International Journal of Advanced Science and Research
ISSN: 2455-4227
Impact Factor: RJIF 5.12
www.allsciencejournal.com
Volume 2; Issue 6; November 2017; Page No. 99-102

A survey of data mining in cloud computing

Chanda Monga, Minakshi Sharma, Etti Sharma
Department of Computer Science, Dev Samaj College for Women, Ferozepur, Punjab, India

Abstract
Data Mining is the development of raw data in to potentially useful information. So It is considered an important process for finding new, valid, useful and understandable forms of data. Cloud Computing gives the new development in internet services that rely on clouds of servers for handling the tasks. Cloud computing is a inventive technology that can support a large range of applications. Cloud computing is basically providing software, platform and infrastructure as a service over the internet and every customer can access these software, platforms and infrastructure from a remote area. With the rapid development of internet and raise in commerce allows the companies to protect, store, retrieve and analysis of their important data through cloud services with data mining. When the concept of Data mining is used in cloud computing then extraction of structured information from unstructured or semi-structured web data sources is done. Here in this paper trying to work together with these two major domains Cloud and Data Mining. Some of the key features are used for the distributing the data in certain things for the user understandable language. As the cloud storage implemented in different servers for the security reasons data mining concept is used for the efficiency of the each part of the data is in a secure state.

Keywords: cloud computing, data mining, data mining in cloud computing, databases and security

1. Introduction
Data mining is very effective tool for analyzing the data from different views and getting useful information from the raw data, Classification of data, and to find correlation of data patterns from the dataset. On the contrary challenges to be faced like data storage and transfer approaches need to deal with excessive amount of data. The Management of data resource and dataflow is becoming the main bottleneck. Huge data set has become a major challenge. The internet is becoming an essential tool in everybody's life, both professional and personal, as its user are becoming more in number. The most innovative concept of recent year is Cloud Computing. Most of the companies are choosing as an option to build their own IT infrastructure for hosting database or software, instead of having a third party to host them on its large servers, so company's would have access to their data and software over the Internet. The use of cloud computing is gaining popularity due to its mobility, huge availability and low cost. On the other hand it brings more threats to the security of the company's data and information. In recent years, data mining techniques have evolved and become more used, discovering knowledge in database becoming increasingly vital in various fields: business, medicine, science and engineering, spatial data etc. We live in a period of enormous information that has inserted a colossal potential and expanded complexity and risks such as insecurity as well as information overload and irrelevance. Likewise business knowledge and analytics are essential in managing the extent and effect of information driven issues and solutions in the con-temporary society and economy. Investigators, PC researchers, economists, mathematicians, political researchers, sociologists, and different researchers are looking for access to the gigantic amounts of information to extract meaningful information and knowledge.

Fig 1: Cloud Computing Logical Diagram

Worldwide, the measure of crude information is developing exponentially, due partially to the blast of joined mechanisms, Internet administrations, online networking, Polaroid's, sensors, and client created substance. Besides, up to 90 percent of corporate information, incorporating archives, website pages, and email, is unstructured. The sheer volume and unconventionality of data is overwhelming normal database customizing, and this condition is calling for an alternate methodology. Information mining is an alternate enhancement to help ventures to keep tabs on data in their information warehouses. It is the extraction of concealed prescient data from gigantic databases. Allowing associations to make proactive and learning driven choices. The information dissection has been increased with roundabout, programmed information handling. A huge destination of information mining is to find at one time obscure relationships around the information, especially when the information starts from dissimilar databases.

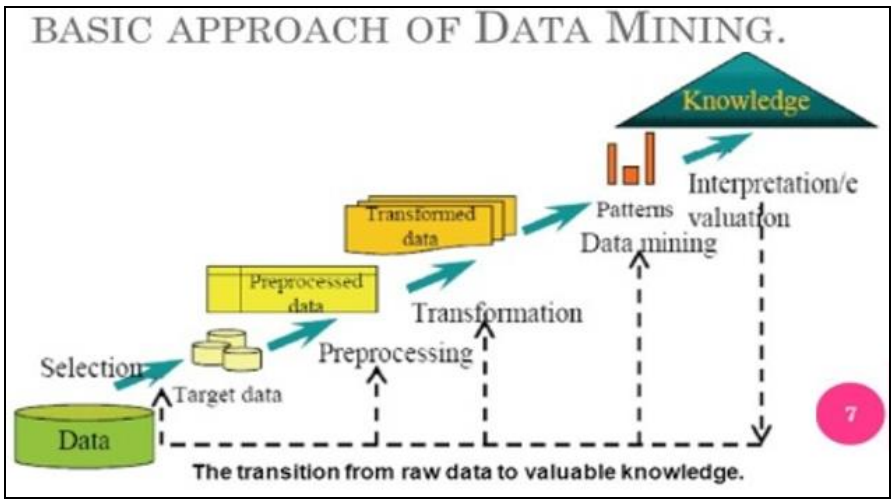

Fig 2: Transition from raw data to valuable data

2. Components of Data Mining

1. Association

Association rule mining is very important segment of data mining. In association, a pattern is discovered. This pattern is based on the relationship between items in the same transaction of discovering regularities/ patterns in data. This is also known as Relation technique. It is may be the most critical model created and broadly examined by databases and information mining community.

