Advances in Computer Science and Information Technology (ACSIT)
Print ISSN : 2393-9907; Online ISSN : 2393-9915; Volume 2, Number 7; April – June, 2015 pp 48 – 52
Krishi Sanskriti Publications
http://www.krishisanskriti.org/ACSIT.html

Integration of Data Mining in Cloud Computing

Chetna Kaushal[1], Aashima Arya[2], Shikha Pathania[3]

[1]Student, DAV University
Punjab, India
[1]+91 9501508898 chetnakaushal3558@gmail.com
[2]Student, DAV University
Punjab, India
[2]+91 9855713601, aashima.arya@yahoo.in
[3]Student, DAV University
Punjab, India
[3]+91 9915500293, shikhapathania02@gmail.com

Abstract : *Data mining is considered as an important process as it is used for finding new, valid, useful and understandable forms of data. The integration of data mining methods in cloud computing provides a flexible and scalable architecture that can be used for efficient mining of huge amount of data from virtually integrated data sources with the goal of producing useful information which is helpful in decision making. This paper provides an overview of the need of integration of data mining in cloud computing to provide efficient and secure services for their users and to reduce the cost of infrastructure and storage.*

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

1. INTRODUCTION

In recent years Internet has become an important tool in our day to day life and activities as the amount of data created by the users using online services is very large. There is hidden information in this data that can be used for making effective decisions. Cloud infrastructure is used in integration with data mining methods to significantly discover useful knowledge.

Cloud computing aims at transforming the traditional approach of computing by providing service of both hardware resources and software applications. These services are delivered over the internet. It gains popularity due to its low cost, mobility and huge availability. It provides unlimited storage and computing power which leads to mine large amount of data.

Data mining methods are used for discovering knowledge in databases. It is used to analyze data from multiple sources and get useful information from data. . Data mining is also used for predicting trends or values, classification of data, categorization of data, and to find correlations, patterns from the dataset. It is necessary in areas of business, science, advertising, marketing, medicine etc.

An integrated approach of data mining and cloud computing is used to obtain fast access to technology and provides a sort of knowledge discovery system that is built of large numbers of decentralized data analysis services.

2. DATA MINING CONCEPT

Data Mining is defined as non-trivial extraction of implicit, previously unknown, potentially useful information from data. It uses statistical, visualization and machine learning techniques to discover and present knowledge in a form which is easily understandable to humans. Data Mining is the process of exploration and analysis of large quantities of data in order to discover meaningful patterns and rules by automatic or semi automatic means. Without automation it is impossible to mine large volumes of data. In large databases, data mining solves the problem to discover the hidden but useful knowledge from data, which can help in the government and enterprises to make decisions so as to get more benefit from it. Data Mining is also known as Knowledge Discovery Databases-KDD.

2.1. Knowledge discovery process (KDD)

The various steps in the KDD [1] process are explained below 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.

- Knowledge Presentation- To present the extracted or mined knowledge to the end user various visualization and knowledge representation techniques is used.

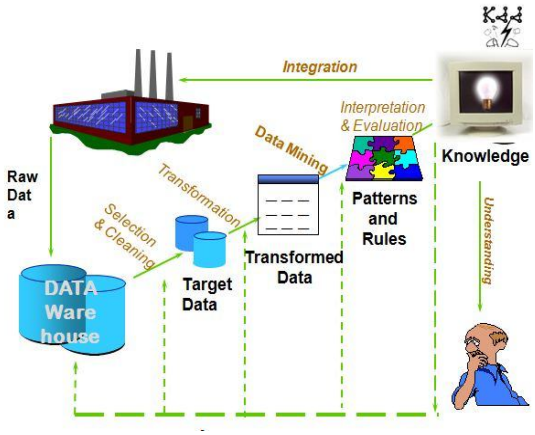

Figure 1. Steps of KDD or data mining process

2.2. Components of Data Mining

- **Databases, data warehouses or other repository information-** A set of databases such as data warehouses, spreadsheets and other kinds of information repositories where data cleaning and integration techniques may be employed.
- **Databases or data warehouses server-** This component fetches data based on user's request from a data warehouse.
- **Knowledge Base-** The domain knowledge is employed for finding interesting and useful patterns.
- **Data Mining Engines-** The functional modules that are used to perform tasks such as classification, association, clusters analysis etc.
- **Pattern Evolution Module-** Interestingness measures are used to focus search towards interesting patterns.
- **Graphical User Interface-** This component or module allows users to interact with the system by specifying a data mining task or a query through a graphical interface. It is an interface between the end user and the data mining system.

