IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710


# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 9, Issue 12, December 2020

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 12, December 2020||

# Efficient Segment Based Data Access from Cloud Storage by Using Data Mining Techniques for Secured

**V. Tamilarasi[1], N. Marudachalam[2]**

M.Phil Research Scholar, Department of Computer Science, Presidency College (Autonomous),Kamarajar Salai, Chepauk Chennai, India[1]

Associate Professor, Department of Computer Science, Presidency College (Autonomous),Kamarajar Salai, Chepauk Chennai, India[2]

**ABSTRACT**: In the recent years, most of the companies use cloud computing for storing large amount of data in secured file form. This technique provides lot of benefits to user at low cost and can be retrieval when required. The problem arise not directly is uncontrolled by the user when all people use at one time. The new proposal "Co-ordinate matching" matches the database efficiently and in a well organized manner and also evaluate among the multi-keyword technique to retrieve finally and generally worn the plain text Information Retrieval (IR).By going through the previous papers about existing system, various modules and clustering were found. There are important five modules of existing system, namely preprocessing module, processing module, query retrieving module and so on. In my proposed work there are preprocessing module deals the removing errors and normalizing the data. The processing module deals the procedure of work, and final module gives the various types of the report of visualization technique. In searching the keyword for ten clusters with the time limit (200000.00)ms. The data mining technique was proposed to modify the existing scheme, it was found to be the better and quick access. Even for retrieving the information for the fifteen clusters upon the same time limit (200000.00)ms. It also reveals that it is the technique for the secured preservation of efficient group based data.

**KEYWORDS:** Co-ordinate Matching, Information Retrieval, Data Mining, cluster, Multi-keyword Technique

## I.INTRODUCTION

A large amount of data is available on the internet which is distributive, heterogeneous, dynamic and more complex in nature. Each day has to be dealt with direct advertisement by using data mining technique through which concerned data become more successful at low cost. A huge amount of data could not be administered by conventional data storage system it is critical for traditional analytic tools to analayze for traditional analytic tools to analyze the large amount of data. Cloud Computing information are being centralized in to the cloud as emails, personal health records, company data, and government documents ect.Cloud computing involves the conveyance of computing infrastructure resources as a service of the user. An infinite storage is provided to the users at limited setup and low cost. The organizations are motivated to outsource their data on cloud.There are large number of cloud service providers namely VMware, Microsoft, Google, Salesforce.com, Rackspace and Amazon. This Papers to provide a cluster based search scheme which permits the end user to retrieve the desired documents from the cloud. My Proposed system is to produce the desired results using less comparison and less time.In this paper to summarize, we proposed a multi-keyword technique by using cluster search over encrypted cloud data. In the experimental analysis of synthetic data file we proved the result on proposed search scheme on cloud.

**Theoretical background:**

The cloud model has five characteristics, including deployment models and service models. The cloud computing essential characteristics are broad network access, on-demand self-service measured service, rapid elasticity, resource pooling.

The deployment models of cloud computing are hybrid cloud, community cloud, public cloud and private cloud[15]

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 12, December 2020||

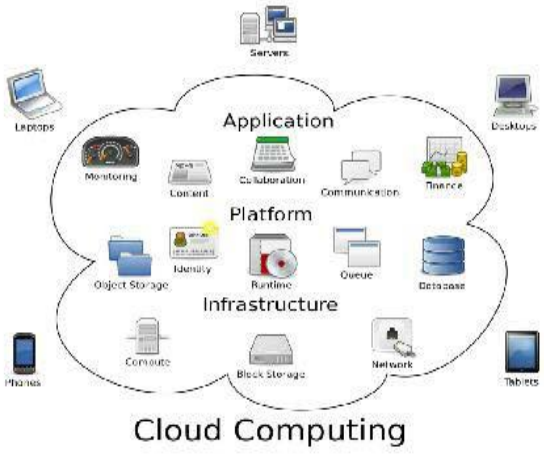



Fig. 1. Cloud Computing [15]

**Cloud Methodology:**

There are three types of cloud services infrastructure as a service, Platform as a service, Software as a Service.

**Infrastructure as a service:**

Delivers compute infrastructure as a utility service, typically in a virtualized environment.

