Data retrieval from cloud with cache implementation

K.Venkata Raju [#1], P. Dinesh[*2], M.Sudheerraja[#3]

[1]Assoc.professor at Department of Computer Science Engineering,
KL University, India
Email: kvraju@kluniversity.in

[2] M.Tech student at Department of Computer Science Engineering,
KL University, India
Email: puttadinesh@outlook.com

[3] M.Tech student at Department of Computer Science Engineering,
KL University, India
Email: sudheerraja@outlook.com

***Abstract*— Cloud has the capability to store large volumes of data in it, consists of both historic and present data. When a user wants to find particular information (record) or data in the huge amount of information either the person have to search manually or by using a help. This is the limitation for using data mining concept in the cloud environment**

***Keywords*— data mining, cache memory, k-means, clustering.**

I. INTRODUCTION

Everyone has an opinion on what is cloud computing. It has the ability to rent a server or a thousand servers and run a geophysical modelling application on the most powerful systems available anywhere. It can be the ability to rent a virtual server, load software on it, turn it on and off at will, or clone it ten times to meet a sudden workload demand. It can be storing and securing immense amounts of data that is accessible only by authorized applications and users [1]. Now a day`s multinational companies are creating their own public cloud to store their information such as the project information , their employment information and their pay scales moving to store their information in cloud as a backup purpose. The cloud computing is providing different technologies, among them storage as a service, which can store large volumes of data in it.

A cache is an area of local memory that holds a copy of frequently accessed data that is otherwise expensive to get or compute [3][4]. The caching memory stores a little volume of data init by using least recently used, least frequently used, most recently used and most frequently used out of these techniques if the data is found it will displays the information and sends backs to the user.

If any data is not found it will searches from the storage servers by using a mining technique and displays the data to the client. At the same time the storage server makes a copy of data into the cache memory for the future purposes. If we need the same information again it will searches the data in cache memory and cache memory having a copy in it easily send the data to the user

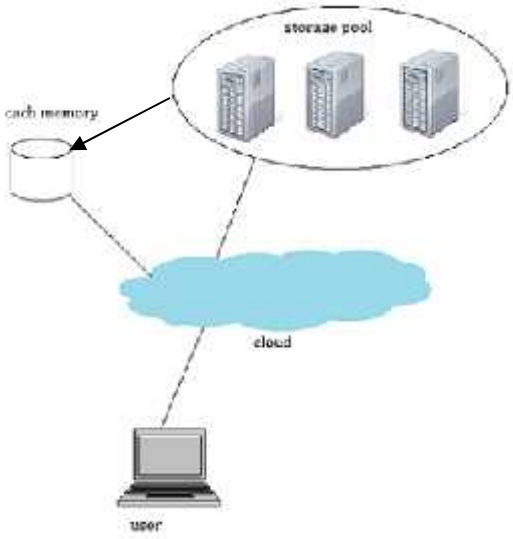

Fig 1: our proposed architecture

II. STORAGE-AS-A-SERVICE:

Storage-as-a-service is the next form of service available in cloud computing. Where Software-as-a-service and Platform-as-a-service are providing applications to customers, but it simply offers the hardware so that user can put whatever data they want onto it. It allows the service providers to rent their storage to the different kind of users[1]. Many cloud storage providers are active on the market, offering various kinds of services, basic cloud storage services are generally not designed to be accessed directly by users but rather incorporated into custom software using application programming interface (API)[2]. Example of such basic cloud storage services are Amazon S3

There are multiple cases for cloud storage services used by both companies and individuals. This includes on-demand storage capacities accessible from various locations (e.g. from mobile and local devices), backup facilities without the need to maintain hardware devices or appropriate software tools, and synchronization features allowing the customers to always have access to the latest version of their data independent of the device [2]. The following cases shall present the potential benefits of such storage services.

