BIG DATA: APPLICATIONS, PROCESSING AND STORAGE MANAGEMENT

Lavi Chadha*1, Jaya*2, Dr. Suman Madan*3

*1 Research Student, Jagan Institute of management studies, Rohini, Sector-5, Delhi, India.
*2 Research Student, Jagan Institute of management studies, Rohini, Sector-5, Delhi, India.
*3 Associate Professor, Jagan Institute of management studies, Rohini, Sector-5, Delhi, India.

ABSTRACT

With the ever-increasing demand for smartphones, IoT Devices, and other mobile computing devices, a large quantity of data is generated and transferred every day, which leads to a rapid increase in data traffic. This increase in data is generally seen in major sectors in the IT field i.e., Social media traffic and e-commerce traffic which is involved in rapid transactions every second all over the world. Handling such a massive amount of data, which is generated every second is a very difficult task, which in turn makes it difficult for companies to analyze it and make meaningful predictions by extracting useful information for increasing user satisfaction and staying competitive in today's business world. Big data has been proven useful in industries like telecommunications, manufacturing among others. It is difficult to manage such a huge amount of data in various forms such as datasets, therefore we need new tools and methods that can be used for big data analytics. The Big Data, Analytics, and Decisions and framework incorporates the big data analytics tools and methods into the decision-making process. Our main aim with this paper is to present how big data affected different industries and big data analytical tools.

Keywords: Big data, big data analytics, big data applications, big data analytics tools, Big Data Storage, decision making, traditional data analytic method.

I. INTRODUCTION

In the era of modern technology, it is almost impossible to imagine a society without data Storage. Inferable from the persistent and quickly developing advancement of information assortment and capacity innovations, the extent of accessible logical information is truly expanding nearly at a dramatic rate [1]. For an instance let’s assume if our daily transactions are not recorded it will be a major setback for the companies/organizations which use this valuable data which includes details like name, address, product availability, sales details and employee recruited, etc. With the tremendous increase in storage capabilities and data collection methods, it is convenient and easier for organizations to access all the data. Every day or to be precise every second new data is generated which needs to be stored so that it can be analyzed and used to extract value. As the technology is getting better day by day it is easier and comparatively cheaper to store huge/large data[1][10]. It is mandatory for an organization to thoroughly use the available data to stay ahead or at least on the same level as their Competitors.

APPLICATIONS

According to my research and Knowledge through various platforms and all the available information regarding Big Data below I have stated my take on Big Data uses in Various Sectors and how it is used in these places [7].

1. Big Data and Internet of Things (IoT): IoT addresses one of the primary markets of big data applications. In view of the great assortment of articles, the uses of IoT are consistently developing. These days, there are different Big Data applications supporting strategic endeavours. Indeed, it is feasible to follow vehicle positions with sensors, remote connectors, and GPS. In this way, such information-driven applications empower organizations not exclusively to direct and oversee workers yet additionally to advance conveyance courses [2].

2. Importance of Big Data in the Banking Sector: Generally, banks use Analytics to examine or we can say to differentiate the business transaction from fraudulent activities. With the help of analytics, it makes it possible for the bank to define normal activity based on the customer's history and it will make them distinguish it from what it means to fraud, and these analytics systems suggest immediate actions such as stopping the unusual transactions.

3. Income-tax Department using Big Data to Inspect bank accounts after Demonetization: After demonetization, it is necessary to catch hold of the people who are keeping black money in their accounts in
huge amount, so the government and Income tax department must use some tools or technology which helps them in finding discrepancies. So, it is said that the Income-tax department used Big Data to find out such accounts that are making huge transactions without paying taxes, and we know that the Income-tax department can’t find out such accounts manually because it will take a lot of time. These Big Data mechanisms won’t be only used to gather details regarding the money being deposited by the individuals but they will also be used to collect information like paid taxes by the individual in previous years.

4. Importance of Big Data in the Healthcare Sector: Currently, the data generated in the process of medical care are rarely processed in real-time nor are they collected and used for data analysis. But if it is collected regularly will yield better results such that it could change an era of medical knowledge discovery and medical care quality improvement. Free available public health data and google maps have been used to create visual data that allows the faster identification and more accurate analysis of healthcare information that is being used in tracking the spread of chronic disease [2].

