INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

DISTRIBUTED FRAMEWORK FOR DATA MINING AS A SERVICE ON PRIVATE CLOUD

RUCHA V. JAMNEKAR[1], PROF. G. D. GULHANE[2], DR. HEMANT R. DESHMUKH[3]

1. Student of Master of Engineering in (CSE), IBSS college of Engineering and Technology, Amravati, India.
2. Assistant professor Department of (CSE), IBSS College of Engineering and Technology, Amravati, India.
3. Head of the Department of (CSE), IBSS College of Engineering and Technology, Amravati, India.

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: Data mining research faces two great challenges: i. Automated mining ii. Mining of distributed data. Conventional mining techniques are centralized and the data needs to be accumulated at central location. Mining tool needs to be installed on the computer before performing data mining. Thus, extra time is incurred in collecting the data. Mining is 4 done by specialized analysts who have access to mining tools. This technique is not optimal when the data is distributed over the network. To perform data mining in distributed scenario, we need to design a different framework to improve efficiency. Also, the size of accumulated data grows exponentially with time and is difficult to mine using a single computer. Personal computers have limitations in terms of computation capability and storage capacity. This paper provide a solution for distributed data mining on Hadoop framework using an interface to run the algorithm on specified number of nodes without any user level configuration. Hadoop is configured over private servers and clients can process their data through common framework from anywhere in private network. It is observed that the framework is helpful in processing large size data in less time as compared to single system.

Keywords: DAAS – Data storage as a service, DMAAS- Data Mining as a Service, FDMPC- Framework for Data Mining as a Service on Private Cloud

Corresponding Author: MS. RUCHA V. JAMNEKAR

Access Online On:

www.ijpret.com

How to Cite This Article:

Rucha V. Jamnekar, IJPRET, 2015; Volume 3 (9): 1167-1172

PAPER-QR CODE

INTRODUCTION

Cloud infrastructure gives benefit of huge availability of resources in low cost. The integration of data mining techniques with cloud computing can allow the users to extract and mine useful information from a cloud based storage and mining service. We propose a cloud based data mining model which provides the facility of mass data storage along with distributed data mining facility. Cloud model provides access to the data that can be turned into valuable patterns through data mining techniques. Public cloud is accessible through internet and brings more threats to the security of the organization's data. The data must be protected from interception. Thus, we propose a secure private cloud based mining framework for mining data securely. As the data size grows, the performance of data storing and mining gets degraded when implemented on a single system. Multi node setup of computers can enhance the performance of mining when the data size is very large. Cloud computing can be exploited for compute-intensive and data intensive applications. Data mining algorithms are both compute and data intensive, therefore cloud based tools can provide an infrastructure for distributed data mining. This paper is intended to use cloud computing to support distributed data mining. Data to be mined can either be chosen from cloud data server or can be uploaded from private computers on the network. Service oriented architectures implemented on private network can also help to resolve these issues[3] [4]. Grid based methods for data mining are already proposed for knowledge discovery. Here, we used cloud based tools and techniques to provide service oriented framework for data mining. For optimizing the performance, we have used two services in the framework:

i. Data Storage as a Service

ii. Data Mining as a Service

II. FRAMEWORK IMPLEMENTATION

2.1 Data Mining as a Service

Hadoop[1] is a popular open-source implementation of MapReduce for the analysis of large datasets. In our framework, K-Means algorithm is provided as a service on private multinode setup. Map Reduce[2] is implemented as two functions, Map( ) which applies a function to cluster the data on and returns local results based on local nodes. Reduce( ), collects the results from multiple Maps and gives consolidated final clusters. Both Map( ) and Reduce( ) can run in parallel, on multiple machines at the same time. To manage storage resources across the cluster, Hadoop uses a distributed file system. Hadoop supports MapReduce which is a

distributed programming model intended for processing massive amounts of data in large clusters, MapReduce can be implemented in a variety of programming languages, like Java, C, C++, Python, and Ruby. MapReduce is mainly intended for large clusters of systems that can work in parallel on a large dataset. Data mining as a service is implemented by leveraging the advantages of Hadoop.

2.2Data Storage as a service

The virtually integrated data sources in the private cloud model are created through Owncloud [8]. It is an open source file synchronization solution. Owncloud is used to store the data from different sources on private network. Administrator has full control over the Owncloud Data. New users can be created, deleted and permissions can be granted through Owncloud. Other users can not access Owncloud data. They can mine local data files from any computer on the network. We have proposed Data Storage as a service model for an academic institute where there are different level of users who can mine data. Data can be input from various users like teachers, staff and students and kept on cloud server. Mining can be performed by all users.

3. Algorithm Selected: K-Means

K-Means follows a simple way to classify a given data set and assign them to K clusters. Here, K is fixed a priori. The main idea is to define k centroids, one for each cluster. Algorithm runs in following sequence [5]:

- Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
- Assign each object to the group that has the closest centroid.
- When all objects have been assigned, recalculate the positions of the K centroids.
- Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups.

From the given data set, the k-Means clustering algorithm reads some sample and computes the centroids of the cluster. Each mapper reads these centroids via a centroid server running on master node. Each reducer can write the computed centroid via the same server. We have run it on Hadoop cluster.

