USER ACCESS POLICIES FOR ENHANCEMENT OF PRIVACY IN CLOUD COMPUTING USING DATA MINING

Mandeep Kaur[1], Er. Harinderpal Singh[2]

[1]M.Tech Student, [2]Assistant Professor, CSE Department,

Adesh Institute of Engineering and Technology Faridkot, (India)

ABSTRACT

Data mining in the cloud is the process of extracting structured information from unstructured or semi structured web data sources. To use the full potential of cloud computing, data is transferred, processed, retrieved and stored by external cloud providers. However, data owners are very skeptical to place their data outside their own control sphere. Their main concerns are the confidentiality, integrity, security and methods of mining the data from the cloud. In this thesis we will be enhancing the security by using single cloud provider and dividing single cloud into different zones. We will segregate data by creating virtual partitions of data for saving and allowing user to access data in his partitions only. Each user will have the rights according to the role of the client i.e. role based access policies. Cloud is divided into multiple zones.

Keywords:* *Cloud Computing, Data Mining, User Access Policies.

I. INTRODUCTION

Cloud computing is the convenient and well located on-demand network access model ,that requires minimal management efforts for rapid network access to resources who so ever is ready to take the access. In cloud computing, word cloud is regarded as a metaphor for "the Internet," hence this phrase cloud computing constitutes of the meaning "a type of Internet-based computing," where multifarious services — such as servers, storage as well as applications are delivered to computers of any of the organization and devices via the Internet. It is a developing strategy that provides tremendous merits in economic fields, such as to curtail time to market, efficient and flexible computing capabilities, and infinite computing power. Popularity of cloud computing is burgeoning ubiquitously in distributed computing era. To access the complete potential of cloud computing, data is migrated, processed, retrieved and stored by external cloud providers. Howbeit, data owners are very skeptical to place their data outside their own environmental control sphere. Their important aspects are the confidentiality, integrity, security and methods of data mining from the cloud.

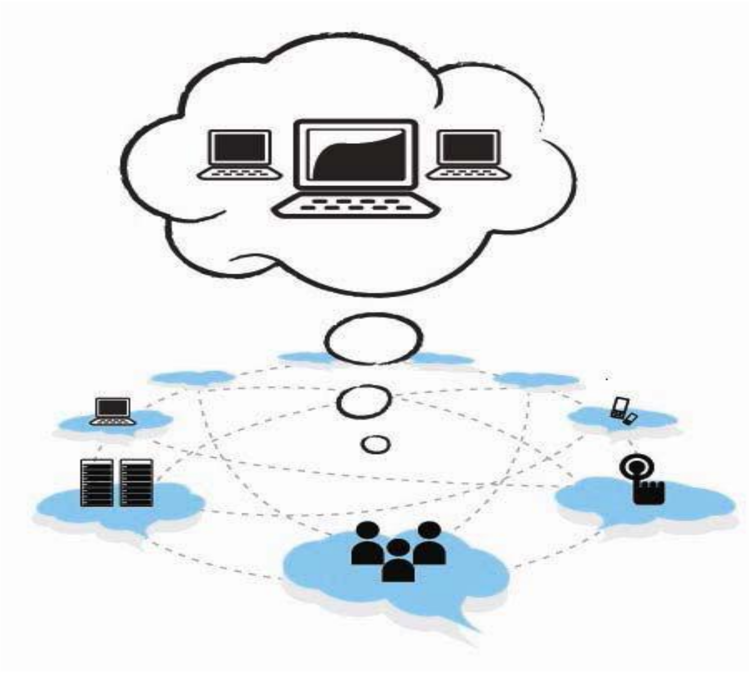


Fig.1.1 Cloud Computing Structural Design [1]

1.1 Data Mining

To extract hidden and predictive information from huge databases, the process is known as data mining. Hence it is the access to provide mining focus on their data warehouses. It predicts future aspects and behaviour, and it also allows the business to be proactive, and helps in providing knowledge- driven decisions.

1.2 Data Mining in the Cloud

When important and useful information is extracted from unstructured or semi structured web data sources, the process is known as data mining in the cloud. So, Cloud providers use this data mining to help the clients in providing a better service. There are several data mining tools like SAS, PAS and IaaS that are used in the field of cloud computing to extract or to retrieve vital information from the databases. There are various growing cloud computing providers for instance; Amazon Web Services, Windows Azure, Open Stack. The main effects of data mining tools being delivered by the Cloud are:

- Paying for the tools that the customer needs helps in curtailing the cost because of the reason that he has not to pay for the data mining tools that he needs – that and he doesn’t have to pay for complex data mining suites that he is not using exhaustive;
- As the customer need not to maintain a hardware infrastructure, because of the reason that he can apply data mining via a browser. This results in curtailing of the hardware costs and has to pay for cloud computing services only.

