International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 04, April 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

BIG DATA:AN ANALYSIS OF TECHNIQUE AND TECHNOLOGY

Neena Chaudhary[1], Vikas Chaudhary[2]

[1]*Department of Computer Science, Baddi University of Emerging Science and Technology, Baddi, Solan, Himachal Pradesh (India)*

[2]*Department of Anthropology, Panjab University, Chandigarh (India)*

ABSTRACT

Big data is a large amount of dataset that includes the large number of quantities of data, like-scientific analytical data, financial data, institutional data, real time data, social media analytics and data management capabilities. big data is described by the dimensions volume, variety and velocity and another types of data. Hadoop is a programming framework which uses the map-reduce.

Keywords: Big Data; Hadoop; parameters.

I. INTRODUCTION (*DEFINITION*)

The growth of technology today day by day change and grow large amount of data like structure and unstructured form and different resources. This type of data is very difficult to manage and difficult to process that contains the billions records of millions of people information contain in many website like-social media, websites like social media, web sales, audios, text, messages, image etc. Google, yahoo, Facebook like a need of big amount of data because data is unstructured form. Google contain large amount of information so that need of data analytics process methodology contain massive data set.

II.BIG DATA PARAMETERS

The data is very big so used various sources like:

1.1 Variety

1.2 Volume

1.3 Velocity

1.1 Variety: Variety means the data is several type like structure and unstructured or semi-structure types. Several variety of data is includes in Table 1.

International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 04, April 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Table:1

Variety	Data
Time(minutes)	Number of People
30-35	30
35-40	60
40-45	110
45-50	50
50-55	30
55-60	10
60-65	10
Total	300

In Table 2.The number of children per house hold in a sample of 30 households. Draw a frequency diagram in the form of a bar chart to illustrate information like that-

Table:2

Number of households	Number of children
1	4
2	7
3	11
4	4
5	3
6	0
7	1
Total	30

Solution:

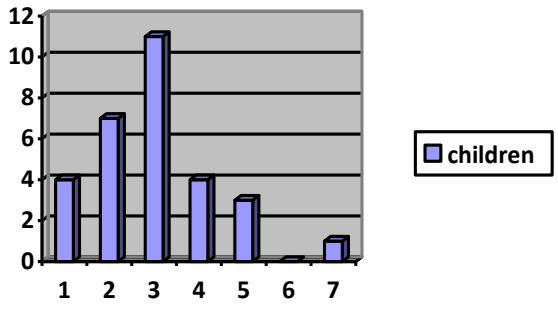

Figure:1

1.2 Volume: Volume represents size of the data. Data size represented different form like for milk which is measured in liters.

Example like-

International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 04, April 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Table:3

Data	Volume
Milk	Liter
Weight	Kilogram
Image	Pixel

The size of Data is represented in terabytes and petabytes.

1.3 Velocity: The speed of a given object with respect to direction. Speed of wind travelling with same speed such as 40 miles per hour from South dispersed in different directions are been termed under velocity as tabulated below:

Table:4

Speed of wind (miles/hour)	**Direction**
40	North
40	North-east
40	North-west
40	East

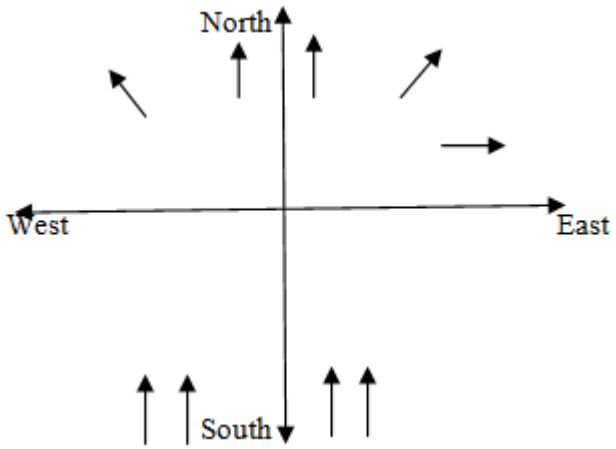

Figure:2

III. LITERATURE SURVEY:

Over the last many years, there are many researchers has completed their work successfully on big data. Hundreds of articles have appeared in the general business press (For example Forbes, Fortune, Bloomberg, Business week, The Wall street journal, The Economist)[1]. National Institute of Standards and Technology [NIST] said that Big Data in which data volume, velocity and data representation ability to perform effective analysis using traditional relational approaches [15]. In March 2012, The Obama Administration announced that the US would invest 200 Million Dollars to launch a big data research plan [2].