5. Importance of Big Data in Manufacturing and natural resources: Increasing demand for natural resources including oil, minerals, and so on has led to an increase in the volume, complexity, and velocity of data that is very much challenging to handle. In the natural resources industry, big data permits predictive modelling to carry decision making that has been utilized to ingest and integrate a large amount of data from the geospatial data, graphical data, text, and temporal data. Big data has been used in resolving today’s manufacturing challenges and to gain competitive advantages among the other benefits.

6. Importance of Big Data in the Transportation Sector: In recent times, a huge amount of data from location-based social networks and high-speed data from telecom companies have affected travel behaviour. But the research to understand travel behaviour has not progressed as quickly as it was supposed to have progressed[2].

7. Importance of Big Data in the Education Sector: There are various uses of big data in education like it can be used to collect information about student’s performance in various subjects and to make decisions on how to increase the performance of each student depending upon which areas they are lacking. It can also be used to check teacher’s effectiveness by collecting data about student’s performance in a particular subject that has been taught by that particular teacher, to improve the quality of education amongst students [4].

II. LITERATURE REVIEW

During late years critical endeavours have been made to create and appropriately arrange new flexible and effective data analysis tools for the processing of data contained inside scientific publications and other research. With the aim of automatic recognition and data mining scientific publications, Jung (2017) proposed a novel semantic explanation system [9]. The framework has been applied for information processing of research papers related to chemical engineering and communication technology based on relevant domain dictionaries. Accordingly, the framework was able to precisely recognize and separate significant destinations and relations inside the content, in this way making logical materials more conceivable.

Here we have listed few authors who have researched various factors of big data, big data storage & management, big data analytics, and analytical tools.

<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Title</th>
<th>Key Points</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020</td>
<td>Pavel Maksimov, Tuomas Koiranan</td>
<td>Application of novel big data processing techniques in process industries</td>
<td>Unstructured data analysis via cognitive search platform, Data acquisition and service configuration</td>
<td>Intercedes Online</td>
</tr>
<tr>
<td>Year</td>
<td>Authors</td>
<td>Title</td>
<td>Conference/Book Details</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td></td>
</tr>
</tbody>
</table>

To extract useful information from Big Data, different models, programs, and technologies have been presented, they try to give accurate and more trustworthy results for Big Data applications, but somehow, in the current environment, it's time-consuming and difficult to choose one among the various technologies. Various factors should be considered such as compatibility with technology, deployment complexity, efficiency, reliability, support, and security issues [7]. There are so many surveys on big data but most of them are focusing on algorithms and approaches used to process Big Data rather than technologies.
III. BIG DATA ANALYTICS TOOLS AND METHODS

With the advancement in technology and the increased amount of data flowing in and out of the organization, there is a need to analyze and process such a huge amount of data. Having a huge amount of data in the form of a dataset, one cannot process such a huge amount of data like petabytes of data using a traditional database management system [1]. Therefore there arises a need for new tools and methods that are required for big data analytics, as well as architecture for storing such a huge amount of data.

The Big Data, Analytics, and Decisions (B-DAD) framework consolidates the huge information examination apparatuses and techniques into the dynamic cycle [3]. The framework maps/combines the different big data storage technologies, management, and processing tools, analytics tools and methods, and visualization and evaluation tools to the different phases of the decision-making process i.e. which helps in making the decision or to generate valuable information from such a huge amount of data [5].

IV. BIG DATA STORAGE AND MANAGEMENT

One of the major tasks that an organization has to deal with is where and how this huge amount of data once is collected. The traditional method of data storage is in the form of relational databases, data marts, and data warehouses. But for big data various types of Non-Relational databases such as NoSQL, were developed for storing and managing the non-relational data[3]. NoSql Databases separates data storage and data management, such databases focus on the high performance of the data storage.

Hadoop is a framework for performing big data analytics which gives dependability, versatility, and reasonability by giving execution to the MapReduce paradigm.

Hadoop consists of two main components: HDFS for big data storage and map-reduce for big data analytics. Where the HDFS holds a very large amount of data, and the files are stored across multiple machines. These files are stored redundantly so that whenever data is lost or a file is corrupted, we can have a copy of data in some other machine and we can recover data from that particular machine or location [5].

MapReduce is a framework that is utilized for composing applications that can interact with colossal measures of information proficiently. The MapReduce algorithm contains Map and Reduce. The map takes a bunch of data and converts it into another arrangement of data, where singular components are broken into key sets. Reduce takes input from the map and combines all the data rows which are present in the map into a smaller set of rows [3].

