International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 12 Issue 3 –JANUARY 2015.

Overview of the Application of Cloud Computing in Data Mining

K.Ramya[1], B.Bharathi[2]

[1,2]*Research scholar, Sathyabama University, India*

***Abstract*— Cloud computing has gained its limelight because of its on-demand and elastic service. Many services are offered to the cloud users by the cloud providers by employing powerful data centres. One of the powerful services rendered by cloud is data storage. Data mining is the process of sorting through large amounts of data and picking out relevant information. It helps for extracting hidden useful information from large data warehouses. It helps in predicting future trends and behaviours to help businesses for taking knowledge based decisions. In this work, an overview of cloud computing and data mining is presented along with the scope of cloud computing in data mining.**

Keywords- Cloud computing, data mining, data warehouse.

I. INTRODUCTION

Cloud computing has gained its limelight because of its on-demand and elastic service. Many services are offered to the cloud users by the cloud providers by employing powerful data centers. One of the powerful services rendered by cloud is data storage. Cloud users can put an end to the problem of data and memory management. This makes sense that cloud users provide all their data to the cloud provider, in order to save space and cost. However, data outsourcing introduces the issue of security that the confidential or sensitive data can be misused. So, such data has to be kept private and confidential. In order to forbid the successful implementation of cloud, a cloud provider is expected to provide a strong privacy preserving policy to the cloud users. Cloud's service can be categorised into Software as a Service (SaaS) which is the service made available by the cloud such that the cloud users can be able to access the applications over network. Platform as a Service (PaaS) makes sense that the cloud user can build anything with the platform (e.g. Operating System) provided by the cloud. Infrastructure as a Service (IaaS) is a type of service in which the supporting equipments such as storage, servers, hardware and much more are provided to the cloud user for some cost. Here, the cloud follows the policy of pay-as-you-go model. A cloud can be of four types and they are as follows. Public Cloud is the cloud that offers resources to the general public over network and the users will be charged for what they used and there is no minimal fee.

Private Cloud is the type of cloud that is meant for a single party or an organization. Here, the users can be from a single organization. Community cloud is the cloud that allows its infrastructure to be shared by different organizations with same focal point.Hybrid cloud can be claimed as the combination of public, private and community clouds. In this cloud, the service provider can utilize the third party cloud service provider, aiming at increasing the flexibility. Some of the advantages of using cloud are its inexpensiveness that is the infrastructure is not needed to be built but can be rented. Increased data storage is encouraged by cloud, which means that tera and petabytes of data can be placed in a cloud without any struggle as the cloud is based on the principle of elasticity. In spite of all these advantages, cloud has still got many challenges such as data security, data recovery and data management etc.,

II. DATA MINING

Data mining is the process of sorting through large amounts of data and picking out relevant information. It helps for extracting hidden useful information from large data warehouses. It helps in predicting future trends and behaviors to help businesses for taking knowledge based decisions. The modern technologies of computers, networks, and sensors have made data collection and organization much easier. However, the captured data needs to be converted into information and knowledge to become useful. Data mining is the entire process of applying computer-based methodology, including new techniques for knowledge discovery, to data. Data mining identifies trends within data that go beyond simple analysis. Through the use of sophisticated algorithms, non-statistician users have the opportunity to identify key attributes of business processes and target opportunities.

2.1 Techniques of Data Mining

Some of the most important techniques of data mining are listed below.

2.1.1 Classification

It is a data mining technique used to map data instances into one of the various predefined categories. It can be used to detect individual attacks but it has high rate of false alarm. Various algorithms like decision tree induction, Bayesian

networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques are used for classification techniques. The classification algorithm has been then applied to audit data collected which then learns to classify new audit data as normal or abnormal data.

2.1.2 Association rule mining

Association describes relationship between various data records .Association rule mining is one of the most popular techniques within data mining. It acts as a sensor which provides source data for meta-learning like techniques which are at higher level of processing. Association rule mining is a slow process and can be replaced by other techniques like classification, clustering etc. An association rule has two parts, an antecedent (if) and a consequent (then).Association rules are created by analyzing data for frequent if/then patterns and support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database and confidence indicates the number of times the if/then statements have been found to be true. These rules are used for analysing and predicting the customer behavior.

2.1.3 Clustering

In this technique, data points are clustered together based on their similarity factors and is often nearness according to some defined distance. Clustering [8] is an effective way to find hidden patterns in data that humans might miss. It is useful for ID as it can cluster malicious and non malicious activity separately. k-means is a clustering algorithm used to cluster observations into different groups of related observations without having prior knowledge about their relationships. Here data is divided in k clusters where k is provided as input.

2.1.4 Feature Selection

In this process of machine learning, a set of features from available data is selected and a learning algorithm is trained using selected features for creating classification model. Extraction of features is must as it is not feasible toapply all the available features to learning algorithm. It is also called as feature reduction or variable selection technique.

2.1.5 Support Vector Machine (SVM)

It is the technique which maps network connections to the hyper plane. It attempts to separate data into multiple classes using hyper-plane. SVM algorithm can be modified to operate in the supervised learning domain.

2.1.6 Fuzzy logic

Fuzzy logic techniques are being used in computer security since 90's. It allows greater complexity for IDS while it provides some flexibility to the uncertain problem of ID. Most fuzzy IDS require human intervention to determine fuzzy sets and set of fuzzy rules.

2.1.7 Meta learning

It is the techniques where new rules are derive from several rule sets which are collected over a period. A meta-rule set [5] relates any two given sets by describing rules that expired, changed, remain unchanged or appeared new.