An IDC Reports predicts that from 2005 to 2020, the global data volume will grow by a factor of 300, from 130 Exabyte's to 40,000 Exabyte's, representing a double growth every two years[9]. IBM estimates that everyday 2.5 quintillion bytes of data are created out of which 90% of the data in the world today has created in the last two years. It is observed that social networking sites like Facebook have 750 Million users, LinkedIn has 110 million users and Twitter has 250 million users [17]. From industry, government and research community, Big

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Data has led to an emerging research field that has attracted tremendous interest. The broad interest is first exampled by coverage on both industrial reports and public media for example: The economist, New York Times [12]. Mobile Phones becoming best way to get data on people from different aspect, the huge amount of data that mobile carrier can process to improve our daily life [13]. In Year 2005, it would appear from this graph that the amount of data was practically increased. However, Consider exponential growth in data from 2005 year, when enterprise system and user level data was flooding into data warehouse [11].

IV.TECHNIQUE

- Text mining
- Machine learning
- Statistical programming
- Schema-less database
- Map reduce
- Hive
- Pig
- Storage technology

V.TECHNOLOGY

- Hadoop
- NOSQL
- Text mining
- Machine learning
- Statistical programming
- Apache
- MongoDB
- Tera Data
- Green plum
- Cassandra couch DB

Hadoop:

It is a most popular programming framework.it used to large-data set in a distributed computing system. In a past experience the small programming framework performance is very small but hadoop is a popular and used computing environment. Google's Map reduce developed Hadoop that is a software framework where an application break down into various components.

The Current Apache Hadoop ecosystem consists of the Hadoop Kernel, Map reduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper . MapReduce is a programming framework for distributed computing system which is created by the Google in which divide and conquer method is used to break down the large complex data into small units of code . MapReduce have two types which are :

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Map ():- The master node takes the input, divide into smaller subparts and distribute into worker nodes. A worker node further do this again that leads to the multi-level tree structure. The worker node process the m=smaller problem and passes the answer back to the master Node.

Reduce ():- The, Master node collects the answers from all the sub problems and combines them together to form the output.

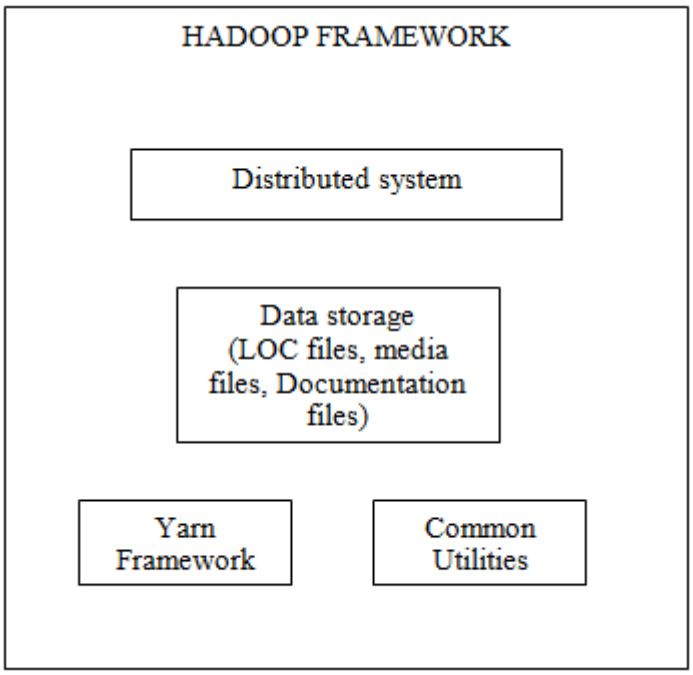

Figure:3

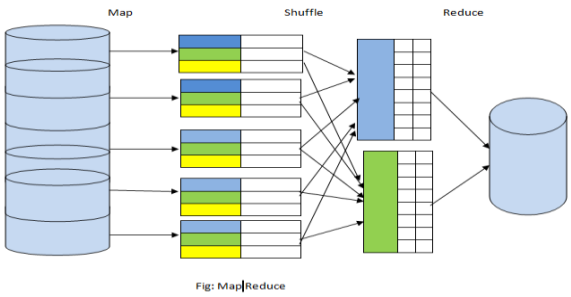

Figure:4

HADOOP FILE SYSTEM (HDFS)

