A STUDY ON INTEGRATED APPROACH OF DATA MINING AND CLOUD MINING

B. KAMALA

Department of Computer Applications, Sri Sai Ram Engineering College, Chennai – 44.

Abstract:Data Mining and Cloud Computing are the rising trends in the current world of information technology Data mining is a process of extracting information from the raw data and Cloud computing provides scalable and flexible infrastructure which provides everything as a service. By integrating data mining and cloud computing (Integrated Data Mining and Cloud Computing– IDMCC) provides agility and quick access to technology. The result of such integration should be strong and capacitive platform that will be able to deal with the increasing production of data, or that will create the conditions for the efficient mining of large amount of data from various data warehouses with the aim of creating helpful information or the production of new knowledge. This paper deals with the study of how data mining is used in cloud computing with a case study in a healthcare system.

Keywords: Cloud computing, Data Mining, Integrated Data and Cloud Computing – IDMCC, Healthcare systems

I. INTRODUCTION

The increasing ability to generate large quantities of data brings potentials to discover and utilize valuable knowledge from data. Data mining has been a successful tool to analyze data from different angles and getting useful information from data. It can also help in predicting trends or values, classification of data, categorization of data and to find correlations, patterns from the dataset. The global economic recession and the shrinking budget of IT projects have led to the need of development of integrated information systems at a lower cost. Today, the emerging phenomenon of cloud computing aims at transforming the traditional way of computing, by providing both software applications and hardware resources as a service. Enterprise IT infrastructure incurs many costs ranging from hardware costs and software licenses/maintenance costs to the costs of monitoring, managing, and maintaining IT infrastructure. The recent advent of cloud computing offers some tangible prospects of reducing some of those costs; however, abstractions provided by cloud computing are often inadequate to provide major cost savings across the IT infrastructure life-cycle. [10] Cloud infrastructure can be effectively used for exhaustive and demanding operations with data that is typical for processes of data mining. It is necessary to have available scalable data warehouses and scalable computing resources that are capable to accept. The scalable warehousing and computing resources capability provides the efficient way of storing and analyzing the large amounts of data. [9] Using an integrated approach based on Data Mining and Cloud Computing may be a solution to obtain the liveliness and quick access to technology. In addition, the solution may offer new opportunities to improve practices and attain innovation. The remainder of this paper is organized as follows: Section 2 deals with Scope of Data Mining, their parameters and steps involved in data mining. Section 3 discusses the approach of cloud computing. Section 4 describes the integrated approach of data and cloud computing. Section 5 deals with a case study on Health Care Domain in IDMCC.

II. SCOPE OF DATA MINING

Data mining can be defined as "the process that attempts to discover patterns in large data sets". The overall goal of the data mining process is to extract information from a large data set and convert it into an understandable structure for future use. Data mining is the process of discovering or finding new, valid, understandable and potentially useful forms of data. The form of data refers to a discovered regularity among the data variables. If the detected regularity applies to all data, then it is about discovered model, and the regularity of data can be correlated with the extent of data – it is a pattern or template. It is carried out over large volumes of data in order to pull new information out of them that will be the basis for making better business decisions. It is also used to find knowledge, and knowledge is represented through certain patterns. [2] Association rule is the most often used method in data mining, which finds out the association between data and various objects by finding the potential dependence among data. Classification and clustering can be used to sort out things by characterizing the common significance among different things. [7] The disadvantage of data mining in centralized database, generally have the several following points: network traffic is considered less, mining efficiency is low and the degree of spatial complexity is high. The most classic classification data mining is classification methods based on distance, classification methods

based on decision tree, Bayesian classification and so on. [5]

Data Mining parameters

The data mining parameters are as follows:

- 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.

Steps involved in Data Mining process

The steps involved in the data mining process are,

- Definition of the business problem – the problem statement is defined and it determines the required data.
- Data preparation – this step includes transformation and sampling of data, evaluation of data.
- Modeling – includes the selection of suitable mining technique, building and evaluation of the models.
- Implementation – involves the interpretation and use of results.

The following figure explains the different steps which comprise the overall data mining process.

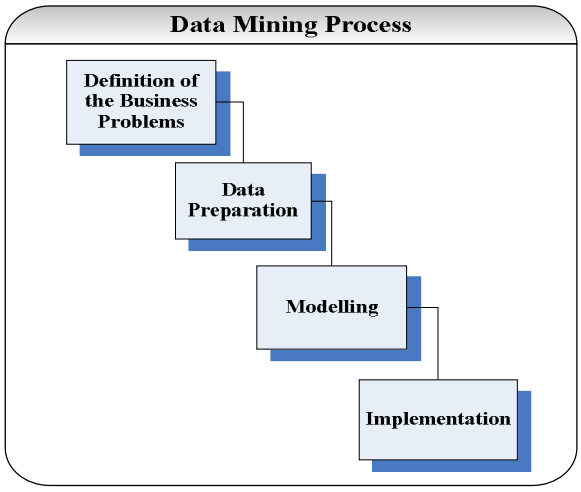

Figure 1 Steps in Data Mining Process

III. WHAT IS CLOUD COMPUTING?

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, services) that can be rapidly provisioned and released with minimal management effort or service provider interaction Cloud computing is the delivery of computing and storage capacity as a service to a community of end-recipients. Cloud computing entrusts services with a user's data, software and computation over a network.