Provides enormous potential for extensibility and scale

**Platform as a Service:**

Delivers a platform or solution stack on a cloud infrastructure

Sits on a top of the Iaas architecture and integrates with development and middleware capabilities as well as database, messaging and queuing functions.

**Software as a service:**

Delivers the application over the internet for intranet via a cloud infrastructure.

Built on underlying Iaas and Paas layer[15]

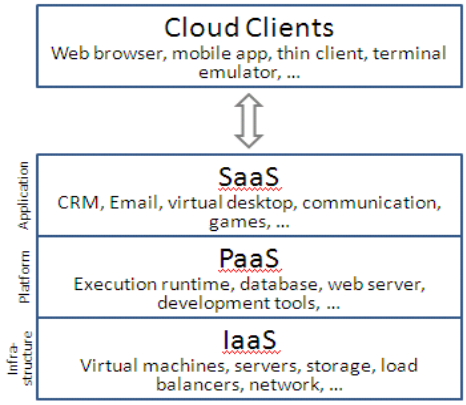



**Fig 2 Layers of Cloud Computing**

**Data mining method:**

It is a very large collection of data. Thousands of individual records are now being compiled data with centralized keywords in warehouse and reorganized globally to make use of the results of powerful statistical relationship and machine learning methods to examine data more comprehensively. Data mining is the art and science of using more powerful algorithm then traditional query tools and records and matching the key by using data mining. Then traditional query tools such as sql to extract the r6ws and columns for powerful information.

It is the term given to the computer process of data preparation, Information extraction and analysis themselves.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 12, December 2020||

**Clustering Analysis:**

Cluster analysis is the task of grouping a set of objects in such a way that objects in the same group to retrieve same document at a time. It is a process where a set of data or objects converted into a set of meaningful sub-classes called cluster. This technique data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval and bio infomatrics.a cluster is used at a time by 20person to retrieve the same 15 document is called cluster. No time limit was found there are several clustering algorithms like k-means clustering, grid based clustering-medioad clustering hierarchal clustering.

**Challenges in document clustering:**

In the first stage, The main challenge is to determine which feature of a document is considered discrimination. In which case of a document model is required. However most clustering approaches the vector space models for representing each document as a vector. Document clustering has been around for decades, but there are certain concerns that necessitate being deal with end for rapid and efficient clustering.

There are large collection of files therefore the computation message digest algorithm is used to encrypt and decrypt the file.

**Data mining in cloud computing:**

Data Mining allows you to perform some basic data mining tasks leveraging a cloud-based Analysis Services connection. Cloud provides the powerful capacities of storage and computing the user can save the document for the free space in data mining techniques. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources. The data mining in cloud computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure service for their users. Cloud provider for a data mining and natural language system, Leading cloud computing providers Amazon web services, Windows Azure, Open stack.

**Effects of Cloud Computing On Information Retrieval:**

In “Search Engine” as an information Retrieval system designed to help find information stored on a computer system[7].The traditional research still have some problems to solve; limited storage capability and computing capability, damaged or senescent hard disk, expensive servers, high cost for maintenance, slow fault detectable velocity etc.

The research of keywords based retrieval from the cloud storage developed vector space model for the efficient ranking of document/files and k-medioad clustering technique works well for grouping documents.

**Relevance Ranking:**

**Ranking** of query is one of the fundamental problems in information retrieval (IR), the scientific/engineering discipline behind search engines. Given a query *q* and a collection *D* of documents that match the query, the problem is to rank, that is, sort, the documents in *D* according to some criterion so that the "best" results appear early in the result list displayed to the user. Ranking in terms of information retrieval is an important concept in computer science and is used in many different applications such as search engine queries and recommender systems. A majority of search engines use ranking algorithms to provide users with accurate and relevant results.

**Boolean Models**

Boolean Model or BIR is a simple baseline query model where each query follow the underlying principles of relational algebra with algebraic expressions and where documents are fetched unless they completely match with each other. Since the query either fetches the document (1) or doesn’t fetch the document (0), there is no methodology to rank them.

**Vector Space Model**

Since the Boolean Model only fetches complete matches, it doesn’t address the problem of the documents being partially matched. The Vector Space Model solves this problem by introducing vectors of index items each assigned with weights. The weights are ranged from positive (if matched completely or to some extent) to negative (if unmatched or completely oppositely matched) if documents are present.

