A Survey on Big Data Storage

Gagandeep Kaur, Maninder Kaur, Vijay Paul Singh

M.Tech CSE student, Assistant Professor, Research Scholar

Doaba institute of engineering & Technology affiliated by PTU Jalandhar, Punjab, India

Panjab University Chandigarh, India

gagandeep.v21@gmail.com, maninderecediet@gmail.com, vpaul9678@gmail.com

Abstract: In the 21st century digital data generated from various sources and growing exponentially. Generated data from various sources are in the different format such as structure, unstructured & semi-structured and huge in size (GB, PB, TB etc). This types of data are complex in nature which cannot be stored by traditional databases. Traditional databases get slow, expensive and gives degrade performance, in the form storing, processing & analysis. To handle Big data Hadoop is introduced by Apache, which can store different type of data in seconds and it is quite better than traditional storage databases. In this paper, we survey on traditional databases storage vs hadoop.

Keywords: HDFS, MapReduce, Hadoop, Processing, Storing.

I. INTRODUCTION

Massive volume of data sets are being collected by various websites and other domains in various formats. This large and complex collection of data sets is called Big Data. Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them [1]. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy [3]. Generally there are three form of data: Data can be structured, unstructured or semi structured. Structured- In form of tables[14]. Unstructured- In form of videos, images, etc[15]. Semi-structured- Data present in tag forms like in xml, html, etc[14].

Areas where BIG DATA can be used are Banking and Securities- securities fraud early warning, tick analytics, card fraud detection, archival of audit trails, enterprise credit risk reporting, trade visibility, customer data transformation, social analytics for trading, IT operations analytics, and IT policy compliance analytics, among others. Big Data providers specific to this industry include: 1010data, Panopticon Software, Streambase Systems, Nice Actimize and Quartet FS [17].

Communications, Media and Entertainment- Create content for different target audiences Recommend content on demand Measure content performance Spotify, an on-demand music service, uses Hadoop big data analytics, to collect data from its millions of users worldwide and then uses the analyzed data to give informed music recommendations to individual users. Amazon Prime, which is driven to provide a great customer experience by offering, video, music and Kindle books in a one-stop shop also heavily utilizes big data [16].

Healthcare Providers-The healthcare sector has access to huge amounts of data but has been plagued by failures in utilizing the data to curb the cost of rising health care and by inefficient systems that stifle faster and better healthcare benefits across the board. This is mainly due to the fact that electronic data is unavailable, inadequate, or unusable. Additionally, the healthcare databases that hold health-related information have made it difficult to link data that can show patterns useful in the medical field. Some hospitals, like Beth Israel, are using data collected from a cell phone app, from millions of patients, to allow doctors to use evidence-based medicine as opposed to administering several [18].

Education- From a technical point of view, a major challenge in the education industry is to incorporate big data from different sources and vendors and to utilize it on platforms that were not designed for the varying data [19].

Manufacturing and natural resources - Increasing demand for natural resources including oil, agricultural products, minerals, gas, metals, and so on has led to an increase in the volume, complexity, and velocity of data that is a challenge to handle. Similarly, large volumes of data from the manufacturing industry are untapped. The underutilization of this information prevents improved quality of products, energy efficiency, reliability, and better profit margins [20].

Government- Big data is being used in the analysis of large amounts of social disability claims, made to the Social Security Administration (SSA), that arrive in the form of unstructured data. The analytics are used to process medical information rapidly and efficiently for faster decision making and to detect suspicious or fraudulent claims. The Food and Drug Administration (FDA) is using big data to detect and study patterns of food-related illnesses and diseases. This allows for faster response which has led to faster treatment and less death [21].

Insurance -Big data has been used in the industry to provide customer insights for transparent and simpler products, by analyzing and predicting customer behavior through data derived from social media, GPS-enabled devices and CCTV footage. The big data also allows for better customer retention from insurance companies [22].

When it comes to claims management, predictive analytics from big data has been used to offer faster service since massive amounts of data can be analyzed especially in the underwriting stage. Fraud detection has also been enhanced.

Retail and wholesale trade- Big data from customer loyalty data, POS, store inventory, local demographics data continues to be gathered by retail and wholesale stores [23].
In New York’s Big Show retail trade conference in 2014, companies like Microsoft, Cisco and IBM pitched the need for the retail industry to utilize big data for analytics [24].

Transportation- Some applications of big data by governments, private organizations and individuals include:

- Governments use of big data: traffic control, route planning, intelligent transport systems, congestion management (by predicting traffic conditions)
- Private sector use of big data in transport: revenue management, technological enhancements, logistics and for competitive advantage (by consolidating shipments and optimizing freight movement) [25].
- Individual use of big data includes: route planning to save on fuel and time, for travel arrangements in tourism etc.

Energy and Utilities:- Smart meter readers allow data to be collected almost every 15 minutes as opposed to once a day with the old meter readers. This granular data is being used to analyze consumption of utilities better which allows for improved customer feedback and better control of utilities use [26]. In utility companies the use of big data also allows for better asset and workforce management which is useful for recognizing errors and correcting them as soon as possible before complete failure is experienced.

