OVERVIEW ON DATA MINING WITH CLOUD COMPUTING

[1]**Anupama Chowdhary**
Principal, Keen College,Bikaner (Rajasthan),India

Abstract – In this paper, certain aspects of data mining and cloud computing are discussed. Data mining is used for finding new, valid, useful and understandable forms of data. When data mining methods are assimilated with cloud computing it delivers a flexible and scalable architecture. This architecture can be used for efficient mining of huge amount of data on clouds. Cloud data is basically a virtually integrated data from different data sources or data warehouse. The goal of data mining is to producing useful information helpful in decision making, finding hidden data, fraud detection, identifying criminal suspects, prediction of potential terrorists and many more.

Index terms – data mining, cloud computing, clustering, predictive analytics, machine learning, decision management.

I. INTRODUCTION

In human history this age is known as information age. In past 30 years data has become critical to all aspects of human life. It has changed our education, entertainment, the way we experience people, business, and the wider world around us. It is like oxygen of our rapidly growing digital data. Data is created, captured, and replicated rapidly on our planet. In just the past 10 years society has witnessed the transition of analog to digital. What the next decade will bring using the power of data is virtually limitless.IDC (International Data Corporation) forecasts that by 2025 the global “datasphere” will grow to 163ZB (that is a trillion gigabytes). That's ten times the 16.1ZB of data generated in 2016 [1].

As data base have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. Thishas been aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees(1960s) and support vector machines (1990s). Data mining isthe process of applying these methods to data with the intention of uncovering hidden patterns in large data bases.

Cloud computing is a type of computing that depends on sharing computing resources rather than having local servers or personal devices to handle applications. In cloud computing, the word cloud is used as a metaphor for "the Internet" and a standardized cloud-like shape was used to denote a network on telephony schematics. So the expression cloud computing means "a type of Internet-based computing," where different services such as servers, storage and applications are delivered to an organization's computers and devices through the Internet.

II. DATA MINING CONCEPTS

The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. It involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating [2].

Making fact-based decisions is not dependent on the amount of data you have. Success will depend on how quickly insights are discovered from all that data and use those insights to drive better actions across entire organization.So data mining is used in combination with predictive analytics, machine learning and decision management.

Predictive analytics

It helps assess what will happen in the future?

Machine learning

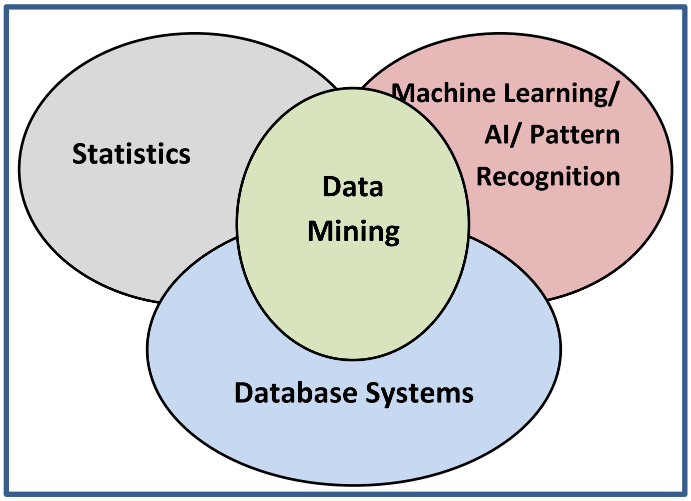
It uses algorithms to build analytical models, helping computers “learn” from data.

