3rd National Conference on Green Technology and Science for Sustainable Development
© 2020 IJSRST | Volume 5 | Issue 6 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
International Journal of Scientific Research in Science and Technology

A Review on Big Data

A. A. Dande*, Shivaji. S. Kathale, Rahul. R. Raimule

Computer Science and Engineering, Anuradha Engineering College, Sant Gadgebaba Amravti, Chikhali, India

ABSTRACT

This research paper is based on Big data overview is pool of large and complex data sets so it becomes difficult to process data using database management tools. With the fast evolution of data, data storage and networking collection capacity, Big Data are rapidly growing in all science and engineering domains. The analysis of big data can be difficult as it often involves collection of mixed data based on different patterns or rules. The challenges include in big data are capture, storage, search, sharing, analysis, and visualization. The trend to larger data sets is because of the extra information drawn from analysis of a single large set of related data, compared to separate smaller sets with the same total amount of data. Big Data mining is the ability of extracting useful information from huge streams of data or datasets, that because of its velocity, variability and volume. Big Data processing model and also Big Data Mining.

Keywords : Big Data, Hadoop, Architecture

I. INTRODUCTION

The Big Data, as the name suggests, is something related to data, where big implies large or huge. To put simply, Big Data refers to large amounts of data (in terms of volume) that cannot be digested (processed) with traditional data processing applications in an effective way. As the data gets bigger, it also becomes more complex, and it requires more advanced and robust mathematical and statistical techniques to get what we want from data.Here, lets try to understand the Introduction of Big Data with an example, Rewind back to 1940s, no computers, no cell phones, no internet, no digital life, so no data, right? Well, there was data, but it was not digital. There was no internet banking that time but there were banks, and banks had customers, and customer made transactions that were recorded, not digitally but on papers.

Fast forward to 1990s, technology kicks in, computers and cell phones came into the market, income statements and balance sheets that were done on papers and stored in registers which had data of roughly 500 customers were now being done on excel and saved in drives that can store more than thousands of customers data. Here in the introduction to big data, we are going to learn that as data increased exponentially, organizations equipped themselves with more firepower to handle data more effectively. Now, on one single day, 2.5 quintillion bytes (2,500,000 Terabytes) of data is generated.

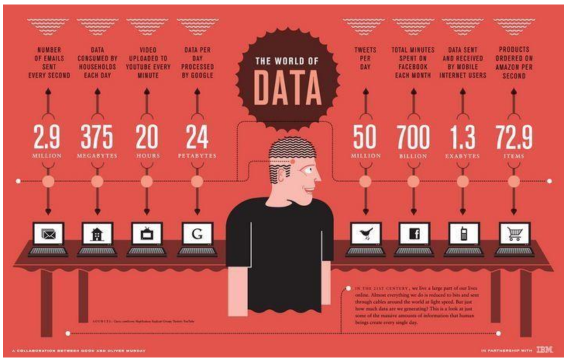
THE WORLD OF DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS
DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES
TWEETS PER DAY
50 MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS
G
IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.
A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY
IN PARTNERSHIP WITH IBM

Fig 1 : The world of big data

II. DEFINATION OF BIG DATA

Big Data is data that is too large, complex and dynamic for any conventional data tools to capture, store, manage and analyze. . Traditional tools were designed with a scale in mind. For example, when an Organization would want to invest in a Business Intelligence solution, the implementation partner would come in, study the business requirements and then would create a solution to cater to these requirement Big Data could be of three types:

Types of Big data :

- Structured
- Semi-Structured
- Unstructured

Structured Semi-Structured Unstructured

Fig 2 : Types of big data

Structured *:* The data that can be stored and processed in a defined format is called as Structured Data. Data stored in a relational database management system (RDBMS) is one example.

It is easy to process structured data as it has a fixed schematic SQLis often used to manage such kind of Data.

Semi-Structured : Semi-Structured Data is a type of data which does not have a formal structure of a data model, i.e. a table definition in a relational DBMS, but nevertheless it has some organizational properties like tags and other markers to separate semantic elements that makes it easier to analyze. XML files is examples of semi-structured data.

Unstructured : The data which have unknown form and cannot be stored in RDBMS and cannot be analyzed unless it is transformed into a structured format is called as unstructured data. Text Files and multimedia contents like images, audios, videos are example of unstructured data.. The unstructured data is growing quicker than others, experts say that 80 percent of the data in an organization are unstructured.

III. ARCHITECTURE OF BIGDATA

This architecture is designed in such a way that it handles the ingestion process, processing of data and analysis of the data is done which is way too large or complex to handle the traditional database management systems.