**Tf-IDF:**

Term Frequency - Inverse Document Frequency (tf-idf) is one of the most popular techniques where weights are terms (e.g. words, keywords, phrases etc.) and dimensions is number of words inside corpus.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 12, December 2020||

The similarity score between query and document by can be found by calculating cosine value between query weight vector and document weight vector using cosine similarity. Desired documents can be fetched by ranking them according to similarity score and fetched top k documents which has the highest scores or most relevant to query vector

## II. LITERATURE SURVEY

Many users are currently working on various issues of mining in cloud environment or without cloud environment.So,here we focus on the various scheme on efficient keyword based retrieval from the cloud storage proposed by different researcher including the fundamentals they used.
Authors of [6] described that in recent days, due to the fast development of the information technology, the amount of data increasing every day. Large amount of data carries a lot of information. The difficulty of the organizations or individual may increase for finding relevant knowledge from the large amount of the data. At another side Cloud computing platform have a characteristics like high virtualization and high availability which can perform dynamic resource scheduling and allocation, but this requires a need of data mining concepts. In this paper authors mentioned that by combining existing data mining technology with cloud computing technology is a quite feasible way for achieving efficiency. The authors also described the process of data mining, structure of cloud computing in business module and the structure of data mining platform based on cloud computing .By using simple block diagram. Distributed computing have a two feature like parallel computing and distributed storage. Parallel computing and distributed file storage are provided by cloud computing platform, so it is a good solution on both the two levels. Due to development in cloud computing which brings new developing direction in data mining platform. Due to this new generation of data mining platform is possible.

Authors of [5] described that Recently, Cloud computing is widely used technology. It is highly recommended of security of the data and data privacy in cloud computing, because many organization use a cloud platform to store the day, among of them some are personal which may kept secret, so the data to be stored on cloud platform in encrypted form. This may require additional processing and the care should be taken that, the burden on overall system should not be increased. In this research paper authors gives a detail review of various techniques by listing some limitations also, which helps user for storing data in secured manner and efficiently access of the data from cloud server. After that authors of [5] introduce a very efficient and secure system which decrease the burden of the system and also decrease complexity and enhance overall performance of the system. But authors doesn't shows any simulation results.

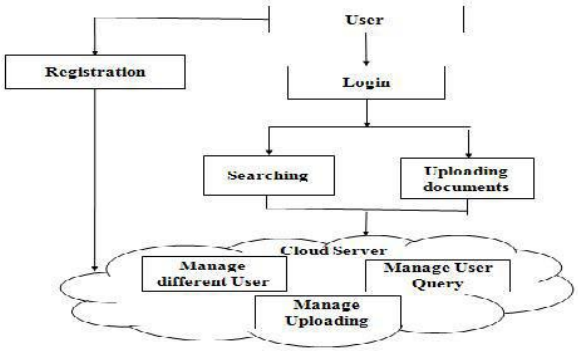
Authors of [16] proposed a scheme for securely keyword based searching mechanism. The user send a query and Cloud server retrieve matched file which contains the keyword and send back to client. They used privacy preserving mechanism to protect their data from unauthorized access. They compared proposed cryptographic primitive with the existing primitive cryptographic mechanism. The authors proposed data security but it is applicable to single keyword search. The proposed scheme name order-preserving symmetric encryption (OPSE) they guaranteed that this scheme works efficiently as compared to other similar approaches available. They also shows the practical experimental resultfor displaying efficiency of proposed scheme.

Authors at [1]proposed a model where they focus on file searching based on multiple keywords. They argued that The traditional keyword based search supports Boolean search which shows weather file may contain the keyword or not, without any relevance of datafiles.And the ranked based file retrieval using a single keyword having a poor result. Authors also mentioned that ranking on server side which is based on order-preserving encryption (OPE) breaks privacy of data. So, the authors provide scalable system with minimize information leakage. Their model prevent overload by working at user side for ranking files, where consume less bandwidth. They perform analysis which shows efficiency of their proposed solution.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

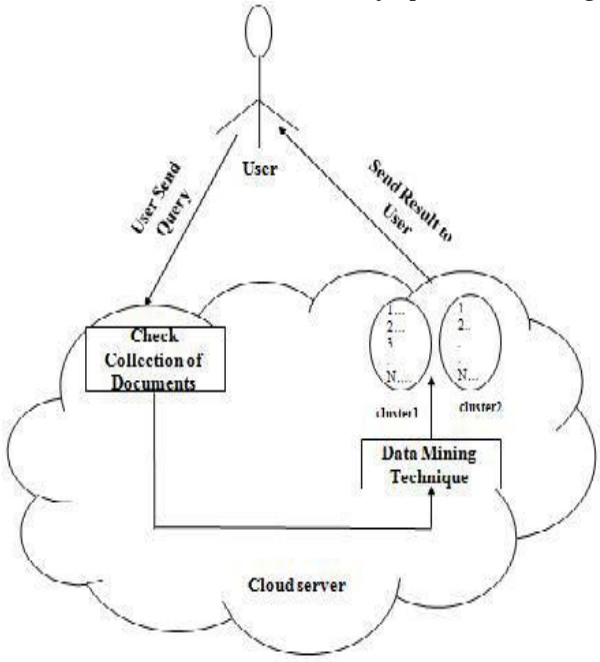
||Volume 9, Issue 12, December 2020||

## III.PROPOSED WORK




The cloud computing has three entities like cloud server, data user, data owner. The data owner has collection of the file that he wants to upload the file on the cloud server. The document must be secure in third person where hacking the file. If not understand the file for encrypt/decrypt format. The data owner expects the cloud server is muliti-keyword based data retrieval as a service to the data user. The user is one of the member to easily to get the files use for mulikeyword techniques is used. The cloud server manage the different user and query to uploading the documents themselves.

The below figure 3 the proposed multikeyword technique from the cloud server. The user helps to send a query on server then server first check the collection of file which are already uploaded in an organization




Note that we apply the k-medioad clustering technique to form cluster which gives accurate result based on the multi-keyword. After the user query is converted in to vector by using Euclidian distance is calculated between the query vector by using TF-IDF scoring. The cluster and the document having the minimum distance is given to the user as a result.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 12, December 2020||

A small collection of document with TF-Idf

| | V1 | V2 | V3 | V4 | V5 | V6 |
|---|---|---|---|---|---|---|
| Data1 | -9.18475 | -1.9469 | -1.1222 | 3.6305 | 1.1941 | 1.4102 |
| Data2 | -2.38570 | 3.7649 | -0.5288 | 1.1173 | -0.6212 | 0.0.286 |
| Data3 | -5.72886 | 1.0742 | -0.5513 | 0.9113 | 0.1769 | 0.5410 |
| Data4 | -7.11669 | -10.266 | -3.2299 | 0.1524 | 2.9563 | 3.0507 |
| Data5 | -3.93184 | 1.9464 | 1.3885 | 2.9381 | 0.5463 | -1.2254 |
| Data6 | -2.37815 | -3.9465 | -2.9323 | 0.9402 | 1.0351 | -0.4506 |
| Data7 | -2.23692 | 2.6877 | -1.6385 | 0.1492 | -0.0403 | -0.1288 |
| Data8 | -2.14141 | -2.3382 | -0.8712 | -0.1269 | 1.4262 | -1.2559 |
| Data9 | -3.17213 | -3.3888 | -3.1172 | -0.6008 | 1.5210 | 0.5591 |
| Data10 | -6.34616 | -7.7204 | -4.3381 | -3.3722 | -1.7088 | -0.7233 |
| Data11 | 0.80970 | 2.6569 | 0.4884 | -1.6711 | -0.2756 | 0.1272 |
| Data12 | -2.64876 | 0.0665 | -1.5251 | 0.0512 | -0.3317 | 0.7642 |
| Data13 | -8.17784 | -2.6986 | 5.7252 | -1.1113 | -1.0426 | 2.5923 |
| Data14 | -0.34183 | 0.9674 | 1.7157 | -0.5945 | -0.4676 | 1.0068 |
| Data15 | -4.33857 | -4.8568 | -2.8136 | -1.4533 | -1.2889 | -0.3494 |
| Data16 | -4.07207 | -2.9744 | -3.1225 | -2.4559 | 0.4080 | 0.4953 |
| Data17 | -0.22985 | 1.5634 | -0.8018 | -0.6500 | 0.4943 | -0.7615 |
| Data18 | -4.41413 | -1.4174 | -2.2683 | -0.1861 | 1.4226 | -0.7518 |
| Data19 | -4.94435 | 4.1107 | -0.3145 | -0.0881 | 0.0567 | -1.1367 |
| Data20 | 0.1888 | -0.5928 | 1.5949 | 0.4418 | -0.0486 | -0.0486 |

