An Efficient Data Mining Technique on Cloud Storage

[a]B.Chithra, [b]Ch.Srivalli, [c]S Archana
[a]Assist Professor (c), Dept of Mathematics, UCS, Osmania University, Hyd, India
[b]Sr.Asst Professor , Department of CSE, Aurora's Scientific Technological & Research Academy, TS, India
[c]Asso. Professor , Department of CSE, Aurora's Scientific Technological & Research Academy, TS, India

Abstract

Cloud Computing has become a main source for the data processing, storage and distribution. The storage of the data is simple and free to use. In data mining the data which is used as data security in a parallel computing platform. The some of the key features are used for the distribution of the data in certain things for the user understandable language. As we implemented the cloud storage 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. According to this concept we use data effectiveness from some so the supports of the Anazon EC2 map reduce platform. This paper explains how through cloud computing, the process of data mining is facilitated. Through data mining, information that is potentially useful can be retrieved from raw data. People often face the need for targeted advertising, whereby data mining techniques give businesses greater efficiency, hence helping to lower costs. In the sector of cloud computing, data mining has become of great importance. The facilitation of data mining through cloud computing will enable users to obtain useful information via virtual warehouse of integrated data, helping to lower expenses of storage, technical staff, and purchase of infrastructure.

KEYWORDS: cloud computing, data mining, data mining in cloud computing

----

I. INTRODUCTION

The importance of the internet in our personal as well as our professional lives cannot be overstated as can be observed from the immense increase of its users. It therefore comes as no surprise that a lot of businesses are being carried out over the internet. Cloud computing may be one of the greatest advancements in information technology over the recent past. Cloud computing entails the use of hardware and software computer resources delivered over the internet as a service.

Data mining has been an effective tool to analyze data from different angles and getting useful information from data. Classification of data, categorization of data, and to find correlation of data patterns from the dataset. On the other hand, challenges as data storage and transfer approaches need to deal with prohibitive amount of data. The management of data resource and dataflow is becoming the main bottleneck. Large data set has become a major challenge and data intensive computing is now considered as the "fourth paradigm" in scientific discovery after theoretical, experimental, and computation science.

The internet is becoming an increasingly vital tool in everybody's life, both professional and personal, as its user and becoming more numerous. The most revolutionary concept of recent year is Cloud Computing. Many companies are choosing as an alternative to building their own IT infrastructure
to host database or software, having a third party to host them on its large servers, so company's would have access to its 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

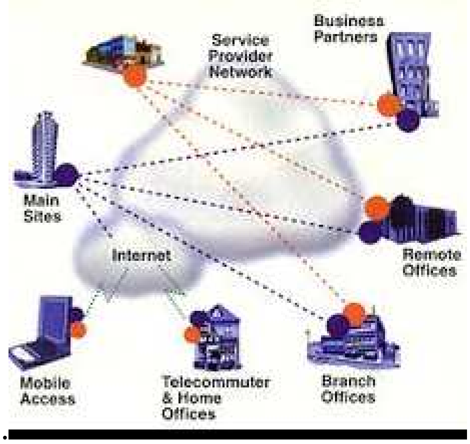


fig-1: Transferring data from one server to another server through the data mining.

II. KEY ASPECTS OF CLOUD COMPUTING SERVICES

Cloud computing allows companies to deliver services in new ways using technologies and techniques that would otherwise be unaffordable. According to Gartner Inc., based in Stamford, Conn., cloud-based services in messaging security controls are projected to account for 60% of revenue in 2013. As more cloud-based services emerge, we also see a significant increase in competition with new and existing vendors.

In 1960's-1990's John McCarthy who has opined the modern-day characteristics of cloud computing, later in 1990's telecommunication companies who previously offered primarily dedicated point to point data circuit, began offering virtual private network (VPN) services with comparable quality of service, but at a lower cost. By switching traffic as they saw fit to balance server use, they could use overall network bandwidth more effectively. They began to use the cloud symbols to donate the demarcation point between what the provider was responsible for cloud computing extends this boundary to cover server as well as the network infrastructure.

The major reason is that several types of security can be more efficiently implemented through the cloud. With cloud computing, emails can be filtered faster and viruses can be intercepted before they are sent out to thousands of customers. The second reason is that there are some security aspects a company is able to do in the cloud that it does not have the ability to do own its own.

