Survey of Data Mining in Cloud Computing Environment

Prof. Y.S. Chouhan (Director Admin), Asst. Prof. Makhan Kumbhkar

Abstract: Data mining is very important process for finding new, valid, useful and understandable form of data. data mining methods in cloud computing provides scalable and more flexible model that can be used for efficient mining of large amount of data from virtually integrated data sources with the goal of producing useful information which are providing support for decision making. This paper basically provides overview of data mining in cloud computing for the purpose of efficient and secure service for their users and with the help of cloud computing also reduce the cost of infrastructure and storage.

Keywords: Cloud computing, Data Mining, Knowledge Discovery Database (KDD).

1. INTRODUCTION

in our day to day life internet is very important tool for life and other activities as the amount of data is created by the users using online services is very large. there is huge amount of hidden information in this data that might be helpful for decision making.to discovering useful information cloud infrastructure is used in integration with data mining methods. Cloud Computing aims at transforming old approach of Computing by new services of both hardware and Software resources and software applications. these services generally delivered with the help of internet. its gives benefits to user due to its low cost, mobility and large availability. its also provide unlimited storage as well as computing power which leads to mine large amount of data. Data mining methods are used for discovering knowledge large dataset. it is used to analyse data from dataset or different sources and get useful information from data. data mining is also used for predicting trends or values, classification of data etc. it is necessary in areas of business, science, advertising, marketing etc. An integrated approach of data mining and cloud computing is used to obtain fast access to technology and also provides knowledge discovery system that build of large numbers of distributed data analysis services.

2. DATA MINING CONCEPT

Data Mining, by its simplest definition, automates the detection of relevant patterns in a dataset. it uses machine learning, statistical and Visualization techniques to discover and present knowledge which is easily understandable to users or humans. mining is the process which we explore and analysis of large quantities of data in order to discover useful pattern and rules. Now a days without automation that is impossible to mine large volumes of data. in huge amount of dataset, the data mining is best solution to discover hidden pattern which is helpful for government to take decisions so as to get more benefit. KDD is also Data mining.

2.1. Knowledge discovery process (KDD):

KDD is the automatic extraction of non-obvious, hidden knowledge from large volumes of data.

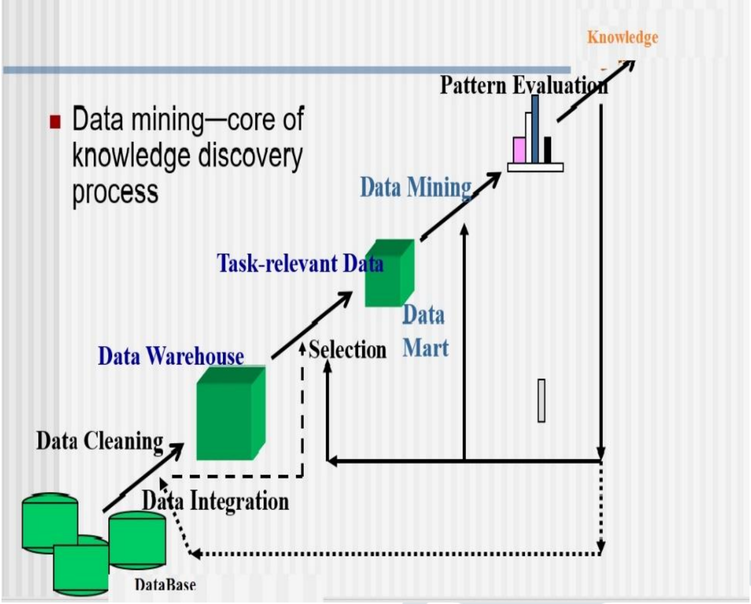

Fig (1) KDD Process

The various steps in the KDD [1] process is explained above and shown in Figure 1.

Data Integration-The data in integrated from a combination of multiple sources of data.

Data Selection and cleaning-The data relevant for analysis is retrieved from the database and noise and inconsistent data is removed.

