A SURVEY OF BIG DATA ANALYTICS

Nirali Honest[1] and Atul Patel[2]

[1]Smt. Chandaben Mohanbhai Patel Institute of Computer Applications,CHARUSAT, Changa
[2] Smt. Chandaben Mohanbhai Patel Institute of Computer Applications,CHARUSAT, Changa

ABSTRACT

Due to the arrival of new technologies, devices, and communication means, the amount of data produced by mankind is growing rapidly every year. This gives rise to the era of big data. The term big data comes with the new challenges to input, process and output the data. The paper focuses on limitation of traditional approach to manage the data and the components that are useful in handling big data. One of the approaches used in processing big data is Hadoop framework, the paper presents the major components of the framework and working process within the framework.

KEYWORDS

Data Mining, Big Data, Data analytics, Hadoop Framework.

1. INTRODUCTION

The term big data is used for data that go beyond the processing power of traditional database systems. The general characteristics of the big data is that, the size of data is too big, the generation of data is too fast and most of the times the data is not directly in the form suitable for the database systems. As the term data differs from the big data, so also the processing required handling the big data differs from the conventional computing techniques. The digitization process is tremendously fast [1] and due to that the production of data is almost in digital form and the data generated is increasing in size exceeding Exabyte. In accordance to data generation the computer systems are much faster than the old systems, yet analyzing [2] of large scale data is a critical factor.

2. TRADITIONAL APPROACH OF DATA MINING

The process of data mining includes the operations like selection, preprocessing, transformation and evaluation of data [12] in the discovery of knowledge. The first task in data mining is data input which includes collecting, selecting, preprocessing the data. Preprocessing includes cleaning and filtering the data to make it useful for further activities. After the data is in cleaned and reduced from various data mining methods like clustering, classification, association rules, and sequential patterns can be applied for data analysis. Most of the methods cannot be applied to big data because of the following reasons,

- They are designed to work with a single machine with all the data in the memory. Most of the methods are not for huge and complex data.

- Most of the methods cannot produce the analysis dynamically based on the input.
- Most of the methods work with the same format of input.

After applying the methods evaluation and interpretation are applied to generate the output. They provide the mechanism to measure the results. The output can be measured for the operators like number of errors, accuracy of results, computation speed, computation cost, response time, utilization of memory, etc. The knowledge generation becomes for complex and need to be more versatile for handling the big data.

2.1. Big Data

Big data involves the data produced by different devices and applications. Below are some of the fields which can generate big data.

- **Social Sites Data:** Social media such as Facebook, Twitter, etc. carry information, suggestions, invitations, etc. posted by several people across the world. The responses for their campaigns, advertising mediums, etc are also known.
- **Search Engine Data:** Search engines retrieve lots of data from different databases.
- **Medical History Data:** Hospitals can generate medical history of patients for various health issues.
- **Online Shopping Data:** Shopping of various products online can help to know the preferences and product perception of the customers on different products at different intervals.
- **Stock Exchange Data:** The stock exchange data holds information about the shares of various companies. These data given an insight on the decisions taken by shareholders for the trading activities.
- **Vehicle Booking Data:** Booking of vehicles like train, bus, flight, cab, etc. can generate the data of booking a vehicle based on model, size, distance and availability of a vehicle.
- **Aviation Data:** Audio and Video data recordings, the performance information of the aircraft, etc.

2.2. Characteristics of Big Data

Big data is really critical to handle as it is emerging as one of the fastest technologies in current era. The importance of big data is in analytical use which can help in generating informative decision to provide better and quick service. The big data has three characteristics, known as data volume, velocity and variety, it is also known as 3Vs [3] , which means that the size of data is large, the data is generated very fast, and the data exists in heterogeneous formats which can be among structured data, semi structured data and unstructured data captured from different sources. The common concept of 3Vs is given below,

Volume

The main attraction of big data analytics is to process large amounts of information. The volume presents the major challenge for the traditional approaches of data analytics. It motivates the use of parallelism and distributed approach in computation. Most of the organizations and firms have large amount of data but they don't have the capacity to process it.

Velocity

The speed at which the data generated in an organization defines the velocity of data, which is cumulatively increasing. With the increase in use of Internet with different devices the services have become faster and the services are increasingly instrumented, which gives rise in the rate of data velocity. Those who are able to quickly utilize that information, for example, by suggesting options, the company can take advantage of selling of more products. The smart phone era increases again the rate of data inflow, as consumers hold the source of data that can be in more than one form. The data should be streamed, for feasible storage space and for applications that require immediate response to the data. As with velocity of input data, velocity of output data also matter, the results may impact on decision making.

