PRESENT AND FUTURE ACCESS METHODOLOGIES OF BIG DATA

Praveen Kumar[1], Payal[2], Bhawna Dhruv[3],
Seema Rawat[4] , Dr. Vijay S. Rathore[5]

[1]Research Scholar , NIMS University Jaipur (India)
[2,3]M.Tech Student, Department of CSE , Amity University Noida (India)
[4]Assistantp professor Department of CSE , Amity University Noida (India)
[5]Professor & Director, Shri Karni College Jaipur (India)

ABSTRACT

Big Data is a whole new concept in information technology driven computing world that manages large and complex datasets through certain data mining processing tools and functional models. As we all know the process of data mining is knowledge driven discovery process through which we can extract relevant information about a particular matter of subject. Big size datasets have numerous velocity and volume so it demands research techniques to extract useful information. As the complexity of data increases, there is more challenging work for us to implement big data methodologies. This research paper highlights the concept of Big Data used in datasets, its present scenario place in industry, hidden truths behind development of big data and what will be the future of big data in coming years.

Keywords: Big Data; Data Mining; Transformation; Analysis.

I. INTRODUCTION

Mobile Gateways have now become the gateway to achieve real time data from different areas. The huge data that a mobile career processes so as to improve our CDR i.e. call data record for billing. In today's world people as well as machines, small or large are equally connected to each other. Such vast connected components generate huge data and information that requires to be extracted so as to improve quality of life. For example- When we get ready and move to office everybody, for calculating the total time & completing the optimization, the system requires to process any new details like weather, traffic and other schedules. We need optimization under tight time constraints also[3,5].

By considering the growth of Big Data in information technology compute system, there are certain conferences and workshops which are held to deliver the mischievous opportunities of Big Data as Knowledge Data Driven Conference on "Big Data Mining" and some workshops on Heterogeneous Streams through Big Data Mining and analysis models and algorithm efficiency of big data[21]. All these events contribute towards growth of data knowledge in industry. Now the big concern is how to implement the system models of big data but it is definitely true that the big data is going to be the most innovative and trendsetter technology in future. There exist certain Big Data challenges like efficiency of algorithmic design and capturing capabilities of system models.

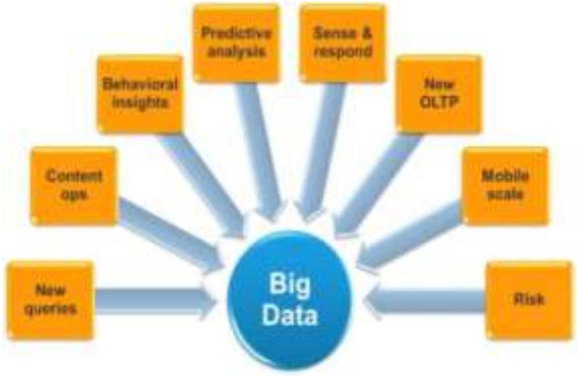


Fig.1. Big Data Paradigm

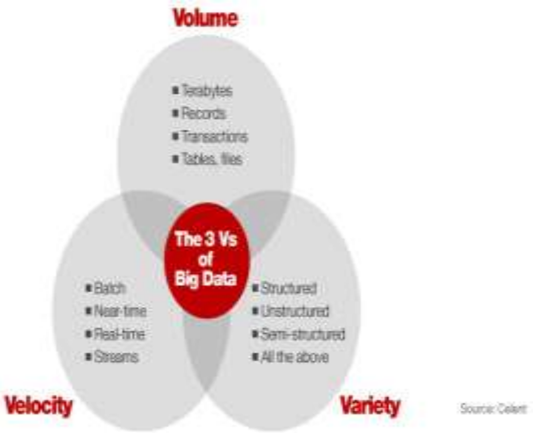
II.BIG DATA MINING

The term Big Data came into being in 1998 and was given by John Mashey with a title Big Data and the NextWave of infrastructure[8,16]. Big data mining has proved to be very important since the beginning. The use and processing of vast amount of data has given rise to the word Big Data. Usama Fayyad in his talk at KDD came out with some amazing statistics. For example: The number of queries that Google faces per day is more than 1 billion. There are 800 million updates on facebook and 250 million tweets on twitter per day. The data being used every day is in the order of zetta bytes and is increasing by 40% each year. We need huge and new algorithms to handle this data.