Data Transformation-This step involves consolidation and transformation of data into forms appropriate for mining e.g., by performing aggregation of summary of data.

Data Mining- This is the most important step and it is done by use of intelligent patterns from data.

Pattern Evaluation-Evaluation includes identification of patterns that is interesting

2.2. Components of Data Mining:

a) Data Sources

The actual sources of data are Database, data warehouse, World Wide Web (WWW), text files and other documents. so, we need huge volumes of historical data for data mining to be successful. Organizations usually store data in database or data warehouses. Data warehouse

Data warehouses may contain one or more databases, text files, spreadsheets or other kinds of information repositories. Sometimes, data may reside even in plain text files or spreadsheets. World Wide Web or the Internet is another big source of data.

b) Database or Data Warehouse Server

The database or data warehouse server contains the actual data that is ready to be processed. Hence, the server is responsible for retrieving the relevant data based on the data mining request of the user.

c) Data Mining Engine

The data mining engine is the core component of any data mining system. It consists of a number of modules for performing data mining tasks including association, classification, characterization, clustering, prediction, time-series analysis etc.

d) Pattern Evaluation Modules

The pattern evaluation module is mainly responsible for the measure of interestingness of the pattern by using a threshold value. It interacts with the data mining engine to focus the search towards interesting patterns.

e) Graphical User Interface

The graphical user interface module communicates between the user and the data mining system. This module helps the user use the system easily and efficiently without knowing the real complexity behind the process. When the user specifies a query or a task, this module interacts with the data mining system and displays the result in an easily understandable manner.

f) Knowledge Base

The knowledge base is helpful in the whole data mining process. It might be useful for guiding the search or evaluating the interestingness of the result patterns. The knowledge base might even contain user beliefs and data from user experiences that can be useful in the process of data mining. The data mining engine might get inputs from the knowledge base to make the result more accurate and reliable. The pattern evaluation module interacts with the knowledge base on a regular basis to get inputs and also to update it.

2.3. Data-Mining Methods:

a). Decision tree induction: A decision tree is a structure that includes a root node branches and leaf node. Each internal node represents a test on attribute, each branch denotes the outcome of test each leaf node holds a class label. The top most node in tree is root node. The main goal is to predict the output of continuous attribute but data mining is less appropriate for estimating tasks.

b). Rule: It is represented by set of IF-THEN riles. First of all, how these rules are examined and next is how these rules are built and can be generated from data. Expression for rule is IF condition THEN conclusion.

The goals of prediction and description can be achieved through various data-mining methods which are following:

Regression is learning a function that maps a data item to a real-valued prediction variable.

Classification is learning a function that classifies or maps a data item into one of several predefined classes.

Clustering is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data.

2.4. Applications of Data Mining:

Major application areas for data mining are as follows:
Fraud detection: This is used for monitoring credit card fraud, watching over millions of accounts. It is used to identify financial transactions that might indicate money laundering activity.

Investment: Numerous companies use data mining for investment, but most do not describe their systems. One

exception is LBS Capital Management. Its system uses expert systems, neural nets, and genetic algorithms to manage portfolios.

Marketing: In marketing, the primary application is database marketing systems, which analyse customer databases to identify different customer groups and forecast their behaviour.

Telecommunications: The telecommunications alarm sequence analyzer (TASA) offers pruning, grouping, and ordering tools to refine the results of a basic brute-force search for rules. Large sets of discovered rules can be explored with flexible information-retrieval tools supporting interactivity and iteration.