Decision management

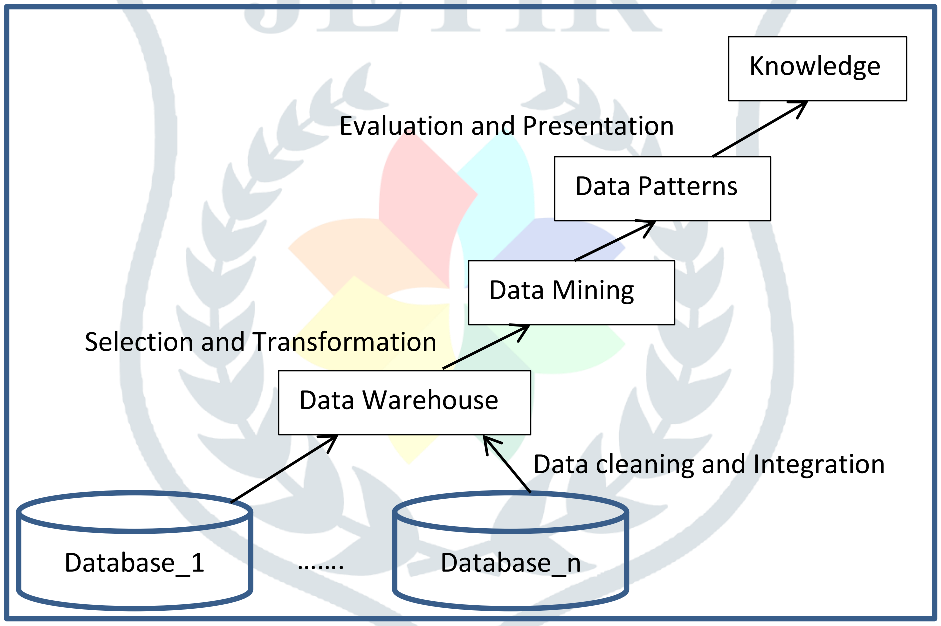
It turns those insights into actions that are used in your operational processes. So while the same approaches can still be applied today – they need to happen faster and at a larger scale, using the most modern techniques available.

Data mining

It looks for hidden patterns in data that can be used to predict future behavior. Businesses, scientists and governments have used this approach for years to transform data into proactive insights. The idea of data mining is derived from machine learning, artificial intelligence, pattern recognition,statistics, and database systems. Data Mining is a seven step process.



1. **Data cleaning:** remove noisy and inconsistent data.
2. **Data integration:** combines multiple data sources.
3. **Data selection:** retrieves the data relevant to the analysis task from the database.
4. **Data transformation:**transformation of data into forms appropriate for mining.
5. **Extract data patterns:** intelligent methods are applied to extract data patterns.
6. **Pattern evaluation:** identifies the truly interesting patterns.
7. **Knowledge presentation:** presents the mined knowledge to user.


Knowledge
Evaluation and Presentation
Data Patterns
Data Mining
Selection and Transformation
Data Warehouse
Data cleaning and Integration
Database_1
…….
Database_n

III. CLOUD COMPUTING CONCEPTS

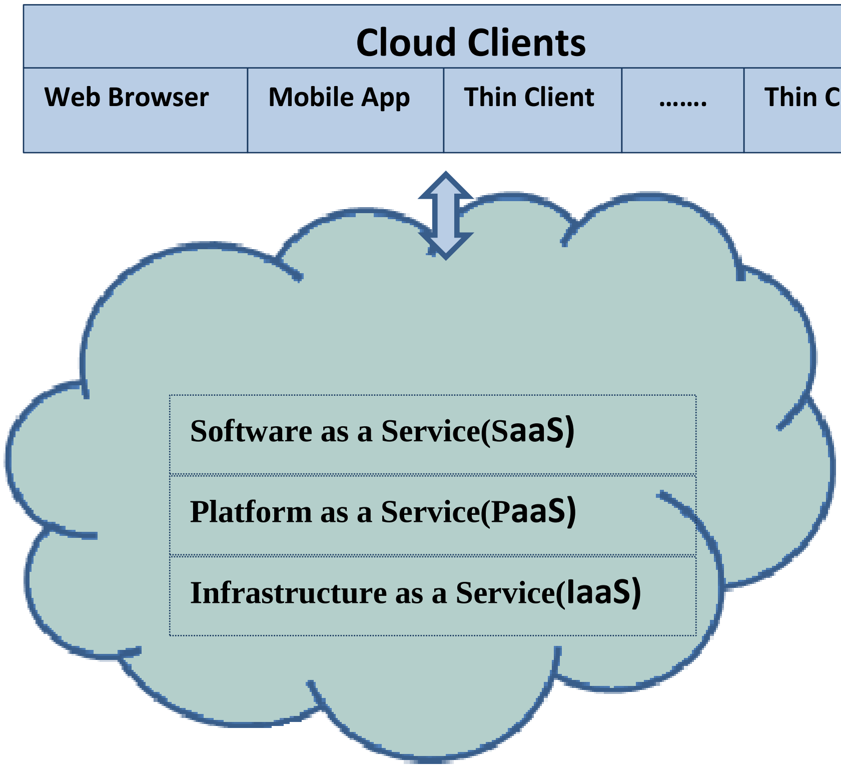
Cloud computing service model

According to National Institute of Standards and Technology there are three types of cloud services and are stacked one upon the other [3].

