DATA MINING CHALLENGES WITH BIG DATA

Prema Gadling*

Mr. Avinash Wadhe**

ABSTRACT—

The large amount of data being generated and that data stored is growing in an exponential manner, it is often used in part to the continuing advances in computer technology. While "big data" has become a highlighted buzzword since last two year, "big data mining", i.e., mining from big data, has almost without any delay followed up as an come into existence, interrelated research area. Big data is set to offer companies tremendous insight. But with terabytes and petabytes of data pouring in to organizations today, traditional architectures and infrastructures are not up to the challenge. IT teams are burdened with ever-growing requests for data, ad hoc analyses and one-off reports. Decision makers become frustrated because it takes hours or days to get answers to questions, if at all. This paper put up an overview of big data mining and discusses the related challenges and the new opportunities. The discussion resolve into a review of data mining challenges with big data for processing and managing big data as well as the endeavor expected on big data mining. We address broad issues associated to big data mining and/or big data, and point out challenges and issues topics as they shall at proper time flesh out.

Keywords- Issues, Challenges, Big data mining, Data mining, Big data.

* M.E (CSE) 2nd Semester, Department of Computer Science & Engg., G.H Raisoni College of Engineering, Amravati University

** M-TECH (CSE), Department of Computer Science & Engg., G.H Raisoni College of Engineering, Amravati University

1. Introduction

A. Data Mining

Data mining is a term from computer science. Data mining is about finding new information in a lot of data. The information obtained from data mining is hopefully both new and useful. Sometimes it is also called knowledge discovery in databases (KDD).

The overall goal of the data mining process is to take out information from a data set and modulation it into an apprehensible structure for furthermore utilization.[1] The process of data mining : The data mining process involves a series of steps to define a business problem, gather and prepared the data, build and evaluated mining model, and apply the models and disseminated the new information.

Figure 1: The Data Mining Process

B. Big Data

Big Data [2] is a unique term used to render identify the datasets that due to their large size and complexity. The Big Data is nothing but a data, available at heterogeneous, autonomous sources, in extreme large amount, which get updated in fractions of seconds. This data is nothing but the big data, which is so called due to its complexity. ***Doug*** Laney was the first one talking about 3V's in Big Data Management.

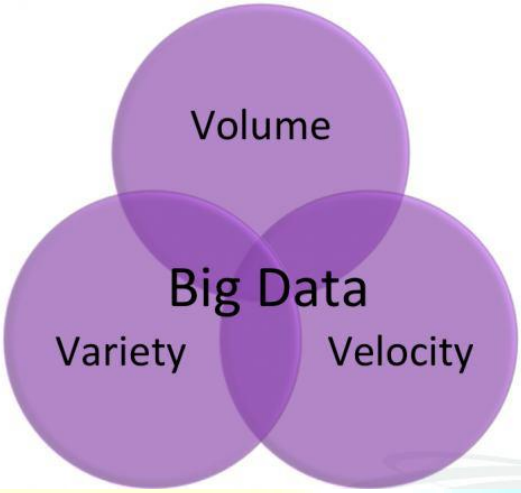

Figure 2: The three Vs (volume, variety, and velocity)

Because of these characteristics, there are currently a number of issues and challenges in addressing these characteristics going forward.

Velocity – how fast the data is entering the system.

Variety – includes all types of structured and unstructured data

Volume – the potential data capacity of terabytes to petabytes.

The features of Big Data [14] are:

1) It is huge in size.
2) The data keep on changing time to time.
3) Its data sources are from different phases.
4) It is free from influences, guidance or control of anyone.
5) It is too much complex in nature, thus hard to handle.

C. Big Data And Data Mining

The term Data Mining, finding for the exact useful information or knowledge from the collected data, for future actions, is nothing but the data mining. The following example of big data would be, the readings taken from an electronic microscope of the universe.

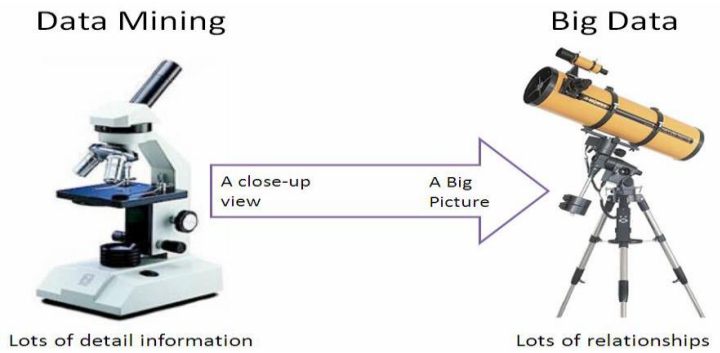

Figure 3: Data Mining with Big Data

So, collectively, the term Big Data Mining is a close up view, with lots of detail information of a Big Data with lots of information.