There are various opinions on what is cloud computing. It can be the ability to rent a server or a thousand servers and run a geophysical modeling application on the most powerful systems available anywhere. It can be the ability to rent a virtual server, load software on it, turn it on and off at will, or clone it ten times to meet a sudden workload demand. It can be storing and securing immense amounts of data that is accessible only by authorized applications and users. It can be supported by a cloud provider that sets up a platform that includes the OS, databases, and scripts with the ability to scale automatically in response to changing workloads. Cloud computing can be the ability to use applications on the Internet that store and protect data while providing a service — anything including email, sales force automation and tax preparation. It can be using a storage cloud to hold application, business, and personal data. And it can be the ability to use a handful of Web services to integrate photos, maps, and GPS information to create a mashup in customer Web browsers. [1] The following figure describes cloud services framework which has Infrastructure, Platform and Applications as its main services classification. Storage and computing can be classified in IaaS, Business and development in PaaS and Software as a Service and on demand web services in Application services. [11]

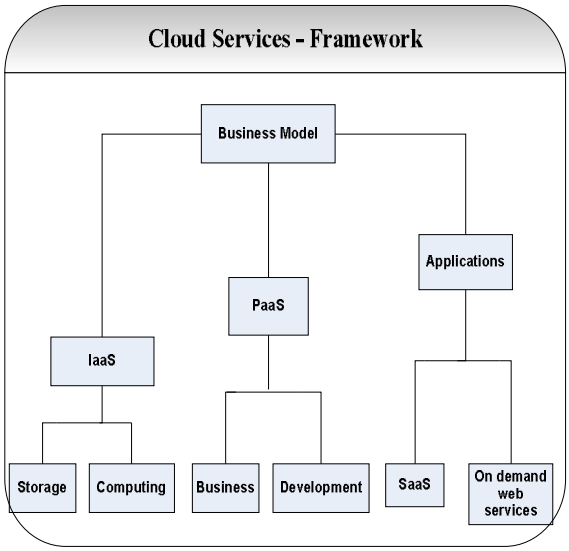

Figure 2 Cloud Services - Framework

3.1 Cloud - Advantages

The use of cloud computing has significant advantages such as

- Cost reduction
- Great storage capacity
- Scalability
- Needless software installation and maintenance
- Accessibility of on-demand services or applications from anywhere
- Elasticity and
- Pay-as-you-go model and energy saving

1.IDMCC – Integrated Data Mining and Cloud Computing

The integrated approach of data mining and cloud computing and mining is the process of extracting structured information from unstructured or semi-structured web data sources. It also facilates analyzing and extracting the useful information from various fields such as finance, banking, healthcare, genetics, marketing etc. The application of this technology should enable that with just a few clicks one can collect the information about the end user of the application entirely. Cloud should enable everyone to use this potential providing everything in the form of service. Data mining in cloud computing allows organization to centralize the management of software and data storage with assurance of efficient, reliable and secure services for their users. It provides technology that can handle large amount of data which cannot be processed efficiently at reasonable cost using standard technologies and techniques. It also allows the users to retrieve meaningful information from virtually integrated data warehouse that reduces the cost of infrastructure and storage. [3] Through cloud computing, mass data storage and distribution of computing, massive data mining environment for cloud computing provides new ways and means to effectively solve the distributed storage of massive data mining and efficient computing. Extension of cloud computing will drive the internet and technological achievements in the public service is to promote the depth of information resources sharing and sustainable use of new methods and new ways of traditional data mining. [6] The data mining in cloud computing allows organizations to centralize the management of software and data storage with assurance of efficient, reliable and secure services for their users. As cloud computing refers to software and hardware delivered as services over the internet, in cloud computing data mining software is also provided in this way. The main effects of data mining tools being delivered by the cloud are,

- The customer only pays for the data mining tools that he requires
- The customer doesn't have to maintain a hardware infrastructure

Data mining in cloud computing is the process of extracting structured information from unstructured or semi structured web data sources. The data mining in cloud allows organizations to centralize the management of software and data storage with assurance of efficient, reliable and secure services for their users. The following figures describes the overlap between cloud computing features and the data mining functionalities in the integrated data mining and cloud computing environment.