The administration of Big Data revolves around three V's that are as follows[36]:

- Velocity- It deals with the batch processing, heterogeneous streams and real-near time processing analysis through which a user extracts useful information through KDD.
- Volume- It manages transaction processing, numerous file records, relational tables whose size are in terabytes and this bigger size continues to increase.
- Variety- It consists of structure, unstructured and semi-structured sensor based all kinds of data.

In addition to those 3V's, there exists 2 more V's i.e. Variability and Data Value. These all V's as seen in fig.2. merged to form the definition of big data in 2012. This analysis has certain demands that big data methodologies should be cost efficient and innovation driven.



Source: Celent

Fig.2. 3 V'S of Big Data

There are many applications of Big Data which are as follow [17; 2]:

- Business: costumer personalization, churn detection
- Technology: reducing process time from hours to seconds
- Health: Determine DNA of each person so as to improve the lifestyle of people.
- Smart cities: cities focused on sustainable economic

2.1 Global Pulse: "Big Data for development"

To understand the relevance of big data mining. We represent the work that global pulse has been doing with the help of big data to improve life in different countries[17,24]. Global Pulse was introduced in 2009 and is an innovative lab that works on mining the big data. Strategy of this lab is as follows:

1) To detect the vulnerabilities innovative techniques are used for studying real time digital data.

2) For analyzing the real time data, they assemble free open source technology.

3) To establish global pulse at a very huge level. Global Pulse has described the opportunities which big data offers. They are as follows:

- Real Time Awareness: Design programs in such a way that reality is being represented.
- Real Time Feedback: To check which policy or program has failed and do the needful changes.
- Early Warning: Adopt fast responses at crucial time.

III. PRELIMINARY AREAS OF BIG DATA

This research paper basically discusses about the present scenario and future of Big Data Mining. The position of Big Data is very strong as different conferences and workshops are already taking place via numerous organizations and research institutes. As we look deep into feasibility of data mining model concepts, there is a long duration of time needed for execution of these models. Researchers have to provide insight towards infrastructure of Big Data Analysis so that time duration can be shortened. As figure 3 shows taxonomy of Big Data which directly connected to software development, Big Data infrastructure and Data Analysis. Software development is further casted into databases and open source development system. Data mining approaches and visualization tools come under the category of Data Analysis.

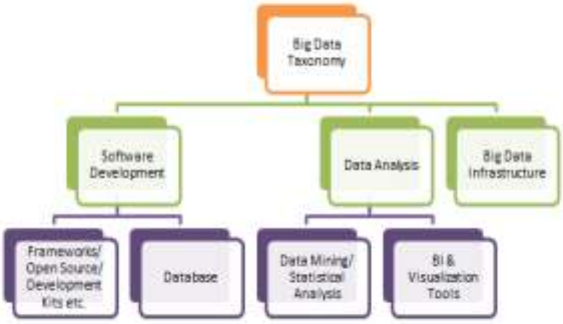


Fig.3. Big Data Taxonomy

As we stated earlier that data mining task of Big Data is very challenging as it contained multi-valued data, relational tables, defined from numerous sources that resulted into degradation of taxonomy of Big Data. There are some statistical analysis that contains rich semantics of structured type of data which is further used for extraction of Knowledge.[6,19] This representational knowledge can be represented using Big Graph Mining. There are social networking websites such as twitter, Face book which are using mining methodologies to reduce execution time of queries.