2.3. Data-Mining Methods

The two primary goals of data mining tend to be prediction and description. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest, and description focuses on finding human-interpretable patterns describing the data. The goals of prediction and description can be achieved through various data-mining methods [2] described here.

- *Regression* is learning a function that maps a data item to a real-valued prediction variable. Eg: Estimating the probability that a patient will survive given results of a set of diagnostic tests, predicting consumer demand for a new product as a function of advertising expenditure.
- *Classification* is learning a function that classifies or maps a data item into one of several predefined classes. E.g.: Automated identification of objects of interest in large image databases and classifying of trends in financial market.
- *Clustering* is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data. The categories can be mutually exclusive and exhaustive or consist of a richer representation, such as hierarchical or overlapping categories. Eg: Discovering homogeneous subpopulations for consumers in marketing databases.
- *Change and deviation detection* focuses on discovering the most significant changes in the data from previously measured values.
- *Dependency modeling* consists of finding a model that describes significant dependencies between variables. Dependency models exist at two levels: (1) the *structural level* of the model specifies (often in graphic form) which variables are locally dependent on each other and (2) the *quantitative level* of the model specifies the strengths of the dependencies using some numeric scale.

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 analyze customer databases to identify different customer groups and forecast their behavior.

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 [3] is a general term used to describe a new class of network based computing that takes place over the Internet. It is a new concept that defines the use of computing

as a utility, that has recently attracted significant attention. National Institute of Standards and Technology (NIST) [4] defines Cloud Computing as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Cloud computing is a computing paradigm shift where computing is moved away from personal computers or an individual application server to a "cloud" of computers. Users of the cloud only need to be concerned with the computing service being asked for, as the underlying details of how it is achieved are hidden. This method of distributed computing is done through pooling all computer resources together and being managed by software rather than a human.

The computing paradigm shift [5] on the last half century through six distinct stages is:

Stage 1: people used terminals to connect to powerful mainframes shared by many users.

Stage 2: stand-alone personal computers became powerful enough to satisfy users' daily work.

Stage 3: computer networks allowed multiple computers to connect to each other.

Stage 4: local networks could connect to other local networks to establish a more global network.

Stage 5: the electronic grid facilitated shared computing power and storage resources.

Stage 6: Cloud Computing allows the exploitation of all available resources on the Internet in a scalable and simple way.

The characteristics of cloud computing are:

- On-demand self-service
- Resource pooling
- Broad network access
- Pay as per use service
- Rapid elasticity and flexibility

3.1. Basic Cloud models

The basic models [6] of providing cloud computing services are shown in Figure 2.

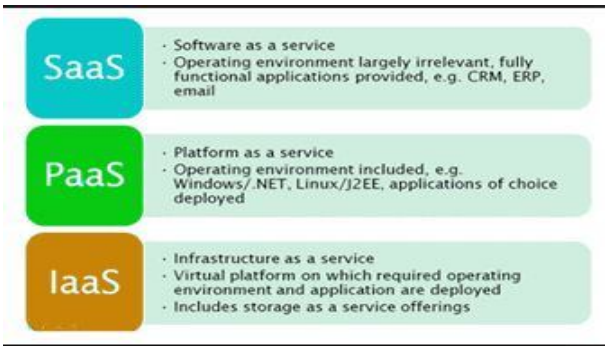

Figure 2. Basic Cloud Service models

- IaaS (Infrastructure as a Service) delivers computer infrastructure, typically a platform virtualization environment, as a service. Rather than purchasing servers, software, data center space or network equipment, clients instead buy those resources as a fully outsourced service.
- PaaS (Platform as a Service) deliver a computing platform where the developers can develop their own applications.
- 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. Cloud Deployment models

The deployment models [7] of cloud computing are:

1) Private cloud

Private cloud is cloud infrastructure operated solely for a single organization, whether managed internally or by a third-party, and hosted either internally or externally.

