Big Data

Shweta Kumari Rajput, Dinesh Nanda, Ankush Preet, Akshil Patel

MAD 309, Cégep de la Gaspésie et des Îles

***Abstract*-**In today's era, we have enormous amount of data available on hand to decision makers. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Big data basically refers to these datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. This paper aims to analyze some of the different analytics methods and tools which can be applied to big data so that we can store that information for future research work or for any startups and much more.

Index Terms- Bigdata, Datamining, Analytics, Decision making.

I. Introduction

Big data basically refers to huge volume of data that can not be stored & process using the traditional computing system within the given time frame. In this modern era, huge volume of data gets accumulated within a blink of an eye either through social networking sites or airline communication system or online delivery services or mobile data or government online forms and many more. We don't need huge computers to deal with such data, people work with cloud & endless network of servers with a very powerful algorithms. In this way, we can analyze over a million pieces of data in minutes.

For example, in future, we can use big data of traffic to develop a car that can be driven completely accident free all by itself or we can use big data of DNA to determine the perfect treatment. In this way, curing genetic diseases like cancer would become much easier..

II. BIG DATA ANALYTICS

Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in single data set. Consequently, some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they don't know before.

Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data samples reveals and leverages business change. However, larger the sets of data, the more difficult it becomes to manage. Naturally, business benefit can commonly be derived from analyzing larger and more complex data sets that require real time capabilities, however, this leads to a need for new data architectures, analytics methods and tools.

III. Characteristics

There are four main features of Big data: **Volume**, **Velocity**, **Variety** & **Veracity**. The **volume** of data is its size and how big it is. **Velocity** refers to the rate at which data is changing or how often it is created. **Variety** includes the type and nature of the data and also different kinds of ways of analyzing the data. Data volume is the primary attribute of big data. One of the things that can make big data really big is that it is coming from a variety of sources than even before, including social media.

Using these sources for analytics means that common structured data is now joined by unstructured data, such as text and human language, and semi-structured data, such as extensible markup language (XML). There is also data, which is hard to categorize since it comes from audio, video and other devices. Furthermore, multi-dimensional data can be drawn from a data warehouse to add historic context to big data. Thus, with big data, variety is just as big as volume.

Moreover, big data can be described by its velocity or speed. This is basically the frequency of data generation or the frequency of data delivery. The leading edge of big data is streaming data, which is collected in real-time from the websites.

IV. HOW BIG DATA WORKS?

The sources for big data generally fall into one of three categories:

Streaming data

This category includes data that reaches your IT systems from a web of connected devices, often part of the IoT. You can analyze this data as it arrives and make decisions on what data to keep, what not to keep and what requires further analysis.

Social media data

The data on social interactions is an increasingly attractive set of information, particularly for marketing, sales and support functions. It's often in unstructured or semistructured forms, so it poses a unique challenge when it comes to consumption and analysis.

Publicly available sources

Massive amounts of data are available through open data sources like the US government’s data.gov, the CIA World Factbook or the European Union Open Data Portal.

After identifying all the potential sources for data, consider the decisions you'll need to make once you begin harnessing information. These include:

How to store and manage it

Whereas storage would have been a problem several years ago, there are now low-cost options for storing data if that's the best strategy for your business.

How much of it to analyze

Some organizations don't exclude any data from their analyses, which is possible with today's high-performance technologies such as grid computing. Another approach is to determine upfront which data is relevant before analyzing it.

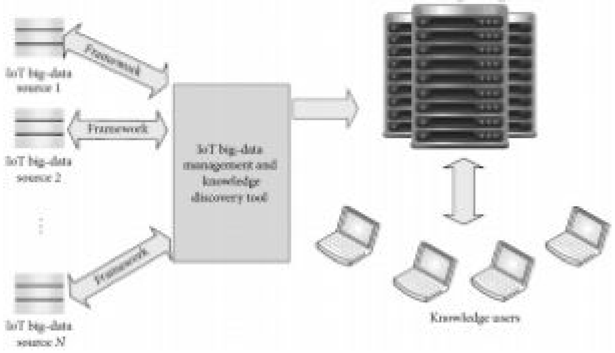
How to use any insights you uncover

The more knowledge you have, the more confident you'll be in making business decisions. It's smart to have a strategy in place once you have an abundance of information at hand.

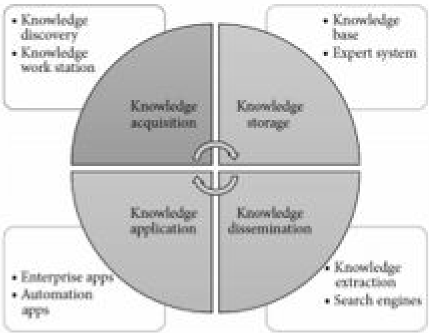
V. IoT FOR BIG DATA ANALYTICS

Internet has restructured global interrelations, the art of businesses, cultural revolutions and an unbelievable number of personal characteristics. Currently , machines are getting in on the act to control innumerable autonomous gadgets via internet and create Internet of Things. Thus appliances are becoming the user of the internet , just like humans with the web browsers . Internet of things is acting the attention of recent researcher for its most promising opportunities and challenges. It has an imperative economic and societal impact for the future construction of information, network and communication technology. The new regulation of future will be eventually, everything will be connected and intelligently controlled. The concept of IoT is becoming more pertinent to the realistic world due to the development of mobile devices ,embedded and ubiquitous communication technologies, cloud computing and data analytics. Moreover, IoT presents challenges in combinations of volume, velocity and variety. In a broader sense, just like the internet , Internet of Things enables the devices to exist in a myriad of places and facilitates applications ranging from trivial to the crucial.