Hadoop file system is simple a file system applications that lots amount of data stored in a HDFS file system. HDFS file system is a distributed system that the numbers of nodes and files are distributes each system .it is run on hardware devices and low cost. The performance of HDFS file system is higher than traditional computing system. Some examples of file system like –

- Tightly coupled system
- Loosely coupled system

Tightly coupled system:

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Tightly coupled system is a single system. The primary memory that is shared by all the processor. In this system, communication between the processor is done by message passing and interconnecting the processors. In software Engineering the term tightly coupled system is used to define software that works only in a single part of a particular type of system and the software depends on other software. It is further explained by an example as, an operating system considered to be tightly coupled as it depends on software drivers to correctly install and activate the system's peripheral devices.[18]

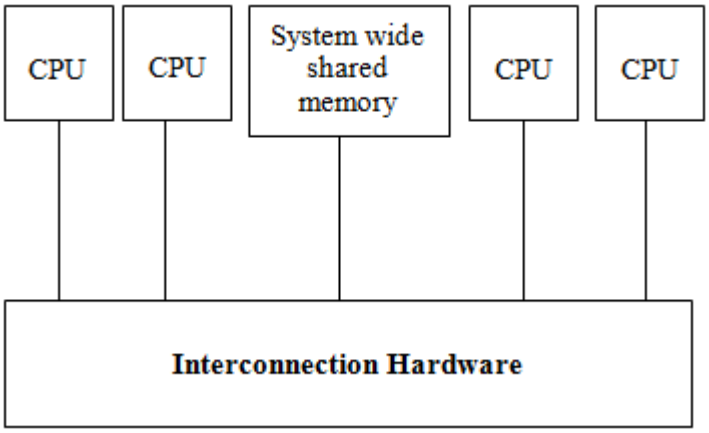

Figure:5

Loosely coupled system: Loosely coupled system is a distributed system.it does not share memory and each processor has its own local memory. It is easily communicated in each other's network.

Components like connectors, nodes in a loosely coupled system can be interchanged with another implementations which provides the same work. Thus Components are less constrained to the same platform, operating system, language or build environment.[19]

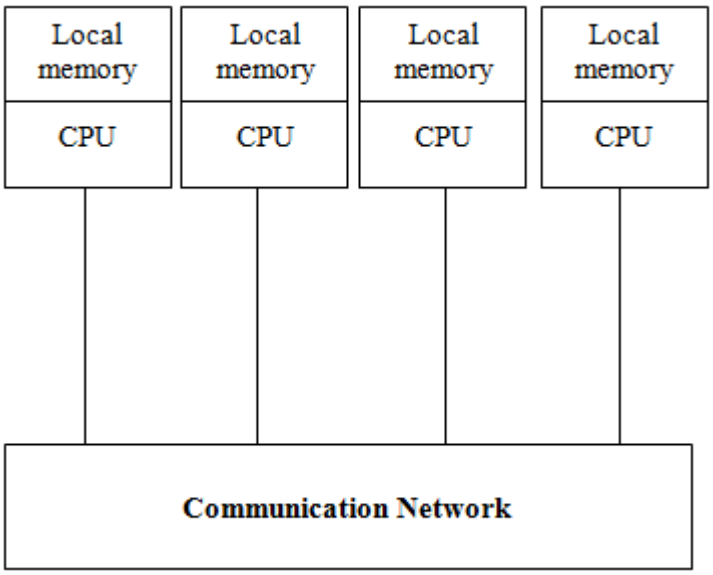

Figure:6

NOSQL: A NOSQL means non-relational database. It is a storage database and easily retrieval of data which is modeled in means and other relation like a tabular form.

