ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH ARTICLE

Available Online at www.ijarcs.info

# Big Data: An Introduction

Bhimendra Pal Singh and Alka Agrawal
Department of Information Technology
Babasaheb Bhimrao Ambedkar University Lucknow, UP, India

***Abstract***-We are living in data world and the data exist everywhere in different form. Data, that is beyond to storage capacity and data beyond to processing power that data called as the big data. Big data is collection of huge amount of data which summarized a technique to storing, processing, distributing and analyzing large size of dataset. The term data is increasing per day rapidly in form of volume as Megabytes, Gigabytes Terabytes, Petabytes, etc and volume is co-related to the velocity which describes fetching the data. To storing the different type of data there is used data centre which has set of data from different field. The big data is divided in three parts as structured data, unstructured data and semi-structured data.

***Keywords***- Big data, Structured data, Unstructured data, Semi-structured data.

## 1 INTRODUCTION

Big data relatives to capability of system at higher level and also it relative to capability of organization .Big data is a platform to providing accurate analysis which may lead to more concrete decision making resulting in greater operational efficiency, cost reduction and reduced risk for the business[1]. Big data can start from any point so that it has not perfect definition. Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on hand database management tools or traditional data processing application like relational database management system(RDBMS), database management system(DBMS), SQL (Structured Query language) file. Traditional system including RDMS are not capable to handling big data and challenges multiple level including capturing, storing, analyzing, sharing, searching, transferring and even visualizing the data. So that all data that is not a fit for traditional data. In market there are lots of companies which are handling to big data like Amazon, Microsoft, IBM, etc [2]. The big data is measured by 3V as **volume**, **velocity** and **variety** [3]. It is given by IBM. Data has been a backbone of any enterprise and will do so moving forward. Storing, extracting and utilizing data has been key to many company's operation.


Figure 1- Big Data

## 2 PARAMETERS OF BIG DATA

### *2.1 Volume*

Volume refers to scaling of data which is rapidly increasing in form of GB, TB, PB, etc .Let assume a file it might be of few kilobytes, a audio file that may be of few megabytes and video file that may be of few gigabytes, now imagine amount of data that might have if we put thousands file together, this is called as volume. For example Facebook Company has most amounts of data generated everyday almost 600+ terabytes per day [4].

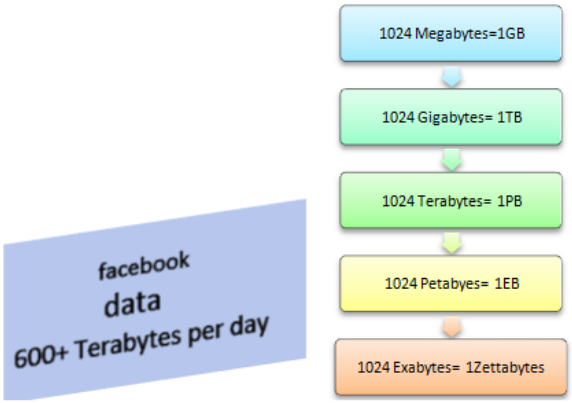


Figure 2.1-Volume

### *2.2 Velocity*

Velocity defines analyzing the streaming data that is how fast data process and in velocity there is fetching of data and sending that one towards data centre that means created velocity problem. Example Facebook which have1.65 billion active user approximate[5] and 70 thousands data in one minute produced that is Facebook generated data and processed data which follows to how fast data process.

Figure 2.2-Velocity

### *2.3 Variety*

It defines the different type data in any form. Earlier data came in form excel or RDBMS in form of row and column but in now days data come in form of photos, audio file, video file, documents, etc .For example from facebook data comes in form of post, like, comments, message, uploading pictures as from twitting from twitter, and also audio call, chatting from whatsApp and lots of resources which have data in different forms which is defined to the variety.

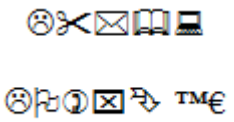

Figure 2.3-Variety

**Types of Big Data**

### *3.1 Structured data*

The data stored in table form having row and column referred as the structured data in big data and best example is RDBMS which is the most suitable for structured data. In this type of big data traditional RDBMS deal adopted. This type of data follows to defined data with repeating pattern.

### *A Table in RDBMS*

Table 1-RDBMS

*Column*

*Attributes*

*ROW*

| Roll Number | Student Name | Phone Number | City |
|---|---|---|---|
| 100 | ABC | 123456 | PUNE |
| 101 | LMN | 789456 | PUNE |
| 102 | XYZ | 987452 | PUNE |

### *3.2 Unstructured data*

In big data the unstructured data defined as variety of data and this data follow to variety and this data can be images file, video file, text file, audio file and so on.
For example WhatsApp messenger which has uploading image file, audio file, video file, text message, voice calling and more features which reflected to unstructured data.


Figure 3.2-Unstructured Data

### *3.3 Semi-Structured Data*

This data deals with structured data and unstructured data called as the semi-structured data and semi-structured data is like log files let assume a yahoo mail user login to mail and there will be generated a log file and it will be store in yahoo server and this methodology refers to semi-structured data. There is no restriction for one account for one user so that log files generated huge data.
In table a user has different login account as yahoo, Gmail, Hotmail, and Facebook for example if user access yahoo account, there are 5 account of user of yahoo and he is accessing 5 times so that he generated 20 log files and these log files will be store on yahoo server similarly for other and total log files generated 55.

Table 2-Semi Structured Data

| Account Name | Login(No. of A/c)(x) | No. of Times (y) | Log Files Generated(L) |
|---|---|---|---|
| Yahoo | 5 | 4 | 20 |
| Gmail | 4 | 4 | 16 |
| Hotmail | 5 | 3 | 15 |
| Facebook | 2 | 2 | 04 |
| Total | | | 55 |

$$L=x*y$$

Where x is number of accounts of an user, y is number of accessing of login account and L is generated log files.

## 4 CONCLUSIONS

The term big data exists everywhere around the in different form and requirement of the big data is increasing per day because any organization want to store data of every person in database and this database may be official or unofficial so that big data has a platform describing the big data in form of storing, capturing, analyzing, processing, sharing, transferring the data and these all features can be represented in different type of big data as structured big data, unstructured big data and semi-structured big data. This paper described to Big data and its types in various forms.

## 5. REFERENCES

[1] http://www.tutorialspoint.com/hadoop/hadoop_tutorial.pdf
[2] http://www.tutorialspoint.com/hadoop/hadoop_tutorial.pdf
[3] http://dashburst.com/infographic/big-data-volume-variety-velocity/statistics/?_e_pi_=7%2CPAGE_ID10%2C3284169911
[4] http://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/?_e_pi_=7%2CPAGE_ID10%2C2629924621
[5] http://zephoria.com/top-15-valuable-facebook-statistics/?_e_pi_=7%2CPAGE_ID10%2C3284169911