Knowledge acquisition from IoT data is the biggest challenge that big data professional are facing. Therefore, it is essential to develop infrastructure to analyze the IoT data. An IoT device generates continuous streams of data and the researchers can develop tools to extract meaningful information from these data using machine learning techniques. Understanding these streams of data generated from IoT devices and analysing them to get meaningful information is a challenging issue and it leads to big data analytics. Machine learning algorithms and computational intelligence techniques is the only solution to handle big data from IoT prospective.



Knowledge exploration system have originated from theories of human information processing such as frames, rules, tagging, and semantic networks. In general, it consists of four segments such as knowledge acquisition, knowledge base, knowledge dissemination, and knowledge application. In knowledge acquisition phase, knowledge is discovered by using various traditional and computational intelligence techniques. The discovered knowledge is stored in knowledge bases and expert systems are generally designed based on the discovered knowledge. Knowledge dissemination is important for obtaining meaningful information from the knowledge base. Knowledge extraction is a process that searches documents, knowledge within documents as well as knowledge bases. The final phase is to apply discovered knowledge in various applications. It is the ultimate goal of knowledge discovery. The knowledge exploration system is necessarily iterative with the judgement of knowledge application. There are many issues, discussions, and researches in this area of knowledge exploration.



Bio-inspired Computing for Big Data Analytics

Bio-inspired computing is a technique inspired by nature to address complex real world problems. Biological systems are self organized without a central control. A bio-inspired cost minimization mechanism search and find the optimal data service solution on considering cost of data management and service maintenance. These techniques are developed by biological molecules such as DNA and proteins to conduct computational calculations involving storing, retrieving, and processing of data. A significant feature of such computing is that it integrates biologically derived materials to perform computational functions and receive intelligent performance. These systems are more suitable for big data applications.

Huge amount of data are generated from variety of resources across the web since the digitization. Analyzing these data and categorizing into text, image and video etc will require lot of intelligent analytics from data scientists and big data professionals. Proliferations of technologies are emerging like big data, IoT, cloud computing, bio inspired computing etc whereas equilibrium of data can be done only by selecting right platform to analyze large and furnish cost effective results.

V. TOOLS USED IN BIG DATA

1. Apache Hadoop

Apache Hadoop is a software framework that can store big amount of data in a cluster. This framework runs in parallel on a cluster and it allow us to process data among all nodes. Hadoop

Distributed File System (HDFS) is the storage system of Hadoop which splits big data and distribute to nodes in a cluster. As HDFS is using the cluster it provides high availability and greater efficiency across plenty of servers which are inexpensive.

2. Microsoft HDInsight

It is a Big Data solution from Microsoft powered by Apache Hadoop which is available as a service in the cloud. HDInsight uses Windows Azure Blob storage as the default file system. This also provides high availability with low cost.

3. NoSQL

Data could be in structured and unstructured format; to handle the large amount of structured data the traditional SQL can be used and for the unstructured data we need NoSQL. NoSQL databases store unstructured data with no particular schema. Each row can have its own set of column values. NoSQL gives better performance in storing massive amount of data. There are many open-source NoSQL DBs available to analyse big Data.

4. Hive

This is a distributed data management for Hadoop. To access big data this supports HiveSQL. This runs on top of Hadoop and used for Data mining purpose.

5. Presto

Facebook has developed and recently open-sourced its Query engine (SQL-on-Hadoop) named Presto which is built to handle petabytes of data. Unlike Hive, Presto does not depend on MapReduce technique and can quickly retrieve data.

VI. TECHNIQUES USED IN BIG DATA

- **a/b testing**

  Imagine that you are a CEO of Amazon, and trying to work out whether rearranging your website into a new format affects conversion rate (i.e. the proportion of visitors to Amazon who become customers):

In this case we would conclude that Layout B is superior to Layout A.However, such a simple approach suffers from two possible errors that statistics students will be very familiar with:

- **Type I error** — or falsely concluding that your intervention was successful (which here might be falsely concluding that layout B is better than Layout A). Also known as a false positive result.
- **Type II error** — falsely concluding that your intervention was not successful. Also known as a false negative result.

The hypothesis

A hypothesis is a formal statement describing the relationship you want to test. A hypothesis must be a simple, clear and testable statement (more on test-ability below) that contrasts a control sample (e.g. Layout A) with a treatment sample (e.g. Layout B).