![Figure 1. Big Data Analytics using map() & reduce function.](image)

V. BIG DATA ANALYTICS AND DECISION MAKING

As per [3][13],The main purpose of collecting and storing a huge amount of data is to process such a huge amount of data so that it can yield valuable information for the decision-makers. A large amount of data from various sources like social media etc. provides benefits to organizations. This is possible only if such a huge amount of data is analyzed properly using tools and technologies. The first phase of the decision-making process is the intelligence phase, where information that can be utilized for settling on choices is gathered from internal and external data sources[5]. The next phase in the decision-making process in the design phase,
where conceivable game plans are created and investigated utilizing different ways. Finally, the last phase in the decision-making process is the implementation phase, where the proposed solution from the previous phase is implemented. Once the data is properly collected, analysis can give very useful information which is essential for the organization to make decisions, so that they can improve the quality of working of the organization [3].

Technologies used to store and analyse Big data:
1. Apache Hadoop
2. Microsoft HDInsight
3. NoSQL
4. Hive
5. Sqoop
6. PolyBase
7. Big data in EXCEL
8. Presto

VI. BIG DATA PROCESSING

As per [3], there are four necessities for big data processing. The first requirement is quick data loading. Since the disc and network traffic meddles with the query executions during data loading, it is important to diminish the data loading time. The second requirement is fast query processing. To fulfill the prerequisites of heavy workloads and real-time requests, numerous queries are response-time critical. Consequently, the data placement structure should be fit for holding high query speeds as the measures of queries increases. Furthermore, the third requirement for big data processing is efficiently utilizing storage space. Since the user activities increase rapidly they can request scalable storage capacity and computing power, limited disk space requires that data storage be very much overseen while preparing, and issues on the best way to store the data so that space utilization is maximized. At last, the fourth requirement is the strong adaptivity to highly dynamic workload patterns. As big data sets are dissected by various applications and clients for various purposes the system ought to be profoundly versatile to startling elements in data processing, and not explicit to certain workload patterns. Map Reduce is a parallel programming model, propelled by the “Map” and “Reduce” of functional languages, which is reasonable for big data processing. It is the center of Hadoop and plays out the data processing and analytics functions. As indicated by EMC, the MapReduce paradigm depends on adding more systems or resources, instead of expanding the storage capacity of a single system; all in all, scaling out and not scaling up [3]. The crucial thought of MapReduce is separating an undertaking into stages and executing the stages in equal to reduce the time expected to finish the errand. The main period of the MapReduce work is to plan input esteems to a bunch of key/esteem sets as yield. The "Guide" work appropriately parcels enormous computational assignments into more modest undertakings and allows them to the proper key sets. In [12], the Duplicate-Divergence-Different properties enabled dragon Genetic (DDD) algorithm dependent on the k-DDD anonymization algorithm. The proposed DDDG algorithm permits the protection safeguarding in the big data by altering the MapReduce techniques with the proposed DDDG algorithm. The execution of the proposed anonymity model is analyzed with the measurements like information loss (IL) and the classification accuracy (CA). The database from the UC Irvine dataset is utilized for the simulation. The simulation results show that the proposed DDDG algorithm accomplished the least IL of 0.0191 and the highest CA with the estimation of 0.8977 than the current algorithm for k value of 2. Consequently, unstructured data, like content, can be mapped to an organized key/esteem pair, where, for instance, the key could be the word in the content and the worth is the quantity of events of the word [7]. This yield is then the contribution to the "Lessen" work. Reduce at that point plays out the assortment and blend of this yield, by joining all qualities that share something very similar key worth, to give the end-product of the computational undertaking. The MapReduce work inside Hadoop relies upon two distinct nodes: the Job Tracker and the task, running on a node. Then again, the Task Tracker node runs the positions and communicates results back to the Job Tracker. That correspondence between hubs is frequently through documents and indexes in HDFS, so between nodes correspondence is limited. The crucial thought of MapReduce is separating a task into stages and executing the Task Tracker nodes. The Job Tracker centers are the ones that are responsible for passing on the mapper and
Big data brings extraordinary freedoms for changed areas however as we know everything brings its difficulties. Big data brings big freedoms and extraordinary potential for different areas; It is known that big data examination can give immense horizons of chances in various applications and regions, similar to clinical consideration, retail, telecom, manufacturing, etc Big data is difficult to oversee. It requires a genuine limit, the board, blend, coalition, cleansing, planning, inspecting, etc. big data expands challenges because of extra volumes, velocities, and variety of data and sources which must be managed [2]. Variety means that Big Data analyses regularly unite heaps of various kinds of data from various sources, in manners not done previously. Now and again the data are organized with unmistakably characterized rules and rationale, and on different occasions, they are unstructured (like streams of comments from Twitter or Facebook). Therefore, the data are normally muddled and require a considerable measure of cleanup. Velocity means that Big Data comes in quickly and changes rapidly. I have worked with enormous data stock items that get over 25 billion lines of new data every day! This kind of data investigation requires measures that can consolidate recently created data rapidly and productively. With all of the issues that stood up to standard data with the board, big data significantly grows these difficulties as a result of additional volumes, speeds, and combinations of data and sources which should be overseen, then again, it likewise presents remarkable difficulties in tackling such huge expanding volumes of data[9]. Progressed data investigation is needed to comprehend the connections among highlights and investigate data. For example, data examination empowers an association to extricate important understanding and screen the examples that may influence emphatically or adversely the business. Other data-driven applications need likewise ongoing examination, similar to route, informal communities, Finance, biomedicine, stargazing, clever vehicle frameworks. In this manner, progressed calculations and effective techniques for data mining are expected to get precise outcomes, screen the progressions in different fields, and anticipate future perceptions. Nonetheless, big data investigation is as yet trying for some reasons: the perplexing idea of Big Data, the requirement for versatility and execution to examine such colossal heterogeneous data sets with continuous responsiveness [7]. These days, there are different scientific procedures including data mining, representation, factual examination, and AI. Numerous examinations tackle this territory by either upgrading the pre-owned procedures, proposing new ones, or testing the mix of different calculations and advancements. Hence, Big Data pushed the advancement of framework designs, equipment well as programming. Notwithstanding, we need scientific progression to confront Big Data difficulties and stream handling. One of the issues is how to ensure the practicality of reaction when the volume of data is huge? In the accompanying sub-areas, we investigate instances of the challenges experienced while applying current logical arrangements: Machine learning, profound learning, steady methodologies, granular processing.

