DATA MINING: YESTERDAY TO TOMORROW

Ashish Pant[1], Smrati[2]

[1,2]*Computer Science and Information Technology/MJPPRU, (India)*

ABSTRACT

We live in the age where large amount of data is being collected every day. Analyzing and processing this data is the biggest need. This data is very huge and cannot be processed easily so it is also called as big-data. Big-data is fancy term to define this huge amount of data, which cannot be manage by current data mining techniques. Terabytes and petabytes of data is being collected every day in the Internet, and this is the biggest issue of having that big data. For example the President of U.S.A. Mr. Obama came to India then it that this information must be published in lots of websites, social networking sites and at-least there are 25-30 videos uploaded for this . We also have social networking sites like Facebook and Twitter which collect lots of data every day. Earlier we were using relational database but our data has been changed from structured to unstructured and semi-structured so this is not a useful idea. Processing and maintaining this huge amount of data is the biggest issue. Computer scientist considered this data as the kitchen sink as if some useful thing is wasting in sink is hard to find. Lots of techniques have been emerged out for processing this huge amount of data. Big-data mining is the capability of extracting useful information from the huge data sets which due to its size and variability could not be manage before.

Keywords: Big-Data, Hadoop, Parallel Processing, Architecture, Deep-Learning.

I. INTRODUCTION

Recent year have witnessed the massive increase in collecting the data from various devices and formats, from independent or current technology. This data flood have out-paced to analyze, understand and analyze these datasets. Consider the webpages were billion in 2008 and and now it is up to 5-10 times more than that. This dramatic increase is due to the increase in the social networking, Moore's Law[19] and Metcalfe's law. The massive increase in the technology is also a reason of increasing such huge amount of data. The WAP have increase the capability to use internet from the mobile phone and this have increased the huge amount of data .It is being seen that Internet have massively increase this huge data-sets. So every time when the data is increasing in this huge amount then these challenges is being facing 1) Algorithm used to process this data 2) System capabilities and 3) Business models.

Many conferences have been occurring for discussing this data sets. Here in my paper issues have been discussed and challenges faced during the processing the huge data. Big companies use large severs to store these data and have the super computers for processing the data, so the data can be easily managed there, but small firm have lots of problem in managing this data.

II. DATA WE ARE INTERESTED

So we might say that we are interested in the processing and analyzing data from database systems, data warehouse data and transactional data .It can also be applied to other form of data (e.g. data streams .ordered/sequence data graph and network data, spatial, textual and multimedia data) and also from WWW.So database systems is a software to store and retrieve that data. An example a relational database is the collection of tables and SQL to retrieve that data. Now second the data warehouse is repository of information collected from various sources of unified schema and unified in a single site.

Transactional database consists of transaction such as customer purchase, a flight booking .It includes the UID and list of items making up the transaction. Also we are interested in analyzing and processing the data from various other sources.

III. TECHNOLOGY USED

So the question comes which techniques may be used to solve this retrieval problem from database systems or data servers. As we know the first approach to solve any problem is statistics. So statistical approaches is also applied to solve it. It also takes approaches from various other domains as machine learning pattern recognition, database and data warehouse system and information retrieval ,visualization ,algorithm ,high performance computing and also various other domains. Hadoop is a java based programing platform that support the processing of large data sets in distributed computing environment.

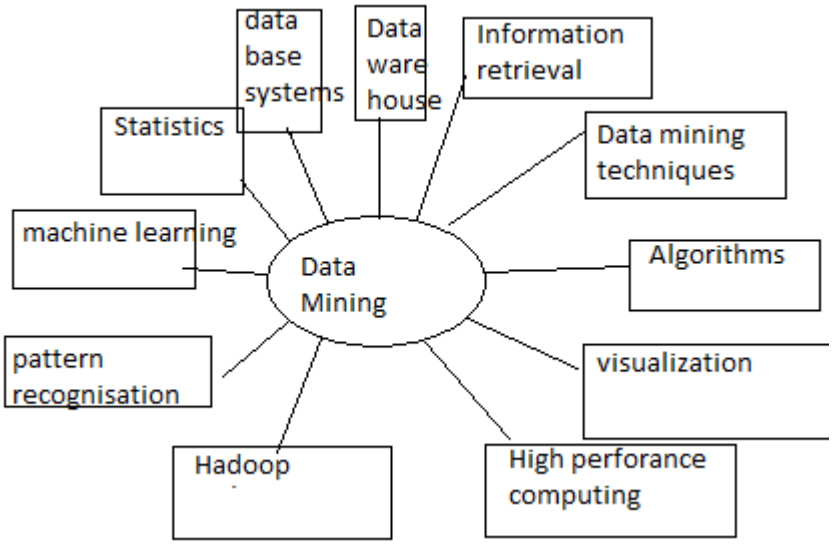

Fig. 1 Mining Data From various fields

IV. VARIOUS APPROACHES OF DATA MINING

The main approaches to solve the data mining problem are as below

1) **Statistical:** It studies the collection analysis, interpretation and representation of data. Data mining have the inherent connection with it. A statistical model is the set of mathematical function that describe the behavior of objects in target class in term of random variable used to model data and data classes. As an

example in data mining task like data characterization and classification statistical model of target class are build. In other words such statistical models can be the outcome of data mining tasks.