To form a hypothesis, we re-phrase "does an SMS system improve repayment" into two statements, a null hypothesis and an alternative hypothesis:

Null hypothesis (H0) : The null hypothesis usually states that there is **no difference** between treatment and control groups. (To put this another way, we're saying our treatment outcome will be statistically similar to our control outcome)

Alternative hypothesis (H1): The alternative hypothesis states that **there is a difference** between treatment and control groups. **b)Machine learning**

Machine learning is based on algorithms that can learn from data without depending on rules-based programming. Big data,is the type of data that may be supplied into the analytical system so that a ML model could learn.

A quick example: preventive machinery maintenance. We use big data from sensors (temperature, humidity, pressure and vibration readings for each machinery part that come every second) to train, test and retrain a ML model. The role of the model is to identify hidden patterns that lead to machinery failure and check newly incoming data against the identified patterns. As a final step – the analytical system may trigger alerts to the maintenance team if the model identifies a match with a pre failure condition pattern.

c)regression analysis

Regression analysis is used to estimate the strength and direction of the relationship between variables that are *linearly* related to each other. Two variables *X* and *Y* are said to be *linearly* related if the relationship between them can be written in the form:

Y= mX+b   where

where

m is the *slope,* or the change in *Y* due to a given change in *X*

b is the *intercept,* or the value of *Y* when *X* = 0

As an example of regression analysis, suppose a corporation wants to determine whether its advertising expenditures are actually increasing profits, and if so, by how much. The corporation gathers data on advertising and profits for the past 20 years and uses this data to estimate the following equation:

$$Y = 50 + 0.25X$$

VII. BENEFITS OF BIG DATA

- It can provide ideas from huge amounts of data from multiple sources that include those that come from external third-party sources, the internet, social networks, those already stored in company databases etc.
- Real-time forecasting and monitoring of occasions that may affect the performance or operations of the businesses.
- Capability to locate, get, extract, change, analyze, and blend data with different tools.
- Identification of important information that can improve the quality of decision making.
- Ability to mitigate risks by optimizing complex decisions about unplanned events more quickly.

- Identification of the causes of failures and problems in real time.
- Full understanding of the potential of data-driven marketing.
- Generation of offers to customers based on their purchasing habits.
- Improvement of the commitment of the client and increase of his loyalty.
- Revaluation of the risk portfolio quickly.
- Customization of the customer experience.
- Adding value to the interactions with online and offline customers.

VIII. REFLECTIONS

Dinesh- Through my research, i found out, some people have misconceptions about big data,like, the size of data does not needs to be in GB or TB or PB to be called as big data. For example, id we attach 25 MB file to gmail, the file will not be attached because gmail does not allow files larger than 10 MB to get attached, so in that case, that 25 MB file can be called big data.

Akshil -Big data refers to data sets that are too large or complex for traditional data-processing application software to adequately deal with. Data with many cases offer greater statistical power, while data with higher complexity may lead to a higher false discovery rate. Big data challenges include capturing data, data storage, data analysis,search, sharing, transfer, visualization, querying, updating, information privacy and data source.

Ankush - The big data problems mean that data is growing at a much faster rate than average speeds. And it is the result of the fact that storage cost is getting cheaper day by day, so people as well as almost all business or scientific organizations are storing more and more data. Social activities, scientific experiments, biological explorations along with the sensor devices are great big data contributors. Big data is beneficial to the society and business but at the same time, it brings challenges to the scientific communities.

Shweta- After reviewing enormous research paper, i came to know that big data is very trending in every field like health, education,insurance,media, manufacturing, international development and many more. Even every countries' government are using the big data analytics. Hence, we can say that big data is very useful in our real life to analyze,observe and fetch important data.

IX. CONCLUSION

High volume of big daily are being produced daily,and furthermore those data's are having details which should be extracted and utilized. Hence, advanced big data analytics techniques can be applied to enhance businesses and decision making. In this research paper, those techniques have been discussed. We believe that big data is playing a significant role in this era of data overflow and can provide unforeseen insights and benefits to decision makers in various areas.

References

[1] Nada Elgendy and Ahmed Elragal,Big Data Analytics: A Literature Review Paper,German University in Cairo (GUC), Cairo, Egypt

[2] https://www.sas.com/en_ca/insights/big-data/what-is-big-data.html#dmusers

[3] https://www.oracle.com/big-data/guide/what-is-big-data.html

[4] https://en.m.wikipedia.org/wiki/Big_data

[5] :http://thesai.org/Downloads/Volume7No2/Paper_67-A_Survey_on_Big_Data_Analytics_Challenges.pdf

[6] https://www.youtube.com/watch?v=obC_Ot02bGI&t=323s

[7] Top big data tools used to store and analyse data-https://bigdata-madesimple.com/top-big-data-tools-used-to-store-and-analyse-data/

[8] Many Benefits of Big Data Analytics for Your Company-https://www.dataversity.net/many-benefits-big-data-company

[9] 7 Limitations Of Big Data In Marketing Analytics-https://marketingland.com/7-limitations-big-data-marketing-analytics-117998