International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 05, Issue 03, September 2024
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

Data Mining Techniques Based on A Cloud Computing

Walid Qassim Qwaider
Business Administration Dept, College of Science and Humanities in Ghat
Majmaah University, Saudi Arabia
w.qwaider@mu.edu.sa

Abstract - Data mining is an essential process as it is used to find and discover new, correct, useful and understandable forms of data. Cloud computing has become a multi-use technology for data processing, storage, and distribution, offering a wide range of applications and infrastructure such as the Internet service that customers can access from anywhere. The massive volume of data can have stored in low-cost cloud data centers. Both data mining and cloud computing technologies help business organizations maximize profits and reduce costs in different ways. The focus of this paper on data mining is the technique of knowing the previously unknown relationship and new patterns in extensive data that can even predict future decisions by using some useful algorithms and techniques such as classification, prediction, clustering, summarization, base assemblies, regression analysis, time series analysis

Keywords: *Cloud Computing, Data Mining, Data Mining Techniques, Clustering, k-Means Algorithm.*

I. INTRODUCTION

Our current era has characterized by great torrents and general data, making it impossible for analysts to extract meaningful information by resorting only to traditional approaches to preliminary data analysis. With the presence of large amounts of data stored in databases and data warehouses, the need to develop robust tools for data analysis and extraction of information and knowledge has increased. Hence, data mining has emerged as a technique aimed at extracting knowledge from vast amounts of data [1].

The term cloud refers to the net or the Internet. In other words, we can say that the cloud is something that exists remotely. Clouds can provide services over any network in public or private networks on WAN, LAN, or LAN networks VPN or VPNs [2]. Cloud computing also processes, configures, and accesses Internet applications. It stores data on Internet, infrastructure, and applications [3]. Cloud computing technology services have broadly divided into three services, and customers have the right to choose one or more of the desired functions.

Clustering is a technical process of placing data in similar clusters, a branch of data mining. The aggregation algorithm divides a data set into several groups since the similarity between points within a particular pool is higher than the similarity between points within two different clusters. k-means clustering is a method of vector quantification, formerly in signal processing science, which is best known for its use in cluster analysis during data mining [5]. The purpose of this algorithm is to divide the number of elements (n data) into some k sections in which each feature has included in the part with the nearest central point. The central location represents the basis on which the data is divided and categorized, clustering. The result of the classification is a division into Furonic regions [6].

This paper has arranged as follows. In section II, we introduce the cloud computing, types of cloud services, advantages and disadvantages of cloud computing. Section III Then, we present the Data mining definition, data mining technique, clustering, k-means clustering algorithm. Section V discusses the data mining techniques based on a cloud computing; the part presents Some of the data mining tools has delivered on a cloud, such as Weka4WS, Ricardo, BC-PDM, Rapid Miner, ESOM-Maps, Mahout. Finally, the conclusions.

II. RESEARCH METHODS

Cloud computing " Is a process through which tasks are distributed over a large number of computer resources gathering so that these communities can carry out the necessary applications by accessing computing power, storage space and information service when required. A new, user-driven model to flexibly access hardware and software resources through the Internet and allow companies to lower costs and increase performance. It is due to the presence of a database that provides services, and the cloud can be considered as a key and unique point of access to receive all requests from customers around the world [7]. Cloud computing technology services have broadly divided into three services, and customers have the right to choose one or more of the desired services [8]. Figure 1 shows the cloud-computing services are [9]:

1) Software as a service (SaaS): Software as a service (SaaS): Is the highest level in the cloud where software applications or data for the library have hosted on the Internet. This level of cloud computing is the most easily accessible by non-profit organizations and libraries because it requires relatively little development and training from within the organization to obtain and operate it

International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 05, Issue 03, September 2024
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

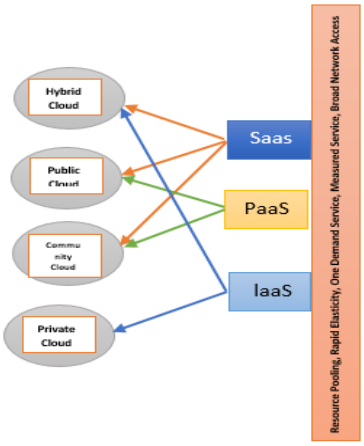


Figure 1. Cloud Computing Architecture

2) Platform as a service (PaaS): Platform as a service (PaaS): The platform or computing environment as a service (PaaS) is the next level of the cloud. It has often used by organizations that develop or modify their software applications. The computing environment supports software development processes, including prototyping, development, testing, deployment, and hosting of software.

