CLOUD-POWERED DATA MINING

Author 1: Shareeba Firdose
Masters Of Computer Application's
St.Francis College
Bangalore, India
fshareeba@gmail.com

Author 2: Mohammed Fahad Ahmed
Masters Of Computer Application's
St.Francis College
Bangalore, India
mdfahad1608@gmail.com

Author 3: Tayuib Saqlain
Masters Of Computer Application's
St.Francis College
Bangalore, India
tayuibsaqlain69@gmail.com

ABSTRACT

Data Mining is considered as a main process as it is used for verdict new, valid, useful and comprehensible forms of data. The unification of data mining forms in cloud computing determine a flexible and scalable design that can be used for effective mining of monumental amount of data from virtually joined data beginnings with the goal of bearing useful information that is helpful hesitation making. This chapter provides an survey of the need of integration and data mining in cloud computing to support efficient and secure aids for their uses and to reduce the cost of infrastructure and depository.

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

I. INTRODUCTION

In recent years, the internet has emerged as a pivotal tool in our daily lives, profoundly impacting various activities, given the colossal volume of data generated through online interactions. This data harbors are concealed insights that can profoundly inform effective decision-making processes. Seamlessly integrating cloud infrastructure with advanced data mining techniques has ushered in a transformative era of unearthing valuable insights. Cloud computing departs from traditional computing paradigms by furnishing not only hardware resources but also software applications via the internet. Its appeal stems from cost-efficiency, mobility, and extensive accessibility, offering boundless storage and computing capabilities that facilitate the exploration of substantial datasets.

The essence of data mining lies in its capacity to extract knowledge from vast databases. It enables the analysis of data from diverse sources, extracting meaningful insights that drive informed conclusions. Beyond this, data mining fuels predictive modeling, data classification, categorization, and the identification of correlations and patterns within datasets. Its pertinence spans various domains, encompassing business, science, advertising, marketing, and medicine, among others.

An integrative synergy between data mining and cloud computing has culminated in swift technological accessibility. This synergy forms the bedrock of a knowledge discovery system, comprised of decentralized data analysis services. By harmonizing these two dynamic fields, rapid access to insights is facilitated, empowering enterprises and individuals to harness the collective power of distributed data resources.

II. DATA MINING CONCEPT

Data Mining refers to the intricate extraction of implicit, previously undisclosed, and potentially valuable information from datasets. Employing an amalgamation of statistical analyses, visualization techniques, and machine learning methodologies, it unveils and presents insights in a comprehensible manner for human understanding. This multifaceted process entails the exploration and scrutiny of substantial data volumes, leading to the identification of meaningful patterns and rules via automated or semi-automated approaches. The sheer scale of data necessitates automation, as manual analysis would prove infeasible.

Within extensive databases, data mining serves as the solution for unearthing concealed yet impactful knowledge. This knowledge holds the potential to guide governmental bodies and enterprises in making astute decisions, thereby maximizing their gains. Often referred to as Knowledge Discovery in Databases (KDD), data mining orchestrates a process that resonates with the unveiling of hidden treasures, propelling innovation and strategic actions.

A. 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.

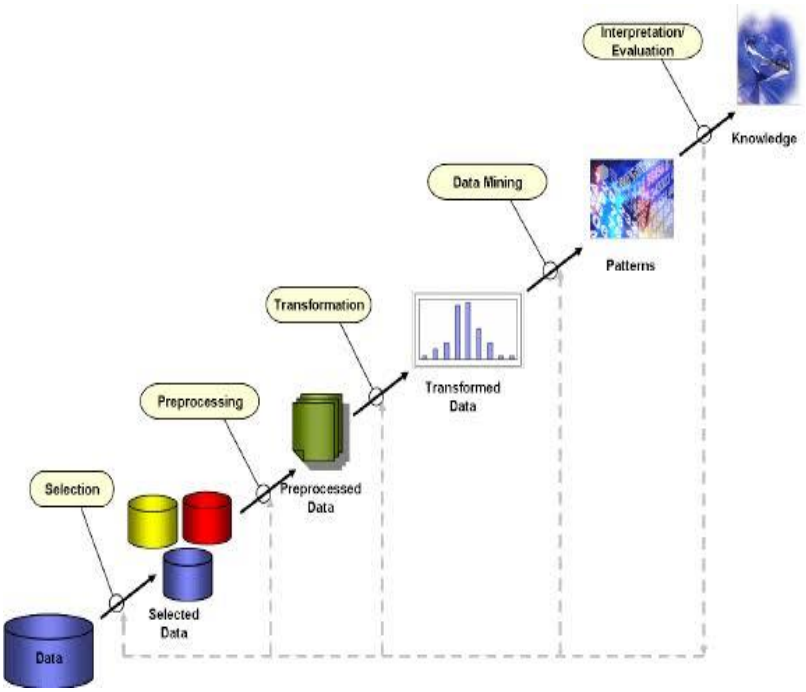

Figure 1. Steps of KDD or data mining process

B. Components of Data Mining

Components of the data mining framework include:

- **Information Repositories:** This set encompasses databases, data warehouses, spreadsheets, and other data repositories. Here, data cleaning and integration techniques can be applied to refine and unify the data.
- **Database/Data Warehouse Server:** This pivotal component retrieves data from a data warehouse in response to user queries, facilitating the extraction of pertinent information.
- **Knowledge Base:** Leveraging domain knowledge, this element serves as the foundation for uncovering compelling and valuable patterns within the data.
- **Data Mining Engines:** These functional modules execute critical tasks like classification, association, and cluster analysis. They embody the computational prowess driving the data mining process.
- **Pattern Evolution Module:** Employing measures of interestingness, this module hones the search, guiding the focus towards patterns that possess significance and relevance.
- **Graphical User Interface (GUI):** Acting as a bridge between end users and the data mining system, the GUI empowers users to interact effortlessly. Through this graphical interface, users can articulate data mining tasks or queries, facilitating seamless communication with the system. This framework synergizes these components into a cohesive ecosystem, orchestrating the intricate process of data mining while providing users with an intuitive means of engagement.

C. Data Mining Methods

The two primary goals of data mining likely to be prediction and description. Prediction involves utilizing some variables or fields in the database to predict unknown or future principles of other variables of interest, and Description focuses on verdict human-interpretable patterns interpreting the data. The goals of prediction and description can be achieved through various data-mining methods [2] depicted here.

- **Regression:** Regression involves the process of learning a function that establishes a connection between input data and a continuous, real-valued outcome. For example, it can be used to predict the likelihood of a patient's survival based on the results of various diagnostic tests, or to forecast sales figures by considering advertising expenditures.
- **Classification:** Classification is the task of developing a function that assigns input data to distinct predefined classes or categories. For instance, it plays a vital role in automated tasks such as identifying specific objects within large image databases or categorizing trends in the economic market.
- **Clustering:** Clustering is a descriptive technique where the objective is to identify distinct groups or clusters within a dataset. These groupings can either be mutually exclusive and comprehensive, or they can offer a more intricate representation, possibly through hierarchical or overlapping categories. For instance, clustering can be used to uncover similar consumer segments within marketing databases.
- **Change and Deviation Detection:** Change and deviation detection are concerned with identifying significant changes in data patterns when compared to previously observed norms. This can be valuable in recognizing anomalies or shifts in data distribution that might indicate unusual or noteworthy events.
- **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.

D. Applications Of Data Mining

Major application areas for data mining are as follows:

- **Fraud Detection:** Data mining plays a pivotal role in monitoring credit card transactions to uncover instances of fraud, effectively scrutinizing millions of accounts. Its application extends to the identification of financial activities that could be indicative of money laundering schemes.
- **Investment:** Within the realm of investment, data mining is widely utilized by various companies, albeit with limited disclosure about their specific systems. Noteworthy among these is LVS Capital Management, which employs a system incorporating expert systems, neural networks, and genetic algorithms to proficiently manage investment portfolios.

- **Marketing:** In the field of marketing, data mining finds significant utility through database marketing systems. These systems diligently analyse customer databases to delineate distinct customer segments and predict their behavioural trends, facilitating targeted marketing strategies.
- **Telecommunications:** The Telecommunication Alarm-Sequence Analyzer (TASA) introduces an array of refinement tools like pruning, grouping, and ordering to enhance the outcomes of the foundational brute-force search for rules. This toolset empowers the exploration of extensive sets of derived rules, supported by adaptable information-retrieval mechanisms that foster interactivity and iterative analysis.
- **Healthcare and Medical Research:** Data mining assumes a pivotal role within medical research, orchestrating the analysis of patient records, clinical trials, and genetic data. Its significance is underscored by its contributions across disease diagnosis, facilitation of drug discovery, optimization of treatment planning, and anticipation of patient outcomes.

III. CLOUD COMPUTING CONCEPT

Cloud Computing [3]is a general term used to describe a new class of network based on computing data that takes place over the internet. It is a new concept that defines the user computing and has a utility, that has recently attracted significant attention. National Institute of Standard and Technology (NIST) [4] defines Cloud Computing model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, server, storage, applications and services) that can be rapidly provisioned and released with minimal management efforts on service provided interaction.