2) **Machine learning:** It tells how computer learn based on data typical are of interest is to automatic learn to recognize complex programs and make intelligent decisions. For e.g. A typical machine learning problem is to program the computer so it can recognize hand pattern code on mail after learning from set of examples and output the result according to it. These are of 4 types and are below.

- **Supervised learning**: The supervision is done on the basis of labelled data. For e.g. Postal code recognition problem a set of handwritten postal code and machine readable transactions are used to supervise the classification model.
- **Unsupervised learning:** In this data is not labelled. We used clustering to discover classes within data. eg. This algorithm take a set of images with the handwritten digits. Suppose it find 10 cluster of data.10 cluster may correspond to 10 digits from 0-9.As not labelled so it won't tell the semantic meaning of cluster formed.
- **Semi supervised:** This use the labelled as well as unlabeled data to learn a model.
- **Active learning:** This technique takes user for the active learner. An active learning ask the user to label the example which is from the set of unlabeled examples or synthesized by learning program.

3) **Pattern recognition:** Pattern recognition is the study of methods and algorithms for putting data objects into categories. While classical pattern recognition techniques are rooted in statistics and decision theory, the machine learning paradigm is commonly used to design practical systems.

Along with this there exist other techniques to retrieve the data from these techniques Hadoop is the emerging field, which is a java based programming framework that support processing of large data sets in distributing computing environment. As we already know that it is a framework which works on java. It act as the extensible for recovering the failures of data storage and processing in distributed systems. Apache Hadoop is the open source software framework [13] for processing and analyzing a large data sets on clusters of hardware. The main components of Hadoop is distributed file systems (HDFS) which is useful for large files and map reduce which is the heart of Hadoop. HDFS is the high bandwidth cluster storage. Map-reduce performs two different tasks (1)collection of data where it is transfer to another set of data (2)After transformation it is broken in (key, value) pair. In the future we used to see that data mining is being using the Deep learning algorithm for processing that data.

V. HADOOP WITH OTHER TECHNOLOGIES

5.1 Hadoop with Parallel Database [20]

In earlier days of Hadoop there are various problems. But today there are various data management techniques to reduce the performance gap. In this sense Hadoop is studied in the sense of similarities and difference with parallel database .The parallel database techniques like job optimization, data layout and index are focus in this discussion. The researcher found that it is useful to use parallel database in combination with Hadoop map reduce. Hadoop and map-reduce affected from row oriented layout. So other data layout techniques are propose for Hadoop map-reduce .A good source of indexing has propose for map reduce.

5.2 Hadoop and Data Warehouse [6]

As the advertisement of Hadoop is unrestrained, the practitioners are easily effected by diversity of opinion like Hadoop is becoming a new data-warehouse .But it not really what it seems. There are lots of difference between Hadoop and data warehouse. This context explore when to use data warehouse and when to use Hadoop. Let us consider a firm uses Hadoop to preprocess raw click generated by customer using their website. The data warehouse set the customer preference with marketing campaign and recommendation engine to offer investment suggestion and analysis of customers. So data-warehouse is used as source in complex Hadoop jobs. This bring the advantage of these systems in parallel. Choosing data-warehouse and Hadoop depends on the requirement of organization.

VI. OPEN-SOURCE TOOLS FOR MINING THE DATA

1)To solve the problem of mining the big data the companies are using a lot of open-tools which help companies to manage their data .Some of these open source project are:

a) **Apache Hadoop [3]:** Apache Hadoop is a set of algorithms (an open source software framework for distributed storage and distributed processing of very large data sets (Big Data) on computer cluster build from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are commonplace and thus should be automatically handled in software by the framework.

b) Apache Hadoop related projects [18]: Apache-Mahoot, Apache-Hive, Apache-, Avro, Apache-HBase, Apache- Zookeeper , Cascading, scribe and many others.

c) Apache s4: Abstract—S4 is a general-purpose, distributed, scalable, partially fault-tolerant, pluggable platform that allows programmers to easily develop applications for processing continuous unbounded streams of data.

d) Storm: We traced the problem to a sloppy implementation detail of Hadoop. It turns out that Hadoop sometimes shells out to do various things with the local file system. When you shell out in Java, the process gets forked. Forking a process causes the child process to reserve the same amount of memory for itself as the parent process is using (to fully understand what's happening, you need to learn about memory overcommit and the copy-on-write semantics of forking in Linux). This means that the Hadoop process which was using 1GB will temporarily "use" 2GB when it shells out. So this led to the development of storm. It was developed by Nathan Marz at twitter.

In data mining there are many other open source tools than Hadoop. Some of them are Apache Mahout [4], Moa[5],R[17], Vowpal Wabbit. There are many graph mining tools that are being developed by various companies today .Some of them are Pegasus, Graph Lab. Graph lab is high level graph parallel system build without using map-reduce. Graph lab compute the dependent records which are stored as graph in the large distributed data graph.