III. Cloud Computing in Data Mining

It is predicted that cloud computing has a good scope in the application of data mining algorithms. The reason for the statement is that cloud computing is capable of managing a huge range of datasets at a least cost. The basic idea is that the data can be distributed to several participating nodes, such that the computational load can be balanced effectively. The major issue with this case is the security of data. This makes sense that the data should be kept private and confidential. To achieve this, the data to be processed must be converted to multidimensional arrays, so that the data can be analysed by certain tools. This considerably reduces the storage cost of data.

The main theme of this work is to distribute the datasets to the cloud and the operations are distributed among the clusters. In this case, the middle level tuples are generated by the mapping functions and are processed by reduction functions. This idea is already been executed in several works. In [4], a classification model which is a mid-point of K-Nearest Neighbour (KNN) and Bayes' scheme is proposed. Another approach that is developed using the Map Reduce model aims at addressing the progressive sequential pattern mining problem, which intrinsically suffers from the scalability problem. Two Map/Reduce jobs are designed; the candidate computing job computes candidate sequential patterns of all sequences and updates the summary of each sequence for the future computation. Then, using all candidate sequential patterns as input data, the support assembling job accumulates the occurrence frequencies of candidate sequential patterns in the current period of interest and reports frequent sequential patterns to users.

Gao et al. introduces in an experimental analysis using a Random Decision Tree algorithm under a cloud computing environment by considering two different schemes in order to implement the parallelization of the learning stage. The first approach was that each node built up one or more classifiers with its local data concurrently and all classifiers are reported to a central node. Then the central node will use all classifiers together to do predictions. The second option was that each node works on a subtask of one or more classifiers and reports its result to a central node, then the central node combines work from all local nodes to generate the final classifiers and use them for prediction.

In the authors propose a scheme namely FD-Mine, which can exploit cloud nodes for pattern discovery. This scheme identifies the frequent very fast. As an added advantage, it provides security to the data also. The scalability and execution time of the system is appreciable.

IV. CONCLUSION

In this paper, an overview of cloud computing and data mining is presented. This is followed by the scope of cloud computing in data mining. Several data mining algorithms are migrating to exploit clod computing for effective scalability and responsiveness.

REFERENCES

[1] Abell´o, A., Romero, O.: Service-Oriented Business Intelligence. In: Aufaure, M.A., Zim´anyi, E. (eds.) eBISS 2011. LNBIP, vol. 96, pp. 156–185. Springer, Heidelberg(2012)

[2] Alonso, G., Casati, F., Kuno, H., Machiraju, V.: Web Services: Concepts, Architectures and Applications. Springer, Heidelberg (2004)

[3] Buyya, R., Yeo, C., Venugopal, S., Broberg, J., Brandic, I.: Cloud computing and emergingit platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems 25(6), 599–616 (2009)

[4] Castellanos, M., Dayal, U., Sellis, T., Aalst, W., Mylopoulos, J., Rosemann, M., Shaw, M.J., Szyperski, C. (eds.): Optimization Techniques 1974. LNBIP, vol. 27. Springer, Berlin (1975)

[5] Chang, F., Dean, J., Ghemawat, S., Hsieh, W.C., Wallach, D.A., Burrows, M., ChandraT., Fikes, A., Gruber, R.E.: Bigtable: A distributed storage system for structured data.ACM Trans. Comput. Syst. 26(2) (2008)

[6] Chang, J., Luo, J., Huang, J.Z., Feng, S., Fan, J.: Minimum spanning tree based classificationmodel for massive data with mapreduce implementation. In: Fan, W., Hsu, W., Webb, G.I., Liu, B., Zhang, C., Gunopulos, D., Wu, X. (eds.) ICDM Workshops, pp. 129–137. IEEE Computer Society (2010)

[7] Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Communicationsof the ACM 51(1), 107–113 (2008)

[8] d'Orazio, L., Bimonte, S.: Multidimensional Arrays for Warehousing Data on Clouds. In: Hameurlain, A., Morvan, F., Tjoa, A.M. (eds.) Globe 2010. LNCS, vol. 6265, pp. 26–37. Springer, Heidelberg (2010)

[9] Duer, W.: CRM, Customer Relationship Management. MP editions (2003)

[10] Foundation, T.A.S.: Hadoop, an open source implementing of mapreduce and GFS (2012), http://hadoop.apache.org

[11] S. Selvakani and R.S. Rajesh, "Genetic Algorithm forFraming Rules for Intrusion Detection" IJCSNS InternationalJournal of Computer Science and Network Security, Vol. 7 No. 11, November 2007.

[12] Wei Li, "Using Genetic Algorithm for Network IntrusionDetection", Department of Computer Science andEngineering, Mississippi, State University, Mississippi State, Ms 39762.

[13] ZoranaBankovic, Jose M. Moya, Alvaro Araujo, SlobodanBojanic and Octavio Nieto-Taladriz, "A Genetic AlgorithmbasedSolution for Intrusion Detection", Journal ofInformation Assurance and Security 4 (2009) 192-199.

[14] RenHui Gong, Mohammad Zulkernine, PurangAbolmaesumi, "A Software Implementation of a GeneticAlgorithm Based Approach to Network Intrusion Detection", Proceedings of the Sixth International Conference onSoftware Engineering, Artificial Intelligence, Networking andParallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks(CNPD/SAWN '05).

[15] Tamas Abraham, "IDDM: Intrusion Detection using DataMining Techniques", Information Technology Division,Electronics and Surveillance Research Laboratory, DSTOGD-0286.

[16] TheodorosLappas and KonstantinosPelechrinis, "DataMining Techniques for (Network) Intrusion DetectionSystems", Department of Computer Science and Engineering,UC Riverside, Riverside CA 92521.