RESEARCH ARTICLE OPEN ACCESS

Distributed Framework for Data Mining As a Service on Private Cloud

Shraddha Masih[*], Sanjay Tanwani**
*Research Scholar & Associate Professor, School of Computer Science & IT, DAVV, Indore, India
**Professor and Head, School of Computer Science & IT, DAVV, Indore, India

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.

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. We propose a cloud based data mining model which provides the facility of mass data storage along with distributed data mining facility. 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. Data to be mined can either be chosen from cloud data server or can be uploaded from private computers on the 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

I. 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. Cloud model provides access to the data iii.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.

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.

III. Experimental Setup

Interface is designed in Netbeans-8 with Java 7. Apache Hadoop 2.3.0 is used for creating multi node setup and for running mapreduce jobs. Single system configuration is:

CPU	Core2duo 2.93GHz
RAM	2 GB
Operating System	Ubuntu 64-bit
Hard disk	140 GB

Multi node setup of 2 and 3 computer systems is implemented through Apache Hadoop. The setup includes Single node setup, 2-node setup and 3- node setup through which the system resource increases two to three folds. This setup can be accessed from any other node in the network.

Analyst can apply clustering on the data that is available through i. Owncloud and ii. User's own System files. The framework can be accessed from any node on the private network. Then the analyst can provide useful patterns to the administrator of the institute by applying the distributed data mining solution.

3.1 Input File Format:

We are currently working on numeric data. The file assumed is of the marks of students. The K Means clustering will group the students into K clusters each having nearest objects.

```
223 37 222
111 159 198
414 440 271
430 203 325
278 274 194
287 488 94
138 182 205
349 152 474
488 60 340
315 278 40
341 83 369
414 229 7
124 34 146
317 141 463
353 17 182
474 398 406
374 242 34
321 32 173
161 89 243
114 191 305
236 3 134
287 489 173
```

3.2Algorithm 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:

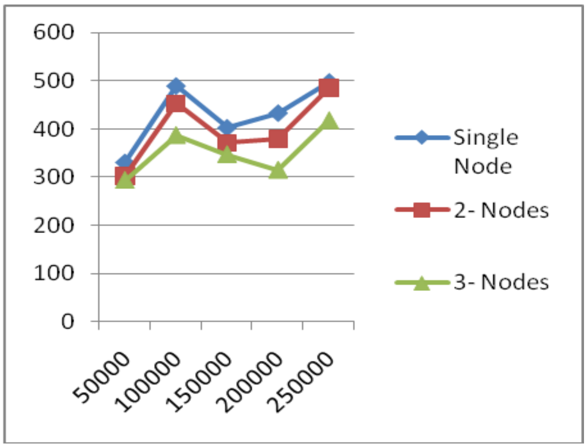
i. Enter into the framework of Data Mining as a Service.
ii. Two types of user can access the framework Admin and other users. Admin can access data from Owncloud.
iii. Admin can select the file to mine from Owncloud whereas other users can upload a file from their system.
iv. Input file size to get 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.

IV. Experimental Results

Data for clustering was selected from files of different sizes and the execution time was noted for single node, 2-node and 3- node setup. For small sized files with number of records below 50,000, the performance on single node was best. But as the number of records grow, the performance gets degraded on single node and it is observed that for very large datasets, the performance on multi node setup is better.

Time taken for clustering using K-means(in secs)					
No. of Records / No. of Nodes	50K	100K	150K	200K	250K
Single Node	330.285	489.013	402.222	432.707	497.381
2- Nodes	302.241	453.499	371.475	379.407	486.088
3- Nodes	293.908	386.768	347.156	314.811	417.281

Table1: Execution time for different nodes



Time plot (in secs.) with performance measured on single node, 2- nodes and 3- nodes

V. 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 can not 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. Required files can be selected and can be provided to K-Means algorithm which runs in a distributed fashion through Map Reduce.