There are several approaches for NOSQL like-

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

- Column
- Document
- Key-value
- Graph
- Multi-value database
- Relational database

Characteristics of NOSQL Database-

NOSQL Avoids-

- Overhead of ACID Transactions
- Complexity of SQI query
- DBA presence
- Database schema

PROVIDE-

- Easy and frequent changes to Database
- Horizontal scaling
- Fast development

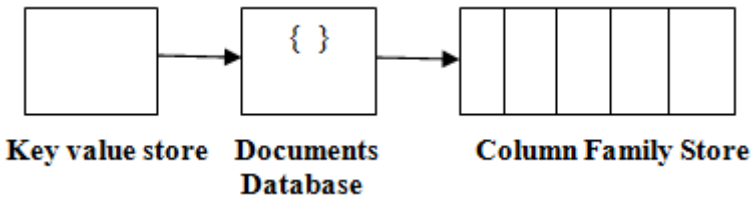

Figure:7

SQL:

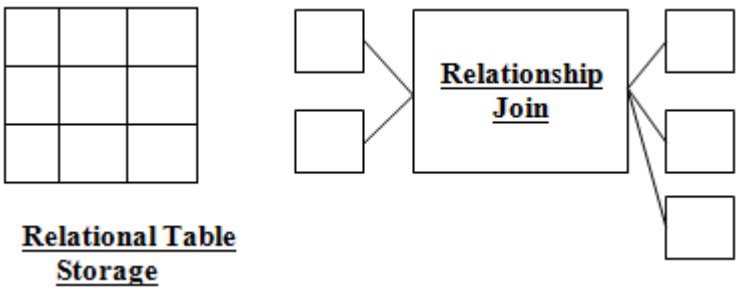

Figure:8

Text-Mining :

Text-Mining also referred to Text Data like messages, email, fax etc.It is a natural language stored high-quality information from Customer support, Technical support,Advertising & Marketing, Human resources and Competitor.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

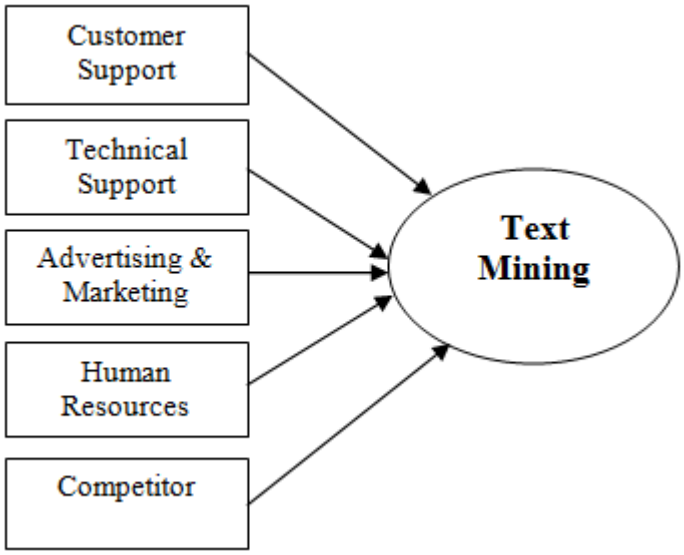

Figure:9

Machine Learning : Machine Learning is actually a lot of things. The field is quite vast and is expanding rapidly. Firstly the machine is trained then particular work is started.Without user input ,machine does not proceed.

Example: Neural Network

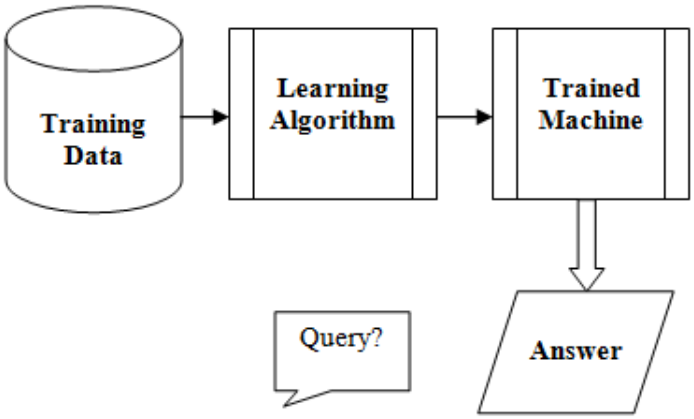

Figure:10

PIG: PIG is a high level scripting language that is used with Apache Hadoop. PIG is used with workers to write complex data transformation without Java.

The need of New Technology :

- RDBMS are the Square peg for all round holes.
- ACID properties compliant data stores are not required by all data-like atomicity, consistency, isolation, durability.
- To implement ACID, tradeoff limits scalability of traditional system is required.