Variety

Mostly the data is in the unordered and unstructured format which requires processing. The data may be generated from diverse sources. The data can be in the text, image, audio, video, etc. formats. None of these data are readily in the acceptable formats for integration into an application. A general characteristic of big data processing is to take unordered and unstructured data and extract ordered meaning, for consumption either by an individual or an application. In the course of generating processed data from the source data there can be loss of information. Based on the nature of data the storage can be fixed to make it simpler and efficient, like using relational database, XML, Graphs, etc. selecting the right approach to provide enough structure to organize data is an important part of big data.

Apart from the 3Vs, more components were added to explain big data [4][5] like, value, venue, vagueness, veracity, validity, vocabulary and variability. The report of IDC [6] indicates that the marketing of big data is about $16.1 billion in 2014. Another report of IDC [7] forecasts that it will grow up to $32.4 billion by 2017. The reports of [8] and [9] further pointed out that the marketing of big data will be $46.34 billion and $114 billion by 2018, respectively.

Observing the above information it becomes mandatory to have an insight of big data with the knowledge of tool selection.

3. TOOL SELECTION

Big data is not about data, it involves various tools, techniques and frameworks use to manage the data. There are many big data platforms available with different characteristics, selection of the platform [10] depends on the capability of the platform and various dimensions as listed in table 1.

Table 1: Dimensions of big data

Dimension	Description	Issues in selection of dimension
Big data solution	Can be implemented as software, appliance or cloud based. Can be implemented as a hybrid solution, also.	Data locality Privacy and regulation Human resources Project requirements
Data transfer	Can be implemented as processing by transporting the data or processing without transporting the data.	Cost of time or money to transfer it. Size of data to transfer Locality of data, especially with rapidly updating data.
Data Structurization	Can be implemented as performing the data acquisition and cleaning by ourselves or make your data available for the marketplace.	Cost of time and money for data cleaning. Quality of data Selection of market place

Technologies used to handle big data play an important role in data analysis which leads in the accuracy of decision making resulting in cost reduction, faster services, considering calculative risks, and gaining operational efficiencies. To manage and process the large volume of data selecting correct infrastructure is very important. The technologies can be used in capturing and storing the big data and analyzing big data.

3.1. Technologies used to handle the Big Data

NoSQL database

For capturing and storing the data, NoSQL database are commonly used. MongoDB provides facilities for capturing and storing data. Other database like CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper etc, are also used. NoSQL provide support for cloud computing architectures which gives the benefit in reduction of cost, increase in speed of computing and increase in efficiency. It also provides the facility to generate patterns and trends without need for additional infrastructure.

Massively Parallel Processing and MapReduce

For big data processing, systems like Massively Parallel Processing (MPP) and MapReduce are commonly used.MPP includes multiple processors, and each processor works on different parts of the program and has its own operating system and memory to utilize. MPP processors communicate using a messaging interface. Data paths are interconnected to allow message passing among processors. Usually partitioning of common database and assigning work among processors is quite complex process.

MapReduce is a computational approach that involves breaking large volumes of data down into smaller batches, and processing them separately. MapReduce is a programming model, Google has used successfully is processing its big data sets [11]. The computation is done in terms of map and a reduce function. A cluster of computing nodes which are built on commodity hardware will scan the batches and aggregate their data. Then the multiple nodes' output gets merged to generate the final result data.

MPP has many things in common with MapReduce. In MPP, as in MapReduce, processing of data is distributed across many compute nodes, these separate nodes process their data in parallel and multiple output sets are assembled together to produce a final result set. But, for a variety of reasons, MPP and MapReduce are used in rather different scenarios as listed in table 2.

Table 2: MPP and MapReduce characteristics

MPP	MapReduce
MPP gets used on expensive, specialized hardware tuned for CPU, storage and network performance.	MapReduce is generally deployed to clusters of commodity servers that in turn use commodity disks.
MPP products are queried with Structured Query Language (SQL).	MapReduce's native control mechanism is Java code.
Loading of data is slower.	Loading of data is faster.
Querying is easier.	Forming maps and reducing is complex.
For structured data MPP is good in data refining and transformations.	For semi-structured and unstructured data refining and transformations is faster depending on file type and transformation logic.
When it is required to call for fast, iterative queries and analytics on huge amount of structured and multi-structured data it performs good.	When it is required to capture, store and refine unstructured and semi-structured data in its native format it performs good.