1. Infrastructure as a Service (IaaS):
- In a virtualized environment delivers computer infrastructure as a utility service such as virtual machines, servers, storage, load balancer, network etc.
- Provides enormous potential for extensibility andscale
- It is on the bottom of the stack.

2. Platform as a Service (PaaS):
- Delivers a platform or solution stack on a cloudinfrastructure such as execution runtime, database, webserver, development tools etc..
- It integrates with development and middleware capabilities as well as database, messaging andqueuing functions.
- It sits on a top of the IaaS.

3. Software as a Service (SaaS):
- Delivers the application over the Internet or Intranet via a cloud Infrastructure such as CRM, E-mail, Virtual desktop, Communication, Games etc.
- Built on underlying IaaS and PaaS Layer.
- It at the top of the stack.


Cloud Clients
Web Browser
Mobile App
Thin Client
…….
Thin Client
Software as a Service(SaaS)
Platform as a Service(PaaS)
Infrastructure as a Service(IaaS)

Cloud computing deployment models

Basic cloud computing deployment models are private clouds, public clouds, and hybrids clouds.

1. **Private Clouds:** Private clouds are data center architectures owned by a single company that provides flexibility, scalability, provisioning, automation, and monitoring. These clouds offer the greatest level of security and control. The main drawback with a private cloud is that all management, maintenance and updating of data centers is the responsibility of the company. Situations when a private cloud is an apparent choice
   - Organization's business is its data and its applications. Therefore, control and security are paramount.
   - Organization's business is part of an industry that must conform to strict security and data privacy issues.
   - Organization is large enough to run a next generation cloud data center efficiently and effectively on its own.
2. **Public Clouds:** Public cloud is basically the internet. Service providers use the internet to make resources, such as applications and storage, available to the general public, or on a 'public cloud. Data is stored in the provider's data center and the provider is responsible for the management and maintenance of the data center. It reduces lead times in testing and deploying new products but companies feel security could be lacking with a public cloud. Examples of public clouds include Amazon Elastic Compute Cloud (EC2), IBM's Blue Cloud, Sun Cloud, Google App Engine and Windows Azure Services Platform.Situations when a public cloud is an apparent choice
   - Organization's standardized workload for applications is used by lots of people, such as e-mail.
   - Organization need to test and develop application code.
   - Organization has SaaS (Software as a Service) applications from a vendor who has a well-implemented security strategy.
   - Organization needs incremental capacity (the ability to add computer capacity for peak times).
   - Organization is doing collaboration projects.
   - Organization is doing an ad-hoc software development project using a Platform as a Service (PaaS) offering cloud.
3. **Hybrid Clouds:** In hybrid clouds companies maintain their private cloud and rely on the public cloud as needed. For instance during peak periods individual applications, or portions of applications can be migrated to the Public Cloud. Situations when a hybrid cloud is an apparent choice
   - Organization wants to use a SaaS application but is concerned about security. Its SaaS vendor can create a private cloud just for this organization inside their firewall. They provide the organization with a virtual private network (VPN) for additional security.
   - Organization offers services that are tailored for different vertical markets. Organization can use a public cloud to interact with the clients but keep their data secured within a private cloud.

IV. DATA MINING WITH CLOUD COMPUTING

There are three major aspects of Data Mining with Cloud Computing namely – Perform Data Mining on cloud data, Provide Data Mining tools via clouds (SaaS), Security.

Perform Data Mining on cloud data

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 [4]. Data mining in cloud may perform following tasks [5]:

· Analyze Key Influencers
· Detect Categories
· Fill From Example

- Forecast
- Highlight Exceptions
- Scenario Analysis
- Prediction Calculator
- Shopping Basket Analysis

Some categories of distributed and parallel data mining algorithms for Cloud Computing that perform above listed tasks.

- **Association Rule Learning Algorithms:** Apriori algorithm and its variants are commonly used to find rules associated with frequently co-occurring items, used for market basket analysis, cross-sell, and root-cause analysis. Apriori algorithm
  1. Scans the database time after time to finds all the frequent item sets
  2. It will consume a lot of time and memory space when scanning the database
- **Classification Algorithms:** Decision tree, AO algorithm, and neural network commonly used to propose a classification model which can map the data item in the database to one of the given categories such as yes/no, more/moderate/less etc.
- **Clustering:** K-means and its variants, Hierarchical clustering and density based clustering algorithms are used to find clusters of objects such that the objects in a cluster will be similar to one another and different from the objects in other clusters.
- **Stream Data Mining Algorithms:** There is continuous arrival of new data in stream data mining so
  - multiple passes are not possible
  - storage of continuous data stream is a great challenge for storage devices
  - there is less time to access data sometime only once

To generate pattern or knowledge from stream data, algorithms with different techniques are needed. Parallel association rule mining algorithm based on Apriori algorithm, parallel k-means clustering algorithm and MapReduce parallels Naive Bayes text classification algorithm and many other algorithms are developed and improved by scholars.