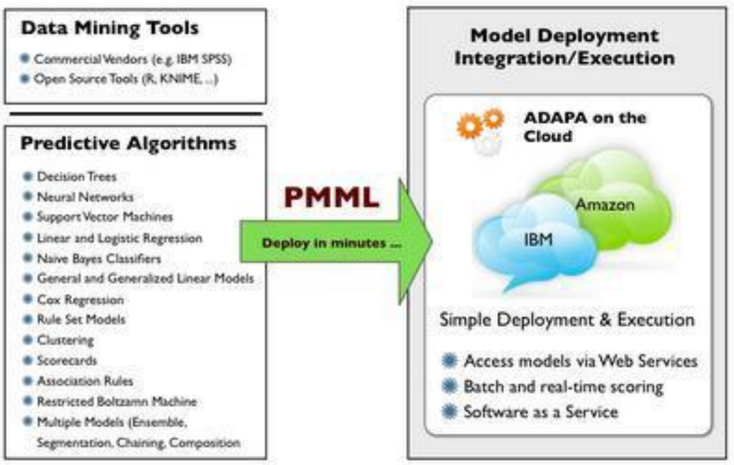

Data Mining Tools
Commercial Vendors (e.g. IBM SPSS)
Open Source Tools (R, KNIME, ...)
Predictive Algorithms
Decision Trees
Neural Networks
Support Vector Machines
Linear and Logistic Regression
Naive Bayes Classifiers
General and Generalized Linear Models
Cox Regression
Rule Set Models
Clustering
Scorecards
Association Rules
Restricted Boltzamn Machine
Multiple Models (Ensemble, Segmentation, Chaining, Composition
PMML
Deploy in minutes ...
Model Deployment Integration/Execution
ADAPA on the Cloud
Amazon
IBM
Simple Deployment & Execution
Access models via Web Services
Batch and real-time scoring
Software as a Service

Fig. 1.2 Data Mining in Cloud Computing [3]

1.3 Data Mining Parameters Include

1. Association – Peeping for the template where one event is allied to another event and also to find rules associated with repeatedly co-occurring articles, used for market basket analysis, cross-sell, and root cause analysis. Also functional for product bundling, in-store placement, and defect analysis.

2. Sequence or path analysis – Peeping for the template where one event leads to another later event.

3. Classification - Looking for new patterns. For predicting a specific outcome such as response / no-response, high / medium / low value customer, likely to buy / not buy; it is the most frequently used procedure.

4. Clustering – To find and visually documenting groups of particulars not formerly acknowledged. Useful for exploring data and finding natural groupings. Members of a cluster are more like each other than they are like members of a diverse cluster. Common instances include finding new customer segments and life sciences discovery.

5. Forecasting - Discovering patterns in data that can lead to reasonable predictions about the future. This area of data mining is known as predictive analytics.

6. Regression- Procedure for prophecy an unremitting numerical conclusion in the manner that customer lifetime worth, house value, process yield rates.

1.4 Applications of Data Mining

1. Student Management
2. Airline Reservation
3. Hospitals
4. Forecasting
5. Biometrics

II. PROPOSED METHODOLOGY

In order to achieve the data mining on mass data, a large number of distributed and parallel data mining algorithms have been proposed. A detailed parallel data mining algorithms, which not only include four major categories of distributed data mining algorithms in association rule learning, classification, clustering, streaming data mining, but also include related research works such as distributed systems, and privacy protection. Data mining algorithms often need to traverse the training data to obtain the relevant statistical information for solving or optimizing the parameters of model. But frequent access on large-scale data requires a lot of compute time.

- Various data analysis techniques are available now a day that are successfully extract valuable information from a large volume of data. These analysis techniques are being used by cloud service providers.
- Attackers can use these techniques to extract valuable information from the cloud.
- By distributing data on different clouds it introduces performance overhead when client needs to access all data frequently, e.g. client needs to perform a global data analysis on all data. The analysis may have to access data from multiple locations, with a degraded performance.
- By simply using In single cloud provider can having the following main issues: Less Security; Loss of data; No privacy; Cost of maintenance is high.

Uploading data on distributed cloud providers: - Although this scenario will protect the client's data as the data will be distributed to the different cloud providers. But it will increase the cost to the client as purchasing different cloud will increase the cost. But using only single cloud also has the issues. So by using single cloud and then dividing the single cloud into multiple zones overcomes the problem of cost and privacy.

