An Overview of Big Data Management and its challenges

1 Dr Shahnaz Fatima, 2 Prakrati Tiwari
1 Asst. Prof.-III, 2 Student MCA(2nd year)
1,2 Amity Institute Of Information Technology, Amity University Uttar Pradesh

Abstract:

Big Data is used for referring huge data sets i.e. which has high Volume, Velocity, Variety and even has a complex structure which is beyond the abilities of traditional software database tools so that it could be captured, stored, managed and analysed. Thus, a large amount of gathered data which was earlier not significantly important is now used to its fullest due to the newly available and designed tools of Big Data. Consequently, now Big Data is tremendously important to the researchers of data mining and the organizations as the results which are obtained from large volumes of data are better. Due to the arrival of the Big data tools the accuracy in the business analysis and the predictions is improving. This paper introduces the readers with the idea of Big Data and its various different sources. Certain challenges of Big Data are even conferred to. Then a brief description is given on how to manage Big data by making use of certain Big Data tools and finally the conclusion is made.

Keywords: Data, Big Data, Challenges, Big Data tools, Management.

1. Introduction:

Even though the term “Big Data” is recently in trend and is being talked about quite often however, it is an ages old terminology which is used for eventual analysis by “the act of collection and storage of large amount of information”. In the early 2000s, when Doug Laney, analyst of an industry articulated this innovative definition of big data then this concept gained its momentum which was described as the five Vs.

![](image)

**Figure 1: 5 V’s of Big data[1]**

The first one is Volume which describes about the huge amount of data. Different organizations work by collecting data and information from a diverse range of sources, which may include machine-to-machine data, social media, business transactions, or information from the sensor[2]. Storing such huge amounts of data would have been a problem in the past – but the emergence of new technologies (such as Hadoop) has eased up the burden. There is an estimation made upon the amount of data that will be created by 2021 i.e. the amount of data could be 43 Trillion Gigabytes (which is 40 Zettabytes).

Second is the Variety which states that there is no limit on the types of formats that the data comes in and ranges from numeric information in traditional or old-styled databases, structured databases to the stock ticker, audio, email, text documents, video data and the data of financial trades[3]. Every month on YouTube amount of hours that the videos are watched are 4 Billion+ and similarly the content which is shared on Facebook every month is 30 billion pieces.

Third is the Velocity which focusses on the analysis of data streaming. Data streaming is at a speed which is unprecedented in nature and thus it must be dealt with an appropriate method. In addition to data management, the companies require that the information or data should flow quickly – possibly as close to real-time[3]. The necessity to deal with such torrents of data in real time is driven by smart metering, sensors and RFID tags. Like in every trading session of the “New York Stock Exchange” 1 Terabyte of trade information is produced. In addition to this there are...
almost 100 sensors in modern cars which monitors the items such as the tire pressure and the fuel level.

Fourth is the Value, states that it’s always good to have accessibility to the big data but it’s of no use unless the extracted data can be turned into certain valuable data. So it can be clearly said that the utmost essential V of the “Big Data” is ‘Value’[4]. For the collection and leverage of the big data it’s very important that business cases should be made for any such attempt. Without any clear understanding of the costs and benefits it’s very easy to drop in the initiatives of big data which is just a buzz trap.

Fifth is the Veracity which incorporates the uncertainty linked up to the data. Confering to a study, 1 out of 3 business leaders don’t have faith on the info which they utilise for making judgements. Additionally to this the data which flows can be extremely inconsistent in nature with period leaps because of increase in velocity and the varieties of data[4]. Is something which is trending on social media? On a daily basis, event-triggered and seasonal peak of loads of data can be of a challenge to achieve which becomes extra challenging with the data that unstructured in form. The US economy is costed 1.3 Trillion US Dollars for one year just because of poor data quality.

2. **Challenges of Big Data:**

a. **Dealing with the growth of the data:**

The simple storage and the analysis of the big data information is the challenge which is most common. Most of the data does not exist in a database which means it is unstructured data. The analysis and the searching of videos, photos, documents, audio, and the other data which is unstructured can be tough[5].

b. **Integration of the disparate sources of data:**

Data integration challenges are led by the variability which is related to the big data. Email systems, applications of enterprises, documents created by the employees, social media streams, etc. are some of the varied sources of big data collection[5]. It can become extremely difficult and challenging in the integration and reconciliation of that data in order to utilise the same for report creation.

c. **Recruitment and retaining the talent of big data:**

Organisations require Big data experts so that the development, management and execution of the applications which generate the insights becomes easy. This resulted in the increased demand for the experts in big data and consequently the salaries of big data have been improved intensely[5].

d. **Data Validation:**

The idea behind integration of data is related closely to the validation of data too. Organizations get similar data pieces from a variety of systems but the data does not agree always in those dissimilar systems[5].

e. **Organisational barrier:**

Challenges of the big data do not always arise from the technological aspects but issues can even be created by people. Things are needed to be done in a different manner to subsidize on the chances that big data offers. That kind of changes can be very challenging for big organizations[5].

f. **Security of big data:**

Organisations having storage of the Big data is majorly concerned on its security aspects also. So, the APTs commonly known as the Advanced Persistent Threats or the hackers can be attracted and may target some of big data storages[5].