3) Infrastructure as a service (IaaS); is the foundation or bottom layer of cloud computing, sometimes referred to as "Haas (Service as Hardware)." It involves services such as storage, backups, disaster recovery, Data, and security. In an enterprise, cloud computing allows the company to pay only as much as possible as needed, and to make the Internet as soon as possible, and because this "pay for what you used" model, is similar to the way electricity, fuel, It has sometimes referred to as the computing facility.

A. Advantages and disadvantages of cloud computing

The benefits of cloud computing are summarized as follows [24]:

1) A person can access applications and services over the Internet through the computerized cloud.
2) Cloud computing can have handled and demand generated online at any time.
3) Does not need to install a specific piece of software to access or process the application on the computer cloud.
4) Cloud computing provides development and deployment of Internet tools, environment programming and operating time through a platform as a model the service.
5) The cloud provides resources in a way that offers an independent platform to reach any customer.
6) Cloud computing provides self-service on demand. Resources can have used without interacting with the service provider cloud.
7) Cloud computing is a cost-effective goal because it increases efficiencies and benefits. It is It only requires an Internet connection.
8) Cloud computing offers the budget that makes it more reliable.

The disadvantages of cloud computing are summarized as follows [24]:

1) Fear of hacking server servers and stealing data or selling them by third party service providers. Therefore, you should deal with reliable, credible and transparent companies.
2) Fear of interruption of Internet service in general.
3) The current applications of cloud computing on the Internet have not yet reached the desired level and expected efficiency. There is not, however, an application program on the Internet for the degree of modification to the desired images such as Photoshop.

Data mining emerged in the mid-1990s in the United States, combining statistics and information technologies (databases, artificial intelligence, and machine learning). There are several definitions of this concept, which can have defined as "automated or automated exploration of interesting, and invisible patterns hidden in a given database," or "a process of accurate, intelligent. The interactive and sequential analysis that allows the activity facilitators to use this process Making decisions and doing appropriate work in favor of the activity they are responsible [10]. The institution in which they work, "or" analyzes large amounts of data to create rules, examples, and models that can be used to guide decision makers and predict future behavior. ", Can also be defined as: "Analysis of large pilgrimage collections M of data seen to search for potential relationships and summarize data in new forms to be understood and useful to its user"[11].

Through previous definitions, data mining is a process of extracting or discovering useful and exploitable knowledge through a wide range of data. It helps to explore hidden knowledge and unexpected models, as well as explore new rules that exist in large databases [12]. The following figure explains the different steps which comprise the overall data mining process:

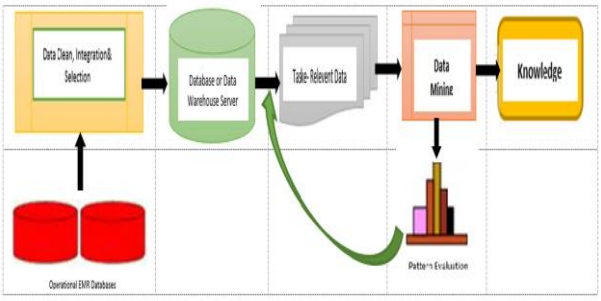


Figure 2. Data Mining Process

A. Data Mining Technique

Data mining models are two types: Predictive models are designed to predict the value of some properties. Such as predicting a potential purchase for the customer. Descriptive models have divided into two types: clustering patterns that allow the assembly of individuals, events, products in clusters, and correlation models that allow for the identification of relationships between them,

figure 3 shows the classification of data mining tasks. There are several tools to explore the data, the most important of which are [11] [13]:

1) Classification: Classification is the interpretation or prediction of an individual's property through other characteristics. This property is generally how. The grading can be done using old statistical methods such as regression and differential analysis or using relatively recent techniques such as correlation forces, case-based conclusions, and neural networks. As examples of the classification methods used as part of knowledge exploration applications that include the classification of financial market trends and the automatic identification of significant objects in an extensive database.
2) Prediction: The prediction is similar to classification or estimation, except that data have classified as predicting future behavior or estimating its future value. The predicted dependent variable is a quantitative variable. Some of the traditional tools used in forecasting are, for example, regression and differential analysis. The new methods include correlation rules, decision tree, neural networks, and genetic algorithms.
3) Regression Analysis (RA): Regression analysis is a statistical method. It can be used in the digital prediction process and modeling the relationship between the independent variable or the dependent variable.
4) Time Series Analysis: A series of events that change with time has usually measured at equal intervals. The analysis of time series means that statistical methods are used to model the sequence of events through time-based data points.
5) Association Rule: The base mining process of the Association consists of first finding the recurrent elements from which strong association rules have established in form A and B. Numerous statistical analysis can be carried out to detect statistical linkage between items A and B.

