ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 4, Issue 3, March 2017

Big Data-An Emerging and Innovative Technology: Survey

[1]Keerthana R, [2] Nandini N [3] Karthik N S [4] Divya L [5]Mrs Sheela devi
[1][2][3][4] UG Scholars, [5]Asst Professer
[1]-[5]Dept of Computer science Engineering, Sri Sairam College of Engineering, Bangalore.

Abstract— This paper is a review of the big data concept, its dimensions, its architecture comparison between the earlier concept and the latest, the storage possible i.e. the databases and origin of big data. The relational database, having rigid schema, has been prevailing since a long time but it is difficult to store the unstructured data in relational database. The unstructured data has mainly text nature or is in the form of logs. Here comes the concept of No SQL databases. Big Data is small data with large data size.

***Keywords*– big data, dimensions, inconsistency, granularity, databases**

I. INTRODUCTION

In the early years of 21st century, due to the rising use of internet, the term 'Big Data' was introduced but suddenly it got a boom lately near 2013, due to some need. This need was the analysis of data. Storage was never an issue. It was the inability of traditional relational databases that led to the evolution of No SQL databases. The traditional databases have rigid schema whereas the No SQL databases have flexible schema without downtime (a situation when system fails to perform primary operations). It was then this data got famous as big data. Big Data is nothing but small data with large data size. The data management tools which are present since decades find it difficult to process complicated data sets. Certain processing applications are also unable to process such voluminous and dynamic data. Such data sets form the big data [2].

The size of digital data in 2011 was roughly 1.8 Zettabytes (1.8 trillion gigabytes) i.e. supporting networking infrastructure has to manage 50 times more information by year 2020[1]. Also, considerations of efficiency, economics and privacy should be planned carefully while including new big data building blocks into existing data and networking infrastructure [1]. After the survey of various meanings of big data, the one mentioned in [3] finally gives the most appropriate definition of big data i.e. data that's too big or too fast or too hard for currently known tools to process. Here, "too big" means that organizations must deal with increasing petabyte-scale collections of data that come from click streams, transaction histories, sensors, and elsewhere. "Too fast" means that not only is data big, but it must be processed quickly — for example, to perform fraud detection at a point of sale or determine which ad to show to a user on a webpage. "Too hard" is a catchall for data that doesn't fit neatly into an existing processing tool or that needs some kind of analysis that existing tools can't readily provide [3].

Big Data is spreading vastly in the industry. Most of the industries want to have the records of not only the work they do but also are eager to know the taste of the consumer. This is going to lead business advantage. Big Data is becoming relative to almost all aspects of human activity from just recording events to research, design, production and digital services or products delivery to the final consumer [4]. The data generated from various sources has various behavior, characteristics and nature. It contains certain useful and useless information but most of the data is in textual format. It is nothing but the unstructured data.

II. DIMENSIONS AND INCONSISTENCIS

Due to the advancement in the social and economic activities and rapid growth in science and industry, there has been a massive production of the digital data. This big data phenomenon will only get intensified and diversified in the years to come [5]. As data grows there is a need to analyze the data which is only possible if we are aware of the various dimensions of the data. A meta-definition based on the size dimension is given in [5]: "big data should be defined at any point in time as 'data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time' [5].The storage of data is at an unprecedented rate. In [5], author has mentioned that data has vast granularity. Data goes through various forms till it become usable, as shown in fig.1. As keenly observed in [5] [6], "what makes big data big is repeated observations over time and/or space." Hence, "most large datasets have inherent temporal or spatial dimensions, or both" [5] [6].

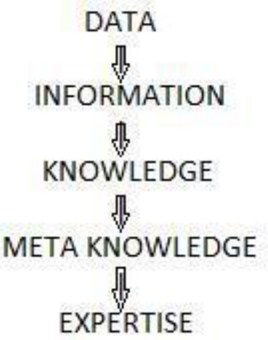

Fig.1 Granularity of big data (Inductive)

Various researchers have considered domain and the technology as the major dimension of big data, but in [5], author has mentioned various dimensions which clearly defines what makes big data, as shown in fig.2. The technologies like machine learning, cloud computing, crowd sourcing, etc. all have big data. Some of the major challenges are privacy and security. The analysis task includes data acquisition, integration, storage, search retrieval, analysis and visualization. Sources of big data include: transactions, scientific experiments, genomic investigations, logs, events, emails, social media, sensors, RFID scans, texts, geospatial data, audio data, medical records, surveillance, images, and videos [5]. Data may be text, video, audio, log files of any type either in semi-structured form or unstructured form. Domain includes transportation, government, communication, media, education, life sciences, manufacturing. The basic objective of it is value creation, improved productivity and scientific discoveries. These all lead to business advantage. In [5], the author has explained this using a flowchart. This forms the Big Data.

