BIG DATA: FRAMEWORKS AND CHALLENGES

Shikha Soni*, Jimmy Singla, Mir Mohammad Yousuf
School of Computer Science and Engg.,
Lovely Professional University Phagwara, India.

Abstract—In today’s world, data is produced and consumed at a fast pace. This has given birth to a domain to collect, store, process and maintain this huge influx of data called big data. Running any form of analysis on such amounts of data is challenging at best and so new and creative techniques were needed to be created. Techniques had to be developed to handle different types of processing like batch processing and stream processing. Each type has its own individual method of function by which it carries out its objectives. In addition to this domain, big data is usually collaborated with other emerging domains like information security, intelligent systems and data science.

Keywords—Big data; Big data analytics; Apache Spark; Apache Hadoop; Cloud Computing; IoT.

INTRODUCTION

Up until 2007, relational databases and data warehouses were used to store the data by an organization. But eventually after the surge of social media, it became harder to store the data in a traditional manner which gave rise to a new concept named “Big Data”. Big Data was introduced for storing large amount of data in an efficient manner. The data being produced due to the rise of technologies can be structured or unstructured depending on the source. It is difficult to process such data as it contains billions of records. The need comes specifically from big companies such as Google, Netflix, Facebook etc. since a huge amount of data gets generated by them on a daily basis. Big Data Analytics is used to process these massive and complex datasets.

HISTORY

Keeping tracks of data dates back to 7000 years ago when the Mesopotamians introduced accounting for recording crop growth. The earliest implementation of data dates back to 1887 when computing machines used for punching holes in paper cards were used to maintain census data by Herman Hollerith. Data projects and processing machines were used in World War II for deciphering codes, therefore helping in reducing the time taken to decipher the codes. The first data center was implemented by US government for handling tax returns and introduced the concept of electronic storage of data.

In 1989, World Wide Web was introduced by British scientist Tim Berners-Lee. The 90s were the years when more and more devices started connecting to the internet. The first supercomputer was introduced in 1995 which could do large number of calculations in a second. The term Big Data was coined for the first time in 2005 by O’Reily media, only after a year after the term Web 2.0 was created.

Big Data refers to a large amount of data which is impossible to manage with traditional data tools. In 2009, the government of India decided to keep a record of photograph, iris scan and fingerprints of all 1.25 billion citizens. This data is stored in one of the biggest biometric databases in the world.

PARAMETERS

The data is characterized by 3 V’s in big data. The 3 V’s are: Volume, Variety and velocity.
A. Volume

It is the amount of data which is being generated from various sources. The sources can be social media, online applications, websites etc. YouTube has 1 billion users, Facebook has 2 billion users, Instagram and Twitter have 700 million and 350 million users respectively. These users help in contributing billions of photos, tweets, posts etc. The large amount of data that is being generated every second and every hour is imaginable.

B. Velocity

In big data, Velocity is the rate at which new data gets generated per second. According to a research, 3.5 billion search results are generated in Google search and over 900 million images are uploaded to Facebook every day. The velocity is very high and big data helps the companies or organizations to hold the explosion of data.

C. Variety

Variety is the type of data that is being generated. There are usually two types of data- structured and unstructured. Examples of structured data include messages, photos, videos etc. Examples of unstructured data include electronic mails, recordings etc. Variety helps in classifying the data into different categories.

FRAMEWORKS

There are different frameworks required to process different kinds of data that is being generated. Data is basically processed in two categories- batch processing and stream processing. The frameworks are used to process the dataset according to the condition that needs to be fulfilled.
A. Hadoop

It is one of many classic and highly developed frameworks in use today. This framework is almost synonymous with big data. Hadoop is used to do batch processing i.e. the data is put in the form of batches for further processing. It has MapReduce and Hadoop Distributed File System (HDFS) as its primary component in the ecosystem. Hadoop is an important framework in data as it has HDFS and MapReduce which are used for another big data storage for processing.

B. Apache Spark

Spark is another major and frequently used big data framework. Unlike Hadoop, Spark works on stream processing i.e. data is processed as soon as it is entered in the system. It is faster and more flexible than Hadoop. Spark uses HDFS as it doesn’t contain a storage layer of its own. Spark consists of MLlib, SparkSQL, GraphX and other components. Spark is a real time framework.