2. Literature Review

The procedure of the research into composite data basically concerned with the revelation of hidden patterns. Sagiroglu, S.; Sinanc, D. (20-24 May 2013),"Big Data: A Review" describe the big data content, its range, methods, samples, advantages and challenges of Data. The critical issue about the Big data is the privacy and security. By this paper, we can terminated that any organization in any industry having big data can take the advantage from its careful analysis for the problem solving purpose. [3]

The questioning the statement is not only to collect and manage the data also how to infusion the useful information from that gathered data. The overall Evaluation describe that the data is increasing and becoming complex.

Determined by real-life particular purpose and charge industrial stakeholders and assign an initial value by internal financial support agencies, handling and mining Big Data have seem to be a challenging still now very prevailing task. At that point the concept of Big Data literally related about data volumes, our HACE theorem recommendation that the key characteristics of the Big Data are

1) High attitude with varied and multifarious data sources
2) independent entity with concentrated and decentralized control, and

3) involution and transcription in data and knowledge affiliation. Such consolidate characteristics suggest that Big Data requisite a "big mind" to consolidate data for maximum values [4].

Cabena, P.; Hadjinian P.; Stadler R.; Verhees, J. & Zanasi, A. (1998) [5] in this paper "Data mining is an drawing from two or more field distributed collectively techniques from machine learning, pattern recognition, statistics, databases, and visual image to address the issue of information extratermination from essential data bases" (Cabena et al., 1998).

Gartner Group Advanced Technologies and Applications Research [6] in this paper "Data mining is the summos of recognize significant new correlations, patterns and trends by movable through transcend amounts of data reserve in repositories, using pattern recognition technologies as well as computable and numerical methods" (Gartner Group, 1995).

A. Torralba, R. Fergus, and W. Freeman [7] in this paper the authorized correlated with high relative amount is the data insufficiency and unintelligible tagged. Unlike most regular collection of data used for machine learning, which were extremely rector and noise free, Big Data is commonly unaccomplished resulting from their different origins. To make things even more difficult, large amount of data may not be tagged, or if tagged or mark with labeled, there obtain powerful labels. Take the 80 million very small image database index as an example, which has 80 million lower solution color images over 79,000 search terms.

J. Chien and H. Hsieh, W. de Oliveira [8] [9] in this paper there is another issues challenge to difficulties related to the high velocity is that data are oftentimes non-stationary, i.e., data allocation is variable over time. Substantially, non-stationary data are commonly dissociated into allotment with data from a very tiny time interval. The presumption is that data enclose in time are piece-wise stationary and may be characterized by a declaratory degree of correlation and, therefore, follow the same allocation.

Priya P. Sharma, Chandrakant P. Navdeti [10] in this paper discusses about the big data security at the environment level along with the probing of built in protections. It also presents some security issues that we are dealing with today and propose security solutions and commercially accessible techniques to address the same. The paper also covers all the security solutions to secure the Hadoop ecosystem.[10]

A. Labrinidis and H. Jagadish, et al. [11] in this paper there are many technical challenges that must be addressed to compose the full exist of in possibility big data and provide a

comprehensive discussion of such challenges based on the notion of data analysis pipeline. Example for Big Data generated, as the knowledge comes to pass from several, lack of relation, existing as an independent entity sources with exponential and acquiring relationships, and keeps gaining.

In this paper " QIANG YANG,XINDONG WU [12] —10 Challenging Problems in Data Mining challenge will remain very important for the data mining community: developing systems which derive understandable patterns and making already derived patterns understandable.

Agrawal, D., Bernstein, P., Bertino, E., et al. [13] in this paper there is the big data mining is a promising research area, still in its infancy. In spite of the limited work done on big data mining so far, we believe that much work is required to overcome its challenges related to heterogeneity, scalability, speed, accuracy, trust, provenance, This paper also provides an overview (though limited due to space limit) of state-of-the-art frameworks/platforms for processing and control big data as well as platforms and libraries for mining big data.

3. Issues In Big Data

We suggest there are three fundamental issue areas that need to be addressed in dealing with big data: storage issues, management issues, and processing issues.

3.1 Issues Related To The Characteristic

Data Volume – Data volume distinguish the quantity of data accessible to an organization, which does not eventually have to own all of it as long as it can access it. So, there will be produced data in order of terabytes everyday and this amount of data is emphatically difficult to be interact in certain way.

Data Velocity – Our traditional system are not capable of enough on performing the analytics on data which is constantly in transitional. Data velocity management is a great amount of more than a bandwidth derivation; it is also an ingest issue (extract transform- load).

Data Variety - Data variety is a valuation of the abundant wealth of the data illustration – text, video, audio, etc. From an analytic perspective, it is most likely the biggest prevention to reality or fact using large volumes of data. transcendental data publications, non-aligned data structures,

and not capable of being made data semantics impersonate significant challenges that can lead to enumeration conurbation.

3.2 Storage And Transport Issues

Consider that a 1 gigabyte per second network has an influential extend in duration transfer rate of 80%, the sustainable bandwidth is about 100 megabytes. Thus, transferring an exabyte would take about 2800 hours, if we accept that a sustained transfer could be maintained. It would continue lacking to convey the data from a collection or storage point to a processing point than it would to actually putting through a prescribe procedure it! Two solutions manifest themselves. The first one is "bring the code to the data", vs. the traditional method of "bring the data to the code." Second, perform triage on the data and transmit only that data which is critical to downstream analysis.

3.3 Management Issues

Management will, perchance, be the most exquisitely problem to address with big data. Due to data volumes and the obligation to have punctual data, pressure is mounting to rather build applications that use the data services without deviation as the data source and perform any inexplorable integration in real time. This reduces not necessary data duplication and increases the timeliness but make a new entire set of issues to do with accordance with the fact, recoverability and accessibility that will require determination. "We are incognizant of any constitution, open source, platform autonomous solution to this problem." As far as we know, this left over true today. To summarize, there is no accurate big data management solution yet.

3.4 Processing Issues

Big Data processing include multidimensional signal analytics that build upon a signals and systems fabric, hence signal processing expertise is sure to play an important role in designing and deploying such large, distributed, fault-tolerant systems. Thus, prevalent processing of data will require capacious parallel processing and give something useful analytics algorithms in order to provide propitious and actionable information.

4. Challenges In Big Data

Data integration – The ability to combine data that is not similar in structure or source and to do so quickly and at reasonable cost. With such variety, a related challenge is how to manage and control data quality so that you can meaningfully connect well understood data from your data warehouse with data that is less well understood.

Data volume – The ability to process the volume at an acceptable speed so that the information is available to decision makers when they need it.

Skills availability – Big Data is being harnessed with new tools and is being looked at in different ways. There a shortage of people with the skills to bring together the data, analyze it and publish the results or conclusions.

Solution cost – Since Big Data has opened up a world of possible business improvements, there is a great deal of experimentation and discovery taking place to determine the patterns that matter and the insights that turn to value. It is crucial to reduce the cost of the solutions used to find that value.

5. Data Mining Issues

There are several effective implementation issues related with data mining:

Human interaction – For data mining difficulties are often not precisely stated, interfaces may be need with both designation and technical experts. Technical experts are used to devised the queries and assist in interpretation the results. Users are needed to coincide training data and craved results.

Over-fitting – When a model is bring into existence that is related with a given database state, it is acceptable that the model also state eventuality database states. Over-fitting occurs when the model does not fit future states. This may be caused by evolution that are made about the data or may somewhat be caused by the insignificance size of the training database. Over-fitting can arise under other division as well, even though the data are not changing.

Outliers – There are repeatedly certain number data entries that do not accord into the derived model. This remains even more of an issue with very large databases. If a model is more useful that includes these outliers, then the model may not conduct well for data that are not outliers.

Social One – Isolate of the key issues in relief by data mining technology is not a enterprise or technological one, but a social one. It is the issue of particularly privacy. Data mining create it attainable to examine routine.

Data integrity – Some other issue is that of data integrity is that an intelligent manner data analysis can only be as commodity as the data that is actuality canvas. A key execution challenge is integrating in disagreement or more than is needed data from alteration sources.

Interpretation of results – Presently, data mining output may necessitate experts to in an accurate manner interpret the results, which might otherwise be meaningless to the round number database user.

Visualization of results – To readily perspective and recognize the output of data mining algorithms, visual perception of the results is useful.

Large datasets – The impose in size datasets belonging to with data mining generate problems when as to an analysis algorithms intentional for small datasets. Sample tabulation and parallelization are dominant tools to attack this scalability problem.

High dimensionality – A conventional database schema serenely self-possessed of many different attributes. The resolution here is that not all attributes may be needed to enumerate a given data mining problem. The use of other attributes may simply increase the overall the quality of being intricate and decrease the efficiency of an algorithm. One solution to this high dimensionality problem is to contract the number of attributes.

Multimedia data – Most previous data mining algorithm objective to descriptive data types (numeric, character, text, etc.). The use of multimedia data such as is establish in GIS databases complicates or declare valid many proposed algorithms.

Relational or Multidimensional databases – An hated manner argument technical issue is whether it is superior in quality to set up a relational database structure or a multidimensional one. In a relational construction, data is stored in tabular array, allowable ad hoc queries. In a marked by several dimensional configuration, on the one hand, places of cubes are developing arrays, with subsets make a new according to condition.

Noisy data – Some attribute value might be no longer valid or wrong. These values are commonly corrected earlier running data mining practical applications.

Irrelevant data – Some property in the database might not be of involvement to the data mining attempt existence germinated.

Missing data – During the pre-processing phase of knowledge discovery in databases (KDD), missing data may be replaced with estimates. This and other approaches to handling missing data can lead to invalid results in the data mining step.

Changing data – Databases cannot be assumed to be static. However, most data mining algorithms do assume a static database. This requires that the algorithm be completely rerun anytime the database changes.

Application – Determining the intended use for the information obtained from the data mining function is a challenge.

Cost – Finally, there is the issue of cost. Effectively integrating sustainable design elements into projects during project development and design phases can minimize building costs. Conversely, if green design elements are considered late in the design process and designers have to "redesign" the entire project, overall costs can increase significantly.This increases pressure for larger, faster systems, which are more expensive (Baazaoui Z et al,2005).

6. Data Mining Challenges

Distributed data – The data to be mined is reserve in divided up structure process environments on heterogeneous platforms. Consequently, development of algorithms, tools, and services is required that make easier the mining of distributed data (Mechitov A. et al, 2001).

Distributed operations – In subsequent more and more data mining preparations and algorithms will be uncommitted on the power system. To facilitate seamless integration of these resources into expansion out data mining systems for complex problem solving, novel algorithms, tools.

Massive data – Amplification of algorithms for mining large, imposing in size and high-dimensional data sets (out-of-memory, parallel, and distributed algorithms) is necessary for relief or supply.

Complex data types – Increasingly composite data sources, structures, and types (like natural language text, images, time series, multi-relational and object data types etc.) are coming into existences. Grid-enabled mining of such data will require the development of new methodologies, algorithms, tools, and grid services.

Data privacy, security, and governance – Operated by automated data mining in distributed environments raises serious issues in expression of data privacy, security, and governance. Grid-based data mining technology will need to address these issues.

User-friendliness – In the end a system must blindfold technological complication from the user. To facilitate this, new software, tools, and infrastructure out growth is needed in the areas of grid-supported workflow management, resource identification, allocation, and scheduling, and user interfaces. [10]

Data Cleaning and preprocessing [15] [16] **–** It is an important step to ensure the data quality and to improve the efficiency and ease of the mining process. Actually Data preprocessing includes data cleaning to remove noisy data and outliers, data reduction to reduce the dimensionality and complexity of the data. In order to improve the efficiency and accuracy of knowledge discovery and data mining process, ensuring data quality is a big challenge.

Post processing – It is the process to refine and evaluate the knowledge derived from mining procedure (Bruha and Famili, 2000).

Tasks and Algorithms – Data mining tasks and algorithms are essential steps of knowledge discovery. For domain specific applications, utilizing the domain knowledge to guide data mining process or improve data mining performance is a challenging issue.

Dirty Data – Here no surprise that dirty data tops the list, because it has been at the top of the list for the past several years [15] in area of data mining as a challenging issue. Many data miners provided input as "How they have tried to overcome the problem and how to provide a clear theme emerges". This can help manage expectations about providing potential results of a data modeling exercise and also create action plans to improve quality of data.

7. Challenging Issues in Data Mining With Big Data.

There are three sectors at which the challenges for Big Data arrive. These three sectors are:

1) Mining platform.
2) Privacy.
3) Design of mining algorithms.

Basically, the Big Data is stored at different places and also the data volumes may get increased as the data keeps on increasing continuously. So, to collect all the data stored at different places is that much expensive. Suppose, if we use these typical data mining methods (those methods which are used for mining the small scale data in our personal computer systems) for mining of

Big Data, and then it would become an obstacle for it. That's why the typical methods are required data to be loaded in main memory, though we have super large main memory.

To maintain the privacy is one of the main aims of data mining algorithms. Presently, to mine information from Big data, parallel computing based algorithms such as MapReduce are used. In such algorithms, large data sets are divided into number of subsets and then, mining algorithms are applied to those subsets. Finally, summation algorithms are applied to the results of mining algorithms, to meet the goal of Big Data mining. In this whole procedure, the privacy statements obviously break as we divide the single Big Data into number of smaller datasets. When we divide the Big Data in to number of subsets, and apply the mining algorithms on those subsets, the results of those mining algorithms will not always point us to the actual result as we want when we collect the results together. While designing such algorithms, we face various challenges. As shown in the figure 4 below, there are blind men observing the giant elephant. Everyone is trying to tell in advances their conclusion on what the thing is actually. Somebody is saying that the thing is a hose; someone says it's a tree or pipe etc.

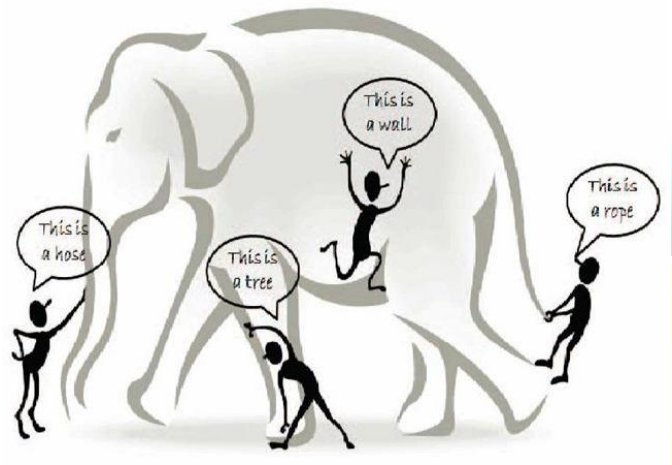

Figure 4: Blind men and the giant elephant

Actually everyone is just observing some part of that giant elephant and not the whole, so the results of each blind person's prediction is something different than actually what it is.[14]

Another challenging issues in big data mining is that fetching the requested information or even uncovering some hidden relationships and patterns between numeral parameters. Analyzing fast and massive stream data may lead to new valuable insights and theoretical concepts [17].

Comparing with the results derived from mining the conventional datasets, unveiling the huge volume of interconnected heterogeneous big data has the potential to maximize our knowledge and insights in the target domain.

Economical, temporary, and unprepared data are any process serving to define shape features for Big Data applications. Being expanded, the complex number of data points is too few for drafting unmistakable conclusions. This is ordinarily a process of complicating of the data dimensionality issues, where data in a high-dimensional multitude (such as more than 1,000 dimensions) do not expose simplify general line of orientation or appropriation. For about machine cognitive process of acquiring skill and data mining algorithms, high-dimensional spare data in a declaratory manner develop the reliability of the models derived from the data. The generic access are to consumption dimension subtraction or tendency selection to shrink the data dimensions or to carefully come under additional principles to assuage the data shortcoming, such as all over under constant matriculation methods in data mining.

Fashioning use of complex data is a major challenge for Big Data Mining applications, because any two parties in a complex network are with possibility of becoming actual interested to each other with a social connection. Such a connection is quadratic with related to the number of client in the network, so a million node network may be constrained to one trillion connections. For a large social network site, like Facebook, the number of active users has beforehand reached 1 billion, and examining such an enormous network is a big challenge for Big Data mining.

Data secrecy has been at all time an issue even from the introductory when data mining was applied to real-world data. The concern has become extremely serious with big data mining that often requires personal information in order to produce relevant/accurate results such as location-based and personalized services, e.g., targeted and individualized advertisements. Also, with the huge volume of big data such as social media that contains tremendous amount of highly interconnected particular to a given individual information, every piece of information about everybody can be mined out, and when all pieces of the information about a person are dug out and put together, any privacy about that someone instantly disappears.

