Big Data: Concept, Handling and Challenges: An Overview

Soumya Shukla
B.E. (last year)
Computer Engineering in
MKSSS' Cummins college of
Engineering, Pune

Vaishnavi Kukade
B.E. (last year)
Computer Engineering in
MKSSS' Cummins college of
Engineering, Pune

Sofiya Mujawar
B.E. (last year)
Computer Engineering in
MKSSS' Cummins college of
Engineering, Pune

ABSTRACT
In this paper, we've presented an overview of the concepts of big data its characterization as well as the various methods of handling big data. We have also discussed the various challenges faced during handling of big data.

Keywords
Big data, Data analytics, Business intelligence, Data mining, Challenges, Techniques.

1. INTRODUCTION
In today's world, every tiny gadget is a potential data source, adding to the huge data bank. Also, every bit of data generated is practically valued, be it enterprise data or personal data, historical or transactional data. This data generated through large customer transactions, social networking sites is varied, voluminous and rapidly generating. All this data prove a storage and processing crisis for the enterprises. The data being generated by massive web logs, healthcare data sources, point of sale data, satellite imagery needs to be stored and handled well. Although, this huge amount of data proves to be a very useful knowledge bank if handled carefully. Hence big companies are investing largely in the research and harnessing of this data. By all the predilections today for Big Data, one can easily state Big Data technology as the next best thing to learn. All the attention it has been getting over the past decade is but due to its overwhelming need in the industry.

Thus, this paper gives an overview of the key concepts in Big Data, some practiced Big Data handling techniques and the challenges posed by the technology itself.

2. CONCEPT
2.1 Big data Characteristics
The three V's" "volume, velocity and variety," definition of big data originally coined by Doug Laney in 2001 to refer to the challenge of data management was quite in place to define big data for a few years. It basically interpreted big data as being a lot of data that is in a scattered form and needs to be processed quickly for proper interpretation. [1]

In August, 2013 the definition was further enhanced to include, "veracity, variability, visualization, and value" which gave a newer perspective to it. [2]

With this new definition, big data now seem to not only describe itself according to its amount, but further was enhanced according to its interpretation and usability.

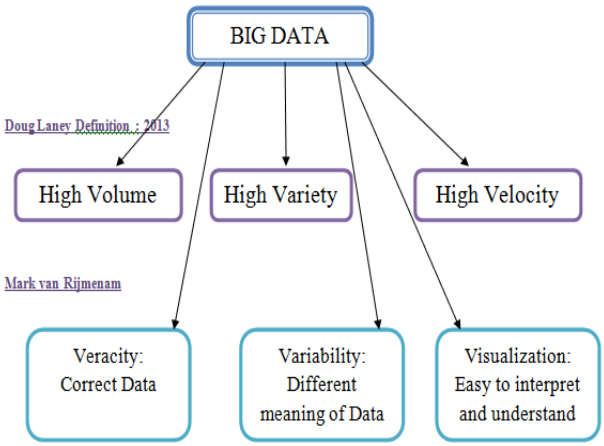

Fig 1: Big data Definition

Big Data mainly involve six aspects as per the mentioned definition.

Volume –Volume defines the quantity of Big Data. The size of this data ranges from terabytes and petabytes, to even Exabytes.

Variety –Variety define data types of Big Data, which includes structured and unstructured data such as text, audio, video, sensor data, posts, log files and many more.

Velocity – As the generation of data is rapid, the process of acquiring, processing and analyzing it requires fast mechanisms. The velocity emphasizes on the real time processing power of big data for enterprise needs.

Veracity– Refers to the requirement of correct form of data as it is relied upon for all further analysis.

Variability- Data can be in the same form but having different semantics.

Visualization- Data should be easy to process and interpret to derive intelligence out of it.

2.2 Unique Features of Big Data:
Data is expanding at an astonishing rate. By 2020, experts say that there will be 4300% of annual data increase. Hence it's not only the size of big data that makes it unique but also its unstructured form that can cause serious issues for handling it. With the way data has been expanding every minute, new technique and analysis tools have been made to handle them. These tools analyze large data sets simultaneously and storage on cloud on secure data centers has made their analysis easy and on the go. Hence, big data is not only unique in its size and form but also in its processing and knowledge discovery. Big data in petabytes are analyzed quickly and give more accurate interpretation of respective queries than ever before.

