The Name of the Title is MASSIVE BIG DATA PROCESSING

ISTIAQUE AHMED, 17-34334-1, Sec D , CSE, AIUB

HOSSAIN,MOHAMMAD SAZZAD, 17-33750-1 , Sec D, CSE, AIUB

Big data is one of the most essential things in modern technology and a modern life. Big data is consists of Five "V"- Volume, Variety, Velocity, Variability and Value. From this term, it can be said that, How important the Big data is?Anyway, the amount of data is growing very rapidly and requires special ways of analyzing it to get information from huge data amounts of data. There are lots of new methods that have been invented and ways of applying them large scale .Using the data proficiently and intelligently can make a huge different in technology. Also there are many important tools are used in big data. The main focus is how efficiently store the important data from the large and huge amount of data.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: data sets, neural networks, gaze detection, text tagging

ACM Reference Format:

ISTIAQUE AHMED and HOSSAIN,MOHAMMAD SAZZAD. 2018. The Name of the Title is MASSIVE BIG DATA PROCESSING. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The main object of this paper was to know, how people are able to handle massive data processing. We can get a simple idea about "Big Data" by hearing these two word, "BIG" and" DATA". Simply Big Data means, a huge data which is stored high Volume, high Velocity, high Verity, high Veracity and high Value. Industry, Educational institute, Hospital, Research Center, Bank etc. Across the world generate massive volume of data. In hospital or clinic ,lots of data are collected annually in the form of patient record. All these data is generated in a very high speed, which attributes to the velocity of a big data. Big data is classified into three main type. i)Structured Data. ii)Unstructured Data. iii)Semi Structured Data. These all data are stored in a database system. Now, we need to know how to store these big amount of data efficiency and it is the purpose of the Big Data Processing. Anyway, Big data popular use cases are related information security and data where has big data optimizations big data tools are being used to remove some of the junk and unnecessary data and store only the best outcome. The main provocation of big data is accumulating and procedure at a specified time. There are lot of big data tools which help the users in saving time, money and business purpose. There are many big data popular tools like Hadoop, Apache Spark, Apache Storm, MongoDB, are Programming Tool and so many lots of tools are popular for data storage and management tools. Anyway, Big data tools are offered lots of advantages . They provide the analytics algorithm and

Manuscript submitted to ACM

models. They help the user to run on big platforms such as Hadoop . Big data helps to integrate with other technologies very easily. Huge amount of data are continuously stored in offline and online, we have to handle these data in a efficient way.

2 RESEARCH METHODOLOGY

People typically tend to think of research methods in terms of qualitative research and quantitative research . These two distinct disciplines or two distinct type of research methods and the research is the domain on the social scientists and anthropologists and the quantitative.Research is the domain of the epidemiologists and economist and what we like to talk. The main challenge of big data is storing and processing the data at specific time . The traditional approach is not efficient in doing that .Hadoop technology and various big data tools emerged to solve the challenge faced in the big data environment.when we write data into what called the Hadoop distributed .We just bring it without any specific rules. So there is a lots of big data tools and all of them help user in some or another way in saving time,money and uncovering business.

2.1 Research Objective

The objective of the research is the answer of some specific term applying procedure of the research.The main aim of research is to find out the truth which is hidden and has not discovered.

1.Examine the specific factor of BIG DATA. 2.Identify what features contribute to improve BIG DATA technology. 3.Identify the level of performance. 4.Identify and explore the underlying factor.

2.2 Research Questions

As we look at the history of data science one of the things that we want to point out at this really early stages we are trying to get our arms around the topic and understand what it is that this is discipline is trying to solve to realize that data science is actually a combination of a things now we probably saw that on the WIKIPEDIA article but we wanted to point it out here. Anyway first of all We need to know what type of data we stored any analysis them.Then we need to analysis how a big data analysis helpful in research. Explain the steps to be followed to deploy a big data solution. Heating up: Self-service BI,Heating up: Mobile dashboards,Cooling down: Hadoop,Heating up: R language,Heating up: Deep neural networks, Cooling down: IoT,Heating up: TensorFlow, Cooling down: Batch analysis etc are most most data analysis trends. We will talk about some question which is now hot topic in this era. i.DATA STREAMING ii.Predictive Analytics iii. In-Memory Databases iv.Big Data Security Solutions v.Edge Computing vi.Prescriptive Analytics

1.WHAT are the main component of big data that determine the data in efficient way?

2.IS It possible to set development path in term of big data?

3.How can improve the development? How to describe the method?

4.what is the Security of big data database systems?

5.what is the Impact of using Big Data tools?