As computer become more prevalent, scientists and technologists explored ways to make large-scale computing power available to more users through time sharing, experimenting with algorithms to provide the optimal use of the infrastructure, platform and application with prioritized access to CPU and efficiency for the end users.

III ESSENTIAL FEATURES OF CLOUD COMPUTING

Resource Pooling and Elasticity

In cloud computing, resources are pooled to serve a large number of customers. Cloud computing uses multi-tenancy where different resources are dynamically allocated and de-allocated according to demand. From the user's end, it is not possible to know where the resource actually resides.

Self-Service and On-demand Services

Cloud computing is based on self-service and on-demand service models. It should allow the user to interact with the cloud to perform tasks like building, deploying, managing, and scheduling. The user should be able to access computing capabilities as and when they are needed and without any interaction from the cloud-service provider. This would help users to be in control, bringing agility in their work, and to make better decisions on the current and future needs.

Quality of Service

Cloud computing must assure the best service level for users. Services outlined in the service-level agreements must include guarantees on round-the-clock availability, adequate resources, performance, and bandwidth. Any compromise on these guarantees could prove fatal for customers.

Peer to peer

It means the distributed architecture without the need of central coordination participants are both suppliers and consumers of resources in contrast other model of client server model.

IV CHARACTERISTICS OF DATA MINING

Agility: Agility improves with users' ability to rapidly and inexpensively re-provision technological infrastructure resources.
Cost: Cost is claimed to be greatly reduced and capital expenditure is converted to operational expenditure.
Multi-tenancy: enables sharing of resources and costs across a large pool of users.
Peak-load capacity: increases highest possible load-levels.
Utilization and efficiency: improvements for systems that are often only 10–20% utilized.
Reliability: improves through the use of multiple redundant sites, which makes cloud computing suitable for business

Virtualization: Virtualization Technology allows servers and storage device to be stored and utilization be increased applications can be easily migrated from one physical server to another. Centralization of infrastructure in location with lower cost such as electricity.

Security: Security could improve due to centralization of data, increased security-focused resources. But concern can persist about loss of control over certain sensitive data, and the lack of security for stored kernels. Security is often as good as or better than other traditional system, in part because providers are able to devote resources to solving security issues is greatly increased when data is distributed over a wider area or greater number of devices and in multi-tenant system that are bring shared by unrelated user. Private Cloud installations are in part motivated by users for desire and control over the infrastructure and avoid losing control of information security.

Scalability and elasticity

Scalability and elasticity via dynamic provisioning of resource on fine=grained, self- service basis near real-time, without user having to engineer for peak loads.

Maintenance

Maintenance of Cloud Computing application is easier because they do need to be installed on each user's computer and can be accessed from different places.

Infrastructure as a Service

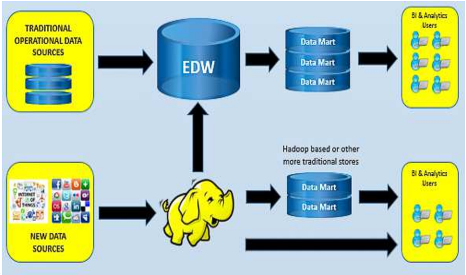
(IaaS) is a cloud model which allows organizations to outsource computing equipment and resources such as servers, storage, networking as well as services, such as load balancing and content delivery networks. The IaaS provider owns and maintains the equipment while the organization rents out the specific services it needs, usually on a "pay as you go" basis. Today, the question is less about whether or not to use IaaS services, but rather which IaaS providers to use.

Cloud Name	Key Feature
Amazon Web Services	Amazon Web Services offers a full range of compute and storage offerings, including on-demand instances and specialized services such as Amazon Elastic Map Reduce (EMR) and Cluster GPU instances, as well as Elastic Block Storage (EBS) and high performance SSDs on the storage side.
Windows Azur	Windows Azure is not a Windows-only IaaS. The compute and storage services

e	offered are typical of what you'll find in other IaaS providers, and administrators used to Microsoft platforms will find working with Windows Azure much easier.
Google Compute Engine	Google Compute Engine is well suited for big data, data warehousing, high performance computing and other analytics-focused applications. It is well integrated with other Google services, such as Google Cloud Storage, Google BigQeury and Google Cloud SQL.
Rackspace Open Cloud	Rackspace offers core cloud computing services with a strong focus on customer service. Rackspace is one of the co-founders of OpenStack, which it uses for its cloud infrastructure, so you can run the same platform in-house if you decide to move to a private or hybrid cloud down the road.