Here; user will create his/her own account at the cloud Provider. Cloud Provider will assign the different privileges to the user depending upon the role of the user. Different access policies for different zones will be implemented over here. If the user has been assigned a role as a Read, then he/she can only read the data from the server. If the policy allows writing the data, then only user can write the data into the server. The file sent by the user is stored into the multiple zones available at the server. If the company tries to perform the mining at the user's data, then proper results will not be available.

2.1 Flow Chart of Proposed Work

1. In the first step, Cloud is divided into multiple zones.

2. User will create his/her own account at the cloud Provider.

3. Cloud Provider will assign the different privileges to the user depending upon the role of the user.

4. Different access policies for different zones will be implemented over here.

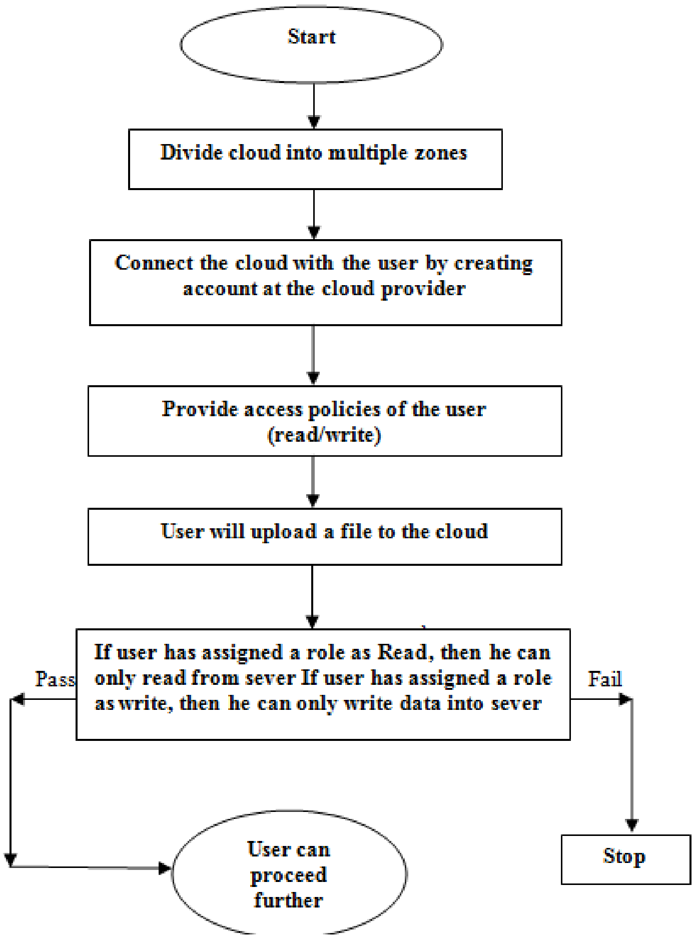
5. If the user has been assigned a role as a Read, then he/she can only read the data from the server.

6. If the policy allows writing the data, then only user can write the data into the server.

7. If the company tries to perform the mining at the user's data, then proper results will not be available.

8. The file sent by the user is stored into the multiple zones available at the server.

Flow chart


Start
Divide cloud into multiple zones
Connect the cloud with the user by creating account at the cloud provider
Provide access policies of the user (read/write)
User will upload a file to the cloud
If user has assigned a role as Read, then he can only read from sever If user has assigned a role as write, then he can only write data into sever
Pass
Fail
User can proceed further
Stop

Distribution of a file in different cloud zones

S.NO	USER NAME	FILE NAME	FILE SIZE	TOTAL PARTITION	PARTITIONED FILENAME	PARITIONED FILE SIZE	CLOUD ZONE
1	CLIENT 1	DEMO.TXT	9MB	3	P1_DEMO.TXT	3MB	ZONE1
					P2_DEMO.TXT	3MB	ZONE 2
					P3_DEMO.TXT	3MB	ZONE 3
2		....	......				

SUMMARY: - The main objective is to partition the data and accessing the role based policies for individual user. As we are using single cloud, hence it reduces the cost. Client 1 uploaded the file (DEMO.TXT) with size 9 MB. The data will be partitioned into three files P1_DEMO.TXT, P2_DEMO.TXT, P3_DEMO.TXT each of size 3 MB. Each file is distributed on three different zones ZONE2, ZONE2, ZONE3 within the cloud. Hence many of the files can be uploaded.

III. OBJECTIVES OF PROPOSED WORK

1. To enhance the security in cloud systems by creating user access policies
2. We will be enhancing the security by using single cloud provider and dividing single cloud into different zones thereby saving a cost of the client and also enhancing the security.
3. We will segregate data by creating virtual partitions of data for saving and allowing user to access data in his partitions only. Each user will have the rights according to the role of the client i.e. role based access policies.
4. Use of virtual partitions and enhanced user access control in cloud system will improve data privacy and thereby fixing the threats in data mining to personal / private data in the cloud systems.