IV. CONTROVERSY ABOUT BIG DATA

The concept of Big Data is very vast with respect to accessibility and deployment of information to users. So there is a great curiosity in terms of controversies about Big Data which aggregates as follows:

- Data analysis and Big Data analysis are two purely different terms with respect to growing tendency and access methods.
- Companies proliferate approximation about Big Data Taxonomy is Hadoop and Map Reduce Algorithm is always best management system but it depends on size of company.
- The formation of data can be swapped in real time environment but what matters the most is its size.
- Big Data contain s huge amount of data so number of variable expands with big size so fake correlations among variables also increased.

4.1 OPEN SOURCE REVOLUTION

The big data phenomenon is directly related to open source software revolution. Large organizations like Twitter and Facebook are investing in Open Source Projects. The infrastructure of big data deals with Apache and Apache Hadoop that are used for Map Reduce Programing and Distributed File System i.e. Hadoop Distributed File System. Apache Hadoop related projects [32]: Apache Pig, Apache Hive, Apache HBase, Apache ZooKeeper, Apache Cassandra, Cascading, Scribe and many others. Apache S4 [26]: platform for processing continuous data streams. S4 is designed specifically for managing data streams. S4 apps are designed combining streams and processing elements in real time. Storm [31]: software for streaming data-intensive distributed applications, similar to S4, and developed by Nathan Marz at Twitter. In Big Data Mining, there are many open source initiatives.

The most popular are the following: Apache Mahout [4]: Scalable machine learning and data mining open source software based mainly in Hadoop. It has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining. [29]. It has implementations of classification, regression; clustering and frequent item set mining and frequent graph mining. It started as a project of the Machine Learning group of University of Waikato, New Zealand, famous for the WEKA software. The streams framework [6] provides an environment for Data mining and running stream processes using simple XML based definitions and is able to use MOA, Android and Storm. SAMOA [1] is a new upcoming software project for distributed stream mining that will combine S4 and Storm with MOA. Vowpal Wabbit [20]: open source project started at Yahoo! Research and continuing at Microsoft Research to design a fast, scalable, useful learning algorithm. VW is able to learn from tera feature datasets. It can exceed the throughput of any single machine network Interface when doing linear learning, via parallel learning. More specific to Big Graph mining we found the following open source tools: Pegasus [18]: big graph mining system built on top of Map Reduce. It allows patterns and anomalies in massive real-world graphs. See the paper by U. Kang and Christos Faloutsos in this issue. Graph Lab [24]: high-level graph-parallel system built without using Map Reduce. Graph Lab computes over dependent records which are stored as vertices in a large distributed data-graph. Algorithms in Graph Lab are expressed as vertex-programs which are executed in parallel on each vertex and can interact with neighboring vertices.

V. FORECAST TO THE FUTURE

There are different types of challenges that arrive from big data management, be it the type of data. [27, 16]. One of the challenges that any practitioner will have to face during the coming years: Analytics Architecture. The mystery behind optimal architecture of data continues and it is difficult to deal with historic data as well as real time issues.. An interesting proposal is the Lambda architecture of Nathan Marz [25]. The Architecture that solves the problem of computing arbitrary functions on arbitrary data in real time by decomposing the problem into three layers is called the lambda architecture.: the batch layer, the serving layer, and the speed layer. It combines the both hadoop

Hadoop
Telemetry
Unstructured
Data Analysis
Machine
Learning
Transforming
Big Data

