Survey on Basic Concepts of BigData

S.Manasa[1] Y.Rama Mohan[2]
[1]M.Tech. Student [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]GPREC Collage, JNTU Anantapuram

Abstract— Big Data, a new technology is a collection of large amount of data sets which include medical data, business data, weather data, organizations data, etc.,. The size of the data will ranges from dozens of terabytes to many petabytes. Because of the presence of such a huge amount of data it becomes tough to handle by traditional data processing applications such as relational data base management systems, desktop statistics and visualization packages. So a massive parallel processing software technique that will run on hundreds and thousands of servers are required. The challenges are listed as storing, retrieving, sharing, searching, transfer, analysis, capture, and visualizing. There are many techniques that are existing to handle and process all these data sets in which some are hadoop, map reduce, hive, pig, zookeeper, Hbase and so on.

Key words: Terabytes, Petabytes, Traditional Data Processing Applications, Hadoop, MapReduce, Hive, Pig, Zookeeper, HBase

I. INTRODUCTION

Big Data is an umbrella term, which will encompasses everything from digital data to health data collected from years and years of paper work issued and filled by the government. The name big data has been generated from large amount of zeroes and ones collected in a single year, month, and day even an hour. Big Data is not just large amount of data that is available from multiple sources, it also refers to the complete process of gathering, storing and analyzing that collected data, and mainly this complete process is being used to make the world a better place. The main concept of big data is nothing but, the thing which is not known to us, we simply Google it. And in a fraction of seconds we will get number of related links which will make a good example for big data. The data that is available from multiple, heterogeneous, autonomous sources, in extreme large amount, will get updated in fraction of seconds. And the main sources of data are Google which will receives over 2 million queries, e-mail users will send over 200 million messages, YouTube users will upload 48 hours of video, Facebook users shares over 680,000 pieces of content, and Twitter will generates 100,000 tweets. In another words big data is simply getting entire world to a single place. The data from some sources are not observable. For example consider the data collected in vast quantity from sensors in metrological and climate systems, patients monitoring systems in hospitals, controlling systems in cars, flying machine, cell towers, and power plants all will collect endless data streams. Mainly the healthcare industry alone will collect data about patient's records, medicines, diseases; doctors, various hospitals and medical billing details, and also insurance companies will collect data regarding each and every claim. All these data are frequently updated each and every second which will finally forms big data. The data base management system concentrations on collection, storage, management, and retrieval of data. Later it has become quite common that businesses to have multiple databases from multiple sources with their own layouts. A data warehouse is a variety of database that focuses on aggregation and integration of data from different sources for exploration and reportage purpose. But many fields in big data focus on extraction of information. For example, business intelligence systems focus on providing historic, present and foretelling views for business making use of it. Data visualization is an essential field in big data that focus on the expansion of methods that can help analysts visualize interesting patterns or relationships in data.

A. *Why BigData*

The key enablers for the growth of big data are:

- Growth of storage capabilities.
- Rise of processing power.
- Ease of use of data.

The main features of big data are it is enormous in size, the data will keep on updating from time to time, and data is collected from multiple and different sources, it is free from influence and control of any one, and it is too much complex in nature and difficult to handle.

B. *Definition:*

Big Data is a collection of large amount of datasets that are ranging from terabytes to petabytes and cannot handle by traditional database management systems such as DBMS or RDBMS. The challenges are storing, searching, retrieving, capture, visualization, transfer and capture.

II. BIGDATA AS 3 V'S

The main 3 V's of big data are:

- Volume
- Velocity
- Variety

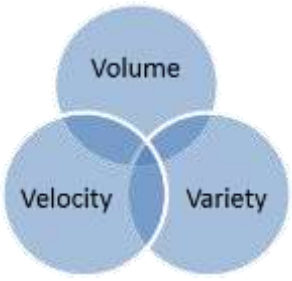

Fig. 1: 3 V's of big data.