Provide Data Mining tools via clouds (SaaS)

Customer will have following advantages by providing data mining tools on SaaS [6]

- Customers have to pays only for the data mining tools that they require.
- Customer's browser will provide data mining tools as required he does not have to maintain a hardware infrastructure.

Security

It is the major problem discussed as it is not clear how safe outsourced data is and when using these services ownership of data is not always clear. There are also issues regarding policy and access of data. Cloud providers use data mining to provide clients a better service [7]. If clients are unaware of the information being collected, ethical issues like privacy and individuality are violated [8][9]. 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 [10][11] to mine data and thus acquire useful information from data. So in general we can state there are security issues for

- Policy and access of data as if customer's data is stored abroad then FOI (Freedom Of Information) policy of which country will be applicable.
- Cloud providers can misuse mined data.
- Attackers outside the cloud my misuse data mining tool such as, analysis of GPS data can be used to create a comprehensive profile of a person covering his financial, health and social status [12], clustering algorithms can be used to categorize people or entities and are suitable for finding behavioral patterns, multivariate analysis identifies the relationship among variables and this technique can be used to determine the financial condition of an individual from his buy-sell records, clustering algorithms can be used to categorize people or entities and are suitable for finding behavioral patterns, association rule mining can be used to discover association relationships among large number of business transaction records [13]. Data in cloud can be effectively secured by encrypting it. Direct access of client can be restricted by using proxy and brokerage services [14].

V. CONCLUSION:

Clients can use virtual machines, servers, storage, load balancer, network, execution runtime, database, web-server, development tools, CRM, E-mail, Virtual desktop, Communication, and Games etc. via clouds as per need, without investing on infrastructure. By using data mining tools cloud providers give better services to clients but secrecy and privacy issues should be kept in mind.

REFERENCES

[1] David Reinsel, John Gantz and John Rydning, "Total WW Data to Reach 163ZB by 2025",on March 2017, sponsored by Seagate Technology LLC.

[2] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.

[3] Peter Mell and Timothy Grance (September 2011). The NIST Definition of Cloud Computing (Technical report). National Institute of Standards and Technology: U.S. Department of Commerce. doi:10.6028/NIST.SP.800-145. Special publication 800-145.

[4] Bhagyashree Ambulkar and Vaishali Borkar, "Data Mining in Cloud Computing", MPGI National Multi Conference 2012 (MPGINMC-2012), 7-8 April 2012, Link: http://research.ijcaonline.org/ncrtc/number6/mpginmc1047.pdf.

[5] Alawode A. Olaide, "On Modeling Confidentiality Archetype and Data Mining in Cloud Computing", African Journal of Computing & ICT ISSN 2006-1781, Vol 6. No. 1, March 2013

[6] 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.

[7] J. Wang, J. Wan, Z. Liu, and P. Wang. Data mining of mass storage based on cloud computing. In IEEE Computer Society, pages 426–431, 2010

[8] L. Van Wel and L. Royakkers. Ethical issues in web data mining. Ethics and Inf. Technol., 6:129–140, 2004

[9] C. Clifton and D. Marks. Security and privacy implications of data mining. In ACM SIGMOD Workshop, pages 15–19, 1996

[10] R. Chow, P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, and J. Molina. Controlling data in the cloud : Outsourcing computation without outsourcing control. pages 85–90, 2009.

[11] L. Li and M. Zhang. The strategy of mining association rule based on cloud computing. In IEEE Computer Society, pages 475–478, 2011.

[12] W. Karim. The Privacy Implications of Personal Locators: Why You Should Think Twice Before Voluntarily Availing Yourself to GPS Monitoring. Washington University Journal of Law and Policy, 14:485– 515, 2004

[13] Y. Liu, J. Pisharath, W. keng Liao, G. Memik, A. Choudhary, and P. Dubey. Performance evaluation and characterization of scalable data mining algorithms abstract

[14] Anuja R.Yeole, Poonam Borkar, “Survey Paper on Data Mining in Cloud Computing”, International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064.