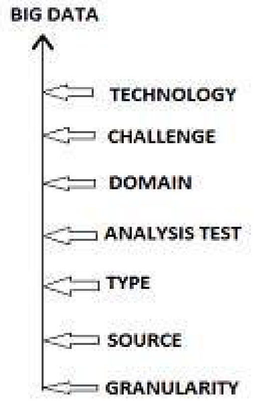

Fig.2 Dimensions of Big Data

In circumstances where big data is produced, acquired, aggregated, transformed, or represented, their inconsistencies invariably find their way into large datasets [5].The major cause of this is conflicting data. The inconsistencies can be a major problem in various fields, such as heuristics, problem solving, research analysis and business prediction analysis [9]. In order to deal with this, one must be aware of the various types of inconsistencies that can arise. These are, as mentioned in [5] [9]:

- ♦ Temporal inconsistency
- ♦ Spatial inconsistency
- ♦ Text inconsistency

Functional inconsistency Temporal inconsistency is basically caused when there is a time interval relation in data. The time interval relationship between conflicting data items can result in partial temporal inconsistency (or) complete inconsistency [9].Temporal inconsistencies have been utilized as problem-solving heuristics in IBM Watson open-domain QA system where temporal reasoning is deployed to "detect inconsistencies between dates in the clue and those associated with a candidate answer"[5]. When there is geometric representations and objects included in the data, it causes for spatial inconsistency. Spatial relations between objects cause the same. Text inconsistency covers the major part. In today's world, lot of textual data is generated from social media, communications, mails, etc. that contribute text inconsistency. Relational databases have certain integrity constraints for the functional dependency of attributes.

Breaching of constraints causes functional dependency inconsistency.

III. CHARACTERISTICS: COMPARISON

When the concept of big data came into existence, the 3Vs Volume-Velocity-Variety given by Gartner, "Big data is 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."[7] Organizations like IBM added one more V i.e. Veracity. Veracity is an indication of data integrity and the ability for an organization to trust the data and be able to confidently use it to make crucial decisions [8]. In [4], author gave the improved Gartner definition as, “Data Intensive Technologies are targeting to process high-volume, high-velocity, high-variety data sets or assets to extract intended data value.

This ensures high-veracity of original data and obtained information. This demands cost-effective, innovative forms of data processing and analysis for enhanced insight, finding relations and processes control. During the whole life cycle of data, the arising demands must be supported by data models, infrastructure services and also the tools that facilitate the states and the stages of the data to obtain it from different domains like sensor networks and supplying information to consumer and various devices [4].”

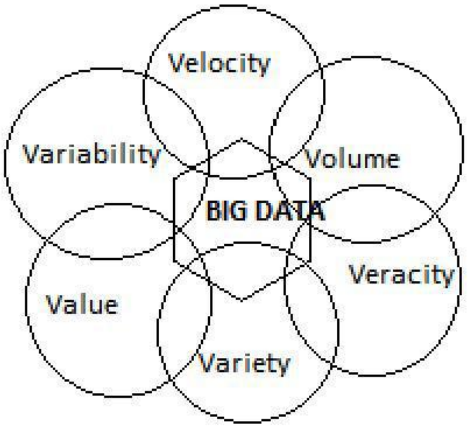

Fig.3 Characteristics of Bi g Data

Thus, this survey leads to find that there are not only 3V main properties of Big Data but also it includes veracity and value. Ad additionally: Data Dynamicity (Variability) and Linkage e [4].Variability is the inconsistency the data can show at times, which can hamper the process o f handling and managing the data effectively [7].The 6V makes the complete characteristics of big data, as shown in fig.3.

According to IBM, users create 2.5 quintillion bytes of data every day. So practically it can be said that, if we measure the data generated in last two years only, it makes around 90% of the total data present in the today’s world [8]. Every hour, Walmart controls approximate ly 1 million customer transactions and much more. These are then transferred into a database that has information around 2.5 petabytes and is working with it [8]. This makes Walmart and IBM count among one of the best companies.