A. *Volume*

The size of the data that is being generating by multiple networking sites. For each and every second large amount of data is been generated by users, which will automatically results in the storage problem due to its large amount of size.

B. Variety

If we want any information about anything we simply Google it and we can obtain different types of related information about we are searching. For example if we want to book a car, there are large numbers of websites which will provide detailed information about each and every thing about different cars. So large amount of variety of data is available from different websites.

C. Velocity

Velocity is nothing but the speed at which these data sets are generated. Every second data is been updating from time to time and is helpful to the user as per user needs and conditions.

III. TECHNIQUES IN BIGDATA

A. MapReduce

MapReduce is one of the programming model and a connected implementation process for storing and processing all collected large amount of data sets with a parallel distributed algorithms on clusters. A map reduce program is composed of two procedures, a map() procedure and a reduce() procedure. A map() procedure will perform all storing and filtering operations and a reduce() procedure will perform a summary operation on the results obtained from map() procedure. It consists of master and slave nodes in its framework. Map reduce system is an infrastructure or can also be called as framework which will provide duplications of functions or components with the intension of increasing reliability of the system and able to keep on functioning to a level of consummation in the presence of faults.

1) Map Step

In this step, the master node will takes all the collected data as input and divides that into number of sub problems and then distributes them to worker nodes or slave nodes. These worker nodes or slave nodes will do the same process again in turn which will leads to a multi-level tree structure. The worker or slave node process the smaller sub problems and sends the answer back to its master node.

2) Reduce Step

Whenever the slave nodes will process the data and sends back the individual results, the master node will collects all those individual results from all the slave nodes and combine them to form a final result as output to the user. Map reduce allows distributed and parallel processing of map and reduce operations. It is beneficial in extensive sort of applications like distributed form based searching; machine learning and statistical machine transformation.

B. Hadoop

Apache hadoop is an open source framework for storing and processing large scale data sets on clusters. It is an apache top level project which is composed of four modules that are hadoop common, hadoop distributed file system (HDFS), hadoop YARN, hadoop map reduce. It is a free java based programming framework that allows parallel processing of large scale datasets in a distributed computing environment. Hadoop will run applications on systems that consists of thousands of nodes and its distributed file system will rapidly transfer data among thousands of nodes and allows the continuous un interrupted processing of system in case of any node failure. Hadoop was mainly inspired by Google's map reduce. The present hadoop framework will consist of hadoop kernel, map reduce, hdfs, and number of related projects like hive, hbase, zookeeper, pig.

1) Hadoop Distributed File System (HDFS)

HDFS is the hadoop distributed file system, and the primary storage used by the hadoop applications. It is a virtual file system that will provide high performance for data access on clusters. The important and main difference between other file systems and hadoop HDFS is, when we move a file on HDFS it automatically splits the file into small pieces and make the replicas of these file and will store on other nodes for providing fault tolerance. It consists of name node and multiple data nodes. The main task of name node is to open, close, rename files and directories. And data nodes are for read and write to the file system. The files and directories are represented on the name node by inodes.

IV. CONCLUSION

As data is being increasing day by day due to the wide range use of social networking sites, the problem will raises like how to store, process, manage, use all these large amount of data. So Big Data a new technology can be used to store, process and manage all these huge amount of data which will takes help of different emerging technologies like hive, pig, map reduce, hadoop. By the use of big data a user can access the past, present and predicted data which has been stored and analyzed from past years and which is mainly useful for business analyses.

REFERENCES

[1] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.
[2] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.
[3] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.
[4] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.
[5] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.
[6] E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," Nature, vol. 489, pp. 49-51, 2012.
[7] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.
[8] S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences," Science, vol. 323, pp. 892-895, 2009.
[9] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinSey Quarterly, 2010.

[10] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," Science, vol. 329, pp. 1194-1197, 2010.
[11] Dhole Poonam B, Gunjal Baisa L, "Survey Paper on Traditional Hadoop and Pipelined Map Reduce" International Journal of Computational Engineering Research||Vol, 03||Issue, 12||