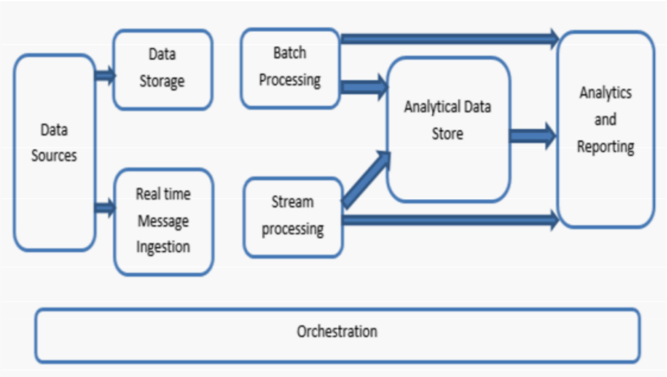

Fig.4 : Architecture of big data

Big Data systems involve more than one workload types and they are broadly classified as follows:

1. Where the big data-based sources are at rest batch processing is involved.
2. Big data processing in motion for real-time processing.
3. Exploration of interactive big data tools and technologies.
4. Machine learning and predictive analysis.

1. Data Sources :

The data sources involve all those golden sources from where the data extraction pipeline is built and therefore this can be said to be the starting point of the big data pipeline. The example are

(i) Datastores of applications such as the ones like relational databases
(ii) The files which are produced by a number of applications and are majorly a part of static file systems such as web-based server files generating logs.

2 Data Storage :

This includes the data which is managed for the batch built operations and is stored in the file stores which are distributed in nature and are also capable of holding large volumes of different format backed big files. It is called the data lake. This generally forms the part where our Hadoop storage such as HDFS, Microsoft Azure, AWS, GCP storages are provided along with blob containers.

3 Batch Processing :

All the data is segregated into different categories or chunks which makes use of long-running jobs used to filter and aggregate and also prepare data o processed state for analysis. These jobs usually make use of sources, process them and provide the output of the processed files to the new files.

4 Real Time-Based Message Ingestion :

This includes, in contrast with the batch processing, all those real-time streaming systems which cater to the data being generated sequentially and in a fixed pattern. This is often a simple data mart or store responsible for all the incoming messages which are dropped inside the folder necessarily used for data processing.

5 Stream Processing :

There is a slight difference between the real-time message ingestion and stream processing. The former takes into consideration the ingested data which is collected at first and then is used as a publish subscribe kind of a tool. Stream processing, on the other hand, is used to handle all that streaming data which is occurring in windows or streams and then writes the data to the output sink. This includes Apache Spark, Apache Flink, Storm, etc.

6 Analytics-Based Datastore :

This is the data store that is used for analytical purposes and therefore the already processed data is then queried and analyzed by using analytics tools that can correspond to the BI solutions. The data can also be presented with the help of a NoSQL data warehouse technology like HBase or any interactive use of hive database which can provide the metadata abstraction in the data store. Tools include Hive, Spark SQL, Hbase, etc.

7 Reporting and Analysis :

The insights have to be generated on the processed data and that is effectively done by the reporting and analysis tools which makes use of their embedded technology and solution to generate useful graphs, analysis, and insights helpful to the businesses. Tools include Cognos, Hyperion, etc.

8 Orchestration :

Big data-based solutions consist of data related operations that are repetitive in nature and are also encapsulated in the workflows which can transform the source data and also move data across sources as well as sinks and load in stores and push into analytical units. Examples include Sqoop, oozie, data factory, etc

IV. WORKING OF BIG DATA

The main idea behind Big Data is that the more you know about anything, the more you can gain insights and make a decision or find a solution. In most cases this process is completely automated – we have such advanced tools that run millions of simulations to give us the best possible outcome. But to achieve that with the help of analytics tools, machine learning or even artificial intelligence, you need to know how Big Data works and set up everything correctly.

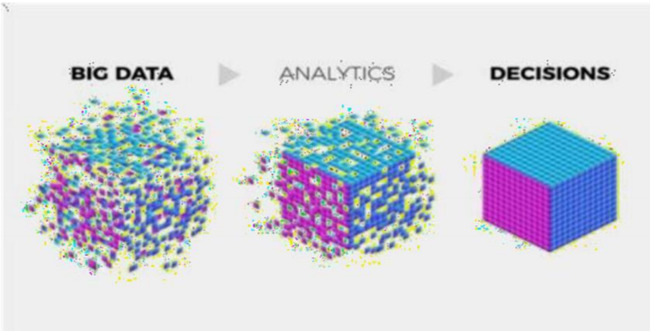

Fig.5 : Working of big data

All the processes should be considered according to the capacity of the system. And this can potentially demand hundreds or thousands of servers for larger companies..

1 Integration

Big Data is always collected from many sources and as we are speaking for enormous loads of information, new strategies and technologies to handle it need to be discovered. In some cases, we are talking for petabytes of information flowing into your system, so it will be a challenge to integrate such volume of information in your system. You will have to receive the data, process it and format it in the right form that your business needs and that your customers can understand.

2 Management

What else might you need for such a large volume of information? You will need a place to store it. Your storage solution can be in the cloud, on-premises, or both. You can also choose in what form your data will be stored, so you can have it available in realtime ondemand. This is why more and more people are choosing a cloud solution for storage because it supports your current compute requirements.