IV. TOOLS USED AND IMPLEMENTATION

4.1 Language

Java is a general-purpose computer programming language that is concurrent, class-based, object-oriented, and specifically designed to have as few implementation dependencies as possible. It is intended to let application developers "write once, run anywhere" (WORA), meaning that compiled Java code can run on all platforms that support Java without the need for recompilation. Java applications are typically compiled to byte code that can run on any Java virtual machine (JVM) regardless of computer architecture. Java is one of the most popular programming languages in use, particularly for client-server web applications, with a reported 9 million developers. Java was originally developed by James Gosling at Sun Microsystems (which has since been acquired by Oracle Corporation) and released in 1995 as a core component of Sun Microsystems' Java platform. The language derives much of its syntax from C and C++, but it has fewer low-level facilities than either of them. However, the overhead of interpretation means that interpreted programs almost always run more slowly than programs compiled to native executables would. Just-in-Time compilers were introduced from an early stage that compiles byte codes to machine code during runtime. Java is platform independent. But as Java virtual machine must convert Java byte code into machine language which depends on the operating system being used, it is platform dependent.

4.2 Simulator

The CloudSim simulation layer provides support for modeling and simulation of virtualized Cloud-based data center environments including dedicated management interfaces for VMs, memory, storage, and bandwidth. The fundamental issues, such as provisioning of hosts to VMs, managing application execution, and monitoring dynamic system state, are handled by this layer. A Cloud provider, who wants to study the efficiency of different policies in allocating its hosts to VMs (VM provisioning), would need to implement his strategies at this layer. Such implementation can be done by programmatically extending the core VM provisioning functionality. There is a clear distinction at this layer related to provisioning of hosts to VMs. A Cloud host can be concurrently allocated to a set of VMs that execute applications based on SaaS provider’s defined QoS levels. This layer also exposes the functionalities that a Cloud application developer can extend to perform complex workload profiling and application performance study. The top-most layer in the CloudSim stack is the User Code that exposes basic entities for hosts (number of machines, their specification, and so on), applications (number of tasks and their requirements), VMs, number of users and their application types, and broker scheduling policies. By extending the basic entities given at this layer, a Cloud application developer can perform the following activities: (i) generate a mix of workload request distributions, application configurations; (ii) model Cloud availability scenarios and perform robust tests based on the custom configurations; and (iii) implement custom application provisioning techniques for clouds and their federation. As Cloud computing is still an emerging paradigm for distributed computing, there is a lack of defined standards, tools, and methods that can efficiently tackle the infrastructure and application level complexities.

4.3 Life Cycle of Clous Sim

1. Begin
2. Initialise cloudsim package.
3. Creation of datacenter.
4. Creation of broker.
5. Creation of virtual machine.
6. Creation of cloudlets.
7. Start simulation.
8. Stop simulation.
9. Output the results.
10. End.

4.4 NetBeans

NetBeans IDE is a free, open source, popular (with approximately 1 million downloads), integrated development environment used by many developers. Out of the box, it provides built-in support for developing in Java, C, C++, XML, and HTML. And this author especially likes the support for editing JSPs, including syntax highlighting, HTML tag completion, JSP tag completion, and Java code completion. NetBeans IDE is available for free downloaded at http://www.netbeans.org Following are step-by-step instructions to help NetBeans IDE greenhorns to get started developing Java applications with NetBeans IDE.

4.5 Getting Started with NetBeans

Assume that you have successfully installed NetBeans on your machine. Start NetBeans from Windows, Linux, Mac OS X, or Solaris. The NetBeans main window appears, as shown in Figure the NetBeans main window is the command center for the IDE. The NetBeans main window contains menus, toolbars, project pane, files pane, runtime pane, navigator pane, and other panes.

The Main Menu- The main menu is similar to that of other Windows applications and provides most of the commands you need to use NetBeans, including those for creating, editing, compiling, running, and debugging programs. The menu items are enabled and disabled in response to the current context.

The Toolbar- The toolbar provides buttons for several frequently used commands on the menu bar. The toolbars are enabled and disabled in response to the current context. Toolbar is faster than using the menu bar. For many commands, you also can use function keys or keyboard shortcuts. For example, you can save a file in three ways: Select File, Save from the menu bar

.• Click the "save" toolbar button ( ).

- Use the keyboard shortcut Ctrl+S.
- TIP: You can display a label known as ToolTip for a toolbar button by pointing the mouse to the button without clicking.

