data Science Foundation White Paper

March 2019

---

www.datascience.foundation

Data Science Foundation
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3641 Email: contact@datascience.foundation Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

ABSTRACT

Nowadays organizations are starting to realize the importance of using more data in order to support decision for their strategies. The size of data in world is growing day by day. Data is growing because of vast use of internet, smart phone and social network. Big data is a collection of data sets which is very large in size as well as complex. Generally size of the data is Petabyte and Exabyte. Traditional database systems are not able to capture, store and analyze this large amount of data. As the internet is growing, amount of big data continue to grow. Big data analytics provide new ways for businesses and government to analyze unstructured data. Nowadays, Big data is one of the most talked topic in IT industry. It is going to play important role in future. Big data changes the way that data is managed and used. Some of the applications are in areas such as healthcare, defense, traffic management, banking, agriculture, retail, education and so on. Organizations are becoming more flexible and more open. New types of data will give new challenges as well.

1. INTRODUCTION

Today the Internet represents a big space where great amount of information are added every day. According to G. Noseworthy (2012), 2.7 Zettabytes of data exist in the digital universe today. He observed in his study that there are 100 Terabytes updated daily through Facebook, and a lot of activity on social networks this leading to an estimate of 35 Zettabytes of data generated annually by 2020. Just to have an idea of the amount of data being generated, one zettabyte (ZB) equals 1021 bytes, meaning 1012 GB. The large quantity of data is better used as a whole because of the possible correlations on a larger amount, correlations that can never be found if the data is analyzed on separate sets or on a smaller set. A larger amount of data gives a better output but also working with it can become a challenge due to processing limitations. Big data includes structured data, semi structured and unstructured data. Structured data are those data formatted for use in a database management system. Semi structured and unstructured data include all types of unformatted data including multimedia and social media content. Some technologies like Hadoop, NoSQL. MongoDB and Map Reduce are required for the analytics of big data. Hadoop, used to process unstructured and semi structured big data. It uses the map-reduce paradigm to locate all relevant data then select only the data directly answering the query. NoSQL, MongoDB, and TerraStore process structured big data. NoSQL data is characterized by being basically available, soft state (changeable), and eventually consistent. MongoDB and TerraStore are both NoSQL-related products used for document oriented applications. The current age of big data poses opportunities and challenges for businesses. Previously unavailable forms of data can now be saved, retrieved, and processed. However, changes to hardware, software, and data processing techniques are necessary to employ this newparadigm. This article is focused to define the concept of Big Data and stress the importance of Big Data Analytics.

Fig 1. Big data size.

2. Big Data Concept

Big Data is a term applied to data sets whose size is beyond the capability of commonly used software tools to capture, manage, and process. The sheer size of the data, combined with complexity of analysis and commercial imperative to create value from it, has led to a new class of technologies and tools to tackle it. The term Big Data tends to be used in multiple ways, often referring to both the type of data being managed as well as the technology used to store and process it. In the past, type of information available was limited. There was a well-defined set of technology approaches for managing information. But in today's world, the amount of data in our world has been exploding. It has grown to terabytes and petabytes. Big data distinct from large existing data stored in various relational databases. It refers to a collection of large data sets which are very complex. Big data can be described using following terms:

- **Volume:** Some small sized organizations may have gigabytes or terabytes of data storage. Data volume will continue to grow, regardless of the organization's size. Many of these companies' datasets are within the terabytes range today but, soon they could reach petabytes or even exabytes. Machine generated data is larger in volume than the traditional data.
- **Variety:** Different types of data are captured. It may be structured, semi structures or unstructured. Refers to the many different data and file types that are important to manage and analyze more thoroughly, but for which traditional relational databases are poorly suited. Some examples of this variety include sound and movie files, images, documents, geo-location data, web logs, text strings, web contents etc (Payal Malik et. al., 2013).
- **Velocity:** The data is arriving continuously as streams of data. It is about the rate of change in the data and how quickly it must be used to create real value (Wei Fan et. al., 2013).
- **Veracity:** If the data coming in large volume is not correct, it can create a problem and is of no use. So, it should be correct