Configuring the cluster 1- First of all generate public key on master node and replicate it on slave nodes so that communication between master and slave nodes can happen. 2- Configure Hadoop on each node by giving the path of JAVA_HOME environment variable in hadoop-

env.sh. Make necessary changes in hdfs-site.xml and core-site.xml file residing in conf directory of Hadoop. 3- Run the script start-dfs.sh found in the directory HADOOP_HOME/bin. It will start the NameNode on master and DataNode on each slave. The framework works in following steps:

i. Enter into the framework of Data Mining as a Service. ii. Two types of user can access the framework Admin and other users. Admin can access data from Owncloud. iii. Admin can select the file to mine from Owncloud whereas other users can upload a file from their system. iv. Input file size to get a prediction about appropriate number of nodes to select for mining. v. As per the prediction, choose the number of nodes on which you want to process the data. vi. Press the process key and wait for results. vii. Output will generate clusters on the basis of K Means algorithm. It will also display the time taken for clustering.

III. **PROPOSED FRAMEWORK**

With time, the volume of data keeps growing and results in data explosion problem. Big Data is a big problem these days. Conventional data mining techniques cannot deal with these problems. This framework is proposed to optimize the mining performance for different large size inputs. After examining the performance of K Means on 1, 2 and 3 nodes, we can predict the optimal number of nodes for different size of data. Framework also ensures the security of data since the data remains on private network only. For the mining of organizational level data, we use Owncloud whereas for mining personal data, individual files can be uploaded from anywhere on the private network. For experimentation of DAAS, we have accumulated academic institutes data files on Owncloud. Folders of users were synchronized with Owncloud. Files to be mined are automatically stored on Owncloud since they are synchronized.

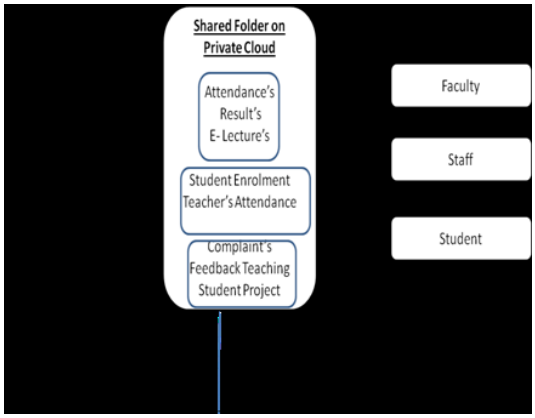

Fig.1 DAAS on Private Cloud of Academic Institution

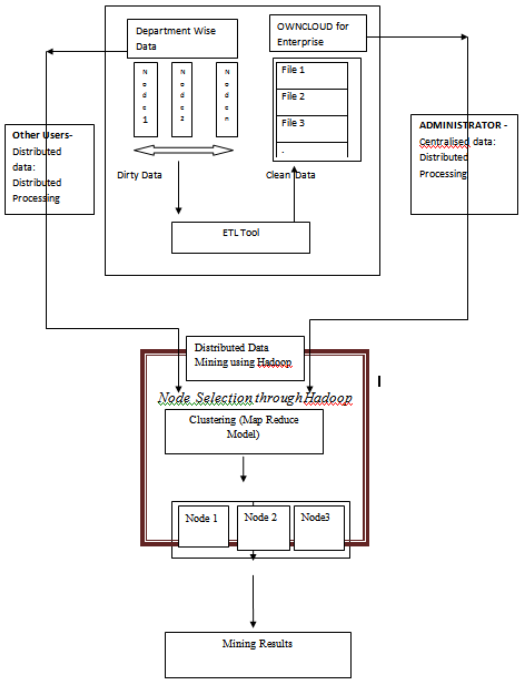

Fig. 6 FDMPC: Framework for Data Mining as a Service on Private Cloud

IV. CONCLUSION

Implementation of distributed data mining technique through private cloud architecture will allow the stakeholders to retrieve meaningful information from the whole organization in a secured way. More and more data is stored by businesses now a days and is available for information extraction and analysis. We have used this model to help a private institute to become more efficient by exploring the capabilities of data mining through private cloud. If the data size is manageable and the results to be obtained are of departmental level, user can directly mine departmental level data from distributed nodes. But when the data size is very large and is of whole enterprise, Owncloud service can be used for storage and through distributing the computation, analysis can be done optimally. On the basis of experimentation, we have proposed a generalized framework that supports Distributed Data Mining and Storage as a service on private network. The advantage of creating a framework is that the user need not configure multi node setup to process Big Data. User can just take the advantage of framework and perform data mining operation. Future Enhancements: Currently the framework is limited to K-Means algorithm only. The framework can be extended for other data mining algorithms also.

V. REFERENCES

1. Hadoop. http://hadoop.apache.org/

2. Dean, Jeffrey, and Sanjay Ghemawat. "*MapReduce: simplified data processing on large clusters*." Communications of the ACM 51.1 (2008): 107-113.

3. M. Cannataro, A. Congiusta, C. Mastroianni, A. Pugliese, D. Talia, P. *Trunfio Distributed data mining on grids: services*, tools, and applications IEEE Trans. Systems Man Cybernet. Part B, 34 (6) (2004), pp. 2451–2465

4. Grid-based data mining and knowledge discovery N. Zhong, J. Liu (Eds.), *Intelligent Technologies for Information Analysis*, Springer, Berlin (2004), pp. 19–45

5. http://home.deib.polimi.it/matteucc/Clusteri ng

6. Zhao, Weizhong, Huifang Ma, and Qing He. "*Parallel k-means clustering based on mapreduce*." Cloud Computing. Springer Berlin Heidelberg, 2009. 674-679.

7. https://bitnami.com/stack/owncloud

8. M. Cannataro, D. Talia ”The knowledge grid” Comm. ACM, 46 (1) (2003), pp. 89–93

9. A. Congiusta, D. Talia, P. Trunfio “*Distributed data mining services leveraging WSRF*” Future Generation Comput. Systems, 22 (2006), pp. 123–132.