2. Sequence Analysis

Sequential pattern Analysis seeks to discover or identify the patterns that are similar, regular events or trends in transaction data over a business period. For example in sales, with the help of historical transaction data, businesses can identify a set of items that customers buy together different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past. Running across examples where one occasion prompts an alternate ensuing occasion.

3. Classification

Classification is a data mining technique which is based on machine learning. Classification is used to classify each item in a set of data into one of a predefined set of classes or groups. The methods of classification make use of mathematical techniques such as decision trees, linear programming, neural network, and statistics. So in short in classification, we build up the software that can learn how to classify the data items into groups. For example, we can apply classification in the application that "given all records of students who left the college, predict who will probably leave the college in a future period." In this case, we divide the records of students into two groups that named "leave" and "stay". At last we can give instructions to our data mining software to classify the students into two separate groups named "leave" and "stay".

4. Clustering

Clustering is that data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The main task of clustering technique is to defines the classes after that puts objects in each class. To make the concept clearer, we can take management of books in the library as an example. In a library, there is a large variety of books on various topics available. The challenge is how to place those books in a way that readers can take many books on a particular topic without hassle. If we are using the clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and tag it with a meaningful name. If the readers want to take books in that topic, they would only have to go to that shelf instead of looking for the entire library.

5. Forecasting

The Forecasting, as its name implied, is one of a data mining techniques that discovers the relationship between independent variables. It also discover the relationship between dependent and independent variables. For example, the prediction analysis technique can be used in the sale to predict profit for the future if we consider the sale is an independent variable, profit could be a dependent variable. we can draw a fitted regression curve that is used for profit prediction that is based on the historical sale and profit data.

3. Cloud Services

There are three types of cloud services

- Infrastructure as a Service
- Platform as a Service
- Software as a Service.

A) Infrastructure as a Service (IaaS)

The consumers of cloud computing uses IT infrastructures (processing, storage, networks and other fundamental computing resources) which are provided in the IaaS cloud. The concept of Virtualization is used in IaaS cloud. This is done in order to integrate/decompose physical resources in an ad-hoc manner to meet growing or shrinking resource demand from cloud consumers. The basic plan of virtualization is to set up independent virtual machines (VM) these machines are isolated from both the underlying hardware and other VMs. An example of IaaS is Amazon's EC2. Convey PC foundation as an utility administration, regularly in a nature. I t is also known as utility computing. Provide enormous scalability.

B) PaaS

PaaS is a development platform that support the full "Software Lifecycle". This "software lifecycle" allows cloud consumers to develop cloud services and applications (e.g. SaaS) directly on the PaaS cloud. So, The difference between SaaS and PaaS is that SaaS host only the completed cloud applications whereas PaaS offers a development platform that hosts both Cloud Computing – Research Issues, Challenges, Architecture, Platforms and Applications. An example of PaaS is Google Approach to lease fittings, working frameworks, space etc. over the web to create provisions sits on a top of the IaaS construction modelling and joins with improvement and middleware proficiencies and database, informing and queuing capacities.

C) Saas

The users of cloud computing use their applications in a hosting environment. This can be accessed through networks from various clients. The networks can be Web browser, PDA, etc. by application users. The users of cloud do not have control over the cloud infrastructure. It often employs multi-tenancy system architecture, namely, different cloud consumers' applications are organized in a single logical environment in the SaaS cloud. The organization of cloud consumers applications is essential to achieve economies of scale and optimization in terms of speed, security, availability, disaster recovery and maintenance. Examples of SaaS include SalesForce.com, Google Mail, Google Docs, and so forth. The delivery of virtualized storage on demand becomes a separate Cloud service - data storage service.

4. Data mining in the cloud

Information mining systems and requisitions are sincerely needed in the distributed computing ideal model. As distributed computing is entering all the more in all degrees of business and experimental processing, it transforms into an incredible issue to be concerned by information mining. The information mining in Cloud Computing grants associations to make the incorporated administration of programming and information space, guaranteeing the effective, solid and secure administrations for their clients. The Microsoft suite of cloud-based administrations presents another specialized sneak peak of Data Mining in the Cloud as "DM Cloud".