6.How can we Clustering data in a proper way?

7.How can we Using algorithm?

8.What is Velocity in big data?

9. What is Variety in Big Data?

10.Where to maintain Volume in Big Data?

11.What is Big Data analyses?

2.3 Article Selection

A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench,An Optimised Method to Allocate and Re-Allocate Aggregated Data in a Data Storage

2.3.1 Keywords and Search String.

Data aggregation INFORMATION STORAGE SYSTEM Discerning Patterns in Complexity

2.3.2 keyword search and Manual Selection.

Hadoop and Spark , Experimental setup, Cluster architecture,Speedup

2.3.3 Final set of Articles.

Title , ABSTRACT , Introduction, Conclusion Results and discussion

3 DISCUSSION

3.1 Massive Data

we are just at the outset of the of the big data era in the future all businesses are going to try understand how they can be become a platform for the collection and the analysis of data and we will be optimizing everything we do in by taking more information and learning from it. So there is a lots of technologies in thick data. is data is not only about the volume of data but it poses other challenges as well like velocity and variety .As a volume 40 zetta bytes of data will be created by 2020 this huge volume of data is either human generated like from social media ,or candy machine generated like though sensors and personal health trackers and can also be generated with organization like credit card details commercial transactions and medical record Processing data in traditional approach usually the data that is being generated out of the organization the finencial instrution such as BANK or stock market and the hospital is given the input to the ETL SYSTEM and then an ETL system extract the data and transform this data that would be convert this data into a proper format and finally load this data on DATABASE. Another challenge is velocity.The speed at which data is coming into system the data needs to be processed with faster speed then there is variety data.There are lots of multinational companies, any other growing company or any sectors have to collect huge amount of data for different purposes. So, they have manage these data very carefully because of some issues. We know that big sectors has big data problem they have to solve it in a user friendly way. Focusing on this there are some vital questions out there. We will talk about some question which is now hot topic in this era.

3.2 Future tech

Big data promises to be transformed ,the effective use of this data cannot only deliver substantial yip and bottom line profits but also improve the performance of existing function and also create opportunities for growth and expansion yet very few organizations are able to reap these benefits the main reason being the lack of the right infrastructure to handle big data leading to poor data management and consequent loss of business revenue of the challenges that businesses face while delivering big data capabilities is the strain that it puts on their existing IT infrastructures due to the huge data influx thereby showing the

systems. Internet of things,these are numerous ways in high analytic can be applied to internet of thing. For example,sensor are used to collect data that can be analyzed to achieve actionable insights tracking customer or product movement etc. Analyzes it and presents it to customer service.So that allows them to gather the rich insights about businesses .Big data popular use cases are related information security and data where has optimisations.Big data tools are being used to remove some of the burdens from the data warehouses.Even the health care industry is looking for pattern and treatment that lead to the best outcomes for patient

3.3 Procedures Of Data

Big data has become an integral part of our lives and offer many benefits and advantages but big data is getting bigger and that should come with bigger responsibility . Data has been an essential part of human evolution for thousand of year. The main challenge of big data is storing and processing the data at specific time.So Hadoop technology and various big data tools emerged to solve the challenge faced in the big data environment. So there is a lots of big data tools and all of them help user in some or another way in saving time, money and uncovering business .Big data is the term use to define large amount of data that can be processed to reveal patterns.Therefore we can term this huge volume of data as big data. Anyway the big data management process is like once the data is hadoop and how it analysis for database. The focus is on what do industries do with the data how can they explore it and that's term called big data because the amount of data is growing very firstly and requires special way of analyzing to get information from huge amount of data. To maintain huge amount of data we need proper functional way to find out the specific data. specific data means, when a person try to find out a data he can find easily. Suppose a banker wants to check customer details, so he can search specific query or using shortest path to find that data.

4 FUTURE RESEARCH DIRECTIONS

Massive data processing is lengthy for its own complex functionality. Every sector has to know future probability to continue this process for next generation . Actually volume, variety, velocity and veracity is the main component for big data. In future we have to focus on some research topic which is more important for next directions.

4.1 Volume Challenges

Hospital, Industries, multinational companies and many other sector deals with data in their professional life. There are lots of data is stored in our daily life. In this era, people are too much dependent on internet. We are dealing with massive data in online, also offline data's are continuously store in a proper way. If anyone want to see some specific data, he or she can find out that query easily by using some method or algorithms. Some algorithms, which are mostly use on massive data processing • Linear Regression • Logistic Regression • Classification and Regression Tress • k-Nearest Neighbors • K-Means Clustering.