There are three types of big data. Structured, unstructured and semi structured. Similar entities are grouped together in structured data. Entities in the same group have the same descriptions. Examples are number, words, figures etc. Relational databases and spreadsheets are examples of structured data. Unstructured data is complicated information. Data can be of any type and does not follow any rule. It cannot be analyzed

with normal statistical methods. For big data, different tools are required. Examples are social media, email, photos, multimedia etc. In semi structured data, similar entities are grouped together. Entities in same group may not have same attribute. Emails, EDI are example of this type of data.

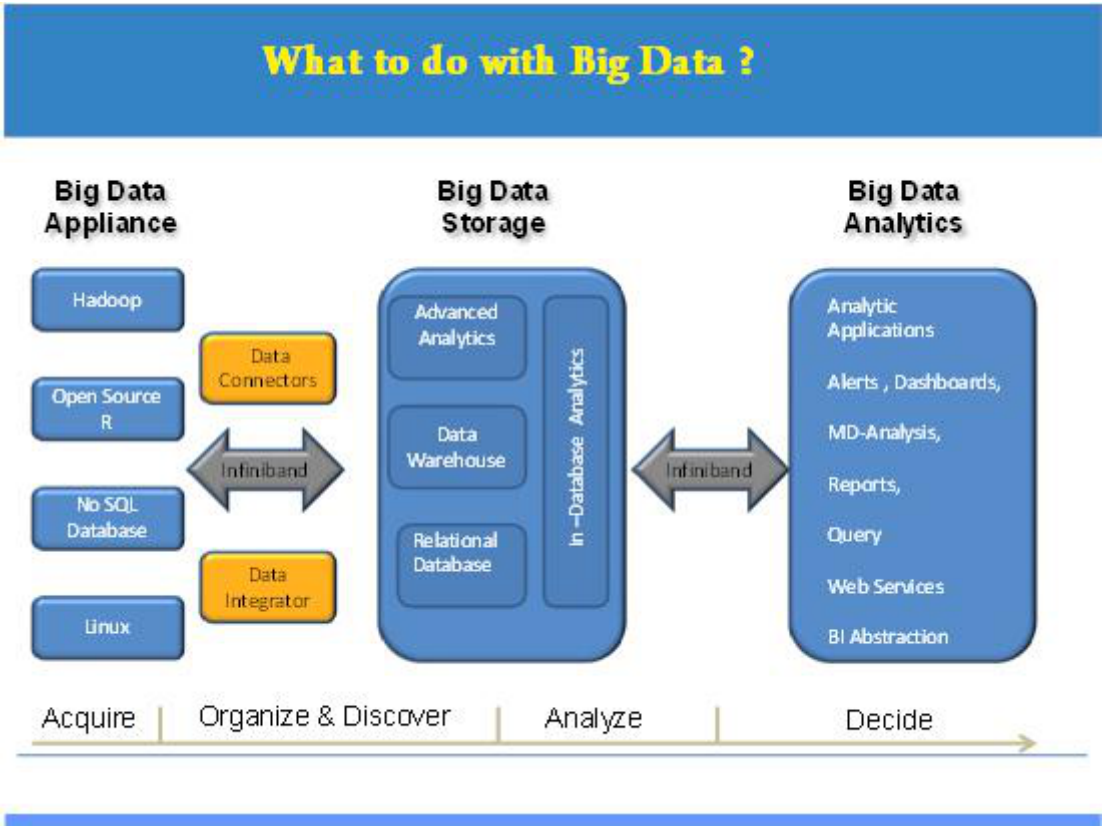

Fig 2. Big data Flow.