**Modules:**
The proposed system can be described into following number of modules

**Data owner modules**
Three different modules are used in cloud computing environment: cloud server, data owner and data user. Data owner first to register in the cloud. The data owner can upload the cloud with multi-keywords. Data owner can retrieve the documents by using multi-keywords.the data user have an authorization to search the retrieved documents using multikeyword.

**Encrypted file modules**
Data owner can upload the file to cloud server. Whole the data is encrypted by server and stored to cloud server.

**Keyword Indexing Module:**
This module multikeyword from document will be extracted and index of such keywords along with the list of documents which contains multi-keyword. This index is also encrypted. Index building is done using MD5 algorithm

**User query module:**
The user searches the document by using multi-keyword. This document is calculated for searched keywords and documents by using multi-keywords they are matched. The result will be displayed on the file

**Document Retrieval Module:**
This file is decrypted file is given to the user if authentication is made between them. whole process done only in encrypted form.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 12, December 2020||

.

**MD5 ALGORITHM:**

**Data owner functions:**

Enter and register in to cloud computing system
Upload the audio and video files etc
Check the above files mounted or not.

**Data user functions :**

Access the particular file to the cloud
Cloud send the result in the best way
Comparing with existing models.

MD5 (technically called *MD5 Message-Digest Algorithm*) is a cryptographic hash function whose main purpose is to verify that a file has been unaltered.

Instead of confirming that two sets of data are identical by comparing the raw data, MD5 does this by producing a checksum on both sets and then comparing the checksums to verify that they're the same.

MD5 has certain flaws, so it isn't useful for advanced encryption applications, but it's perfectly acceptable to use it for standard file verifications.

**Using an MD5 Checker or MD5 Generator**

Microsoft File Checksum Integrity Verifier (FCIV) is one free calculator that can generate the MD5 checksum from actual files and not just text. See our article on how to verify file integrity in Windows with FCIV to learn how to use this command-line program.

One easy way to get the MD5 hash of a string of letters, numbers, and symbols is with the Miracle Salad MD5 Hash Generator tool. Plenty of others exist as well, like MD5 Hash Generator, Passwords Generator, and OnlineMD5.

When the same hash algorithm is used, the same results are produced. This means that you can use one MD5 calculator to get the MD5 checksum of some particular text and then use a totally different MD5 calculator to get the exact same results. This can be repeated with every tool that generates a checksum based on the MD5 hash function.

**More Information on the MD5 Hash**

MD5 hashes are 128-bits in length and are normally shown in their 32 digit hexadecimal value equivalent. This is true no matter how large or small the file or text may be.

Here's an example:

- Plain text: **This is a test.**
- Hex value: **120EA8A25E5D487BF68B5F7096440019**

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 12, December 2020||

When more text is added, the hash translates to a totally different value but with the same number of characters:

- Plain text: **This is a test to shows how the length of the text does not matter.**
- Hex value: **6c16fcac44da359e1c3d81f19181735b**

In fact, even a string with zero characters has a hex value of *d41d8cd98f00b204e9800998ecf8427e*, and using even one period makes the value *5058f1af8388633f609cadb75a75dc9d*.

For example, even though **a** = **0cc175b9c0f1b6a831c399e269772661** and **p** = **83878c91171338902e0fe0fb97a8c47a**, combining the two to make **ap** produces a totally different and unrelated checksum: **62c428533830d84fd8bc77bf402512fc**, which can't be pulled apart to reveal either letter.

However, what's really happening with a decrypt or, or "MD5 reverse converter," is that they create the checksum for *lots* of values and then let you look up your checksum in their database to see if they have a match that can show you the original data.

MD5Decrypt and MD5 Decrypted are two free online tools that serve as MD5 reverse lookups but they only work for common words and phrases.

See What Is a Checksum? for more examples of an MD5 checksum and some free ways to generate an MD5 hash value from files.

## IV. CONCLUSION

The proposed scheme use data mining for group of information retrieved is same time limit for 15cluster so retrieve the data quickly. It has security and mining is produced by takes large data set users are added. To manage everything along the time limit is same for 10cluster (200000.00)ms as same time 15clusters also. Here time and space complexity is maintained. we create another idea like application based cloud server. Here it maintains a specific number of cluster in uniform range. In future it may include flexible range limit with the suitable time and space complexity in cloud computing integrated environments

## REFERENCES

1. D.Pratiba, Dr.G Shobha, Vijay Lakshmi P,"Efficient Data Retrieval from Cloud Storage Using Data Mining Technique", International Journal on Cybernetics and Informatics, Volume 4, No 2, pp.271- 279, 2015.
2. Rohit Handa, Rama Krishna Challa,"A Cluster Based Multi-Keyword Search on Outsourced Encrypted Cloud Data", IEEE 2ndInternational Conference on Computing for Sustainable Global Development, pp.115-120, 2015.
3. R.Kabilan, Dr.N.Jayaveeran,"Survey of Data Mining Techniques in Cloud Computing", International Journal of Scientific Engineering and Applied Science, Volume 1, Issue-8, pp.123-127, 2015.
4. Ms. M.R.Girme, Prof.G.M.Bhandari, "Efficient Ranked Keyword Search for Achieving Effective Utilization of Remotely Stored Encrypted Data in Cloud", International Journal of Application or Innovation in Engineering and Management Volume 3, Issue 6, pp.105-113, 2014.
5. Akshay D Kapse, Piyush K Ingole, "Secure and Efficient Search Technique in Cloud Computing", IEEE Fourth International Conference on Communication System and Network Technologies, Issue 10.1109, pp.743-747, 2014.
6. Zhu Jia, A, Zhang Ping,"Design and Implementation of Data Mining Platform Based On the Cloud Computing", IEEE Workshop on Advanced Research and Technology in Industry Application, pp.163-165, 2014.
7. Mazhen zhong, "Research of Information Retrieval in the Cloud Computing Environment", IEEE 7th International Conference on Intelligent Computation Technology and Automation, Issue 10.1109, pp.476-479, 2014.
8. Harneet Khurana,Kailash Bahl," An Approach to Mine Frequent Itemsets in Cloud Using Apriori and FP-tree Approach", International Journal of Computing and Technology, Volume 1, Issue 7, pp.387-389, 2014.
9. Ms Aishwarya S. Patil, Ms Ankita S. Patil," A Review on Data Mining Based Cloud Computing", International Journal of Research in Science & Engineering E -Volume 1, Special Issue: 1, pp.1-4, 2014.
10. N. Janardhan, T. Sree Pravallika, Sowjanya Gorantla,"An Efficient Approach for Integrating Data Mining Into Cloud Computing", International Journal of Computer Trends and Technology - Volume 4, Issue 5, pp.1291-1294, 2013.
11. Naskar Ankita, Mrs Mishra Monika R," Using Cloud Computing To Provide Data Mining Services", International Journal of Engineering and Computer Science - Volume 2, Issue 3, pp.545-550, 2013.
12. Ms Aishwarya S. Patil, Ms Ankita S. Patil," A Review on Data Mining Based Cloud Computing", International Journal of Research in Science & Engineering E -Volume 1 Special Issue 1, pp.52-55, 2012.
13. Ruxandra-Ştefania PETRE,"Data Mining In Cloud Computing ", Database System Journal – Volume 3, pp.67-71, 2012.
14. Vishal Jain, Mahesh Kumar Madam," Information Retrieval through Multi-Agent System with Data Mining in Cloud Computing ", IJCTA,volume 3, issue 1, pp.62-66, 2012.
15. Bhagyashree Ambulkar, Vaishali Borkar," Data Mining In Cloud Computing", MPGI National Multi Conference, pp.23-26, 2012.
16. Ning Cao et al., "Secure ranked keyword search over encrypted cloud Data", IEEE International Conference of Distributed Computing Systems (ICDCS), Genoa, Italy, pp.253–262, 2010.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details