Cloud computing represents a transformative paradigm shift, relocating computing processes from individual personal computers or dedicated application servers to a collective ensemble known as the "Cloud of Computers." Users of the cloud are primarily focused on the specific computing services they require, while the intricate mechanisms that facilitate these services remain concealed from view. This approach to distributed computing hinges on the consolidation of diverse computer resources into a shared pool, efficiently overseen by software automation, rather than direct human intervention.

The computing paradigm shifts [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 each other.
Stage 4: Local networks could connect to other local networks to establish a more global network.
Stage 5: The electronic grid simplified 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 of self service.
- Resources pooling.
- Board of network access.
- Pay as per use of service.
- Rapid use of elasticity and flexibility.
- Service Models.
- Deployment Models.
- Automatic Updates.
- Green Computing.

A. Basic Cloud Models

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

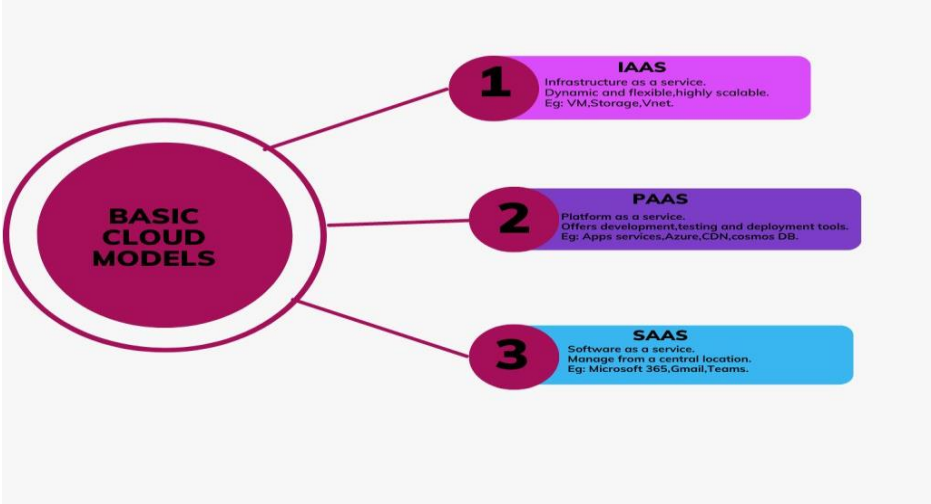

Figure 2. Basic Cloud Service models

- **IaaS (Infrastructure as a Service):** IaaS offers a virtualized computing environment as a service, eliminating the need for clients to invest in physical servers, software, data centre space, or network equipment. Instead, these resources are procured as an outsourced service, allowing businesses to scale their infrastructure up or down based on their needs.
- **PaaS (Platform as a Service):** PaaS provides developers with a complete computing platform as a service, enabling them to focus on building and deploying applications without concerning themselves with the underlying infrastructure. This model offers tools, development frameworks, and runtime environments that streamline the app development process.
- **SaaS (Software as a Service):** SaaS involves delivering software applications to customers as a service. Users can access and use the software through the cloud without needing to install it on their devices. This model offers the advantage of easy scalability, maintenance, and updates, reducing the burden on users to manage the software themselves.

B. Cloud Deployment Models

The deployment models [7] of cloud computing are:

- **Private Cloud:** A private cloud is a dedicated cloud infrastructure designed and managed exclusively for a single organization. This environment can be operated internally or by a third-party provider and can be hosted on-premises or externally. Private clouds offer enhanced control and security, as they cater to the specific needs and requirements of the organization.
- **Public Cloud:** A public cloud provides services over a network that is accessible to the general public. These services can be offered for free or on a pay-as-you-go basis. While the architectural setup might resemble that of private clouds, security considerations can differ significantly. Public cloud services are available to anyone and are delivered by service providers over potentially untrusted networks.
- **Community Cloud:** A community cloud enables multiple organizations with shared interests, needs, and security requirements to utilize the same cloud infrastructure. This setup offers a collaborative environment where resources are tailored to the collective needs of the participating organizations. It allows for a balance between customizability and resource sharing.
- **Hybrid Cloud:** A hybrid cloud is a fusion of two or more distinct cloud deployments (private, community, or public), maintaining their individual identities while being interconnected. This arrangement provides the advantages of different deployment models, allowing organizations to leverage on-premises resources, third-party services, and cloud capabilities. Hybrid cloud also encompasses the ability to link traditional data centre services with cloud-based resources for enhanced flexibility and scalability.