3. **The importance of Big Data**

The main importance of Big Data consists in the potential to improve efficiency in the context of use a large volume of data, of different type. If Big Data is defined properly and used accordingly, organizations can get a better view on their business therefore leading to efficiency in different areas like sales, improving the manufactured product and so forth. Big Data can be used effectively in the following areas:

- In information technology in order to improve security and troubleshooting by analyzing the patterns in the existing logs;
- In customer service by using information from call centers in order to get the customer pattern and thus enhance customer satisfaction by customizing services;
- In improving services and products through the use of social media content. By knowing the potential customers preferences the company can modify its product in order to address a

larger area of people;

- In the detection of fraud in the online transactions for any industry;
- In risk assessment by analyzing information from the transactions on the financial market.
- Enhance 360º View of the Customer: To extend existing customer views by incorporating additional internal and external information sources. Gain a full understanding of customers - what makes them tick, why they buy, how they prefer to shop, why they switch, what they will buy next, and what factors lead them to recommend a company to others.
- Security Intelligence Extension: Lower risk, detect fraud and monitor cyber security in real time. Augment and enhance cyber security and intelligence analysis platforms with big data technologies to process and analyze new types (e.g. social media, emails, sensors) and sources of under-leveraged data to significantly improve intelligence, security and law enforcement insight.
- Data Warehouse Modernization: To integrate big data and data warehouse capabilities to increase operational efficiency.

4. **Difference between traditional and big data analytics**

Big data analytics can be differentiated from traditionaldata-processing architectures. In traditional data, sources are internal and structured. Data integration tools are used to extract, transform and load the data from transactional databases. Then data quality and data normalization occur and the data is modeled into rows and columns. The modeled data is then loaded into an enterprise data warehouse. Big data is data that is too large to process using traditional methods. As the volume of data explodes, organizations will need analytic tools that are reliable, robust and capable of being automated.Traditional data warehouse is not able to handle processing of big data as data is coming from different sources like social media, video etc. This type of data grows at very high speed. The database requirements are very different in the case of big data. With big data analytics data can be anywhere and is in large volume. Big data analytics provides useful information. Hidden patterns are discovered. It focuses on unstructured data. Some technologies like Hadoop, NoSQL and Map Reduce are required for the analytics of big data. In big data analytics, the Hadoop system captures datasets from different sources and then performs functions such as storing, cleansing, distributing, indexing, transforming, searching, accessing, analyzing, and visualizing. So the unstructured data is converted into structured data. The working principle behind Hadoop and all big data is to move the query to the data to be processed, not the data to the query processor. Various languages used in the big data analytics are Java, Oracle JavaScript etc. Big data requires many different approaches to analysis, traditional or advanced, depending on the problem. It depends on the type of that particular problem. Some analytics includes traditional data warehouse concept. But some requires more advanced techniques. The IT techniques and tools to execute big data processing are new, very important and exciting. Big data technologies work faster than traditional data warehousing techniques.

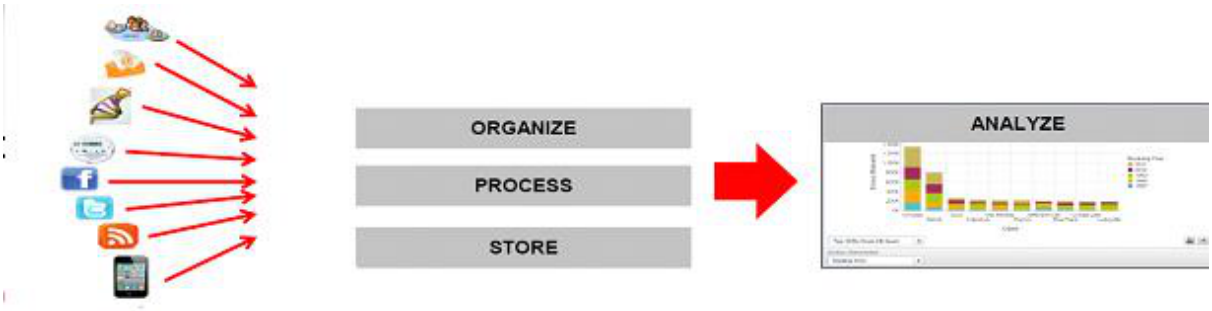

Fig 3. Big data management.