3. **Big data management tools:**

Lot of Big Data tools are flooded in today's market. They are basically used to ensure efficiency in the cost and manage time in the data analytical procedure. Some of them are as follows:

a. **Hadoop**: The library of Apache Hadoop software is a large data structure. Across bunches of computers it permits processing of the big data collections which are distributed in nature. This is basically designed in order to escalate into thousands of machineries from a particular server[6].

b. **HPCC**: LexisNexis Risk Solution developed a big data tool known as HPCC. For the processing of data it delivers utilising only a single programming language, on a single architecture and on a single platform.

c. **Storm**: A well-known open source and free computation system for big data is Storm which offers a processing system that is fault-tolerant and
d. real-time distributed and thus have capabilities of computation in real-time.

e. **Qubole**: An autonomous management platform for Big data is Qubole Data. The data team is easily allowed to focus on the outcomes of the business as it’s a self-optimizing as well as a self-managed tool.

f. **Flink**: A Big data tool which is stream processing as well as open-source is the Apache Flink. It is an accurate, distributed, always-available and high-performing application of data streaming.

g. **Cloudera**: A modern, easiest, fastest and highly secure platform of big data is Cloudera. This tool permits anyone, across different environments to get any sort of data within a platform which is single as well as scalable.

Some of the other tools are Openrefine, Rapidminer, DataCleaner, Kaggle, Hive etc.

4. **Management of Big data**:

Big Data is an terminology which illustrates possibly data that is inconsistent in nature and which is being generated at a fast rate and in diverse structures and extensive volumes. In accordance to the above description, certain tools should be used which can overcome the above mentioned difficulties by managing and operating on the big data in a manner which can help to monitor the processing and the analysis of data. These sort of challenges can be beaten with certain devices and strategies provided by the Big Data Management. The series of Data Science (2012) gives a broader spectrum and goes through all the possible benefits for the organisations and also for all the people to convert in to the big data resources. It has been specified that managing of Big Data should utilise appropriate approaches and tools so that it conceivable for the processing, catching and analysing data which is heterogeneous and massive.

Presently, NoSQL database is the answer to the management of Big Data. Rather than a tool or a technique, NoSQL databases are to a great extent considered as a philosophy. It depicts a procedure of methodologies to make an expert in the management of big data[7]. The way of storing data in the NoSQL databases completely varies from the Old Relational Databases as shown below:

4. **Figure 2: Relational vs Document data model[8]**

For example, there are some NoSQL databases which may use either schema-free or schema-less or even schema’s with flexible approaches. Also, new methods in dealing with storing of data are being used. For example, some systems may use storing system of key value, some may use column families or graph families type as illustrated below:

4. **Figure 3: NOSQL Database types[9]**

As discussed by “DeCandia, Hastorun, Jampani, Kakulapati, Lakshman, Pilchin, Sivasubramanian, Vosshall & Vogels(2007)” in a paper NoSQL uses a distributed structural design in which the information is contained on a few hubs in a manner which is repetitive in nature. By the addition of new servers the system could effortlessly scale out, thus undergo a failure of the server. Data management for such huge amount of data is done through this kind of
database which scales horizontally. High availability aspects of NoSQL databases are improved by the creation of copies of data which are redundant and making them strong against the partition failure. An acronym, ACID (Atomicity, Consistency, Isolation, Durability) are certain properties which can even be considered by the databases of NoSQL. NoSQL is not able to provide a complete dependability across all the distributed nodes as a result of burdening of places on databases. Various databases of NoSQL can even be preserved as the databases that are schema-free. The major advantage behind using this type of strategy is that it permits several applications in upgrading its data structure without rebuilding of the table again. A greater flexibility can be provided to store heterogeneously structured data.

5. Conclusion:

Big Data is very important economically as well as scientifically. Scientifically it is believed that utilisation of larger volumes of the data gives a more accurate result. The volume of the available data constantly keeps on increasing as in real life the creation of the data is made every single second thus, it definitely can never reduce in size. Even the prediction of the Digital Universe of IDC’s study states that digital data would grow 35ZB per year which is 44 folds between the years 2009 and 2021. Thus, important is the fact to identify that much of the exploded data consequently results from the explosion in the devices that are actually located at the network periphery, comprising of the tablet computers, smart phones as well as the sensors that are embedded. New opportunities are created by this data for the analysis of data in many a fields such as healthcare, human genomics, finance, search, Oil and gas, surveillance and even many other fields. There are several challenges faced by Big data and thus to optimally utilize this contemporary detection, the handlers must have an awareness towards these challenges so that measured solution or modification to them can be provided as quickly as possible. Thus, several Big data tools are even discussed in this paper.

References:


