A Novel Integrated Approach for Big Data Mining

Ritu Katarha[#1], Hareram Shah*[2]
[#]Research Scholar, Computer Science & Engg.
Gyan Ganga Instt. of Technology and Sciences, Jabalpur(M.P.)
[*]Asst. Prof., Computer Science & Engg.
Gyan Ganga Instt. of Technology and Sciences, Jabalpur(M.P.)

Abstract: The growing volume of digital internet of things is fuelling it even extra. The rate of data increase is surprising and this information comes at a speed, with diversity which is not necessarily structured. Ability to examine this massive amount of data is bringing a new age of productivity growth, novelty and consumer surplus. Data analysis, organization, retrieval, and modeling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed. Finally, presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge. By integrating data mining and cloud computing in IDMCC provides quick access to technology. The outcome of this integration is strong and capacitive platform that will be able to deal with the increasing production of data, or that will generate the situation for the proficient mining of big amount of data from diverse data storehouses with the intend of creating helpful information or the creation of innovative knowledge. This paper deals with the learning of how data mining is used in cloud computing.

Keywords: Big Data Mining, Integrated Data and Cloud Computing – IDMCC

1. INTRODUCTION

The growing capability to produce huge quantities of data brings potentials to determine and consume valuable knowledge from data. Data mining has been a method to examine data from diverse sources and getting useful information from data. Big data is the term for a compilation of data sets so huge and complex that it becomes complicated to process it using conventional database management tools or data processing applications. The challenges comprise the areas of capture, storage, search, sharing, transfer, analysis, and visualization of this data. Data mining can also help in predicting trends or values, classification of data, categorization of data and to discover correlations, patterns from the dataset [3]. The worldwide economic recession and the reduction budget of IT projects have lead to the require of expansion of integrated information systems at a lesser cost. Nowadays, the promising occurrence of cloud computing aims at transforming the conventional way of computing, by providing both software applications and hardware resources as a service. Enterprise IT infrastructure acquires many costs ranging from hardware costs and software maintenance costs to the costs of monitoring, managing, and maintaining IT infrastructure [1].The current arrival of cloud computing offers some corporeal projection of reducing some of those costs; however, concept provided by cloud computing are frequently inadequate to give main cost savings across the IT infrastructure life-cycle [3]. Cloud infrastructure can be efficiently used for demanding operations with data that is classic for processes of data mining. It is essential to have accessible scalable data warehouses and scalable computing resources that are able to allow. Using an integrated approach based on data mining and cloud computing can be a solution to acquire the quick access to technology. In addition, the solution may offer new opportunities to improve practices and attain innovation.

2. SCOPE OF BIG DATA MINING

Due to the massive victory of a variety of application areas of data mining, the data mining has been establishing itself as the main obedience of computer science and has shown interest potential for the future developments. Constantly growing technology and future application areas are always creates new challenges and opportunities for data mining[8], the usual future trends of data mining includes:-

i. Standardization of data mining languages
ii. Data pre-processing
iii. Complex objects of data
iv. Computing resources
v. Web mining
vi. Scientific Computing Business Data

3. DATA MINING PARAMETERS

The data mining parameters are as follows [6]:

Association– Describes patters where one event is connected to another event.

Path analysis–Describes patterns where one event leads to another later event.

Classification – Finding for new patterns.

Predictive analysis – Discovering patterns in data that can lead to predictions about the future.

Clustering–Finding and documenting groups of facts not previously known.

4. BIG DATA TO BIG DATA MINING

Technology uprising has been facilitating millions of people by producing tremendous data via ever-increased use of a diversity of digital devices and particularly remote sensors that generate nonstop streams of digital data, consequential in what has been called as “big data”. It has been a confirmed incident that huge amounts of data have been being recurrently generated at ever increasing scales [8].

Applying existing data mining algorithms and techniques to real-world problems has been recently running into numerous challenges due to the insufficient scalability of these algorithms and techniques that do not match the three Vs of the rising big data. Present data mining techniques and algorithms are not prepared to meet the new challenges of big data. Mining big data demands extremely scalable approach and algorithms, more efficient pre-processing steps such as data filtering and integration, advanced parallel computing environments, intellectual and efficient user interaction. The speed/velocity demand of big data (especially stream data) processing asks for proportionate real-time efficiency which over again is far ahead of where current DBMSs could reach.