IV. DATABASES

Most of the big data generated from sources like media, mails, communication, etc. being in unstructured form need a storage which is not possible in any Structured Query Language Database for processing. So there came the role of Nasal databases. These store the semi-structured or unstructured data. The Nasal databases come under four categories as mentioned in [10]:

- ***Key Value Databases:*** Key Value Databases use a hash table where there is a unique key and a pointer to a specific set of values; data can only be queried by the key [10].Face book currently uses such type of database as datasets are not related to each other [10]. This facilitates proper analysis of unstructured data as there is no set schema.
- ***Column Oriented Databases:*** These have columns and rows t o store data. The rows can have multiple columns and have a row key. Google uses a distributed data storage system, Big Table, for its package Google Earth. These databases were created to store and process enormous amounts of data in distributed systems, especially versioned data because of its time stamping functions [10]. Examples of such database are H Base and Cassandra.

- ***Document Database:*** In this, the database Maintains a document that stores a record and data re lasted to it. Document Databases use entire documents of structured data files, such as XML or JSON, as datasets [10]. These do not have a schema. Exam plus are Couch base, Monod.

Nasal databases are increasingly considered a viable alternative to conventional databases, as more businesses recognize that its schema less model is a better method for handling the large volumes of semi structured and unstructured data, being captured and processed today [11].According to the CA P Theorem, all the databases must have two of the properties among availability, partition tolerance and consistency.

V. FUTURE RESEARCH

The future research includes handling of big data using some unstructured database. At this stage, the authors have studied about the complete nature of the big data and look forward to study on analysis of big data using Hardtop toolkit using the document type database Monod.

VI. CONCLUSION

Thus, data produced under large scale having a form (Volume-velocity-variety, given by Gartner), which may or may not be easily analyzed but can be used to generate some meaning using some analytical tool is Big Data. The 3V model may have certain additional characteristics namely, value, veracity and variability. The Big Data has various dimensions. The granularity is helpful in giving various forms of data in useful form. The inconsistencies should be pre-analyzed on the basis of dimensions. Thus the unstructured data (mainly) generated from various domains and its sources of various types will be large in size but the useful part can be obtained in the analysis. Data generated by airlines is logs of engine and airline. One airplane dumps the same amount of data as Face book does per day. The aim of storing such data is just to make analysis if there is any accident. There is a need of storing big data into multiple nodes across clusters so that they can be analyzed for business advantage or any need.

REFERENCES

[1] Bash, Kepi, "Considerations for big data: Architecture and approach", Aerospace Conference, IEEE, 2012, pp. 1-7.

[2] Bo Li, “Survey of Recent Research Progress and Issues in Big Data”. Available at: www.cse.wustl.edu/~jain/cse570-13/ftp/bigdata2/index.html

[3] Madden, Sam,”From databases to big data, Internet Computing”, IEEE, Volume 16, Issue 3, 2012, pp. 4-6.

[4] Yuri Demchenko, Cees de Laat, Peter Membrey. “Defining Architecture Components of the Big Data Ecosystem”, Collaboration Technologies and Systems (CTS), IEEE, 2014, pp. 104-112.

[5] Du Zhang, “Inconsistencies in Big Data”, Cognitive Informatics & Cognitive Computing, IEEE, 2013, pp. 61-67.

[6] A.Jacobs, “The pathologies of big data”, Communications of the ACM, Volume 52, Issue 8, 2009, pp. 36-44.

[7] Web link https://en.wikipedia.org/wiki/Big_data.

[8] “What is big data”, Villanova University. Available at: http://www.villanovau.com/resources/bi/what-is-big-data/#.Vgn-xrzh5z0

[9] Arul Murugan R, Anguraj S, Boopathi R, “Big data: privacy and inconsistency issues”, IJRET, Volume 3, Issue 7, 2014, pp. 812-815.

[10] Jagdev Bhogal, Imran Choksi, “Handling Big Data using NoSQL”, Advanced Information Networking and Applications Workshops, IEEE, 2015, pp. 393-398.

[11] Couchbase, “Document Database”. Available at: http://www.couchbase.com/why-nosql/nosql-database