VIII. RESULTS AND CONCLUSION

In this research, we have analyzed various topics of big data. The paper originally characterized what is implied by enormous information to unite the disparate talk on big data. We introduced different meanings of big data,
featuring the way that size is just one component of big data. The paper’s essential spotlight has been on application, analytics to acquire legitimate and important insights from big data. The goal of this paper is to describe, review, and reflect on big data. , which has gained a lot of interest because of its apparent extraordinary chances and advantages. In the data time, we are right now living in, a huge amount of data is created and being used at a very fast rate and we cannot imagine our life without it. Big data applications are everywhere for example Internet of Things (IoT), the Banking Sector, the Healthcare Sector, Manufacturing, and natural resources. We are all surrounded by it Therefore we need to manage it efficiently. The examination can be applied to use business change and improve decision making, by applying advanced analytic techniques on big data and uncovering covered-up experiences and significant information. As needed, the writing was evaluated to give an examination of the big data examination ideas which are being explored, just as their significance to Decision making. Big data was talked about, just as its qualities also, significance. A portion of the big data analytics tools and methods were inspected. Some different advanced data analytics techniques were additionally talked about. By applying such investigation to big data, significant data can be extricated and abused to upgrade decision-making. Thus, a portion of the various territories where big data investigation can support and help in decision-making were inspected. It was tracked down that big data investigation can give huge skylines of chances in different applications and territories, like medical care, retail, telecom, fabricating, and so forth Big data is hard to manage. It requires legitimate capacity, the board, mix, alliance, purging, preparing, examining, and so on With every one of the issues confronted with customary information the board, big data dramatically expands these challenges because of extra volumes, velocities, and assortments of data and sources which must be managed. In this way, the future examination can zero in on giving a guide that can envelop the recently expressed troubles.

IX. REFERENCES

https://doi.org/10.1504/IJCAT.2020.106591

https://doi.org/10.1016/j.jksuci.2017.06.001

https://doi.org/10.1007/978-3-319-08976-8_16

https://doi.org/10.1080/09620214.2019.1690546

[5] Rahul Kumar, Chawda, Ghanshyam Thakur, "Big data and advanced analytics tools", Publisher: IEEE Conference Location: Indore, India,2016. DOI: 10.1109/CDAN.2016.7570890

DOI: 10.1109/CCAA.2017.8229787

DOI:10.1007/978-3-319-44550-2_2

DOI: 10.1016/j.ijinfomgt.2014.10.007