3 Analysis :

Okay, you have the data received and stored, but you need to analyze it so you can use it. Explore your data and use it to make any important decisions such as knowing what features are mostly researched from your customers or use it to share research. Do whatever you want and need with it – put it to work, because you did big investments to have this infrastructure set up, so you need to use it.As we mentioned when we are talking for Big Data we are always talking about the big Vs behind it. When Big Data appeared there were only 3Vs, but now there are more. And there are always adding more and more depending on what you need the Big Data for. We are going to mention some of the Vs in the next part of the article.

V. TOOLS OF BIGDATA

A big data analytics process is not a single activity that encompasses a huge volume of data. Instead it advanced analytics that can be applied to large data, but in reality, several types of different technologies work together to achieve the most value from information. Below are the biggest and important technologies which involve in big data analytics process:

- Data management
- Data mining
- Hadoop in data Predictive analytics.
- Text mining

There 'N' number of Big Data Analytics tools, below is the list of some of the top tools used to store and analyze Big Data. These Big Data Analytics tools can be further be classified into two Storage and Querying/Analysis.

A. Apache Hadoop:

Apache Hadoop, a big data analytics tool which is a java based free software framework. It helps in effective storage of huge amount of data in a storage place known as a cluster. The special feature of this framework is it runs in parallel on a cluster and also has an ability to process huge data across all nodes in it. There is a storage system in Hadoop popularly known as the Hadoop.Distribute File System (HDFS), which helps to splits the large volume of data and distribute across many nodes present in a cluster. It also performs the replication process of data in a cluster hence providing high availability and recovery from the failure – which increases the fault tolerance.

B. OpenRefine:

OpenRefine is introduced as Google Refine. This tool is one of the efficient tools to work on the messy and large volume of data, that all include: cleansing data, transforming that data from one format another, and also to perform extending it with web services and external data. The open refine tool helps explore large data sets easily.

C. Orange:

Orange is famous open-source data visualization and helps in data analysis for beginner and as well to the expert. This tool provides interactive workflows with a large toolbox option to create the same which helps in analysis and visualizing of data. An orange tool has many and different visualizations, that include bar charts, trees, scatter plots, to dendrograms, networks and heat maps.

VI. APPLICATIONS BIG DATA

The Aim of Big Data applications is to help companies make more informative business decisions by analyzing large volumes of data. Big data contains web server logs, Internet click stream data, social media content and activity reports, text from customer emails, mobile phone call details and machine data captured by multiple sensors.

All Organisations from different fields are investing in Big Data applications, for examining large data sets to uncover all hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. In this blog we will we be covering:

- Big Data Applications in Healthcare
- Big Data Applications in Manufacturing
- Big Data Applications in Media & Entertainment
- Big Data Applications in IoT
- Big Data Applications in Government
- challenges-of-big-data-analytics
- Uncertainty of Data Management

Landscape Because big data is continuously incresing, there are new companies and technologies that are being developed every days . A big challenge for companies is to find out which technology works bests for them without the introduction of new risks and problems.

The Big Data Talent Gap: While Big Data is a growing field, there are very few experts available in this field. This is because the Big data is a very complex field and people do not understand the complexity and intricate nature of this field are far few and between. Another major challenge in the field is the talent gap that exists in the industry

Getting data into the big data platform: Data is increasing every single minutes . This means that companies have to take a limitless amount of data on a regular basis. The scale and variety of data that is available today can overwhelm any data practitioner and that is why it is important to make data accessibility simple and convenient for brand managers and owners.

Need for synchronization across data sources: As data sets become more diverse, there is a need to incorporate them into an analytical platform. If this is ignored, it can create gaps and lead to wrong insights and messages.

Getting important insights through the use of Big data analytics: big data is important to companies gain proper insights from big data analytics and it is important that the correct department has access to this information. A major challenge in big data analytics is bridging this gap in an effective fashion

VII. CONCLUSION

We are conclude that A data is expanding every mintues. The expanding data with growing demand and competition it is essential for a professional to remain updated. By efficiently using both the individual and the Organisation can gain in several ways. The analysts get a better understanding of the industry, conveying the same to the workers. A decision can be made based on research paper . The availability of Big Data, low-cost commodity hardware, and new information management and analytic software have produced a unique moment in the history of data analysis and big data is very important for a development of IT industry in future.

VIII. REFERENCES

[1]. Russom, “ Big Data Analytics”, TDWI Research, 2011.

[2]. Sawant, Application Nitin, Architecture." and Himanshu Big Shah. Data "Big Application Data Architecture Q & A. Apress, 2013. 9-28.

[3]. BIG DATA: A Review Munesh Kataria1, Ms. Pooja Mittal2

[4]. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. In: • McKinsey Global Institute Reports, pp. 1–156 (2011)

Author Profile

Mr.A.A.Dande did his M.E. in 2013. He is working as a Asst. Professor in CSE department from last 11 years. His areas of interest are Object Oriented Languages, AI , Machine Learning and Network security. He has published many papers in various National/ International journals and conferences

Mr Shivaji S. Kathale pursuing Bachelor of Engineering in Computer Science & Engineering Department from Anuradha Engineering College of SGBAU Amravati University Maharashtra