C. Apache Flink

Flink is a streaming dataflow engine which facilitates distributed computation. Flink is a combination of both batch and real time processing framework. Flink provides number of application interfaces such as Streaming API, Static data API and SQL-like query Application Programming Interface for Java, Scala and Python respectively. Flink contains an in-house ML and libraries for graph processing. Flink’s features include high performance, low latency, stateful computations, continuous streaming and fault tolerance among many others.

D. Apache Storm

Storm is a real time, distributed computation system. It is used for processing streaming data and has benchmarked at processing one million tuples per second. It is highly scalable and uses Lisp-like programming languages. Storm is applicable for real time analysis and machine learning, and other cases where higher data velocity is present. Storm can integrate with Hadoop ecosystems and work on YARN, providing a solution for real-time stream processing. Storm is fast, scalable, reliable, fault tolerant and easy to operate once deployed.

E. Apache Samza

Samza is a real time distributed stream processing framework. It is built on Apache Kafka for messaging and YARN for resource management. Samza has simple API, is fault tolerant, durable, scalable and pluggable. Samza provides processor isolation as it works with YARN and resource isolation with Linux CGroups. Samza also provides restoration and snapshotting of a processor state as it is built to handle large amount of space per partition.
### Table 1. Types of Processing for various frameworks

<table>
<thead>
<tr>
<th>SNO</th>
<th>FRAMEWORK</th>
<th>TYPE OF PROCESSING</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Apache Hadoop</td>
<td>Batch Processing</td>
</tr>
<tr>
<td>2</td>
<td>Apache Spark</td>
<td>Real Time Processing</td>
</tr>
<tr>
<td>3</td>
<td>Apache Flink</td>
<td>Stream Processing</td>
</tr>
<tr>
<td>4</td>
<td>Apache Storm</td>
<td>Real Time Processing</td>
</tr>
<tr>
<td>5</td>
<td>Apache Samza</td>
<td>Real Time Distributed Stream Processing</td>
</tr>
</tbody>
</table>

### Table 2. Components in various frameworks

<table>
<thead>
<tr>
<th>SNO</th>
<th>FRAMEWORK</th>
<th>COMPONENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Apache Hadoop</td>
<td>YARN, HDFS, MapReduce, Resource Manager</td>
</tr>
<tr>
<td>2</td>
<td>Apache Spark</td>
<td>GraphX, MLlib, SparkSQL, Spark Core, Spark Streaming</td>
</tr>
<tr>
<td>3</td>
<td>Apache Flink</td>
<td>HDFS, Local-FS, Flink Kernel, DatasetAPI</td>
</tr>
<tr>
<td>4</td>
<td>Apache Storm</td>
<td>Map RFS, Message Queue</td>
</tr>
<tr>
<td>5</td>
<td>Apache Samza</td>
<td>Kafka, Samza Task, Samza Task Runner, YARN Client</td>
</tr>
</tbody>
</table>

### CHALLENGES IN BIG DATA

Web applications face big data frequently due to texts and documents, social computing and search indexing. Big data has been involved in domains like healthcare, medical, scientific researches, biochemistry etc. It provides new opportunities along with some challenges.

#### A. Data Analysis and Storage

One of the first challenges is storage medium in speed. Data accessibility should be the highest priority for representation. The data should be easily accessible for analysis. To overcome the disadvantage of slow input-output performance by hard-disk, Phase Change memory and Solid State Drive (SSD) were introduced. However, the techniques discussed above do not have the capability to perform these big data processes. Data diversity is another challenge. While handling big datasets, data selection and data reduction is a mandatory, required task. Ensuring consistency is a major issue. Hadoop and MapReduce help in acquiring huge amount of data in an
unstructured format in significantly short time span. A framework was discussed by Das and Kumar [1] to analyze the data.

B. Computational Complexities

Representation and knowledge discovery are important issue which requires sub fields like presentation, retrieval, authentication and representation to name a few. A lot of hybrid techniques are used to handle issues and/or queries, but every technique depends on the type of issue and/or queries.

As the size keeps increasing in big data, these techniques are not efficient to obtain any meaningful information. Large dataset can be managed by data warehouse and data marts. More computation complexities are required for analyzing large datasets. A mathematical framework is harder to establish when it comes to the domain of big data. Although, to understand complexities, specific data applicable to a particular domain can be applied. One of the basic objectives of research is to minimize complexities and cost processing [2], [3], [4].