Figure 3 (Elena Geanina et. al., 2012) shows management of big data. Data is collected from various sources. It is not just a text data. It contains images, audio or video. It may be social data, machine generated data or documents. This data is unstructured. It is very critical to understand, categorize and analyze this large volume of data. A system is required to organize process and store this data into database so that it is analyzed efficiently.

5. **Big Data challenges**

Big data analytics faces different challenges. These aredescribed as follows:(Malik,P. et. al., 2013).

- **Heterogeneity and Incompleteness:** Machine analysis algorithms expect homogeneous data, and cannot understand nuance. Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain.
- **Timeliness:** There are many situations in which the result of the analysis is required immediately. Given a large data set, it is often necessary to find elements in it that meet a specified criterion The larger the data set to be processed, the longer it will take to analyze. It is difficult to design a structure when data is growing in very high speed.
- **Human Collaboration:** A Big Data analysis system must support input from multiple human experts, and shared exploration of results.
- **Privacy and security:** This is another big challenge preserving inpidual privacy. For example in the healthcare industry, record of inpidual is very personal. But it can be available from multiple sources. So, it is difficult to maintain privacy and security.
- **Data Quality:** A large volume of data is processed. Analyzing which data is important and to capture it is a big challenge.
- **Analysis:** Big data is coming from various data sources. So analytics is a challenge.
- **Skill:** Big data require people with new skill sets. Managing big data effectively requires the right people.

6. **Opportunities with Big Data**

The use of big data (Available at: http://www.mckinsey.com) will become a key basis of competition and growth for inpidual firms. All the companies will use big data. In most industries, established competitors and new entrants alike will leverage data-driven strategies to innovate, compete, and

capture value from deep and up-to-real-time information.

- Big data has opportunities in the field of education. More detailed information for school can be generated. This is beneficial for teacher and parents.
- Almost all sectors like computer and electronic products, insurance, and government will increase their productivity from the use of big data.
- Concept of big data has practical application in the area of healthcare research. In health care, data is coming from medical records, radiology images, human genetics etc. More information is analyzed regarding patient care and disease. Hence studies can be completed faster. Big data will help better future diagnoses and treatment of the patient.
- Use of smart phone and tablet leads to high amount of mobile data traffic. Big Data is important for mobile networks. It is useful to improve network quality, traffic planning, prediction of hardware maintenance etc (Big Data: a new world of opportunities, 2012).
- Various branches of science generates large amount of experimental data. Fulfilling the demands of science requires a new way of handling data (Big Data: a new world of opportunities, 2012).

7. **Big data's big role in Agriculture**

The world's population is increasing very fast. We all know that feeding that population is going to require implementing a number of strategies. Some of those strategies begin in the lab, are implemented in the field, affect harvested crops during transportation and sale, and change consumers' eating habits. Some of those strategies involve creating heartier plants and animals, increasing crop yields, and ensuring safe food supply chains. In each of those strategies, big data analytics is going to play a significant role. Making sense of the data takes a lot of analytical and computing power. Converting this data into information, and understanding what it really means, we will be able to make faster genetic improvements to improve food production and reduce the time to market.

The big data process in agriculture can include farm level data and personal data which are collected from ground and equipment sensors, robotic drones etc. Then service provider aggregates farmer's data, combines other relevant data sets and applies algorithms and finally service provider provides farmer customized solution. Also there are so many opportunities to extract high quality recommendation and information from agricultural data sets like rainfall data, KCC's data (Kisan Call Centers) etc. Precision farming can also be improved by the use of big data analysis

8. **Big Data Analytics**

To turn big data into a business advantage, businesses have to review the way they manage data within data centre. The data is taken from a multitude of sources, both from within and without the organization. It can include content from videos, social data, documents and machine-generated data, from a variety of applications and platforms. Businesses need a system that is optimized for acquiring, organizing and loading this unstructured data into their databases so that it can be effectively rendered and analyzed. Data analysis needs to be deep and it needs to be rapid and conducted with business goals in mind. The scalability of big data solutions within data centers is

an essential consideration. Data is vast today, and it is only going to get bigger. If a data centre can only deals with the levels of data expected in the short to medium term, businesses will quickly spend on system refreshes and upgrades. Forward planning and scalability are therefore important

9. **Big Data Analytics Technologies**

9.1. NoSQL: NoSQL database can handle unstructured and unpredictable data. The data stored in a NoSQL database is typically of a high variety. A NoSQL database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases. Relational and NoSQL data models are very different.