VII. APPLICATION AREAS OF THE MINING

- **usiness intelligence:** It is typical for a business to acquire the better understanding of the commercial context of their organization, such as their customers, the market, supply and resources and competitors. BI technology provide historical, current and predictive views of business management .E.g. include reporting

online analytical processing, business performances management, competitive intelligent, benchmarking and productive analysis.

- **Web Search Engines:** A web search engine result in in form of list. The list may consist of web pages images and files. Search engine differ in web directories are maintain by human editors whereas search engine work on the algorithms and human input Various data mining techniques used in all aspects of search engine range from crawling ,indexing and searching.
- **Geographical analysis: This** category of data includes location and enhance with geographical information in a structured form which is nothing but spatial data. This data requires an understanding of geometry and operations that can be performed on it. The GIS tools are used for spatial framework of Hadoop. It allows user to analyze the spatial data. Some geographical network problem like Distributed Denial of service (DDOS) which are common in security hacking can be analyzed by Hadoop and map-reduce.

S no.	Application areas
1)	Financial data Analysis
2)	Retail Industry
3)	Telecommunication Industry
4)	Biological Data Analysis
5)	Scientific Application
6)	Bio informatics
7)	Software Engineering.

Fig 2.Application area of data Mining

VIII. FUTURE OF DATA MINING

In the future we used to see that lots of algorithm is being developed for data mining in clouds, Hadoop parallel processing. Also we used to see that deep learning algorithm is being used for processing big-data. Also the issue that computer scientist have to face is the hidden big-data which is wasted and cannot be find to be useful. This is sure that the future of data mining is deep learning and lots of work is being done for solving the problem. We can use analytics architecture for mining the data as the deep learning is hard costly to implement, but AI will be definitely the best approach for solving the problem of mining.

1) Analytics Architecture. It is not clear yet how an optimal architecture of an analytics systems should be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz .The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in real time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general, and extensible, allows ad hoc queries, minimal maintenance, and debug gable.

2) **Hidden Big Data**. Large quantities of useful data are getting lost since new data is largely untagged and unstructured data. The 2012 IDC study on Big Data [14] explains that in 2012, 23% (643 Exabyte) of the digital universe would be useful for Big Data if tagged and analyzed. However, currently only 3% of the potentially useful data is tagged, and even less is analyzed.

3) Deep Learning: Implementing deep learning algorithm in mining this data is also the key issue In future we used to see that data mining is using deep learning algorithm. Deep learning is part of a broader family of machine learning methods based on learning representations of data.

Some of the other challenges faced in mining are mining data in cloud computing, Hadoop parallel processing architecture and also due to processing this use kitchen sink we need the processor which are even faster than before, and it will increase the use of Moore's law.

IX. CONCLUSIONS

Big Data is going to continue growing during the next years, and each data scientist will have to manage much more amount of data every year. This data is going to be more diverse and faster. We discuss some insight about this topic and what we consider are the main concerns and the challenges of the future. Big data is becoming the new Final Frontiers for scientific research and for business application .We are at the beginning of a new era where Big data mining will help us to discover knowledge that no-one has discover before. Everybody is invited in this journey.

REFERENCES

[1] SAMOA, http://samoa-project.net, 2013.

[2] C. C. Aggarwal, editor. Managing and Mining Sensor Data. Advances in Database Systems. Springer, 2013

[3] Apache Hadoop, http://hadoop.apache.org.

[4] Apache Mahout, http://mahout.apache.org.

[5] A. Bifet, G. Holmes, R. Kirby, and B. P Fahringer MOA: Massive Online Analysis http://moa.cms.waikato.ac.nz/. Journal of Machine Learning Research (JMLR), 2010.

[6] C. Beckerman and H. Bloom the streams Framework. Technical Report 5, TU Dortmund University, 12 2012.

[7] D. Boyd and K. Crawford. Critical Questions for Big Data. Information, Communication and Society, 15(5):662{679, 2012.

[8] F. Diebold. "Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting. Discussion Read to the Eighth World Congress of the Econometric Society, 2000.

[9] F. Diebold. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012.

[10] U. Fayyad. Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling. http://big-data-mining.org/keynotes/#fayyad, 2012.

[11] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size core sets for k-means, PCA and projective clustering. In SODA, 2013.

[12] J. Gama. Knowledge Discovery from Data Streams. Chapman & Hall Data Mining and Knowledge discovery. Taylor & Francis Group, 2010.

[13] J. Gantz and D. Reinsel. IDC: The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. December 2012.

[14] Gartner, http://www.gartner.com/it-glossary/bigdata.

[15] V. Gopalkrishna , D. Steier, H. Lewis, and J. Guszcza. Big data, big business: bridging the gap. In Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, Big-Mine '12, pages 7{11, New York, NY, USA, 2012. ACM.

[16] Mining Big-Data Current status, and forecast to the future.

[17] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[18] P. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch,and G. Lapis. IBM Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Tata.McGraw-Hill Companies, Incorporated, 2011.

[19] Cramming of transistor inside a chip, by Gordon Moore.

[20] ttps://infosys.unisaarland.de/publications/BigDataTutorial.pdf.