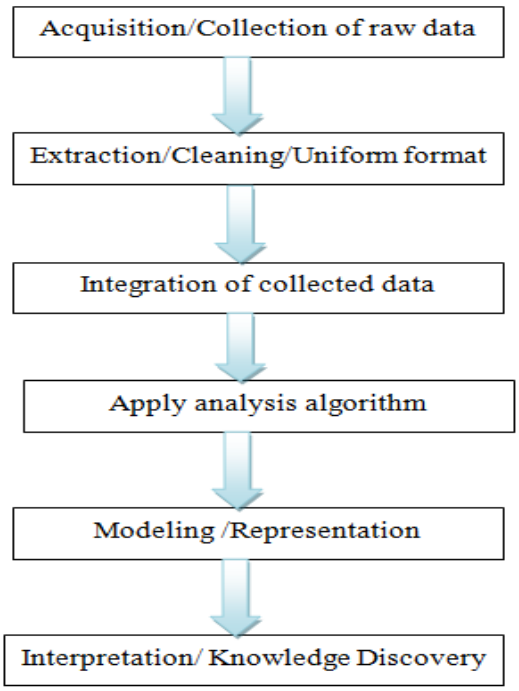

Fig 2: Big data Interpretation Insights

2.3 Big Data as an Opportunity

Companies have worked upon its data analytics and interpretation to open a new horizon of opportunities their way. Big data represents both significant information and a way with which it can be analyzed. This provides an opportunity at every stage of knowledge discovery in big data.

Big data offers an opportunity in many sectors as mentioned below:

- Banking and security
- Communication Media and Services
- Education
- Government
- Healthcare providers
- Insurance
- Manufacturing and natural resources
- Transportation
- Various Traders such as : Retail, wholesalers

Government Sector being the highest among them, offers wide range of opportunities for bigdata analyst and researchers. [3]

2.4 Example of Big data

Ranging from data generated in small enterprises to IT giants, from social network sites to app data on cloud, bigdata is generated in various form every day. An example of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources such as Web, sales, customer contact center, social media, mobile data etc. The data is typically loosely structured data that is often incomplete and inaccessible**.**

3. BIG DATA HANDLING

3.1 Need of Handling Big Data?

The huge amount of data collected has the potential to reveal useful trends and patterns. Hence it needs to be preserved and processed. All this data is being stored in cloud or huge secondary storages like the Hadoop File System. The processing is done using Hadoop or Spark. The major insights from analysis of this data are numerous business intelligence applications, fraud detection, weather forecast, personalized advertising etc. For analysis of this kind various machine learning and data mining tools are used.

3.2 Big Data Handling Techniques:

Handling of Big Data is another major concern. Below are some emerging technologies that are helping users cope with and handle Big Data in a cost-effective manner.

Big data handling can be done with respect to following aspects-

- *Processing Big data*: MapReduce, Hadoop is an integrated framework for processing and storing Big data
- *Analysis and querying of data:* WibiData, PLATFORA, PIG
- *Business Intelligence*: Hive
- *Storage*: Cloud storage, Column-oriented databases, schema-less databases
- *Machine Learning*: Apache Mahout, SkyTree

Some of the various Big data handling techniques defined are illustrated below[5]-

3.2.1 MapReduce

MapReduce is the key algorithm that the Hadoop MapReduce engine uses to distribute work around a cluster.

Mapper function- A map transform function is provided to transform an input data row of key and value to an output key/value:

- map(key1,value) -> list<key2,value2>

That is, for an input it returns a list containing zero or more (key, value) pairs: The output can be a different key from the input. The output can have multiple entries with the same key

Reduce function: A reduce transform is provided to take all values for a specific key, and generate a new list of the *reduced* output.

- reduce(key2, list<value2>) -> list<value3>

3.2.2 Hadoop

Apache Hadoop is an open source framework for distributed storage and processing of large sets of data on commodity hardware. Hadoop enables businesses to quickly gain insight from massive amounts of structured and unstructured data. It is used in maintaining, scaling and analyzing large scale of data. This data can be structured or unstructured.

3.2.3 PIG

Apache PIG is a platform for analyzing large data sets. PIG's language, PIG Latin, lets one specify a sequence of transformation functions like merge, filter, grouping etc. Apart from built-in functions it also provides facility for user-defined functions to do special-purpose processing. PIG's

language allows for query execution over data stored on a Hadoop cluster, instead of a "SQL-like" language.

3.2.4 HIVE

Hive enables traditional BI applications to run queries against a Hadoop cluster. It was developed originally by Facebook, but has been made open source for some time now, and it's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store. It makes Hadoop more useful for BI users.

3.2.3 Column-Oriented Databases

Conventional, row-oriented databases are best fit for online transaction processing with high update speeds, but they fall short on query performance as the data volumes grow and as data becomes more unstructured. Column-oriented databases store data with a focus on columns, instead of rows, allowing for huge data compression and very fast querying.

3.2.4 Schema-Less Databases, or NoSQL Databases

There are several database types that fit into this category, such as key-value stores and document stores, which focus on the storage and retrieval of large volumes of unstructured, semi-structured, or even structured data. They achieve performance gains by doing away with some (or all) of the restrictions traditionally associated with conventional databases, such as read-write consistency, in exchange for scalability and distributed processing.