Workspaces- A workspace is a collection of windows that are pertinent to performing certain types of operations, such as editing, execution, output, or debugging. The workspace windows can be displayed from the Window menu.

The basic steps described are as follows.

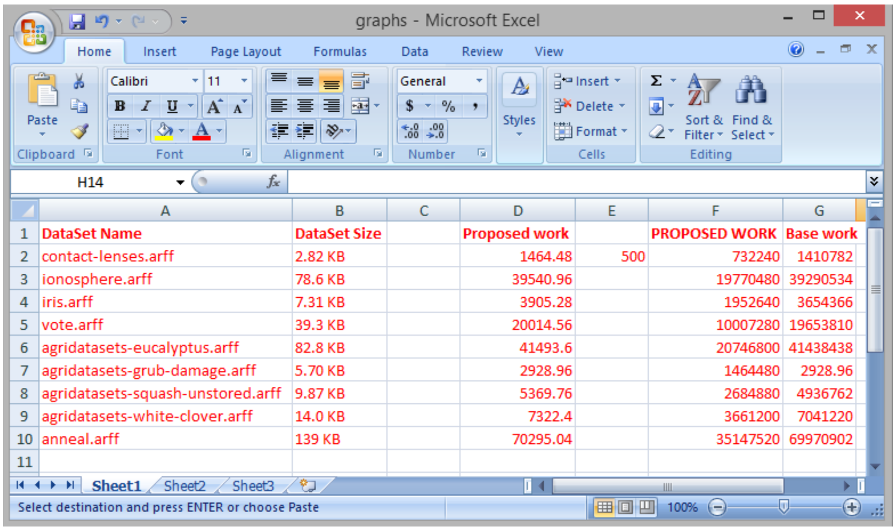
1. Create a new project
2. Mount a directory - specify a location to save project files
3. Add a new class to the project
4. Compile and run a Java program

V. COMPARISON

Comparison will be done on the basis of cloudlets. Data center broker is responsible for the splittion of the zones of a single cloud. As the cloud will be divided into multiple zones, so partitioned data will be assingned to the multiple zones. We are usimg a single cloud so it will result in decreasing the processing cost of attaining different clouds. The execution time will also be decreased because a single cloud zone will have its own datasets file and also different cloud zones will have their own different files. So execution time will be less than that of base work. The privacy will be enhanced because then data will be splitted and assingned to the multiple cloudlet zones. When the user will try to do mining, proper results will not be obtained. Here, we are considering three basics aspects: Processing cost, execution time, and splittion parts.

- **PROCESSING COST**

Cost depends on the number of clouds used. Here also the cost of the clouds and other systems used is less in proposed work as comparative to base work.



	A	B	C	D	E	F	G
1	DataSet Name	DataSet Size		Proposed work		PROPOSED WORK	Base work
2	contact-lenses.arff	2.82 KB		1464.48	500	732240	1410782
3	ionosphere.arff	78.6 KB		39540.96		19770480	39290534
4	iris.arff	7.31 KB		3905.28		1952640	3654366
5	vote.arff	39.3 KB		20014.56		10007280	19653810
6	agridatasets-eucalyptus.arff	82.8 KB		41493.6		20746800	41438438
7	agridatasets-grub-damage.arff	5.70 KB		2928.96		1464480	2928.96
8	agridatasets-squash-unstored.arff	9.87 KB		5369.76		2684880	4936762
9	agridatasets-white-clover.arff	14.0 KB		7322.4		3661200	7041220
10	anneal.arff	139 KB		70295.04		35147520	69970902
11							