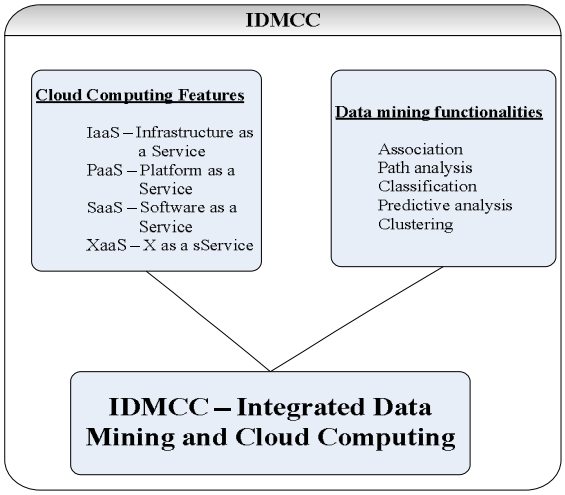

Figure 3 IDMCC Integration

1.1 Advantages of IDMCC

The following are the advantages of the integrated data mining and cloud computing environment. [8]

- Virtual computers that can be started with short notice
- Redundant robust storage
- No query structured data
- Message queue for communication
- The customer only pays for the data mining tools that he needs
- The customer doesn't have to maintain a hardware infrastructure as he can apply data mining through a browser

2. IDMCC in Healthcare – A Case study

Handling and analyzing large volume of data has the opportunity to transform the healthcare industry by giving doctors access to more information about both individual patients and helpful general population health trends. However, it is important for all of that information to be safe and secure, leading to a few more roadblocks than are encountered in the finances and the sciences. There are lots of benefits when the healthcare information is moved in the cloud. When the patient is placed on the ambulance, on the immediate benefits of cloud hosting medical information the data starts being uploaded. The information about the ECG, the blood pressure, the heart rate, the medication, all of that information is being uploaded. All the providers are able to access that cloud information to evaluate and analyze it. The performance and cost savings of operating in the cloud are very high. This network involves using big data analytics, cloud services, and advances in communications including smart phones and social media tools which is used to connect all the stakeholders involved in delivering care. [4] The following table describes that the various challenges and issues that are faced in the healthcare domain and how they can be solved using the proposed integrated data mining and cloud computing and mining approach.

Challenges and Issues in Healthcare	IDMCC Solution
Reporting and Mining Clinical Data to Improve Patient Outcomes	• On-demand scaling allows infrastructure on demand for analytics without capital expenses or delays. • The cloud solution also allows realizing operating improvement while managing cost with the ability to scale down when necessary.
Infrastructure for Patient Care and Claims Data	• A multisourced infrastructure allows a shared pool of computing and storage resources to be available to participating hospitals, practices, clinics and labs on a pay-as-you-go basis. • This solution enhances revenue and captures operational improvements while reducing cost.
Data System Incompatibility	• A multisourced services solution enables providers to bring together different types of data without a large investment and provides the ability to share information and to implement analytical tools.
Storage and Management of High Volume of Images	• Infrastructure and Platform as a Service (IaaS/PaaS) enables cloud-based storage and image sharing • Cloud computing manages and facilitates efficient and secure sharing with radiology specialists and affiliated practices or hospitals. • This solution reduces the need for in-house capacity and related costs and also improves the operational efficiency.

Table 1 IDMCC Solution for the challenges and issues in healthcare

5.1 Applications of IDMCC in Healthcare

Some of the application areas of the integrated data mining and cloud computing is as follows:

- Hospital-based electronic health records (EHRs)
- Community-based health information sharing
- Personal Health Records (PHRs)
- Patient accounting, financial and billing systems
- Enterprise Resource Planning (ERP) systems
- Clinical ancillary systems such as Laboratory Information Management Systems (LIMS) and Electronic Prescribing (E-prescribing)
- Cyclical and seasonal mission requirements
- Statistical and analytical functions requiring large-scale scientific and technical computing
- Episodic requirements which can benefit from rapid, on-demand cloud provisioning for emergency management, outbreak management, and food poisoning

CONCLUSION

This paper presents a review of need data mining services in cloud computing along with a case study on the integrated approach of data mining and cloud computing and mining. The data mining in cloud allows organization to centralize the management of software and data storage with assurance of efficient, reliable and secure services for their users. The implementation of data mining techniques through cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage. This approach also reduces the barriers that keep small companies from benefiting of the data mining instruments. The emergence of cloud computing brings new ideas for data mining. It increases the scale of processing data.

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] Larry H Bernstein, "Cloud computing in Healthcare organizations", A perspective from Science Applications International Corporation (SAIC)

[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] Rahul Sharma, Rohit Sharma, "Excavate the Cloud via Autonomous agents & Data mining", International Journal of Computer, Electronics & Electrical Engineering, Volume 3 – Issue 1

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

[8] Ruxandra-Ştefania PETRE, "Data mining in Cloud Computing", Database Systems Journal vol. III, no. 3/2012

[9] Srinivas. A Et Al, "A Study On Cloud Computing Data Mining", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 5, July 2013

[10] Tingting Hu Et Al, "A Survey of Mass Data Mining Based on Cloud computing", International Conference on Anti counterfeiting, Security and Identification in Communication, 2009

[11] http://skyfollow.com/cloud-services-framework-diagram

★★★