Hadoop Framework

Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Foundation Software



.

FIG-2: Data Flow In the Cloud HADOOP

Advantages of using Hadoop

Apache Hadoop is a core component of any modern data architecture, allowing organizations to collect, store, analyze and manipulate even the largest amount of data on their own terms – regardless of the source of that data, how old it is, where it is stored, or under what format.

Some Aspects Regarding Data mining

Data mining represents finding useful patterns or trends through large amounts of data. Data mining is defined as a "Type of Database Analysis that attempts to discover useful patterns or relationships in a group of data. The analysis uses advanced statistical methods, such as cluster analysis, and sometimes employs artificial intelligence or neural network techniques. A major goal of data mining is to discover previously.

Algorithms: Randomly choose k items and make them as initial centroids.

- For each point, find the nearest centroid and assign the point to the cluster associated with the nearest centroid .
- Update the centroid of each cluster based on the items in that cluster. Typically, the new centroid will be the average of all points in the cluster.
- Repeats steps 2 and 3, till no point switches clusters. Data mining parameters include: 1. Association - Looking for patterns where one event is connected to another event. 2. Sequence or path analysis - Looking for patterns where one event leads to another later event 3. Classification - Looking for new patterns 4. Clustering - Finding and visually documenting groups of facts not previously known
- Forecasting - Discovering patterns in data that can lead to reasonable predictions about the future, this area of data mining is known as predictive analytics.
  Some of the important for data mining technique:

- **Association** - Looking for patterns where one event is connected to another event.
- **Sequence or path analysis** - Looking for patterns where one event leads to another later event
- **Classification** - Looking for new patterns
- **Clustering** - Finding and visually documenting groups of facts not previously known
- **Forecasting** - Discovering patterns in data that can lead to reasonable predictions about the future, this area of data mining is known as predictive analytics.

Data mining is the extraction of hidden predictive information from large database, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouse. Data mining tools predict future trends and behavior, allowing businesses to make proactive, Knowledge- driven decisions. Businesses can make predictions about how well a product will sell or develop new advertising campaigns by using these new relationships reflected by the data mining algorithms. Data mining uses information from past data to analyse the outcome of a particular problem or situation that may arises. Data mining work to analyse data stored that data is bring analysed. That particular data may come from all parts of business from the production to the management. Managers also use data mining to compare and contrast among competitors. Data mining interrupt its data into time analysis that can be used to increase sales, promote new product, or delete that is not value added to the company.

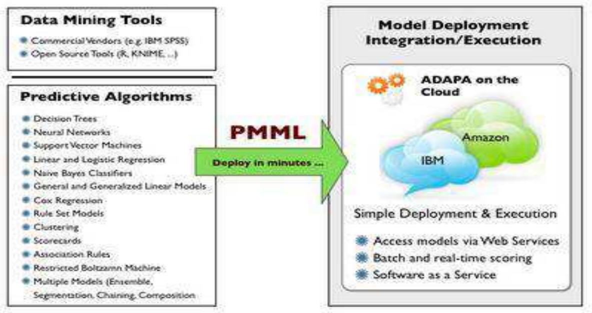

Data Mining Tools
Commercial Vendors (e.g. IBM SPSS)
Open Source Tools (R, KNIME, ...)
Predictive Algorithms
Decision Trees
Neural Networks
Support Vector Machines
Linear and Logistic Regression
Naive Bayes Classifiers
General and Generalized Linear Models
Cox Regression
Rule Set Models
Clustering
Scorecards
Association Rules
Restricted Boltzamn Machine
Multiple Models (Ensemble, Segmentation, Chaining, Composition
PMML
Deploy in minutes ...
Model Deployment Integration/Execution
ADAPA on the Cloud
Amazon
IBM
Simple Deployment & Execution
Access models via Web Services
Batch and real-time scoring
Software as a Service