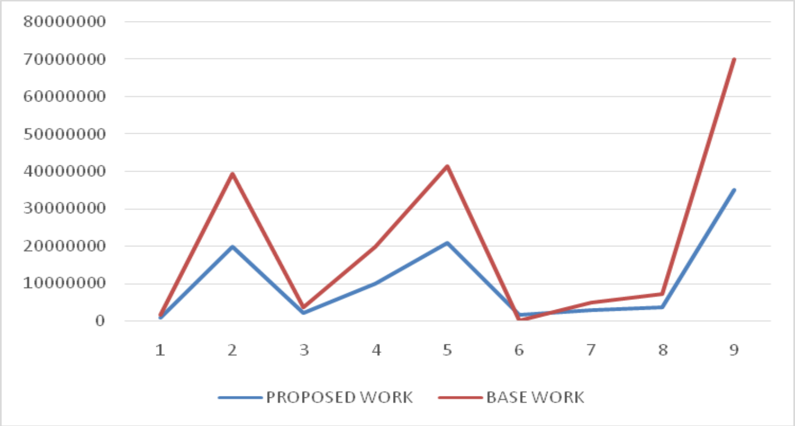
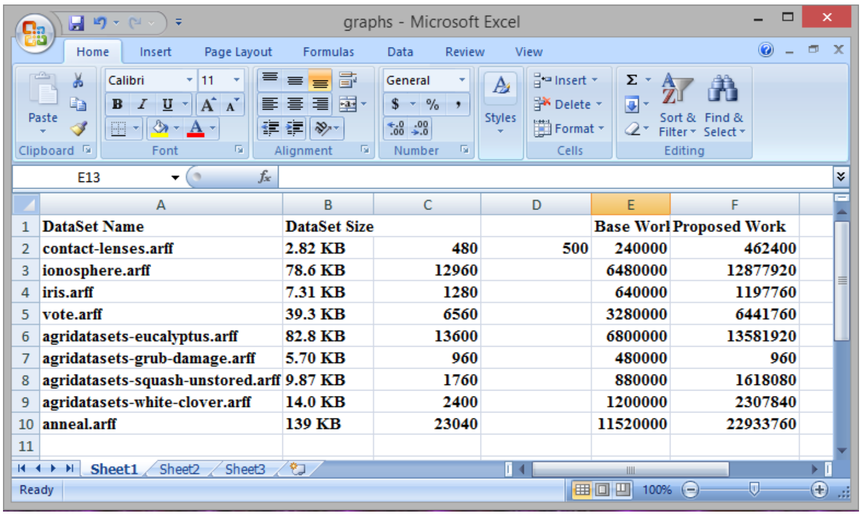


Fig 1.3 Processing Cost

- **Total Execution Time**

The execution time of the base work is more than that of the proposed work considering the number of cloudlets.



	A	B	C	D	E	F
1	DataSet Name	DataSet Size			Base Worl	Proposed Work
2	contact-lenses.arff	2.82 KB	480	500	240000	462400
3	ionosphere.arff	78.6 KB	12960		6480000	12877920
4	iris.arff	7.31 KB	1280		640000	1197760
5	vote.arff	39.3 KB	6560		3280000	6441760
6	agridatasets-eucalyptus.arff	82.8 KB	13600		6800000	13581920
7	agridatasets-grub-damage.arff	5.70 KB	960		480000	960
8	agridatasets-squash-unstored.arff	9.87 KB	1760		880000	1618080
9	agridatasets-white-clover.arff	14.0 KB	2400		1200000	2307840
10	anneal.arff	139 KB	23040		11520000	22933760
11						

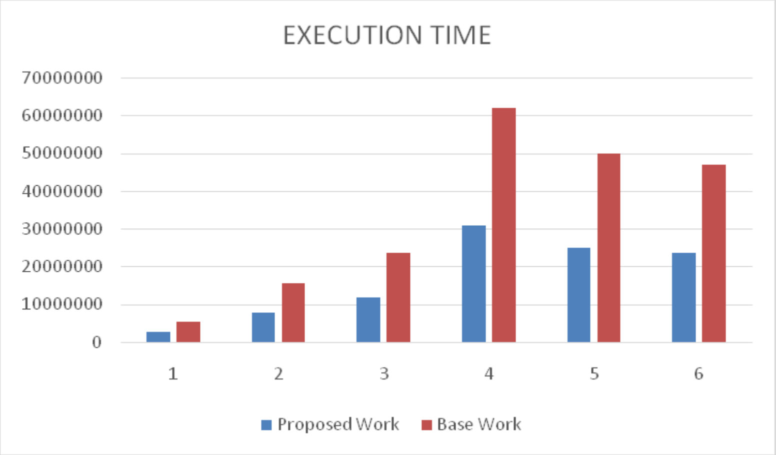
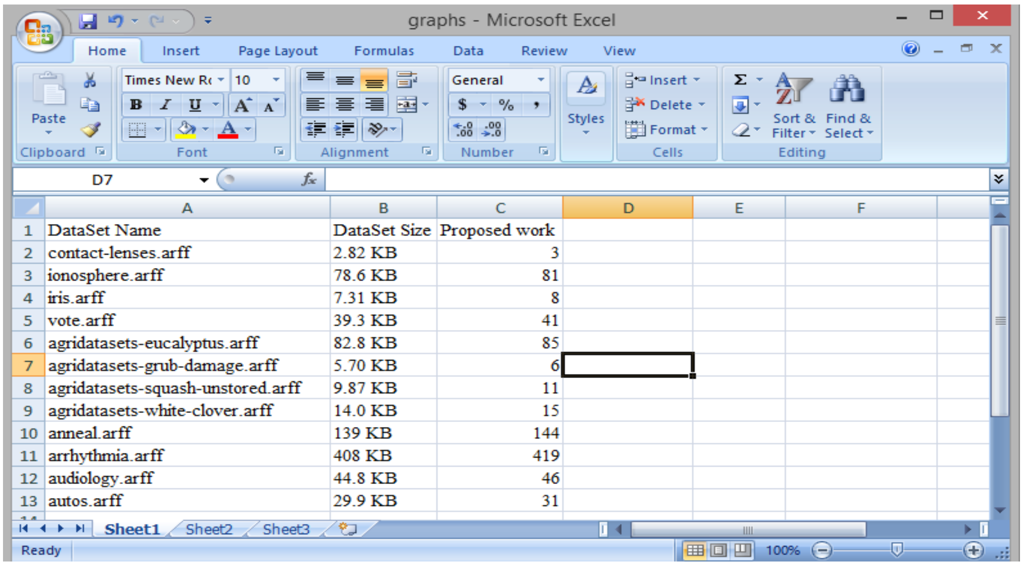