3.2.5 Using cloud for Big data

Most of the above technologies demand cloud, ie , many of the products and platforms mentioned are either entirely cloud-based or have cloud versions themselves. Most cloud vendors are already offering hosted Hadoop clusters that can be scaled on demand according to their user's needs.

Big Data and cloud computing go hand-in-hand. Cloud computing enables companies of all sizes to get more value from their data by enabling faster analytics at minimal costs. This, in turn improves the company's productivity and thus returns the costs invested.

4. BIG DATA CHALLENGES

Big data which is typically of the size petabyte or terabyte is bound to be confronted with many theoretical, technical, technological and practical challenges. Serious research efforts are being invested in order to improve the efficiency of storage, processing and analysis of big data. Following are the various challenges faced while handling big data.

4.1 Data Acquisition and Recording:

It is important to capture the context into which data has been generated and the ability to filter out the noise during pre-processing the data and to compress data. Pre-processing of data is complex and is time consuming thus the real challenge is handling big volumes of unstructured and structured data continuously arriving from a large number of sources. Hence a solution to this would require innovation of new technologies and architectures, designed to efficiently extract value from very large volumes of a wide variety of data, by enabling high velocity capture, discovery and/or analysis.

4.2 Information Extraction and Cleaning:

Often data needs to be transformed in order to extract information from it in order to express this information in a form that is suitable for analysis. Data may also be of poor quality and/or uncertain. Extracting meaningful information from such huge amounts of data of poor quality is one of the major challenges being faced in big data. The accuracy of the results monumentally depends on Data cleaning and data quality verification. Thus cleaning of data and it's quality verification are critical. [6]

4.3 Data Integration, Aggregation and Representation:

Data might not be homogenous and may have different metadata. Thus Data integration requires huge human efforts. Manual approaches fail to scale to what is required for big data, hence the requirement of newer and better approaches arises. Also different data aggregation and representation strategies may be needed for different data analysis tasks. [6]

4.4 Query Processing, and Analysis:

Methods suitable for big data need to be discovered and evaluated for efficiency so that they are able to deal with noisy, dynamic, heterogeneous, untrustworthy data. However despite these difficulties, big data even if noisy and uncertain can be more valuable for identifying more reliable hidden patterns and knowledge compared to tiny samples of good data. [6]

5. CONCLUSION AND FUTURE WORK

Due to the gargantuan increase in the amount of data in various fields, it becomes a major challenge to handle the data efficiently. Thus to come up with plausible solutions to these challenges one needs to understand the concept of big data, its handling methodologies and furthermore improve the approaches in analyzing big data. With the advent of social media the need for handling big data has increased monumentally. If Facebook, Whatsapp, Twitter produce data which keeps increasing exponentially every year (or a few years) then handling such huge data is something to be efficiently dealt with. We will need solutions to such issues without compromising the quality of the results. Hence we attempt to showcase basic concepts of big data that can be used as easy referrals for literature survey of the topic.

6. REFERENCES

[1] Douglas, Laney. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012.

[2] Blog post: Mark van Rijmenam titled "Why the 3V's Are Not Sufficient to Describe Big Data".

[3] Jean Yan, April 9, 2013 "Big Data, Bigger Opportunities Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems.

[4] "Research in Big Data and Analytics: An Overview" International Journal of Computer Applications (0975 – 8887) Volume 108 – No 14, December 2014

[5] Blog post: Thoran Rodrigues in Big Data Analytics, titled "10 emerging technologies for Big Data", December 4, 2012

[6] 2013 IEEE 37th Annual Computer Software and Applications Conference. Elisa Bertino Cyber Center, CERIAS and CS Department Purdue University West Lafayette, Indiana (USA) "Big Data - Opportunities and Challenges Panel Position Paper"

[7] Wie, Jiang, Ravi V.T, and Agrawal G. "A Map-Reduce System with an Alternate API for Multi-core

Environments." Melbourne, VIC: 2010, pp. 84-93, 17-20 May 2010.

[8] 2013 46th Hawaii International Conference on System Sciences. 'Big Data: Issues and Challenges Moving Forward' by Stephen Kaisler, Frank Armour, Alberto Espinosa, William Money.

[9] Mobile Netw Appl (2014) 19:171–209 DOI 10.1007/s11036-013-0489-0 'Big Data: A Survey' by Min Chen, Shiwen Mao, Yunhao Liu

[10] Basic Concepts in Big Data by ChengXiang ("Cheng") Zhai
International Journal of Computer Applications (0975 – 8887) National Level Technical Conference "X-PLORE 14 'Algorithm and Approaches to Handle Big Data' by Uzma Shafaque, Parag D. Thakare, Mangesh Ghonge, Milindkumar Sarode.