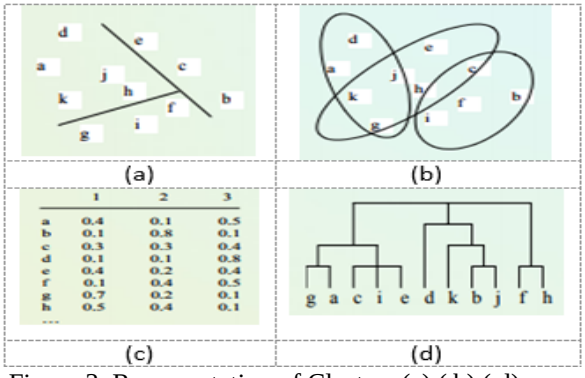


Figure 3. Representation of Clusters (a),(b),(d) are hierarchical and (c) is nonhierarchical.

There are many algorithms used in the data compilation process, and we will examine the most straightforward algorithm, the k-means clustering algorithm [13].

C. K-means Clustering Algorithm

K-groups mean grouping algorithms/sets of different notes relating to each other without any idea about those relationships that exist between them. Some symbols can used to represent objects, where n represents the total number of features used to describe cluster. After this has done, the algorithm chooses k-points in the vector space randomly. These points act as primary centers of the group. After the procedure, all the objects involved have assigned to the center points that are at least the distance from them. Thus, a separate new center has created through the vectors average for all the objects assigned to it [14]. The Lloyd's algorithm, mostly known as the k-means algorithm, used to solve the k-means clustering problem and works as follows. First, decide the number of clusters k. Then:

The algorithm eventually converges to a central point, taking into account that it is not necessarily the minimum of total squares. Where the problem is hypothetical and the algorithm is merely metaphorical, converging to the minimum acceptable. The algorithm's work stops when the assignment from one frequency to the next final shape has not changed. Figure 5 shows the K-Mines synthesis algorithm [14] [23].

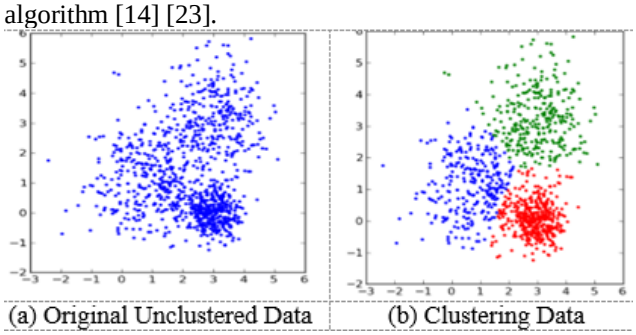


Figure 4. K-Means Clustering Algorithm

III. RESULT AND ANALYSIS

There is a strong need for computer technologies to extract the required data from the cloud computing model. Cloud computing penetrated all business domains to become a significant technical area in the data mining application. "Cloud computing means the new scientific direction in Internet services that rely on zips of applications and servers to handle the tasks required." Data mining in cloud computing is a structured process from unstructured or semi-structured data sources. The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with an assurance of efficient, reliable and secure services for their users. Distributed and parallel data mining algorithms can be used for sharing of resources In cloud computing, software and applications to extract data are also created in this way as follows [17]:

1) *SaaS:* Developers can use applications ready to explore data using specific algorithms that can be accessed directly through a web browser.
2) *PaaS:* Developers can use supporting platforms for their data analytics and storage like Hadoop, Apache Cassandra.
3) *IaaS:* Developers can usually take advantage of the available virtual resources that belong to the

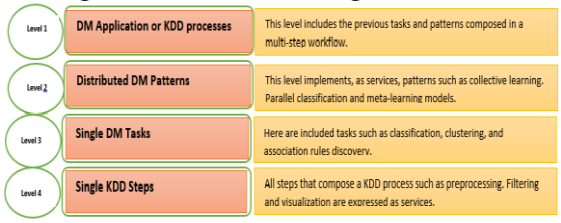
computing infrastructure to perform data analyzes and use data mining tools.