2) Public cloud

A cloud whose services are rendered over a network that is open for public use is called public cloud. Public cloud services may be free. Technically there may be little or no difference between public and private cloud architecture, however, security consideration may be substantially different for services (applications, storage, and other resources) that are made available by a service provider for a public audience and when communication is effected over a non-trusted network.

3) Community cloud

It provides the ability for more organizations to share the same cloud computing structure. Infrastructure supports special communities that have common interests, needs and security requirements.

4) Hybrid cloud

Hybrid cloud is a composition of two or more clouds (private, community or public) that remain distinct entities but are bound together, offering the benefits of multiple deployment models. Hybrid cloud can also mean the ability to connect

collocation, managed and/or dedicated services with cloud resources.

3.3. 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.
- Increased data reliability: Unlike desktop computing, in which if a hard disk crashes and destroys all valuable data, a computer crashing in the cloud does not affect the storage of data.

3.4. Disadvantages of cloud computing

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

4. INTEGRATING DATA MINING IN CLOUD COMPUTING

Data mining methods and application are very essential in cloud computing field. The process of extracting structured information from unstructured or semi-structured web data sources is called data mining. The integration of data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of reliable, secure and efficient services for their users. It is explored that how the data mining tools like SaaS, PaaS and IaaS are used in cloud computing to extract the information. Data mining in cloud is used for analyzing and extracting the useful information in many areas of human activities like banking, medical, marketing etc. With this application one can find the desired information about customer's behavior, their habits, interests and location with just a few clicks of mouse. Cloud provides a benefit for small sized companies to have an opportunity to rent a cloud service for efficient analysis of all the data in the organization which was earlier reserved only for big companies.

Data Mining is preferably used for a large amount of data and related algorithms often require large data sets to create quality models. Cloud providers use data mining to provide clients better service. The use of data mining methods in cloud computing allows the users to extract useful information from virtually integrated data sources that reduces the infrastructure and storage costs.

Cloud Computing signifies the new trend in Internet services that is based on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources. 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 following are the advantages [8] of the integrated data mining and cloud computing environment.

- 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.
- Redundant robust storage.
- Virtual computers that can be started with short notice.
- No query structured data.
- Message queue for communication.

5. CONCLUSION

Data mining integrated in cloud computing is very important characteristic in business to make effective decisions to predict the future trends and behavior. Computing is the serving side, and Data Mining is the side being served. It's not that Data Mining can't be done without Cloud Computing or Cloud Computing only application is Data Mining. They are like cake and icing that are too good and efficient together. Cloud computing relies on clusters of servers that may be located remotely to handle tasks. Data mining is the process of extracting structured information from unstructured or semi-structured web data sources. The data mining in Cloud Computing [9] allows companies to centralize the management of software and data storage, with assurance of cost effective, reliable, secure and efficient services for their users.

6. REFERENCES

[1] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." *AI magazine* 17.3 (1996): 37.

[2] Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann, San Francisco (2006).

[3] Special Publications 800-145 "National Institute of Standard and Technology (NIST)"

[4] http://en.wikipedia.org/wiki/Cloud_computing

[5] Petre, Ruxandra Stefania. "Data mining in cloud computing." *Database Systems Journal* 3.3 (2012): 67-71.

[6] Bhagyashree Ambulkar and Vaishali Borkar, "Data Mining in Cloud Computing", MPGI National Multi Conference 2012 (MPGINMC-2012), 7-8 April 2012.

[7] Dillon, Tharam, Chen Wu, and Elizabeth Chang. "Cloud computing: issues and challenges." *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on*. Ieee, 2010.

[8] B. Kamala ,: "A Study On Integrated Approach Of Data Mining And Cloud Mining", *International Journal of Advances in Computer Science and Cloud Computing (IJACSCC)*, Volume-1,Issue-2,pp 35-38 ,2013.

[9] Nikam, V. B., and Viki Patil. "Study of Data Mining algorithm in cloud computing using MapReduce Framework." *Journal of Engineering Computers & Applied Sciences* 2.7 (2013): 65-70.