VI.CONCLUSION

After a complete understanding I concluded my paper by describing Techniques and Technology by which Big Data can be analyzed. The concept of Big Data generated from multiple sources forming the complexity such as

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

volume, velocity and variety. I have described how multi-national companies completed their work successfully with the help of this technique by citing various examples of United States. Big Data with NATGRID is used in ambitious counter terrorism program to study and analyze the huge amounts of data from intelligence and enforcement agencies to help track suspected terrorists and prevent terrorist attacks.

REFERENCE

[1.] Puneet Singh Duggal, Sanchita Paul, Department of Computer Science & Engineering BirlaInstitute of Technology Mesra, Ranchi, India," Big Data Analysis: Challenges and Solutions", International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV

[2.] Subramaniyaswamy Va, Vijayakumar Vb, Logesh Rc and Indragandhi Vd," Unstructured Data Analysis on Big Data using Map Reduce", 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15).

[3.] Ishwarappa, Anuradha J," A Brief Introduction on Big Data 5V's Characteristics and Hadoop Technology", International Conference on Intelligent Computing, Communication & Convergence(ICCC-2015)

[4.] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , "Shared disk big data analytics with Apache Hadoop"

[5.] Aditya B. Patel, Manashvi Birla, Ushma Nair ,(6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce"

[6.] Neil Raden,"Big Data Analytics Architecture", Hired Brains Inc, 2012

[7.] James Manyika, Michael Chui, Brad Brown, Jacques Bhuhin, Richard Dobbs, Charles Roxburgh, Angela Hungh Byers, "Big Data: The next frontier for innovation, competition and productivity", June 2011.

[8.] Wei Fan, Albert Bifet, "Mining Big Data: Current Status and Forecast to the Future", SIGKDD Explorations, Volume 14, Issue 2.

[9.] Jaradat, Shatha, Nima Dokoohaki, and Mihhail Matskin. "OLLDA: A Supervised and Dynamic Topic Mining Framework in Twitter." In 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 151354-1359. IEEE, 2015.

[10.] Yi, Xiaomeng, Fangming Liu, Jiangchuan Liu, and Hai Jin. "Building a network highway for Big data: architecture and challenges." IEEE Network 28, no. 4 (2014): 5-13.

[11.] Hargittai, Eszter. "Is bigger always better? Potential biases of Big data derived from social network sites." The ANNALS of the American Academy of Political and Social Science 659, no. 1 (2015): 63-76.

[12.] Wang, Ke, Xraobing Zhou, Tonglin Li, Dongfang Zhao, Michael Lang, and Ioan Raicu. "Optimizing load balancing and data-locality with data-aware scheduling." In Big Data (Big Data), 2014 IEEE International Conference on, pp. 119-128. IEEE, 2014.

[13.] Zhao, Dongfang, Zhao Zhang, Xiaobing Zhou, Tonglin Li, Ke Wang, Dries Kimpe, Philip Carns, Robert Ross, and Ioan Raicu. "Fusionfs: Toward supporting data-intensive scientific applications on extreme-scale high-performance computing systems." In Big Data (Big Data), 2014 IEEE International Conference on, pp. 61-70. IEEE, 2014.

[14.] Li, Tonglin, Xiaobing Zhou, Kevin Brandstatter, Dongfang Zhao, Ke Wang, Anupam Rajendran, Zhao Zhang, and Ioan Raicu. "ZHT: A light-weight reliable persistent dynamic scalable zero-hop distributed

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

hash table." In Parallel & distributed processing (IPDPS), 2013 IEEE 27th international symposium on, pp. 775-787. IEEE, 2013.

[15.] Kim, Jaein, Nacwoo Kim, Byungtak Lee, Joonho Park, Kwangik Seo, and Hunyoung Park. "RUBA: Real-time unstructured network Big data framework." In 2013 International Conference on ICT Convergence (ICTC), pp. 518-522. IEEE, 2013.

[16.] www. Searchbusiness analytics.techtarget.com

[17.] www.ebizmba .com/articles/social-networking-websites.

[18.] *http://www.webopedia.com/TERM/T/tight_coupling.html*

[19.] https://en.wikipedia.org/wiki/Loose_coupling