Storage

For storage most commonly Amazon Simple Storage Service (Amazon S3) and Hadoop Distributed File System (HDFS) are used. S3 provides developers and IT teams with secure, durable, highly-scalable object storage. It is easy to use as it provides a simple web service interface to store and retrieve any amount of data from anywhere on the web. There is no setup cost. HDFS is a Java-based file system that provides scalable and reliable data storage, and it was designed to cover large clusters of commodity servers. HDFS has proven good in scaling and forming clusters of servers, which can support billions files and blocks.

Knowledge of these components is very important in managing the big data. Apart from these components, various frameworks are available to operate the big data. One of such frameworks is Hadoop. The next section describes the fundamentals of the framework.

4. HADOOP FRAMEWORK AT A GLANCE

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models.

In traditional approach, an enterprise will have a computer to store and process big data as shown in figure 1. Here data will be stored in an RDBMS like Oracle Database, MS SQL Server or DB2 and sophisticated softwares can be written to interact with the database, process the required data and present it to the users for analysis purpose.

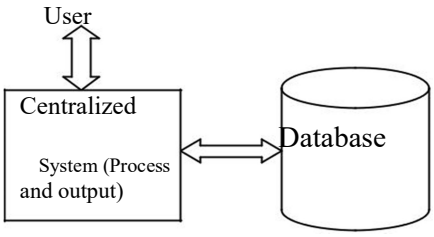

Figure 1: Traditional approach to store and process big data

This approach works fine when we have a smaller amount of data that can be holded by standard database servers and processed in an optimum way by the processor. But when the data is in huge amount, it becomes time-consuming and tedious task to process the data through traditional database server.

Google opted a new approach to tackle the large volume of data using an algorithm called MapReduce. This algorithm divides the input into small parts and assigns those parts to many computers connected over the network, and collects the results to form the final result output. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different CPU nodes as shown in figure 2.

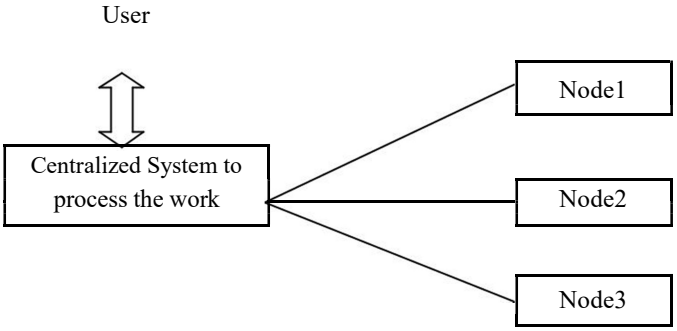

Figure 2: Working of MapReduce algorithm to store and process data in parallel

Hadoop framework is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for huge amounts of data.

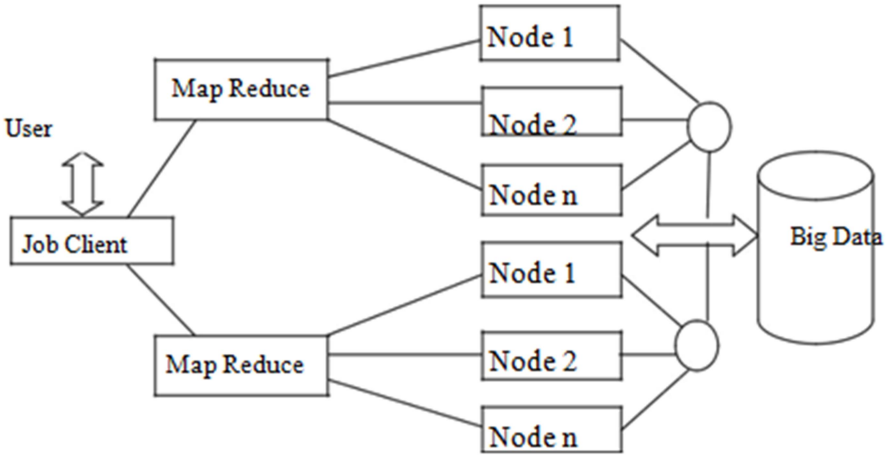

Figure 3: Hadoop framework

4.1. Components of Hadoop

A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers as shown in figure 3. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage. Hadoop framework includes following four modules:

- Hadoop Common contains the Java libraries and utilities required by other Hadoop modules. The libraries provide filesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.

- Hadoop YARN is a framework for job scheduling and cluster resource management.

- Hadoop Distributed File System (HDFS) is a distributed file system that provides high-throughput access to application data. HDFS uses a master/slave architecture where master manages the file system metadata and one or more slaves store the actual data.