A series of cloud computing service platforms have been developed to provide data mining services for the public some platforms are designed to provide data mining in cloud services. Talia et al. summarize data mining in cloud computing in four levels, figure 6 shows the four levels of data mining services [15][16]:

1) *Single KDD steps:* the underlying composition data mining algorithms.
2) *Single data mining tasks:* a separate class of services meant for data mining like classification, clustering, etc.
3) *Distributed data mining patterns:* distributed data mining models like aggregation, parallel classification, and machine learning.
4) *Data mining applications or KDD processes*: Complete data mining application that are based on the elements belonging to all above.

By this design, they designed a Data Mining open service framework based on cloud computing and developed a series of data mining services, such as Weka4WS, BC-PDM, PD Miner, ESOM-Maps. Some of the data mining tools delivered on the cloud are [15] [21]:

1) *Weka4WS:* Weka is a widely used open source data mining toolkit that runs on a single machine. Weka4WS extends the Weka toolkit by implementing a distributed framework that supports data mining in WSRF enabled Grids. Weka4WS integrates Weka and the WSRF technology for running remote data mining algorithms and managing distributed computations as workflows [16]. The Weka4WS user interface supports the execution of both local and remote data mining tasks. On aGrid computing node, a WSRF-compliant Web service is used to expose all the data mining algorithms provided by the Weka library [18].
2) *Ricardo:* is the tool by merging the data mining tool R with distributive frameworks. This system uses declarative scripting language and Hadoop to execute R programs in parallel. This method uses R-syntax that is familiar to many analysts. But, due to the overhead produced by compiling the declarative scripts to low-level MapReduce jobs, Ricardo suffers from long execution times [20].
3) *BC-PDM:* is a SaaS tools, and is based on the MapReduce implementation of cloud computing. Users can use the data from the big cloud by BC-PDM only need to register instead then to buy or deployment, Because it based on cloud computing, so BC-PDM overcome the traditional tools, and can deal with TB level mass data mining.
4) *Rapid Miner:* is another software platform that provides an integrated environment for both machine learning and data mining. The interactive More than 1500 data mining operators are present in Rapid Miner. It can use as a stand-alone application. It also provides efficient multi-layered data view [19].
5) *ESOM-Maps:* is a data mining tool used for clustering, visualization and classification can use as SaaS via a cloud.
6) *Mahout*: a community-based Hadoop-related project, aims to provide scalable data mining algorithm. Its libraries do not provide a general framework for building algorithms. So that, quality of the offered solutions varies that depending on the contributor expertise. That leads to a potential decrease in performance[22]. Mahout mainly focuses on implementing specific algorithms, rather than building execution models for algorithm methods



Figuer 5. Four levels of data mining services

The system can provide overall data mining solution for business decisions and intelligent information processing The order offers a variety of parallel data conversion rules and parallel data mining algorithms, the full support of the production, sales, marketing, financial management, corporate decision-making activities in the field, has broad application prospects. A browser – this implies that he has only to pay the costs that have generated by using Cloud computing. Also, major companies in the field of business intelligence provide business-oriented large-scale data mining services, such as micro-strategy, IBM, Oracle and other companies own the data mining services based on cloud computing platform [15].

VI. CONCLUSION

Data mining techniques based on cloud computing is an Important characteristic of today's infrastructure to make efficient and better knowledge-driven decisions. The concepts of data mining in cloud computing and different types of algorithms that can use for sharing of resources using data mining and cloud computing We also have concluded that clustering serves as one of the best algorithms used for data mining processes owing to its less complexity and ability readily implemented without any hindrance.

K-Mines algorithm is a more efficient algorithm for massive database mining, with cloud computing providing a convenient solution for storing extensive database at the lowest cost. This paper focuses on the implementation of the K-Mines algorithm in the digital cloud environment, and experimental results show that they work well in the cloud.

REFERENCES

[1] Anwaar Ali, Raihan ur Rasool, Arjuna Sathiaseelan, Andrej Zwitter. (2016). Big Data For Development: Applications and Techniques. National University of Sciences and Technology (NUST), Pakistan,arXiv:1602.07810v1.

[2] Ethernet Virtual Private Networks. (2016). Integrated, Scalable Layer 2 and Layer 3 VPN Services, The Broadband Forum. All rights reserved.

[3] Ahmed S, Maria S. (2014). Cloud Computing: Paradigms and Technologies. F. Xhafa and N. Bessis (eds.), Inter-cooperative Collective Intelligence: Techniques and Applications, Studies in Computational Intelligence 495, DOI: 10.1007/978-3-642-35016-0_2, Springer-Verlag Berlin Heidelberg.

[4] Khadir Mohideen. (2015). Survey of data mining techniques (DMT) in Cloud Computing. International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-8, ISSN: 2395-3470.