a. **Copy**: consider the scenario bob loses his laptop, the cloud is providing a new feature named copy. Using this feature of a cloud storage service Bob would be able to solve this problem in cloud. Customer may manually store single files or folders in online by using web browser, or he may use client software provided by the cloud such software has to be locally installed by the customer and may be used for the automatic uploads. His laptop continuously copies all changes to existing data and all new data and stores them in cloud.

b. **Backup:** The backup feature allows recovering any version of a previously stored file or directory over a long period of time, usually many years. Creating backups using cloud services is an automated process of periodically making copies of data, transmitting these copies to and storing them in the cloud so that they may be used to restore the original after a data loss event [2]. It offers software to be installed locally, enabling the customer to select the data to be baked up, to configure the retention period as well as a schedule for the backups. The client software either runs continuously in the background so that newly created or changed files are backup on a regular basis.

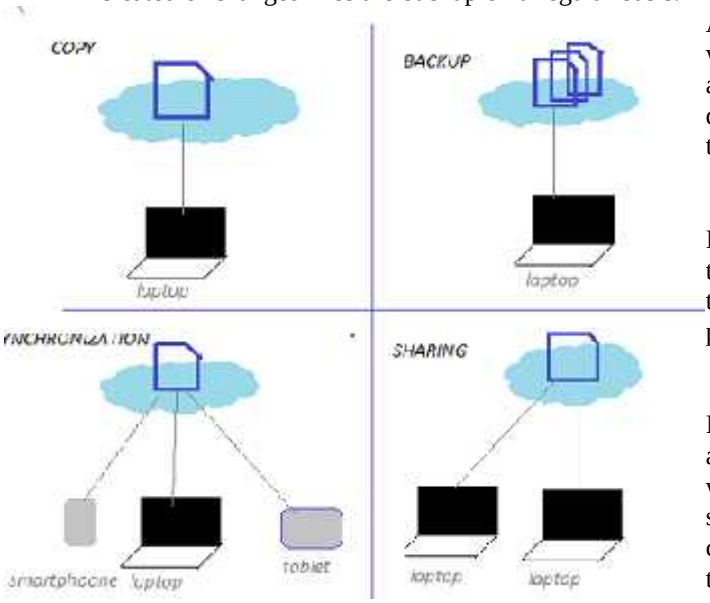

Fig 2: services provided by storage as a service

c. **Synchronization:** Synchronization is the process of establishing consistency among data from different sources [2]. The typical user has a set of devices such as laptops, tablet, and Smartphone and to have all data available on different devices. The client software must be able to detect conflicts that occur if a file has been changed on two devices in different ways. The software should offer a number of choices to the user, either merge the files; overwrite one version, or keeping both versions by applying a renaming scheme.

d. **Sharing:** Data sharing is the process of sharing data with other subscribes or with a closed group of people, users want to collaborate their information among them [2]. Depending on the service, the shared data has a set of fixed or configurable access rights like read, write, upload or delete.

III. INFORMATION HANDLING:

There are different kinds of information that are being handled in the cloud environment they are

a. **General/Usage data**: it contains the readable information that is being stored in our hard disks, laptops example word docs, pdf, etc.

b. **Unique data**: it contains the unique information each and every person has its own unique information, example social security number, ipaddress, etc.

c. **Business information**: it contains the business logical information in which data is stored in the form of a record or in the form of XML sheets. In this data is in the form of records.

Among the above Business information is best suitable for our work which consists of data that is being stored in the form of a table the information may be any sales record, banking details, by using a query based techniques we are obtaining the particular record information with in no time.

IV. EXISTING SYSTEM:

In existing system we are using the manual search by using the find help , it will searches the related information and at the same time it consumes a lot of time for obtaining the particular data

V. PROPOSED SYSTEM:

In the proposed system we are using the caching algorithm and the same time using a data mining algorithm. The cloud will sends the query to both the cache memory and to storage serves at the same time, if the particular records is find in the cache memory or in the storage server it will sends the result to the user with in no time.

VI. CACHE MEMORY:

A cache is an area of local memory that holds a copy of frequently accessed data that is otherwise expensive to get or compute[3]. In our scenario the cache memory is located in between the storage server and the cloud.

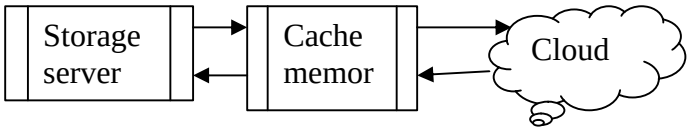

Fig 3: cache memory block

a. **Least frequently used (LFU**): The physical memory which is completely occupied and the substitute resembles the new data that needs to be swapped in to the physical memory, a counter is maintained which keeps the record consisting number of times the data is frequently is obtained[4].

b. **Least recently used (LRU**): In this scenario where in the physical memory is fully occupied with data and there is a new data which needs to be swapped into the physical memory [4]. In this a time stamping is maintained which keeps track of the last time when the data was referred.

c. **Most frequently used (MFU**): In this it also maintains an record of keeping which data is most frequently obtained in the data base.

d. **Most recently used (MRU**): A time stamping is maintained which keeps the track of the most recent time the data was referred in the physical memory. Memory is fully occupied with data and there is a new data which is to be swapped into the physical memory the data which is having lowest preference is removed and it is replaced with the new data.

e. **Not recently used (NRU**): NRU uses a data reference bit which is reset to 0 at regular time intervals. When the data is referenced again the data reference bit is set to 1 to indicate that the page was referenced during the current intervals. Now the data replacement is done by identifying the data which has this bit as 0 [3][4]. If there are multiple such data then one of the data would be swapped out based on some more additional parameters.

VII. DATA MINING:

Technically, data mining is the process of finding correlations or patterns among fields in large relational databases [5]. Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information .Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions, categorize it, and summarize the relationships identified [5].

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally,

it enables them to "drill down" into summary information to view detail transactional data. Data mining analyses relationships and patterns in sorted transaction data based on open-ended user queries.

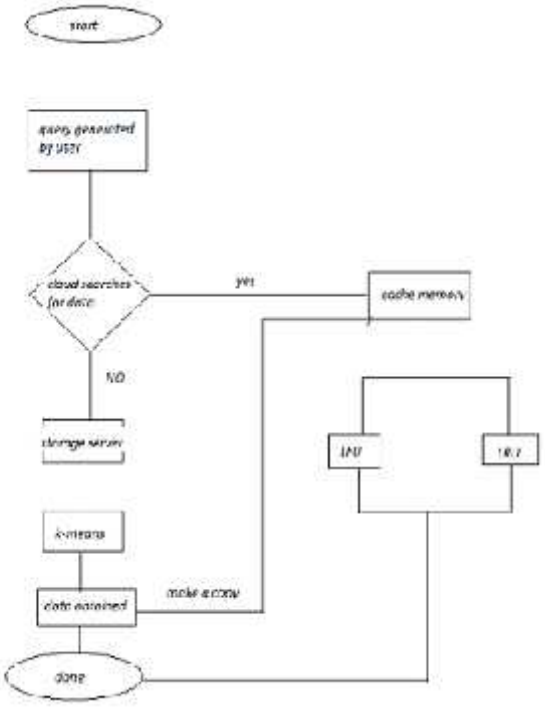

Fig 4: flow diagram for retrieving of data from cloud

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data. Data mining analyses

relationships and patterns in sorted transaction data based on open-ended user queries.

a. **Classes**: Stored data is used to locate data in predetermined groups [5]. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
b. **Clusters**: Data items are grouped according to logical relationships or consumer preferences [5]. For example, data can be mined to identify market segments or consumer affinities.

c. **Associations**: Data can be mined to identify associations [5]. The beer-diaper example is an example of associative mining.

d. **Sequential patterns**: Data is mined to anticipate behaviour patterns and trends [5]. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Out of the above mentioned methods we are using the clustering mechanism out of which K-means clustering algorithm is going to be used

K-means clustering algorithm:

K-means algorithm is fast, robust and easier to understand [7]. It is one of the simplest unsupervised learning algorithm that solve the well known clustering problem the procedure follows a simple and easy way to classify a given data set through a certain number of clusters K fixed a priority[7]. The main idea is to define k-centers, one for each cluster. The centers should be placed in a cunning way because of different locations causes different outcomes. The next step is to take each point belonging to a given data set and associate it to the nearest center [7][8]. The first step is completed and an early grouping is done. At this point we are going to re-calculate the new centroids as bary center of the clusters resulting from the previous method. After all this we have obtained the new k-centroids, a new binding is to be done between the same data set points and the nearest new center [7][8]. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

$$J(v)=\sum_{i=1}^{d}\sum_{j=1}^{Di}(||Pi-Qj||)^2$$

Where,

'$||p_i - q_j||$' is the Euclidean distance between x_i and v_j.

'D_i' is the number of data points in i^{th} cluster.

'd' is the number of cluster centers.

Implementing k-means clustering algorithm:

Requirements: Let $P = \{p_1,p_2,p_3,\ldots\ldots,p_n\}$ the set of data points and

$Q = \{q_1,q_2,\ldots\ldots,q_c\}$ be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate distance between each data point and cluster centers.

3) Assign an data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) For obtaining new center recalculate the step using:

$$J(v)=\sum_{i=1}^{d}\sum_{j=1}^{Di}(||Pi-Qj||)^2$$

Where, 'D_i' represents the number of data points in i^{th} cluster.

5) Calculate distance between each data point and new obtained cluster centers.

6) if no new data point is obtaintained then stop, otherwise repeat from step 3).

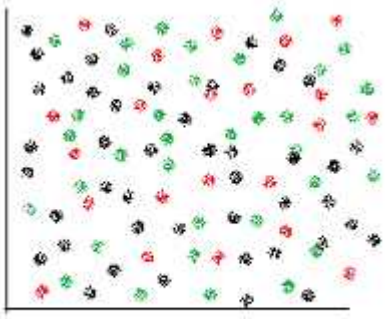

Fig 5: applying k-means algorithm

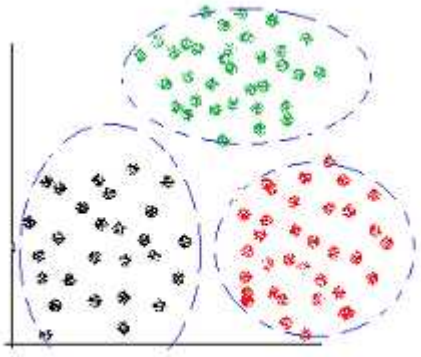

Fig 6: clustered data after applying k-means

VIII. CONCLUSION:

By using the above technique we can retrieve the data with in less time when compared with the previous model. We can extend the paper for a further study of how to keep the fetched information in Cache depending on the utility of the information this proposal can further extended with effective cache utilization algorithms.

IX. REFERENCES:

[1]ANTHONY T. VELTE,TOBY J. VELTE, PH.D. ,ROBERT ELSENPETER " CLOUD COMPUTING A PRACTICAL APPROACH

[2] MORITZ BORGMANN, TOBIAS HAHN, ET.AL "ON THE SECURITY OF CLOUD STORAGE SERVICES"

[3] D M DHAMDHER "OPERATING SYSTEMS" ISBN 0-07-1-061194-7

[4] GALVIN, ET.AL "OPERATING SYSTEM PRINCIPLES" ISBN-978-81-265-0962-1

[5] PIETER ADRIAANS, DOLF ZATINGE "DATA MINING" ISBN 978-81-317-0717-3

[6]TAPAS KANUNGO, DAVID M. MOUNT "AN EFFICIENT K-MEANS CLUSTERING ALGORITHM: ANALYSIS AND IMPLEMENTATION" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL 24, NO, 7. JULY 2002

[7] NOTES BY TAN, STEINBACH, KUMAR GHOSH "THE K-MEANS ALGORITHM"