- Hadoop MapReduce is a YARN-based system for parallel processing of large data sets. In works in two phases, in first phase also known as map function, it takes input data and converts it into a set of data, where individual elements are broken down into key/value pairs, in the second phase, known as reduce function, the output is taken from a map task as input and it combines these data sets into a smaller set of records. The MapReduce framework implements master-slave architecture. The master is responsible for resource management, tracking resource consumption/availability and scheduling the jobs component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves execute the tasks as directed by the master and provide task-status information to the master periodically.

4.2. Working of Hadoop

Hadoop working can be divided into phases, in the first phase, a user or application can submit a job to Hadoop job client with specification of location of input and output files in the distributed file system, jar files containing map and reduce functions and configuring job by setting various parameters related to the job. In the second phase Hadoop job client submits the job and configuration to the MapReduce master called as JobTracker, which distribute the jars/executables to the slaves, it schedules tasks monitor them. In last phase the slaves on various nodes execute task as per MapReduce implementation and output of the reduce function is stored into the output files on the file system.

Advantages of Hadoop

- It is compatible on all the platforms.
- It distributes the data and tasks across the nodes automatically, which allows the users to write and execute the distributed systems quickly.
- The library has the API to detect and handle failures at the application layer which relieves the framework to rely on the hardware for fault tolerance.
- The framework continues to operate smoothly with the addition and removal of servers dynamically.

With the increase in big data, it becomes mandatory to process the data timely and accurately in order to gain value from it. Hadoop is one of the frameworks which provide a good mechanism to handle the distributed processing of data.

5. CONCLUSIONS

In this paper, we presented the significance of big data, limitation of traditional approach in analyzing big data and tools used in handling the big data. We explored the Hadoop framework with its significance and performance. The major issues in processing the big data are computation speed, quality of output, key skills required in operating the tools, compatibility of the tool, security, and privacy.

ACKNOWLEDGEMENTS

The authors would like to thank CHARUSAT University to provide the necessary resources to carry out the work.

REFERENCES

[1] Lyman P, Varian H. How much information 2003? Tech. Rep, 2004. [Online]. Available: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf.
[2] Xu R, Wunsch D. Clustering. Hoboken: Wiley-IEEE Press; 2009.
[3] Laney D. 3D data management: controlling data volume, velocity, and variety, META Group, Tech. Rep. 2001. [Online]. Available:http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
[4] Van Rijmenam M. Why the 3v's are not sufficient to describe big data, BigData Startups, Tech. Rep. 2013. [Online].Available: http://www.bigdata-startups.com/3vs-sufficient-describe-big-data/.

[5] Borne K. Top 10 big data challenges a serious look at 10 big data v's, Tech. Rep. 2014. [Online]. Available: https://www.mapr.com/blog/top-10-big-data-challenges-look-10-big-data-v.
[6] Press G. $16.1 billion big data market: 2014 predictions from IDC and IIA,Forbes, Tech. Rep. 2013. [Online]. Available: http://www.forbes.com/sites/ gilpress/2013/12/12/16-1-billion-big-data-market-2014-predictions-from-idc-and-iia/.
[7] Big data and analytics—an IDC four pillar research area, IDC, Tech. Rep. 2013.[Online].Available: http://www.idc.com/prodserv/FourPillars/bigData/index.jsp.
[8] Taft DK. Big data market to reach $46.34 billion by 2018, EWEEK, Tech. Rep. 2013. [Online]. Available: http://www.eweek.com/database/big-data-market-to-reach-46.34-billion-by-2018.html.
[9] Research A. Big data spending to reach $114 billion in 2018; look for machine learning to drive analytics,ABI Research, Tech.Rep.2013.[Online].Available:https://www.abiresearch.com/press/
[10] Agneeswaran VS, Tonpay P, Tiwary J (2013) Paradigms for realizing machine learning algorithms. Big Data 1(4):207–214.
[11] Dean, J. and Ghemawat, S. 2008. MapReduce: simplified data processing on large clusters. Communication of ACM 51, 1 (Jan. 2008), 107-113.
[12] Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag.1996; 17(3):37–54.

Authors

1) Atul Patel

Dr. Atul Patel is Principal and Dean, at Smt. Chandaben Mohanbhai Patel Institute of Computer Applications – Faculty of Computer Science and Applications, Charotar University of Science and Technology (CHARUSAT) Changa, India. His main research interest includes Wireless Communication and Network Security, Cloud Computing, Data and Web mining. He has published more than 40 articles and research papers in several national and international journals.

2) Ms. Nirali Honest

Nirali Honest is pursuing PhD in Computer Science and Applications from Charotar University of Science and Technology (CHARUSAT). She is working as an Assistant Professor at Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Charotar University of Science and Technology (CHARUSAT), Changa, Gujarat, India. Her research interests include data and web mining.