[5] Mahendiran. A. (2012). Implementated of K-Means Clustering in Cloud Computing Environment. Research Journal of Applied Sciences, Engineering, and Technology 4(10): 1391-1394, 2012 ISSN: 2040-7467.

[6] Sonal M., Kanwal G. (2013). Factors Affecting Efficiency of K-means Algorithm. International Journal of Advancements in Research & Technology, Volume 2, Issue5, 85 ISSN 2278-7763.

[7] Raghavendra K., Pramod K., Arun A, Raghavendra C., Rajkumar B. (2015). The anatomy of big data computing. Raghavendra Kune, Department of Space, Advanced Data Processing Research Institute, Hyderabad, India.

[8] Patidar A., Patidar V. (2015). Analysis of Solving E-Learning Problems (ASELP) using Cloud Computing. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 11.

[9] J. R. Jeba, D. S. Misbha. (2016). A Study of Data Warehousing on Cloud Environment, International Journal of Innovative Research in Science, Engineering and Technology An ISO 3297: 2007 Certified Organization Vol. 5, Issue 7.

[10] Aishwarya S. Pati1.Ankita S. Patil. (2015). A review of Data Mining based Cloud Computing. International Journal of Research In Science & Engineering e-ISSN: 2394-8299, Volume: 1 Special Issue: 1, p-ISSN: 2394-8280.

[11] R.Kabilan, N.Jayaveeran.(2015). A Review of Data Mining in Cloud Computing Environment, International Journal of Innovative Research in Computer and Communication Engineering. Vol. 3, Issue 10. ISSN(Online): 2320-9801.

[12] Petar Ristoski, Heiko Paulheim. (2016). Semantic Web (SW) in data mining and knowledge discovery (KD): A comprehensive survey Web Semantics, 36, 1–22.

[13] Thirunavukkarasu K., Digvijay S., Mohd. A., Ajay S., Singh. (2016). Data Mining Techniques in Cloud Computing: A Survey. International Journal of Recent Trends in Engineering & Research Volume 02, Issue 03; ISSN: 2455-1457.

[14] Juan H., Mercedes P. (2017). Cloud Comuting implementation of the K-means algorithm for hyperspectral image analysis. J Supercomputer 73:514–529 DOI 10.1007/s11227-016-1896-3.

[15] Xia Geng, Zhi Yang. (2013). Data Mining in Cloud Computing. School of Computer Science (SCS) and Telecommunication Engineering , Jiangsu University, Jiangsu Zhenjiang, P.R. China, Atlantis Press.

[16] D. Talia and P. Trunfio, How distributed data mining tasks can thrive as knowledge services Communications of the ACM. 53(2010) 132-137.

[17] Mell, Peter; Grance, Timotht.(2011). The NIST definition of Cloud Computing (CC): Recommendations of the National Institute of Standards and Technology.NIST National Institute of Standards and Technology.

[18] Aris-Kyriakos Koliopoulos, Paraskevas Yiapanis, Firat Tekiner, Goran Nenadic, John Keane. (2015). A Parallel Distributed Weka Framework for Big Data Mining using Spark. IEEE International Congress on Big Data, 978-1-4673-7278-7/15, DOI 10.1109/Bigdata Congress.

[19] https://en.wikipedia.org/wiki/RapidMiner

[20] S. Das, Y. Sismanis, K. S. Beyer, R. Gemulla, P. J. Haas, and J. McPherson. (2010). Ricardo: Integrating R and Hadoop. in Intl Conf. on Management of Data, 2010, pp. 987–998.

[21] Haritha Padmanabhan, Derroll David. (2017). A Survey on Efficiency in Big Data Mining. International Journal of Advanced Research in Computer (IJARC) and Communication Engineering, ISSN 2278-1021.

[22] E. R. Sparks, A. Talwalkar, V. Smith, J. Kottalam, X. Pan, J. E. Gonzalez, M. J. Franklin, M. I. Jordan, and T. Kraska, "MLI: an API for distributed machine learning," ICDM, 2013.

[23] Unnati R. Raval, Chaita Jani. (2016). Implementing & Improvisation of K-means Clustering Algorithm. International Journal of Computer Science (IJCS) and Mobile Computing, Vol.5 Issue.5, pg. 191-203.

[24] Anca A., Florina P., Geanina U., George S., Gyorgy T. (2014). Study on advantages and disadvantages of CC – the benefits of Telemetry Applications in the Cloud. Recent Advances in Applied Computer Science (RAACS) and Digital Services, ISBN: 978-1-61804-179-1.

.