5. ADVANTAGES OF USING DATA MINING WITH CLOUD COMPUTING

Cloud computing collective with data mining can offer powerful capacities of storage and computing and an admirable resource management [4]. Due to the volatile data growth and amount of computation involved in data mining, an efficient and high-performance computing is exceptionally essential for a successful data mining application. Data mining in the cloud computing situation can be considered as the future of data mining because of the advantages of cloud computing paradigm. The major concern about data mining is that the space required by the operations and item sets is very large. But if we merge the data mining with cloud computing we can save an extensive amount of space. This can benefit us to a great scope.

6. DISADVANTAGES OF USING DATA MINING WITH CLOUD COMPUTING

There are convinced subjects associated with data mining in the cloud computing. The key issue of data mining with cloud computing is security as the cloud source has complete control on the underlying computing infrastructure. Exceptional concern has to be taken so as to ensure the security of data under cloud computing environment.

7. PROPOSED WORK

Data mining is a method of extracting information from the raw data and cloud computing provides scalable and flexible infrastructure which offers everything as a service. By integrating data mining and cloud computing provides suppleness and quick access to technology. The conventional database system and RDBMS are not capable to mine big data set and the existing data mining algorithms are not capable to mine big data.

So we suggest the novel approach for data mining using cloud system. IDMCC (Integrated Data Mining and Cloud Computing), this integrated approach uses a hadoop distributed file system, HDFS and Map Reduce, where the data can be easily partitioned over thousands of nodes in a cluster and the Map Reduce framework permits users to describe two functions, map and reduce, to process a big number data entries in parallel. The innovative approach involve in map reduce functionality with efficient data mining approach. Over this design we are implementing our data mining application, which will work with HDFS and map reduce for big data mining.

7.1 A NEW INTEGRATED DATA MINING AND CLOUD COMPUTING - IDMCC

The integrated approach of data mining and cloud computing and mining is the process which can extract structured information from unstructured or semi-structured web data sources. The application of this knowledge should facilitate that with a few clicks one can collect the information about the end user of the application completely. Data mining in cloud computing agree to associations to integrate the management of software and data storage with promise of efficient, secure and reliable services for their users. It gives technology that can handle huge quantity of data which cannot be processed efficiently at reasonable cost using standard technologies and techniques. It moreover allows the users to recover important information from virtually integrated data warehouse that decreases the cost of infrastructure and storage. Expansion of cloud computing will drive the technological and internet achievements in the public service is to support the intensity of information resources sharing and sustainable use of new methods and new ways of conventional data mining.

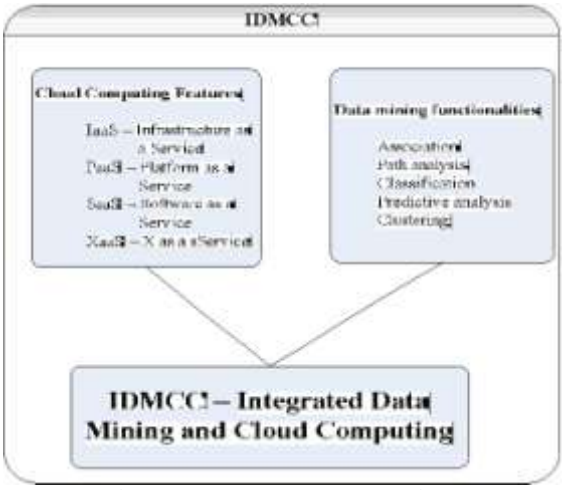

Figure7.1 IDMCC Integration

7.2 ADVANTAGES OF IDMCC

The following are the advantages of the integrated data mining and cloud computing environment.

i. Virtual m/c that can be started with short notice
ii. Redundant robust storage
iii. No query structured data
iv. Message queue for communication

v. The customer only pays for the data mining tools that he needs
vi. The customer doesn't have to maintain a hardware infrastructure as he can apply data mining through a browser

7.3 IMPLEMENTATION OF DATA MINING APPLICATION ON HDFS ARCHITECTURE

After configuration HDFS, we are configuring our application for big data mining. In this step, we are using configured an iso image file of our virtual machine. This setup can be run over any virtual platform; we need not to configure it again for other systems. Over this architecture we configure our Big Data Mining application. This application will process map reducer's huge data according to the filter criteria given to the application. We need not to use upper layer tools of HDFS architecture for further implementation. We are using lamp server which is configured on this architecture for our application for data retrieval.