The relational model takes data and separates it into many interrelated tables that contain rows and columns. But document-oriented NoSQL database takes the data into documents using the JSON format. JSON is JavaScript Object Notation. Another major difference is that relational technologies have rigid schemas while NoSQL models are schema less. Many NoSQL databases have excellent integrated caching capabilities. So, the frequently used data is kept in system memory. NoSQL database types are (Available: http://www.mongodb.com/learn/nosql)

1. **Document database:** Pair each key with complex data structure known as document. Document may contain nested document. This type of database store unstructured (text) or semi-structured (XML) documents which are usually hierarchal in nature.
2. **Graph stores:** Graph database is based on graph theory. It is used to store information about network.
3. **Key value stores:** Every single item is stored as an attribute name together with its value.
4. **Wide column stores:** They are optimized for queries over large datasets and store column of data together instead of rows.

9.2. Apache Hadoop: It is a fast-growing big-data processing open source software platform. Hadoop can handle all type of data like structured, unstructured, pictures or audio. It runs on Linux, OS/X, Windows, and Solaris. Hadoop is scalable, flexible and fault tolerant. It contains HDFS. Hadoop HDFS is scalable, distributed file system written in Java. Figure 4 explains (The Hadoop Distributed File System: Architecture and Design) HDFS architecture. HDFS has master/slave architecture. An HDFS cluster consists of a single NameNode. It manages the file system namespace. The name node is the equivalent of the address router for the big data implementation. In addition, there are a number of DataNodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on. New nodes can be added as needed and added without needing to change data formats The DataNodes are responsible for serving read and write requests from the file system's clients. DataNodes also performs function like block creation, deletion and replication as per the instruction of NameNode (The Hadoop Distributed File System: Architecture and Design). Hadoop creates clusters of machines and coordinates work among them. If any of the clusters fails, then Hadoop continues to operate the cluster without losing data. Map Reduce is a programming model and software framework first developed by Google. It works like a UNIX pipeline.