Fig-3: Data Mining in Cloud Computing

Data mining in a cloud computing: Data mining is one of the fastest growing fields in computer industry that deals with discovering patterns from large data sets. It is a part of knowledge discovery process and is used to extract human understandable information Mining is preferably used for a large amount of data and related algorithms often require large data sets to create quality models .The relationship between data mining and cloud is worth to discuss. Cloud

providers use data mining to provide clients better service. If clients are unaware of the information being collected ethical issues like privacy and individuality are violated. This can be serious data privacy issue if the cloud providers misuse the information. Again attackers outside cloud providers having unauthorized access to the cloud, also have the opportunity to mine cloud data. In both cases, attackers can use cheap and raw computing power provided by cloud computing to mine data and thus acquire useful information from data. The data mining in cloud computing allows organization to centralize the management of software and data storage with assurance of efficient reliable and secure service for their user.

V. **CONCLUSION**

Cloud computing provides storage of data in a server by protecting data by using data mining concept. Actually we are discussing the cloud computing data mining for the advance use of security in data loss purpose. While the data we are storing in cloud is being separated in different servers for a security but the hackers using the cheap and raw cloud computing for the misuse of the software.K-Means algorithm is more efficient algorithm for mining large Databases and Cloud computing provides solution for storing large database with less cost. So, in this paper, we focused the implementation of K-Means algorithm in the Cloud environment and the experimental results shows that it works well in the Cloud.

REFERENCES

[1] "CloudComputingNewWine.pdf."-http://www.cmlab.csie.ntu.edu.tw/~jimmychad/CN2011/Reading/CloudComputingNewWine.pdf.

[2] "Top-Cloud-Computing-companies"-http:www.itstrategists.com/Top-Cloud-Computing-companies.aspx.

[3] "Cloud Computing: Data-Intensive Computing and Scheduling"-By Frédéric Magoulès, Jie Pan, Fei Teng http://books.google.co.in/books?id=06c4jn6kV7AC&printsec=frontcover&dq=cloud+computing+based+on+data+mining+reference+books&hl =en&sa=X&ei=AWXaUdW4GMXprAfDq4DACA&ved=0CDcQ6AEwAA#v=onepage&q&f=false

[4] "Data mining and Analytics"-http://cseweb.http://cseweb.ucsd.edu/~elkan/255/books.html

[5] "Data mining concept by Doug Alexander"-http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/

[6] "Concepts of cloud computing"-http://searchitchannel.techtarget.com/tip/Key-aspects-of-cloud-computing-services

[7] "Different characteristic for the cloud computing"-deca.cuc.edu.cn/.../The-Characteristics-of-Cloud-Computing.pdf

[8] "Grid Computing"-http://www.gridcafe.org/EN/what-is-the-grid.html

[9] "Mainframe Computer"-http://www.businessdictionary.com/definition/mainframe-computer.html

[10] "Utility Computing"-http://www.techopedia.com/definition/14622/utility-computing

[11] Peer to Peer"-http://www.bleepingcomputer.com/glossary/definition125.html

[12] "Cloud Gaming"-http://www.techopedia.com/definition/26527/cloud-gaming

[13] "Key concepts for the cloud computing"-http://new.itstrategists.com/Top-Cloud-Computing-Companies.aspx

[14] "Cloud Storage Infrastructure"-http://www.ibm.com/developerworks/cloud/library/cl-cloudstorage/

[15] Key concepts of the data mining by Lijia Guo"-http://www.casact.org/pubs/forum/03wforum/03wf001.pdf

[16] The Cloud Computing Handbook - Everything You Need to Know about Cloud Computing, By Todd Arias.

[17] http://searchsqlserver.techtarget.com/definition/d atamining.

[18] http://www.ijcaonline.org/volume15/number7/px c387 2623.pdf.

[19] http://www.waset.org/journals/waset/v39/v39- 72.pdf.

[20] http://www.estard.com/data_mining_marketing/data_mining_campaign.asp.
[21] http://dssresources.com/books/contents/berry97.h tml.
[22] http://www.marketingprofs.com/articles/2010/35 67/the-nine-most-common-data-miningtechniques-usedin-predictive-analy