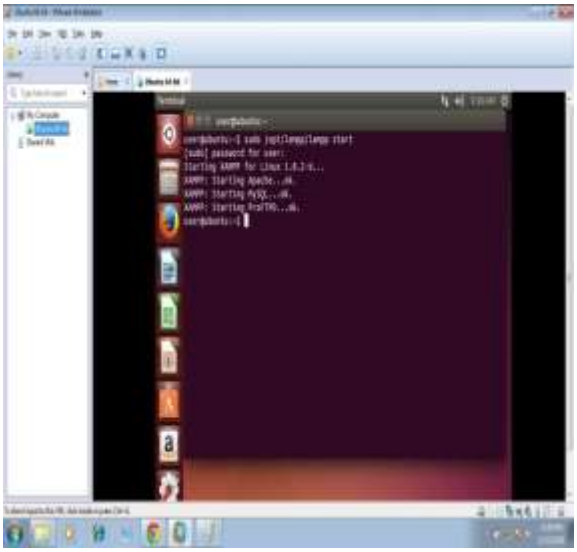

Fig. 7.2 Configured LAMPP and Services

To start and run the mining application, we use browser. Call the application start page by giving the url on address bar. After successfully call and running the index page, the application visualize as the following screen.

Fig. 7.3 Big Data Mining Application

After selecting all options according to need of mining, the final outcome of filter data will be shown on browser as

Fig. 7.4 Big Data Mining Results

7.4 CONCLUSION

This paper presents a review of requirement of data mining services in cloud computing along with a case study on the integrated approach of data mining and cloud computing. The implementation of data mining techniques through cloud computing will allow the users to retrieve significant information from virtually integrated data warehouse that decreases the costs of infrastructure and storage. This approach also reduces the difficulties that keep small companies from benefiting of the data mining instruments. The appearance of cloud computing brings new ideas for data mining. It increases the scale of processing data.

As we have entered an age of Big Data, processing big volumes of data has never been greater. Throughout improved Big Data analysis tools like Map Reduce over Hadoop and HDFS, guarantees faster advances in various scientific disciplines and improving the profitability and success of many enterprises.

REFERENCES

[1] Alawode A. Olaide, "On Modeling Confidentiality Archetype and Data Mining in Cloud Computing", African Journal of Computing & ICT, Vol 6. No. 1, March 2013

[2] Bhanu Bhardwaj, "Extracting Data Through Webmining", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 3, May - 2012

[3] Janardhan. N, T. Sree Pravallika, Sowjanya Gorantla, "An efficient approach for integrating data mining into cloud computing", International Journal of Computer Trends and Technology (IJCTT) - volume4 Issue5–May 2013

[4] Daniel J. Abadi, Yale University, DataManagement in the Cloud: Limitations and Opportunities, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 2009.

[5] Naskar Ankita, Mrs. Mishra Monika R., "Using Cloud Computing To Provide Data Mining Services", International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2 Issue 3 March 2013 Page No. 545-550.

[6] Zeba Qureshi, Jaya Bansal, Sanjay Bansal, A Survey on Association Rule Mining in Cloud Computing, paper of ijetaeISSN 2250-2459, Volume 3, Issue 4, April 2013).

[7] Robert Vrbić, "Data Mining and Cloud Computing",Journal of Information Technology and Applications, Volume 2 December 2012.

[8] B R Prakash, Dr. M. Hanumanthappa, Issues and Challenges in the Era of Big Data Mining, *IJETTCS*-ISSN 2278-6856, Volume 3, Issue 4 July-August 2014.

[9] Apache Hadoop. What Is Apache Hadoop?, 2014. http://hadoop.apache.org/, accessed April 2014.