3. CLOUD COMPUTING CONCEPT

Cloud computing is current buzzword in the market. It is standard in which the assets can be leveraged on per use basis thus reducing the cost and complexity of service providers. Cloud computing promises to cut operational and capital costs and more importantly let IT departments focus on strategic projects instead of keeping data centres running. It is much more than simple internet. It is a construct that allows user to access applications that actually reside at location other than user’s own computer or other Internet-connected devices. There are numerous benefits of this construct. For instance, other company hosts user application. This implies that they handle cost of servers, they manage software updates and depending on the contract user pays less i.e. for the service only. Confidentiality, Integrity, Availability, Authenticity, and Privacy are essential concerns for both Cloud providers and consumers as well. Software as a Service (SaaS) serves as the foundation layer for the other delivery models, and a lack of security in this layer will certainly affect the other delivery models, i.e., PaaS, and SaaS that are built upon IaaS layer. This paper presents an elaborated study of SaaS components’ security and determines vulnerabilities and countermeasures. Service Level Agreement should be Considered very much importance.

The characteristics of cloud computing are:
1. On-demand self-service
2. Resource pooling
3. Broad network access
4. Pay as per use service
5.Rapid elasticity and flexibility

3.1. Basic Cloud models:

1.IaaS (Infrastructure as a Service) delivers computer infrastructure, typically a platform virtualization environment, as a service. Rather than purchasing servers, software, data centre space or network equipment, clients instead buy those resources as a fully outsourced service.

2.PaaS (Platform as a Service) deliver a computing platform where the developers can develop their own applications.

3. SaaS (Software as a service) is a model of software deployment where the software applications are provided to the customers as a service.

3.2. Advantages of cloud computing:

• **Lower computer costs:** There is no need of a high-powered and high-priced computer to run cloud computing web-based applications.

- **Improved performance:** Computers in a cloud computing system boot and run faster because they have fewer
programs and processes loaded into memory.
- **Reduced software costs**: Instead of purchasing expensive software applications, you can get most of what you need for free.
- **Instant software updates:** Another advantage to cloud computing is that you are no longer faced with choosing between obsolete software and high upgrade costs. When the application is web-based, updates happen automatically
- **Unlimited storage capacity:** Cloud computing offers virtually limitless storage.

3.3. Disadvantages of cloud computing

3.3.1. It requires a constant internet connection.
3.3.2. It does not work well with low-speed connections.
3.3.3. Stored data might not be secure in cloud computing

4. CONCLUSION

Today's Business data mining through Cloud Computing is an absolutely necessary for today's business to make decision. Data mining is also helpful for predicting behaviours. This paper also provides an overview of necessity and utility of data mining in cloud Computing environeent.as we know that data mining tools is increasing day by day every day. Data mining is used to discovering interesting patterns from large amounts of data. Mining can be performed in a variety of information repositories. Data Mining using cloud computing has become an area because it now covers almost all business and scientific computing. 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 user. The cloud provider provides a tool for data mining for better service.

5. REFRENCES

[1] Chetna Kaushal et al," Integration of Data Mining in Cloud Computing", Advances in Computer Science and Information Technology (ACSIT), ISSN: 2393-9915; Volume 2, Number 7; April – June, 2015 pp 48 – 52.

[2] Y. S. Chouhan et al," Security Analysing in UNIX for Cloud Computing Environment**",** International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 10, October 2015.

[3] SADHANA RANA. "RISK ANALYSIS IN WEB APPLICATIONS BY USING CLOUD COMPUTING", International Journal of Multidisciplinary Research Vol.2 Issue 1, January 2012, ISSN 2231 5780

[4] Y. S. Chouhan et al," Recent Methodologies for Improving and Evaluating Academic Performance", International Journal of Scientific Research in Computer Science and Engineering, Vol-3(2), PP (11-16) Apr 2015, E-ISSN: 2320-7639**.**

[5] Makhan Kumbhkar et al," Analysis of Cloud Computing in Higher Education", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 6, June 2015.

[6] M. Jensen, J. Schwenk, N. Gruschka, and L. Lo Iacono, "Technical Security Issues in Cloud Computing". IEEE, 2009.

[7] Makhan Kumbhkar at el, "Performance Improvement of Software as a Service and Platform as a Service in Cloud Computing Solution" Vol-1, Issue-6 ISSN: 2320-7639.