C. Scalability and Visualization of Data

There is a natural shift in processing technology which requires number of cores [5]. This development in processor has led to development in parallel computing. Applications like social networks require parallel computing.

The aim during data visualization is to increase adequacy while using techniques in graph theory. Online markets like Amazon, Snapdeal have billions of products to sell to millions of users per month, resulting in large amount of generation of data. A lot of companies use Tableau for visualization of big data. Tableau can transform complex data into pictures. This helps the organization to monitor and visualize the customer feedback, sentiment and search relevance.

It can be observed clearly that the domain of big data has led to many challenges which in turn leads to the development of fields like parallel and cloud computing, scalability and its visualization and distributed processing. A lot of new mathematical models are in demand to tackle and overcome this issue.

D. Information Security

Information Security, by its very definition, means protecting data and that is true in case of big data analysis as well. As with every domain, the security risks in big data is a major concern. A lot of techniques are required to be applied in order to protect data. Techniques like user authentication, data confidentiality, integrity of data and controlled access authorization and finally secure and robust encryption of data. Keeping in mind the above requirements, a standard global framework or policy is required for the protection of data and prevention of information loss.

OPEN RESEARCH ISSUES

Data Science and Big Data Analytics are becoming an important focal point among industries. Data science focuses on extracting from data. Applications of big data include data analysis, statistical learning, signal processing, uncertainty model etc. The open research issues are classified in three categories- Cloud Computing, Quantum Computing and Internet of Things (IoT). More issues related to healthcare can be found in [9].

A. Internet Of Things (IOT)

IoT, Internet of Things is nascent field which is growing rapidly and creating data at an equally rapid pace. By its very definition, IoT is a connection of all the appliances and devices with the internet. This always-connected mode produces a lot of data which needs to be properly stored and worked upon. This is where big data comes in. the integration between big data and IoT is ever present, but analyzing the information at the rate at which it is coming in is a monumental task. For this end, there needs to be a fully functional, developed framework which not
only collects and stores the data but makes it easier to work upon it quickly. A lot of research in this course has been presented by Lin and Chang [10]. Key technologies which have been associated with IoT are in [11].

**B. Cloud Computing**

Thanks to the development in the field of virtualization technologies, supercomputing is now more affordable and easier to access than ever. The usage of virtual system is referred to as cloud computing- one of the most important techniques in big data. Cloud computing helps in harmonizing huge amount of data with the help of accessible computing resources on-demand basis by virtualization technology. One of the benefits of cloud computing technology is paying only for resources which are needed for developing the product.

Cloud computing should support development and data analytics in big data application. The environment shows allowed tools that provide business analysts or data scientists to explore data for processing and extraction of results.

Big data forms frameworks of discussing options in cloud computing. Depending on the need, user can buy services from service providers like Amazon, IBM, SaaS from companies such as Jobscience etc. The major issue is privacy concern related to hosting on public services and storage.

**FUTURE WORK SUGGESTIONS**

Manufacturing and developing powerful systems will provide a significant advantage for the deployment of big data techniques. Getting knowledge through data transformation is not an easy task by any means. Many models like fully set, soft set, neural networks have been fruitful in the representation of data.

Moreover, big data is reduced for including only the main characteristics necessary depending upon the area of application. Therefore, reduction techniques have been introduced. Another issue faced is getting high performance, throughput and being able to store it in the most efficient way possible for use in future. Creating programs or applications to analyze the big streams of data is another challenging task. Also, machine learning tools and techniques are getting popular among the researchers to facilitate results. Researchers in area of big data and machine learning have produced optimization and algorithm implementation. We argue that as each tool has its own advantage and limitation, other efficient tools can be invented for handling the problems inherently in big data.

**CONCLUSION**

In recent times, information and/or data is produced and consumed at a blistering pace. This has given birth to a domain to collect, store, process and maintain this huge influx of data called big data. Running any form of analysis on such amounts of data is challenging at best and so new and creative techniques were needed to be
created. Techniques had to be developed to handle different types of processing like batch processing and stream processing. Each type has its own individual method of function by which it carries out its objectives. In addition to this domain, big data is usually collaborated with other emerging domains like information security, intelligent systems and data science.

REFERENCES