Fig. 1.4 Total Execution Time

- **SPLITTION PARTS -** Data of datasets will be partitioned into several parts according to the size. For instance the file in kb will have less number of partitioned as comparative to the file in mb.



	A	B	C	D	E	F
1	DataSet Name	DataSet Size	Proposed work			
2	contact-lenses.arff	2.82 KB	3			
3	ionosphere.arff	78.6 KB	81			
4	iris.arff	7.31 KB	8			
5	vote.arff	39.3 KB	41			
6	agridatasets-eucalyptus.arff	82.8 KB	85			
7	agridatasets-grub-damage.arff	5.70 KB	6			
8	agridatasets-squash-unstored.arff	9.87 KB	11			
9	agridatasets-white-clover.arff	14.0 KB	15			
10	anneal.arff	139 KB	144			
11	arrhythmia.arff	408 KB	419			
12	audiology.arff	44.8 KB	46			
13	autos.arff	29.9 KB	31			

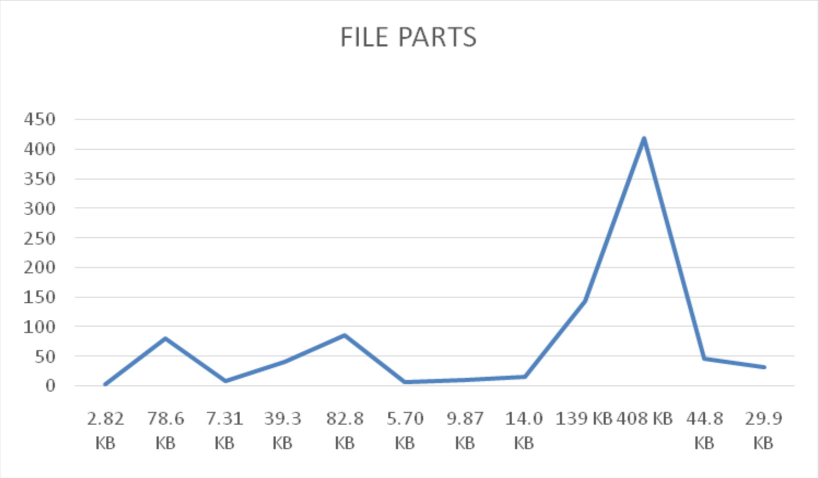
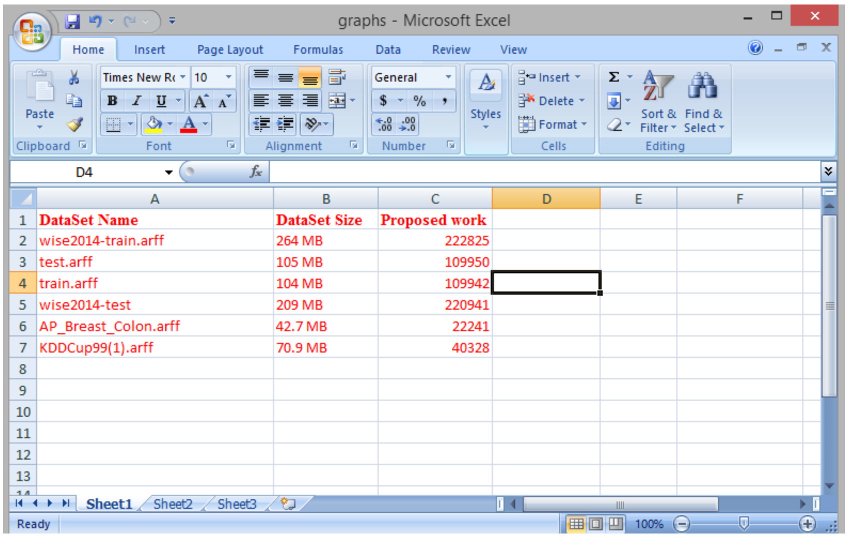