Fig.4. Transforming Big Data

layer, and Storm for the speed layer. The properties defined by the system are: robust and fault tolerant, scalable, general, allows different queries, minimal maintenance, and debug gable. It is important to achieve significant statistical results, and not be fooled by randomness. It is important to achieve significant statistical results, and not be fooled by randomness. The transformation of Big Data is quite complex as it needs difficult implementation methods and requires long duration. Big Data techniques like Hadoop used telemetry methods for analysis of unstructured and semi-structured data. Machine learning approaches are enough efficient to built big data useful. The volume problem of Big Data can be solved by using two techniques as through compression and sampling. But there exists a very difficult challenge of Big Data is that how can we visualize this Big Data in our specified data warehouses. Researchers are working on Merge-Reduce algorithm which can work simultaneously in different environment. It can be used to manage hard machine learning data approaches in real time environment. The recent IDC study on Big Data reveals that 643 exa bytes of digital data form would be beneficial for approximation of big data if and only if when it is analyzed and marked in a managerial manner. But in present scenario, only 3% of big data is useful because of low equipped techniques present.

VI CONCLUSION

In the upcoming era, Big Data emerge as the most powerful concept used in technology world. The amount of data which is handled by today's academicians, they will manage five times more data than today. And through big data mining concepts, it becomes very easy technique to extract useful information in a very short duration of time. Time is the only factor to which we can rely in future and big data can have that much capability to reduce time complexity and make our work efficient and smart.

REFERENCES

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

[2] C. C. Aggarwal," 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. Kirkby, and B. Pfahringer. MOA:MassiveOnlineAnalysishttp://moa.cms.waikato.ac.nz/. Journal of Machine LearningResearch (JMLR), 2010.

[6] C. Bockermann and H. Blom. 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.SIGKDD Explorations Volume 14, Issue 2 Page 4.

[10] B. Efron. Large-Scale Inference: Empirical Bayes Meth-ods for Estimation, Testing, and Prediction. Institute ofMathematical Statistics Monographs. Cambridge UniversityPress, 2010.

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

[12] D. Feldman, M. Schmidt, and C. Sohler. Turning big Data into tiny data: Constant-size coresets for k-means,pca and projective clustering. In SODA, 2013.

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

[14] 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.

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

[16] V. Gopalkrishnan, D. Steier, H. Lewis, and J. Guszcza.Big data, big business: bridging the gap. In Proceed-ings 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.

[17]http://www.intel.com/content/www/us/en/bigdata/big thinkers-on-big-data.html, 2012.

[18] U. Kang, D. H. Chau, and C. Faloutsos. PEGASUS: Mining Billion-Scale Graphs in the Cloud. 2012.

[19] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6, 2001.

[20] J. Langford. Vowpal Wabbit, http://hunch.net/~vw/, 2011.

[21] D. J. Leinweber. Stupid Data Miner Tricks: Over_tting the S&P 500. The Journal of Investing, 16:15{22, 2007.

[22] E. Letouz_e. Big Data for Development: Opportunities & Challenges. May 2011.

[23] J. Lin. MapReduce is Good Enough? If All You Have is a Hammer, Throw Away Everything That's Not aNail! CoRR, abs/1209.2191, 2012.

[24] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson,C. Guestrin, and J. M. Hellerstein. Graphlab: A newparallel framework for machine learning. In Confer-ence on Uncertainty in Arti_cial Intelligence (UAI), Catalina Island, California, July 2010.

[25] N.Marz and J. Warren. Big Data: Principles and best Practices of scalable realtime data systems.Manning Publications, 2013.

[26] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: Distributed Stream Computing Platform. In ICDM Workshops, pages 170{177, 2010.

[27] C. Parker. Unexpected challenges in large scale machine learning. In Proceedings of the 1st International Work-shop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, BigMine '12, pages 1{6, New York,NY, USA, 2012. ACM.

[28] A. Petland,"Reinventing society in the wake of bigdata.Edge.org",http://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data, 2012.

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

[30] R. Smolan and J. Erwitt. The Human Face of Big Data.Sterling Publishing Company Incorporated, 2012.

[31] Storm, http://storm-project.net.

[32] N. Taleb. Antifragile: How to Live in a World We Don't Understand. Penguin Books, Limited, 2012.

[33] UN Global Pulse, http://www.unglobalpulse.org