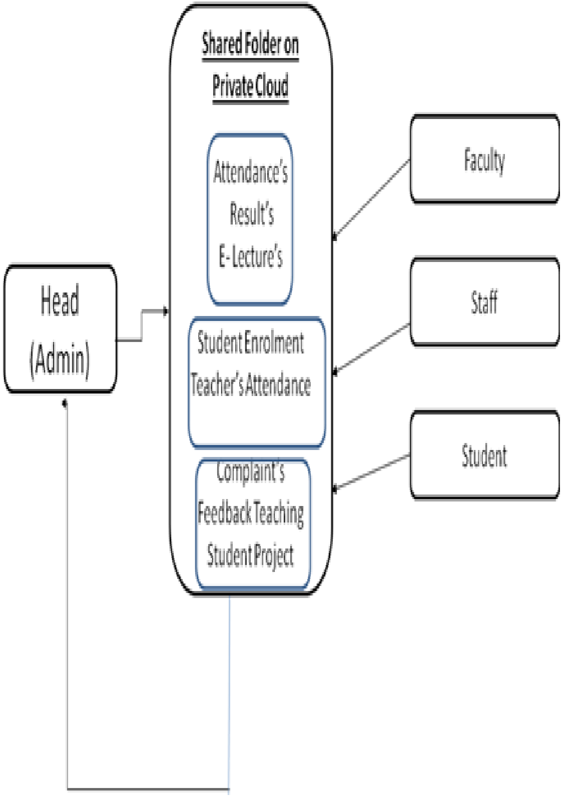
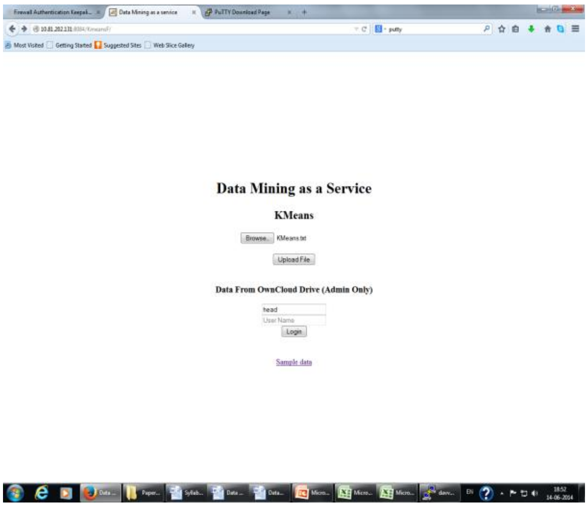


Fig.1 DAAS on Private Cloud of Academic Institution



Data Mining as a Service

KMeans

Browse... KMeans.txt

Upload File

Data From OwnCloud Drive (Admin Only)

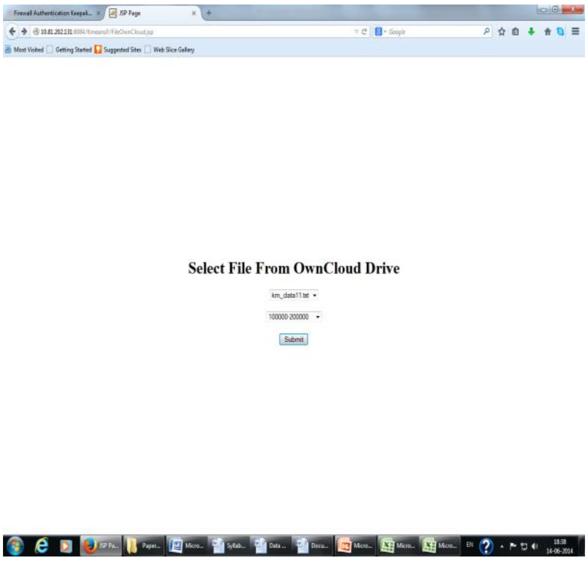
head

User Name

Login

Sample data

Fig 2 Interface for Data Mining as a Service



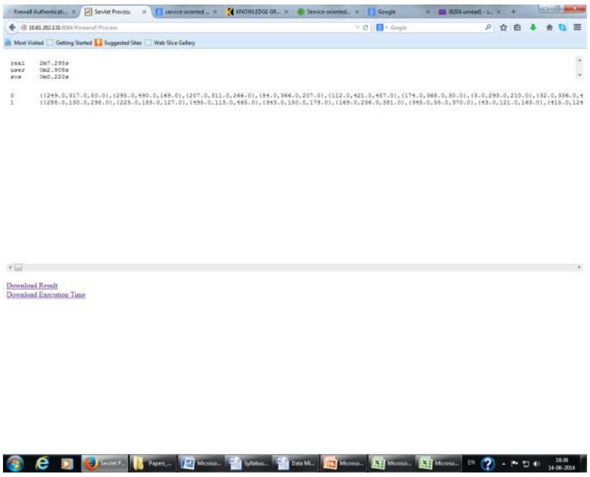
Select File From OwnCloud Drive

km_data11.txt

100000-200000

Submit

Fig 3 Interface for data selection from Owncloud



```
real    2m7.295s
user    0m2.908s
sys     0m0.220s

0       ((249.0,317.0,50.0),(295.0,490.0,169.0),(207.0,311.0,266.0),(84.0,366.0,207.0),(112.0,421.0,457.0),(174.0,368.0,30.0),(3.0,293.0,210.0),(32.0,386.0,4
1       ((288.0,130.0,298.0),(225.0,185.0,127.0),(495.0,113.0,465.0),(343.0,150.0,179.0),(169.0,296.0,381.0),(345.0,55.0,370.0),(43.0,121.0,163.0),(415.0,124
```