Big Data now is recognized as a rich and valuable resource for wide range of organizations; it has been used in scientific researches, business, education, healthcare, transportation, political campaigns, Governments, etc [1]. The financial enterprises for example, understand that information is the foundation of financial service competition [2], and through harnessing the right information from the Big Data, they can take optimized decisions within the acceptable risk levels on the right time to get new opportunities, in order to survive and win the competition.

II. LITERATURE SURVEY

Characteristics of Big Data:

Big data can be described by the following V’s:

- Volume- It directly concerned with size of data and that size GB, PB, TB and so on.
- Variety- Data can be structured, unstructured or semi-structured. Structured- In form of tables (arranged data).Unstructured- In form of videos, images, etc. Semi-structured- Data present in tag forms like in xml, html, etc.
- Velocity- In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.
- Variability- Inconsistency of the data set can hamper processes to handle and manage it.
- Veracity- The quality of captured data can vary greatly, affecting the accurate analysis.
- Validity- refers to how accurate and correct the data is for its intended use.

Value- Alternate attributes of Big data are unimportant on the off chance that you don't get business esteem from the data. Significant esteem can be found in huge information, including understanding your clients better, focusing on them as needs be, enhancing forms, and enhancing machine or business execution.

Hadoop

The need of hadoop in industry, research and academic is quite popular due to store complex and various type of data

Modules of Hadoop

1. HDFS: Hadoop Distributed File System. Google built up its own File System, Google File System (GFS) and based on that HDFS was produced. It expresses that the documents will be softened into blocks and stored in nodes over the distributed architecture [1].

2. Mapreduce: This is a framework which helps programs to do the parallel computation on data using key value pair [1]. There are various research articles on switching from traditional storage to Hadoop, a few papers we have discussed in this paper. Jyoti Nandimath et al. focus on the distributed system through distributed architecture large number of data divided into chunks and store in the form of the block this all task perform using Hadoop[8].

Another Author deploys Hadoop on the centralized architecture, which is not suitable for Big Data processing because it results in high processing cost and low processing performance and quality. MapReduce framework was built as a parallel distributed programming model to process such large-scale datasets effectively and efficiently. This paper presents six successful Big Data software analysis solutions implemented on MapReduce framework [9].

In this paper, an author uses Hadoop for storing different types of data and set block size, memory allocation, CPU allocation, Number of MapReduce jobs, Job scheduling and JVM which should be tuned for improved Hadoop application performance [10].

Digital data continuously growing everywhere. This is the biggest challenge for Big data. To fulfill big data challenges new method and framework require. In this paper author, two case studies were used in the exploration. One is to showcase the performance comparison between Hadoop and DBMS, whereas the other is between Hadoop and a statistical analysis tool. Results clearly demonstrate that Hadoop is superior in processing a large data size. We also derive some recommendations to tune Hadoop optimally[11].

Big data analytics frameworks such as Hadoop [6] and Spark [1], [2] are most popularly used for large scale computation. Due to their popularity, researchers have given significant effort towards improving the performance of these frameworks. Popular MapReduce model can achieve better performance with allocation of more computational nodes, but this solution is not cost effective. For the most part, clients hope to have more computational power with less monetary speculation, in this way execution investigation of different Big Data stages is essential. As preparing power is constrained it is imperative to dissect enormous information
application execution, with the goal that framework head will discover reasons for execution debasement and in this way tune framework parameters to increment huge information applications execution. Likewise, discovering underlying drivers of execution debasement helps in asset arranging stages in future [1],[22].

The Hadoop is inefficient in dealing with the little records. So we will upgrade the execution of the Hadoop using distinctive strategies proposed on various investigation papers. There are various disservices in systems proposed as of not long ago. So we will use the correct framework for improving execution of the Hadoop. Today the union model and SIFM methodologies serves the better method for handling the little records in the Hadoop and the New HAR is likewise significant technique for streamlining the Hadoop [27]. The proposed Strategies are as yet not productive in enhancing the handling of Hadoop in little records. Later on the new Strategy might possibly be presented for dealing with the little records in the Hadoop [13].

A colossal measure of data being produced at a gigantic rate by numerous sources, regularly this information exists in various configurations in this manner making it very hard to process the data utilizing conventional strategies. This paper shows a study of various execution demonstrating systems of Big Data applications. One of the key ideas in execution demonstrating is finding applicable parameters which precisely speak to execution of enormous information stages [12].

III. CONCLUSION

To store unstructured data in a single traditional storage database is very difficult. In this paper, we conclude that Hadoop is a suitable framework for storing and processing not only for structured data but also for Unstructured data. We survey and find that for storing HDFS is suitable and for processing Spark and MapReduce is better. In future, we will set up a large cluster and perform an operation on EDGAR logs datasets and find out an empirical performance of Hadoop.

REFERENCES

[16] Cappella, Joseph N. "Vectors into the future of mass and interpersonal communication research: Big data, social media, and computational social science." Human communication research 43.4 (2017): 545-558.