Fig. 1.5 Splittion Parts (kb)



	A	B	C	D	E	F
1	DataSet Name	DataSet Size	Proposed work			
2	wise2014-train.arff	264 MB	222825			
3	test.arff	105 MB	109950			
4	train.arff	104 MB	109942			
5	wise2014-test	209 MB	220941			
6	AP_Breast_Colon.arff	42.7 MB	22241			
7	KDDCup99(1).arff	70.9 MB	40328			

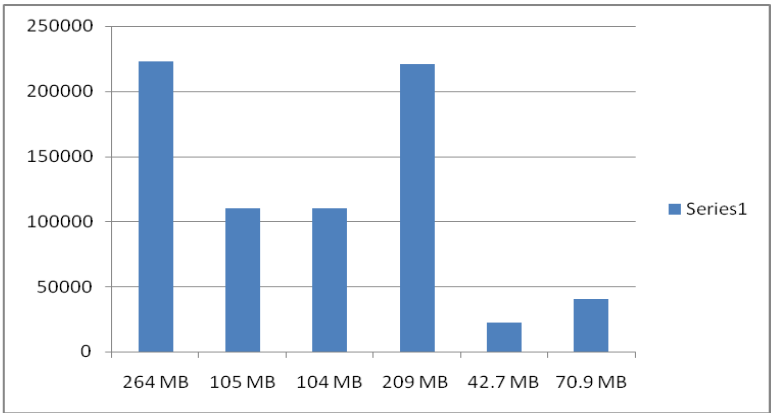


Fig. 1.6 Splittion Parts (mb)

VI. CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

The research work is enhancing the security by using single cloud provider and dividing single cloud into different zones. We will segregate data by creating virtual partitions of data for saving and allowing user to access data in his partitions only. Each user will have the rights according to the role of the client i.e. role based access policies. Cloud is divided into multiple zones. User will create his/her own account at the cloud Provider. Cloud Provider will assign the different privileges to the user depending upon the role of the user. Different access policies for different zones will be implemented over here. If the user has been assigned a role as a Read, then he/she can only read the data from the server. If the policy allows writing the data, then only user can write the data into the server. The file sent by the user is stored into the multiple zones available at the server. If the company tries to perform the mining at the user's data, then proper results will not be available.

6.2 Future Scope

- In the present work we are enhancing the security by dividing single cloud into multiple zones. The data is distributed on the multiple zones within the single cloud only. We are partitioning the data and also assigning user access policies.

- Also, we are considering datasets only. In the new approach data mining will not be carried out in the case of datasets.
- Still there is some hope of improvement. In future, further more files such as text files, CSV files, excel sheets or further any documents should be taken into account.

REFERENCES

[1] A.Raja Rajeswari, R.Sakkaravarthi, “Mitigating Data Mining Attack in Cloud”, International Journal of Innovative Research in Computer and Communication Engineering, ISSN: 2320-9801, Vol. 2, Issue 4, April 2014.

[2] Deepti Mittal, Damandeep Kaur, Ashish Aggarwal “Secure Data Mining in Cloud using Homomorphic Encryption” 2013.

[3] Himel Dev, Tanmoy Sen, Madhusudan Basak and Mohammed Eunus Ali,“ An Approach to Protect the Privacy of Cloud Data from Data Mining Based Attacks”.

[4] Inderjit kaur , Deep Mann : Data mining in the Cloud Computing: Volume 4, issue 3, March 2014.

[5] Introduction to Cloud Computing Architecture‖, Sun Microsystems, 2009.

[6] Jianzong Wang,Zhuo Liu, Peng Wang, “Data Mining of Mass Storage Based on Cloud Computing.”

[7] Mandeep Kaur “Comparative Analysis of Two Fine Grained Data Access Control Techniques in Cloud Computing” volume 5, Issue 5, May 2015

[8] Mr.A.Srinivas, M. Kalyan Srinivas, A.V.R.K.Harsha Vardhan Varma: A Study on Cloud Computing Data Mining, Vol. 1, Issue 5, July 2013.

[9] ORACLE, “Oracle Data Mining Techniques and Algorithm”.

[10] Ruxandra-Ştefania PETRE,“ Data mining in Cloud Computing”, Database Systems Journal vol. III, no. 3/2012.

[11] Shashank Bajpai and Padmija Srivastava “A Fully Homomorphic Encryption Implementation on Cloud Computing” November 2014