C. Advantages Of Cloud Computing

- **Cost-Effective Computing:** Cloud computing eliminates the need for investing in high-powered and costly computers to operate web-based applications.
- **Enhanced Performance:** Cloud-based computers exhibit quicker boot-up and operation times due to the reduced number of loaded programs and processes in memory.
- **Economical Software Expenses:** Instead of purchasing expensive software applications, many essential tools are available for free within the cloud computing environment.
- **Seamless Software Updates:** The convenience of automatic updates is a highlight of cloud computing. This removes the dilemma of outdated software versus expensive upgrades, as web-based applications are regularly updated.
- **Boundless Storage Capacity:** Cloud computing presents nearly limitless storage possibilities, catering to diverse data storage needs.
- **Heightened Data Reliability:** In contrast to traditional desktop computing, where a hard disk crash can lead to the loss of valuable data, cloud-based data remains unaffected even if a local computer crashes.

D. Disadvantages Of Cloud Computing

- **Persistent Internet Connection Needed:** Cloud computing relies on a continuous internet connection for seamless access and operation.
- **Limited Performance on Low-Speed Connections:** Cloud computing's effectiveness diminishes with slower internet connections, impacting its performance.
- **Data Security Concerns:** Cloud computing raises potential security issues regarding the safety of stored data.

IV. INTEGRATING DATA MINING IN CLOUD COMPUTING

Data mining methods and their applications hold a vital role within the domain of cloud computing. Data mining encompasses the process of extracting structured insights from web data sources, whether they are unstructured or semi-structured. By integrating data mining into Cloud Computing, organizations can streamline the management of software and data storage, ensuring users have access to dependable, secure, and efficient services.

This integration explores how data mining tools, such as SaaS, PaaS, and IaaS, operate within cloud computing to extract valuable information. Data mining finds wide-ranging utility across diverse sectors including banking, medical, and marketing. It facilitates the analysis and extraction of pertinent insights, spanning customer behaviour, preferences, interests, and geographical locations where all are readily accessible with a few clicks.

The application of data mining in the cloud domain proves especially advantageous for small-sized enterprises, democratizing the ability to efficiently analyse organizational data. This democratization, which was once exclusive to larger corporations, is now accessible through cloud services.

Notably, data mining's utility shines particularly bright when dealing with vast datasets, as its algorithms often require substantial data to create robust models. Cloud service providers leverage data mining to elevate the quality of client services.

Leveraging data mining methods within cloud computing empowers users to extract valuable insights from virtually integrated data sources, subsequently reducing infrastructure and storage expenses. This convergence of data mining and cloud computing not only cuts costs but also elevates the efficiency of information extraction processes.

Data Mining finds its prime utility in handling substantial volumes of data, as the algorithms associated with it often demand extensive datasets to construct accurate models of high quality. Within the cloud computing landscape, data mining takes centre stage as cloud providers harness its capabilities to enhance the services offered to clients.

By incorporating data mining methodologies into cloud computing, users gain the ability to extract valuable insights from seamlessly integrated data sources. This integration not only yields useful information but also contributes to a reduction in infrastructure and storage expenses.

Cloud Computing represents the contemporary paradigm in Internet services, characterized by the utilization of server clusters, often referred to as clouds, to manage diverse tasks. In the context of cloud computing, data mining encompasses the procedure of extracting structured insights from sources of web data, whether they are unstructured or semi-structured. As Cloud computing pertains to the delivery of software and hardware as services via the Internet, data mining software within this domain follows a similar pattern. It is also provisioned as a service, aligning with the overarching principles of cloud computing.

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.

V. CONCLUSION

The integration of data mining into cloud computing stands as a pivotal factor in enabling businesses to arrive at informed decisions and anticipate future trends and behaviors effectively. In this symbiotic relationship, computing represents the service provider, while data mining assumes the role of the served entity. It's worth noting that Data Mining can exist independently of Cloud Computing, and Cloud Computing is not limited solely to Data Mining. Rather, they complement each other like a well-matched cake and its delectable icing, synergizing to offer remarkable efficiency.

Cloud computing hinges on the utilization of remotely located server clusters to manage a multitude of tasks. On the other hand, data mining involves the systematic extraction of structured insights from unstructured or semi-structured web data sources. This integration leverages the strengths of both domains, culminating in a powerful tool for organizations seeking to optimize their decision-making processes and glean valuable insights from data. 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.

VI. 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), Volume1,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.

[10] Berson, Alex "Data Mining" New York: McGraw-Hill, 1997.