DM Cloud permits you to perform some fundamental information mining assignments leveraging a cloud-based Analysis Services association. The data mining tasks include:

- Analyze Key Influencers
- Detect Categories
- Fill From Example
- Forecast
- Highlight Exceptions
- Scenario Analysis
- Prediction Calculator
- Shopping Basket Analysis

The information mining is utilized within different requisitions, for example, medicinal services, person Administration, math, science, in different site. Utilizing information mining through cloud registering decreases the jumps that keep little organizations from profiting of the information mining instruments. We investigate how the information mining instruments like Saas, Paas and Iaas are utilized within distributed computing to concentrate the data. Individuals utilize this characteristic to manufacture data posting and get data about distinctive themes via seeking in discussions and so forth. The organizations utilize this administration to see what sort of data is gliding on the planet wide for their items or administrations and take activities dependent upon the information displayed. The data recovery commonsense model through the multi-executor framework with information mining in a distributed computing environment has been proposed. It is prescribed that clients might as well guarantee that the solicitation made to the Iaas is inside the extent of combined information warehouse and is clear and straightforward. The work for the multi-executor framework gets to be less demanding through the provision of the information mining calculations to recover serious data from the information warehouse.

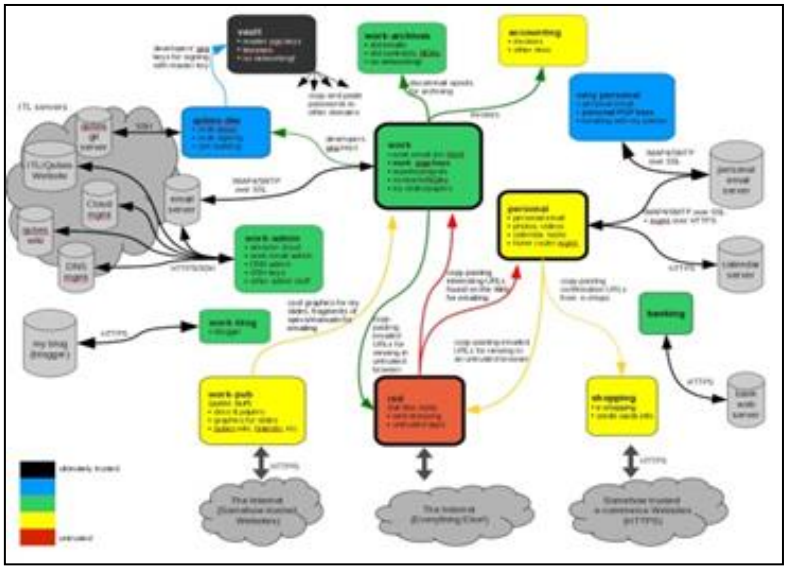

Fig: Job flow in the cloud for Data mining

5. Security for Cloud Computing

Cloud has certain security issues concerning affirmation and classification of information. A client who entrust a cloud supplier may lose access to his information incidentally or forever because of a doubtful occasion, for example, a malware ambush or system blackout. These impossible event can do notable harm to the clients of the cloud computing. Maintaining the Secrecy of client information in the cloud is a huge concern. There is a wide mixed bag of security issues identified with distributed computing however these issues have been classified into 2 general classifications: Security issues visaged by cloud suppliers and security issues visage by their clients. As a rule, the supplier may as well verify that their framework is secure inasmuch as the customer might as well determine that the supplier has taken the right efforts to establish safety to defend their information. Distributed computing could display diverse dangers to an organization than old IT results.

6. Conclusion

Research on classification of data in cloud has already been extensively done; so now it is important to use the result of these researches and analyze the security requirements which are important for keeping data secure. Relying on cloud computing millions of users store their data on a cloud which possess lot many cloud storage risks like unauthorized access, data loss etc. Privacy of data is a major concern in people who use public cloud services, so an approach is proposed to keep data safe and secure also keeping sure only authorized personnel can access data.

It is proposed to implement cloud security aspects for data mining by implementing cloud system. After implementing cloud infrastructure for data mining for cloud system, security measure for data mining in cloud will be evaluated. Threats will be fixed in data mining to Personal/private data in cloud System.

7. References

1. Dillon T, Wu C, Chang E. Cloud Computing: Issues and Challenges, 24th IEEE International Conference on Advanced Information Networking and Applications (AINA). 2010, 27-33, DOI= 20-23.
2. Zhou MQ, Zhang R, Xie W, Qian WN, Zhou A. Security and Privacy in Cloud Computing: A Survey Sixth International, 2010.
3. Agarwal D, Das S, Abbadi A. Big Data and Cloud Computing: Current State and Future Opportunities. ACM 2011; 978-1-4503-0528-0/11/0003.
4. Amazon Kinesis. (N.D.). Developer Resources.
5. Apache S4. (N.D.). Distributed Stream Computing Platform. od2011.pdf
6. Pallavi Roy. Mining Association Rules in Cloud, Research paper, 2012.
7. Kalyani Mali, Samayita Bhattacharya Fingerprint Database Handling Using Cloud Computing With Added Data Mining and Soft Computing Features. International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com Certified Journal. 2013; 3(2):610. ISSN 2250-2459, ISO 9001:2008.