4.2 Technology

We need high power computer technology for massive data processing. High power computing is famous for data science, engineering and business sector. In high power computer synthesizing computer power. HPC machines normally use in large volume of data but small companies try to avoid HPC because of

higher cost, they use hadoop instead of this. Handoop are not too much expensive and easy to run where HPC is difficult for user. Hadoops is more user friendly then HPC.

4.3 Analytics

Dividing data in a equivalent way is more important than others. Analysis the steps of data and handling in a system without getting error is the main motive of data processing. Critical challenges regarding big data analysis is should solve addressed and suggest emerging paths.Analysis some algorithm for big data. (1)K-means clustering algorithm: This is simple and popular unsupervised machine learning algorithm. Used with Big Data sets, pre-clustering or classifying into large categories that other algorithms can refine. (2)Linear Regression Algorithm: This a machine learning algorithm based on supervised learning.

(3)Apriori Algorithm: This is a matching algorithm which is commonly used for transnational matching with a large number of transactions.

(4)Em(Expectation-Maximization)Algorithm:Knowledge discovery cluster algorithm

(5)Association Rule Mining Algorithm: This is referred to as "Market Basket Analysis", that was the original application of this algorithm. The association rule algorithm is a learning algorithm that is for associations that co-occur with a high degree of frequency.

5 VALIDITY THREAD

There are Two types of validity exists • External validity- Truth in real life • Internal validity- Truth in the study Every research has its own functionality. A researcher always try to find out some way which will be blessings for our next generation. Our main focus on this research paper is, how to deal with huge amount of data in a easiest way also using shortest path. These days, Massive Data is rapidly developed into a hot topic around the world. Heavy factories, educational institute, small or large farms and other business sectors has to collect their data. Maintaining data in a proper way is more important than collecting huge data. Construct validity is significantly more challenging for massive data processing. As we can say that, there are number of different measure is mainly called validity thread. Big data problem issue is commonly a big topic for its validity thread.Data scientist are trying to reduce these types of problem for skip those unnecessary thread. Focusing on validity thread is important for massive data processing in our daily life.

6 CONCLUSION

Today Massive data processing is a recent hot topic on data science sector. Data engineers are always trying to search some way which is maintain some specific pattern to search data. Big companies always need to store their data in a secure way. If data loose or hampered by some malware attacks it can be great loss for them. Every person has their personal confidential data, if privacy looses for some reasons it will be big loss. Today, So many high quality software is used by millions of people to go through huge amounts of data. We process the massive data either complicated or simplify our lives. Our research concern is, how to deal with massive data processing and reveal some patterns which is best for finding sequence base data.Data is growing too fast, in this current technology driven decade. Massive amount of data is quite impossible to store in a traditional way. So, lots of Big Data tools and software are using in the data science world gradually. These tools works for time efficiency, analysis tasks and also effectiveness in business. We added some vital questions which is very complicated to solve this Big data problem from root. Some question

needs proper resources for answering them properly. Make a good data visualizations we should go for sequence first.

REFERENCES

(1) https://www.aiche.org/resources/publications/cep/2016/march/ big-data-challenges-and-future-research-directions

(2) https://www.linkeit.com/blog/maximising-high-performance- computing-for-big-data-processing

(3) https://www.researchgate.net/publication/291017284

(4) https://www.researchgate.net/publication/264555968

(5) https://bigdata.ieee.org/publications

(6) www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878

(7) https://www.hbs.edu/faculty/PublicationHow-Big-Data-Is-Different

(8) https://www.nature.com/articles/498255a

(9) https://journalofbigdata.springeropen.com/articles/10.1186/ s40537-020-00388-5

A CONTRIBUTION RECORD

Detail each group member contribution according to the following tables.

A.1 Paper Assessment

Populate the following table with the required information.

Student id& name,	Paper No frm Ref	Paper Title
17-33750-1 HOSSAIN,MOHAMMAD SAZZAD	1,2,3,4,5	Big data processing, Future Direction,High Performance Computing,Algorithms in Big Data, path
17-34334-1 ISTIAQUE AHMED	6,7,8,9	How Big Data is Different, Big Data In Daily Life, Procedural data, Questions for Massive Data

Table 1. Paper collected and read by the group member

A.2 Paper writing contribution

Populate the following table with the required information.

Student id & name	Section No	Section Title
17-33750-1 HOSSAIN,MOHAMMAD SAZZAD	**1,4,5,6**	**Introduction,Future Research Question,Validity Thread, Conclusion**
17-34334-1 ISTIAQUE AHMED	**2,3**	**Research Methodology,Discussion**

Table 2. Section(s) Written in the paper by the group member