Download Result
Download Execution Time

Fig 4 Interface for viewing results

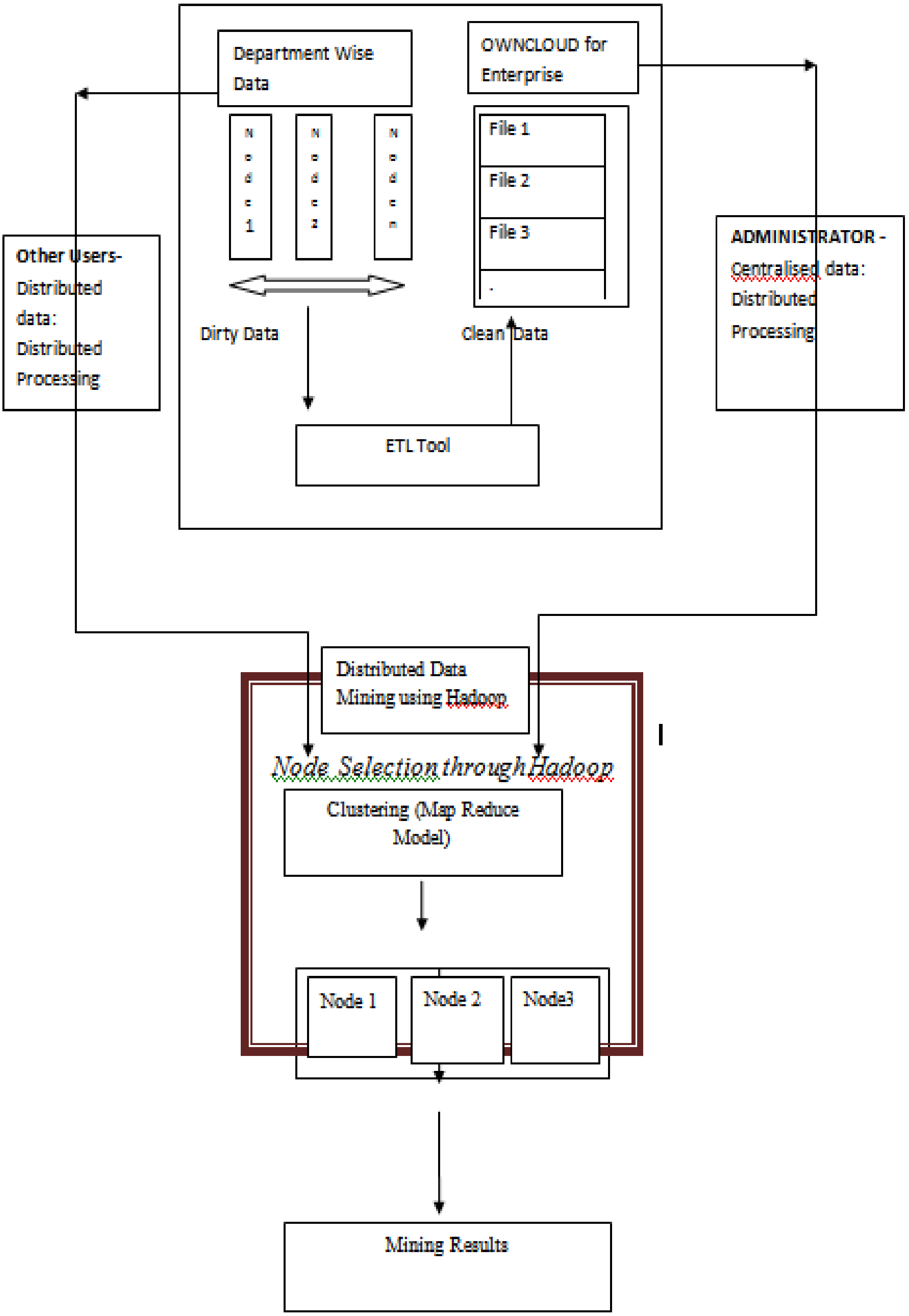

Department Wise Data
OWNCLOUD for Enterprise
Node 1
Node 2
Node n
File 1
File 2
File 3
.
Other Users- Distributed data: Distributed Processing
ADMINISTRATOR - Centralised data: Distributed Processing
Dirty Data
Clean Data
ETL Tool
Distributed Data Mining using Hadoop
Node Selection through Hadoop
Clustering (Map Reduce Model)
Node 1
Node 2
Node3
Mining Results

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

VI. 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.

VII. Acknowledgement

We acknowledge Keshav Kori and Girjesh Mishra studying in M.Tech. IV Semester for technical help.

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/Clustering

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

[7] http://lucene.apache.org/hadoop/

[8] https://bitnami.com/stack/owncloud

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

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

[11] M. Cannataro, D. Talia “*Semantics and knowledge grids: building the next-generation grid*” IEEE Intelligent Systems, 19 (1) (2004), pp. 56–63

[12] Foti, D., et al. "*Scalable parallel clustering for data mining on multicomputers*." Parallel and Distributed Processing. Springer Berlin Heidelberg, 2000. 390-398.