By inter activeness we mean the capability or feature of a data mining system that allows promptitude and adequate user a mutual action such as feedback/interference/guidance from users. Interactiveness is relatively an underemphasized issue of data mining in the past. When

our society is now confronting the challenges of big data mining, interactiveness becomes a critical issue.[17]

8. Conclusion

There are various challenges and issues regarding big data, data mining. We discussed in this paper some insights about the topic, and what we consider are the main concerns, and the main challenges for the future. Big data discloses the limitations of existing data mining techniques, resulted in a series of new challenges related to big data mining. Big data mining is a promising research area, still in its infancy. Additionally this paper also discussed about the various challenges and issues in field of big data mining which is important to do further more effective research in this emerging field.

These challenges and issues are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data Mining.

Acknowledgment

Many thanks for Professor Avinash Wadhe of Computer Science Department of G.H. Raisoni College of Engineering, Nagpur University, for his guidance and support.

References

[1] Wei Fan and Albert Bifet,2013, " Mining Big Data: Current Status and Forecast to the Future", Vol 14,Issue 2.
[2] Bharti Thakur, Manish Mann,2014, " Data Mining for Big Data: A Review",International Journal of Advanced Research in Computer Science and Software Engineering , Volume 4, Issue 5.
[3] Sagiroglu, S.; Sinanc, D.,2013, "Big data: A review," Collaboration Technologies and Systems (CTS), 2013 International Conference on , vol., no., pp.42,47, 20-24 May 2013
[4] A. Jacobs,2009, "The Pathologies of Big Data," Comm. ACM, vol. 52,no. 8, pp. 36-44.

[5] Cabena, P.; Hadjinian P.; Stadler R.; Verhees, J. & Zanasi, A. (1998) Discovering data mining: From Concept to Implementation, Prentice-Hall, Inc.ISBN:0-13-743980-6
[6] Gartner Group (1995), Gartner Group Advanced Technologies and Applications Research. Note http://www.gartner.com
[7] A. Torralba, R. Fergus, andW. Freeman,2008, ``80 million tiny images: A large data set for nonparametric object and scene recognition," IEEE Trans.Softw. Eng., vol. 30, no. 11, pp. 1958_1970, Nov. 2008.
[8] J. Chien and H. Hsieh,2013, ``Nonstationary source separation using sequential and variational Bayesian learning," IEEE Trans. Neural Netw. Learn. Syst., vol. 24, no. 5, pp. 681_694, May 2013.
[9] W. de Oliveira,2007,``The Rosenblatt Bayesian algorithm learning in a nonstationary environment," IEEE Trans. Neural Netw., vol. 18, no. 2, pp. 584_588, Mar. 2007.
[10] Priya P. Sharma, Chandrakant P. Navdeti, (2014), " Securing Big Data Hadoop: A Review of Security" / (IJCSIT) International Journal of Computer Science

and Information Technologies, Issues, Threats and Solution", IJCSIT, 5(2), pp2126-2131, Vol. 5 (2) , 2014, 2126-2131
[11] A. Labrinidis and H. Jagadish,2012, "Challenges and Opportunities with Big Data," Proc. VLDB Endowment, vol. 5, no. 12, 2032-2033,2012.
[12] QIANG YANG,XINDONG WU,2006 —"10 Challenging Problems in Data Mining Research" International Journal of Information Technology & Decision Making Vol. 5, No. 4 (2006) 597–604.
[13] Agrawal, D., Bernstein, P., Bertino, E., et al.,2012,: Challenges and Opportunities With big data Community White Paper Developed by Leading Researchers Across the United States(2012),http://cra.org/ccc/docs/init/bigdatawhitepaper. Pdf
[14] Rohit Pitre, Vijay Kolekar,2014, "A Survey Paper on Data Mining With Big Data" International Journal of Innovative Research in Advanced Engineering (IJIRAE) Volume 1 Issue 1 (April 2014)
[15] Ian H. Witten; Eibe Frank; Mark A. Hall,2014, —Data Mining: Practical Machine Learning Tools and Techniques (3rd Ed.)‖ , Elsevier, 30 January 2011.
[16] Bhoj Raj Sharmaa*, Daljeet Kaura and Manjub,2013," A Review on Data Mining: Its Challenges, Issues and Applications", International Journal of Current Engineering and Technology, Vol.3, No.2 (June 2013), ISSN 2277 – 4106
[17] Berkovich, S., Liao, D.,2012: On Clusterization of big data Streams. In: 3rd International Conference on Computing for Geospatial Research and Applications, article no. 26. ACM Press, New York.