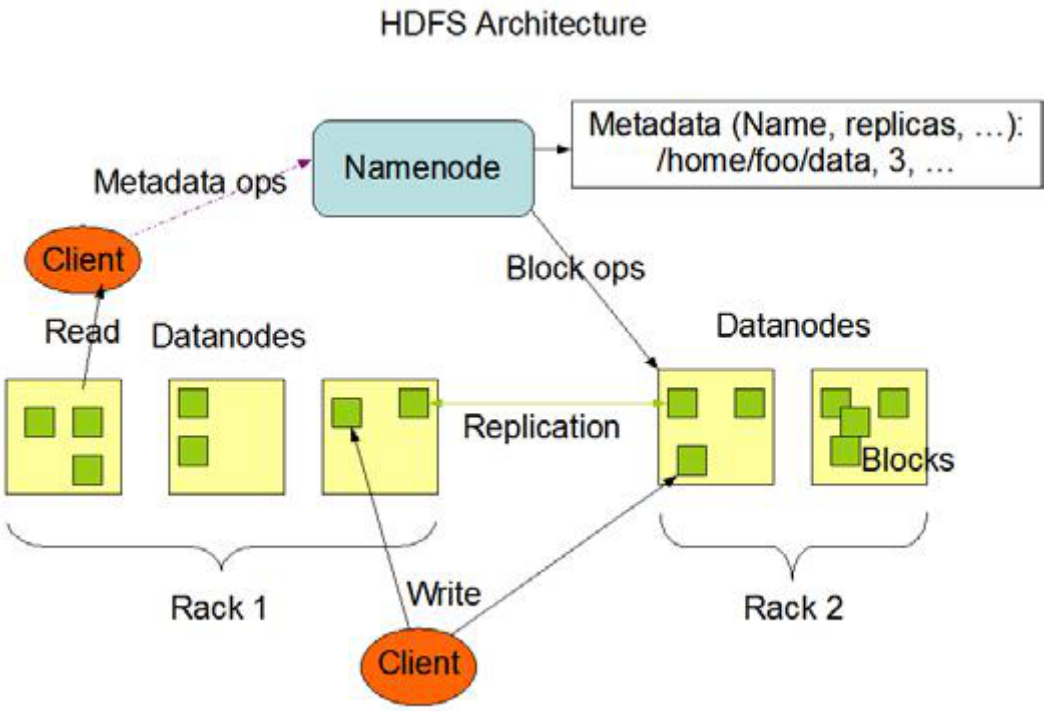

Fig 4. HDFS Architecture.

The job of Map Reduce pides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result. Map Reduce framework consists of a single master JobTracker and one slave TaskTracker per cluster node. The master is responsible for scheduling the job's component tasks on the slave, re-executing the failed task (Available: https://hadoop.apache.org). The slave executes the task as directed by the master.Some of the Hadoop related projects are described as:(http://www.revelytix.com/?q=content/hadoopecosystem)

- **Pig:** It is a Scripting language and run time environment. It allows users to execute MapReduce on a Hadoop cluster. Pig's language layer currently consists of a textual language called Pig Latin.
- **Hive:** It provides SQL access for data in HDFS. Hive's query language, HiveQL, compiles to MapReduce. It also allows user-defined functions.
- **HBase:** A scalable, distributed database that supports structured data storage for large tables. It is column based rather than row based.
- **Mahout:** Library of machine learning and data mining algorithm. It has four types of algorithm.
- **Oozie:** Oozie is a Java Web-Application that runs in a Java servlet-container – Tomcat. It is job coordinator and workflow manager.
- **BigTop:** It is used for packaging and testing the Hadoop ecosystem

10. **Conclusions**

Big data is a relatively new phenomenon. The field of big data is going from different perspective. Big data analytics provide new ways for businesses and government to analyze unstructured data. Research and development is required in this field. There are many technical challenges that must be addressed. Research is required to find new way of handling the data. For many IT decision makers, big data analytics tools and technologies are now a top priority. Big Data is going to play very important role in the future. There is need of analytical software which can handle huge storage as well as processing requirement of big data .To extract more and new value, there will be a focus on developing effective analytics. New big data technologies and tools have been and will continue to be developed.

REFERENCES

- Fan,W. and Bifet,A. (2012). Mining Big Data: Current Status, and Forecast to the Future. SIGKDD Explorations, 14(2), 1-5
- Geanina,E., Camelia, F., Anca and Velicanu,M.(2012). Perspectives on Big Data and Big Data
- Analytics. Database Systems Journal 3(4), 3-13
- Malik, P. and Bose, L. (2013). Study and Comparison of Big Data with Relational Approach. International Journal of Advanced Research in Computer Science and Software Engineering, 3(8), 564-570
- Noseworthy,G.(2012). Infographic: Managing the Big Flood of Big Data in Digital Marketing, Available at http://analyzingmedia.com/2012/infographicinfographic-big-flood-of-big-data-in-digital-marketing/
- Sampada Lovalekar (2014). Big Data: An Emerging Trend In Future . (IJCSIT) International Journal of Computer Science and Information Technologies, 5 (1), 538-541.
- Vorthakur, Dhruba, 2005, The Hadoop Distributed File System: Architecture and Design http://www.enterrasolutions.com/2014/09/big-datas-big-role-agriculture.html https://hadoop.apache.org
- http://www.mongodb.com/learn/nosql
- http://www.revelytix.com/?q=content/hadoopecosystem

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:contact@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation