Data Management in Cloud based Environment using k-Median Clustering Technique

Kashish Ara Shakil
Department of Computer Science
Jamia Millia Islamia
New Delhi, India

Mansaf Alam
Department of Computer Science
Jamia Millia Islamia
New Delhi, India

ABSTRACT
Cloud Computing is the use of computing resources in a new and technologically more advanced manner. It is a fine blend of service oriented characteristics and utility based computing. The growth rate of data in cloud environment is reaching an exponential rate. Therefore there is a need to manage this huge heterogeneous data which can be both unstructured and structured in nature. This data can be managed at different levels in cloud i.e. at the end user level, cloud service provider level and data center level. The proposed approach defines how k medians clustering can be used as an efficient technique for management of data in cloud. The number of data centers is taken as the value of k in the proposed technique. This approach also takes into account the advantage of Cloud data base management system (CDBMS) and applies it to various distributed data centers.

General Terms
Data management and cloud computing

Keywords
CDBMS, Map Reduce, Big table, k-median Clustering, data management

1. INTRODUCTION
Cloud computing is a new generation technology often confused with utility computing, grid computing and distributed computing. Cloud computing is a blend of all these technologies along with a few of its own unique features. It is defined as the use of computing resources which are made available to the users as a service over the internet. It is also considered as the successor of grid computing and promises to fulfill the utility vision of grid computing through its pay per use mechanism. The term cloud comes from the fact that data is not in the users hand within his or her reach but is located far away beyond a person's reach similar to the actual clouds. It is a metered service i.e. users can use the computing resources in a pay per use manner i.e. they pay only for the part of service they are interested in. It also promises to provide several features like on demand self service, broad network access along with an infinite data storage capacity to its users, Figure 1 shows an overview of Cloud system in terms of types of cloud, the various service models and their characteristics which will be discussed further.

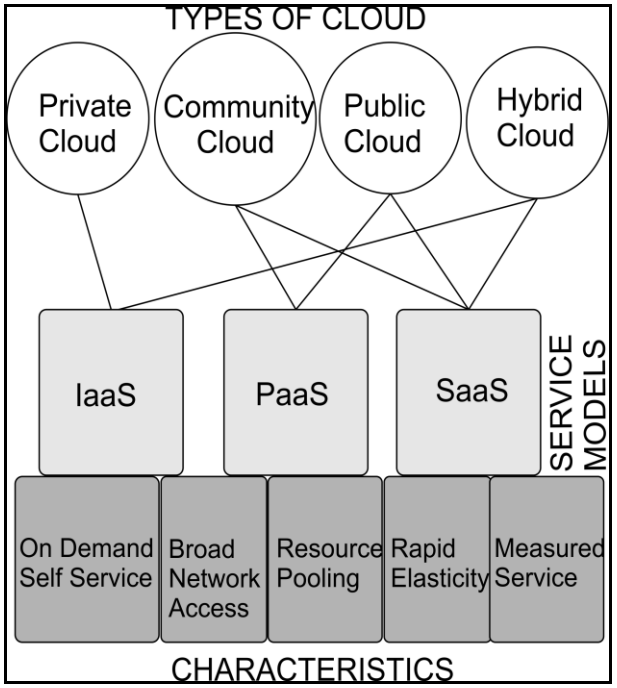

Fig 1: Overview of Cloud system

1.1 Different types of clouds
Different types of clouds are:

1.1.1 Public Cloud
In a public cloud the computing resources are made available to the general public. These services may include applications, servers, storage capacity and various other computing resources. Microsoft, Google and Amazon are some of the public cloud service providers.

1.1.2 Private Cloud
Private cloud is a cloud setup in which the cloud infrastructure is developed for a particular organization and aims to address organizational level cloud goals only. Microsoft and HP are some of the private cloud providers.

1.1.3 Community Cloud
Community cloud is a cloud setup which is provided for two or more organizations having common interests. Thus the cost of the infrastructure is also shared amongst organizations that are part of this community cloud.

1.1.4 Hybrid Cloud
Hybrid cloud is a cloud which is a combination of different types of clouds. Therefore it can be a combination of public and private or public and community.

1.2 Cloud service models
There are also various ways in which a cloud can be deployed. The various cloud service models are:

1.2.1 IaaS
In Infrastructure as a service the computing infrastructure is provided as a service over the internet. This includes virtual machines, servers, storage space etc. Google Compute Engine and Oracle infrastructure as a service are some of the examples of IaaS.

1.2.2 PaaS
In Platform as a service the computing platform itself is made available to the users as a service. It enables application developers to develop and manage their applications on cloud without the concerns of availability of the requisite platforms etc. OrangeScape and OpenShift are some of the examples of PaaS.

1.2.3 SaaS
In Software as a Service applications software's is made available to the users as a service over the internet. It includes services such as virtual desktop, Emails etc. Microsoft Office 365 and PetroSoft are some of the examples of SaaS.

1.3 Characteristics of cloud
Some of the distinguished characteristics of cloud that makes it better than any other technology available at present are:

1.3.1 On Demand Service
Since in cloud computing, computing resources are provided as a service on a utility like basis, therefore the users can demand these services as and when needed without any human interference.

1.3.1 Agility
Cloud computing provides its users with a very agile approach. The users can rapidly deploy their applications without worrying about the initial setup etc.

1.3.2 Measured Service
Cloud computing is a measured service in the terms that the computing resources can be provided in a utility like basis and the users pay only for the part of service they use.

1.3.3 Rapid Elasticity
Cloud computing offers rapid elasticity i.e. the users can allocate and reallocates resources in an elastic manner as per their requirements.

1.3.4 Resource pooling
Since in Cloud Computing, the same resources can be provided to two or more clients as per their requirements. It provides a facility of resource pooling.

1.3.5 Multitenancy
In cloud multiple users share the same set of resources in an elastic and a scalable manner, thereby employing multiple tenants.

Apart from several advantages which cloud computing offers the data storage in cloud is often considered to be insecure [16] because data is held beyond the reach of data owners in the hands of a third untrusted party. But the features offered by cloud are far more lucrative then its limitations and hence, cloud computing is now in the frontier of offering database as a service (DaaS) to its end users. Here data base required for hosting information is made available to its end-users as a service. This database as a service is gradually attracting customers but management of cloud data is very different from management of traditional data. Management of data using traditional methods which are expensive is now slowly becoming obsolete in cloud based environments. Therefore in order to cater to the requirements of management of data in cloud many companies have come up with their own solutions for management of data in cloud based environments[1] such as Big Table[2], Google File system[3], Map Reduce[4], Cassandra[5] and Amazons Simple Storage Service(S3) [6].

The data that needs to be stored and retrieved is rapidly increasing at organizational as well as at individual level [15] thus the owners of this data are now forced to take advantage of several features of cloud computing such as elasticity, flexibility and optimized cost that comes as an easy solution at disposal for them. Management of data in cloud using traditional databases is a difficult and very expensive task. Cloud database management system architecture [13] provides architecture for management of data in cloud and is based on the three schema architecture for database management. There are various data mining techniques available today that will enables its users to cluster web search data [12]. These clustering techniques are also applicable for data management in cloud environment as data available in cloud is of similar nature. In fact cloud computing can also in turn promote the development of newer data mining techniques[14].Thus the proposed approach will enable management of data in cloud using a k median based clustering approach. The choice of this clustering technique provides an edge over other available techniques in terms of speed and optimized cost.

2. RELATED WORK
Cloud dbms (CDBMS) [7] is a database developed by the Bloor group for cloud. It is a distributed database that can handle queries and provide query services across numerous distributed nodes of database at different data centers. Some of the chief characteristics of cloud include rapid elasticity and availability. Therefore a database developed for cloud also needs to guarantee that these requirements are met in terms of database that is made available to the cloud users, CDBMS can address these challenges as well as it also provides an optimized query workload distribution across a distributed system. It has the ability to manage heterogeneous nature of data in cloud environment at varied levels. CDBMs handles query traffic SOA takes care of the transactional traffic .It can handle both local data as distributed data.

CDBMS uses Algebraix Data's technology (A2DB) for deployment. A2DB can reuse results produced from previous queries thereby providing high performances. It can also balance both global and local queries thereby providing global optimization.

Map Reduce [4] is a programming model and an implementation for generating and processing of a very large data set. It is highly scalable in nature and can process large volumes of data. It uses two computation functions map and reduce. The map function takes input and generates output in terms of key/value pair. It is then followed by the reduce function .Map Reduce is used by Google for data mining, web searching and sorting. It is very easy to implement.

Bigtable [2] is a distributed storage system developed by Google for managing large volumes of structured data. It provides the end users with dynamic control of their data's layout, high availability, scalability and varying applicability. Big table is used in applications such as Orkut, Google earth and Google finance. Big table provides its end-users with a data model that can provide them facilities such as dynamic control over data layout and ability to find out properties of locality of data. It is implemented as a distributed, sparse and multidimensional sorted map and data is indexed in it through arbitrary row and column names.

In [11] a discussion about mapping topology of clusters onto cloud environment is done. This provides greater scalability and reliability

3. MOVING DATA MANAGEMENT APPLICATIONS TO CLOUD

3.1 Transactional Data Management

Transactional data management deals with management of day to day transactional data. It can be data in an airline reservation management system or railway reservation management. Transactional data is based on ACID properties that must be guaranteed by a traditional database system. According to [8] transactional database are not well suited for cloud applications due to the following reasons:

3.1.1 No Use of Shared Nothing Architecture

Transactional databases such as Microsoft SQL server and Sybase cannot be implemented through shared nothing architecture. But transactional databases cannot be implemented using a shared nothing approach which can provide facility of rapid scalability.

3.1.2 Difficulty of maintenance of ACID properties across geographical locations

In case of replication of data across varied geographical locations it is difficult to insure that all the ACID properties are guaranteed. In most of the cases Consistency property is not guaranteed. In cases such as Google's big table atomicity property does not hold true.

3.1.3 Risks in storing transactional data on an unsecure third party host

Since transactional data may contain certain critical or sensitive information's such as credit card details and customers name and identity information. Therefore it is very risky to store sensitive information contained in transactional data in cloud.

Therefore, transactional database applications are not well suited for cloud environment.

3.2 Analytical Data Management

Analytical data management aims at supporting a company's decision making activities. According to [8] analytical data management is better suited for cloud environment. The following are the reasons for it [8]:

3.2.1 Shared nothing architecture used in analytical data management

Shared nothing architecture is well suited for analytical data management. Products such as IBM DB2 also use share nothing architecture for their analytical applications.

3.2.2 ACID properties maintenance not required

Since analytical databases do not require updating data frequently therefore it is not necessary to maintain ACID properties in analytical database. There by, making them suitable for cloud environments.

3.2.3 Does not deal with Sensitive data

In analytical data management it is possible to identify probable sensitive data. This data can either be avoided from being kept at a third party location or can either be kept in some encrypted for thus, protecting it from probable intruders.

4. NEED FOR DATA MANAGEMENT IN CLOUD

Management of applications in cloud has varied benefits.

4.1 Agility

One of the biggest benefits of data management in cloud is agility [9]. For example an instance of Amazons Relational database (DB) can be launched instantly within a few minutes without any need for provisioning any hardware or configuring any storage etc. This makes its suitable for use by companies who are looking for optimizing costs of maintaining their databases. Some of the companies such as RedBus use Amazon RDS which enables them to scale up and scale down their applications as per the load and demands thus obtaining large amount of cost benefits.

4.2 Pooling of Resources

Another reason why management of data in cloud has gained momentum is it allows resources to be pooled together [10]. These pooled resources allows resources such as servers, storage space and data centers to be shared amongst multiple users thereby providing the users with benefits such as cost reductions.

4.3 Limitations of Traditional Warehouses

The data is increasing at an astounding rate therefore traditional warehouses cannot manage this data .Also the demands for processing and storage also fluctuates depending upon the load. The traditional infrastructures are not able to handle this. Therefore cloud can be used as away to handle these requirements [10].

4.4 Dynamic resource allocation

Provides organizations with the facility to allocate and reallocate resources as per their requirement by using either public or private cloud [10].

4.5 Cost Reductions

Using cloud to handle data can lead to a massive reduction in the costs incurred in setting up the infrastructure for handling data as well as the cost of human resource engaged in the task.

4.6 Infinite Data Capacity

Cloud can share the workloads amongst multiple nodes thus the databases offered by cloud show infinite data capacity.

5. DATA MANAGEMENT THROUGH CLUSTERING

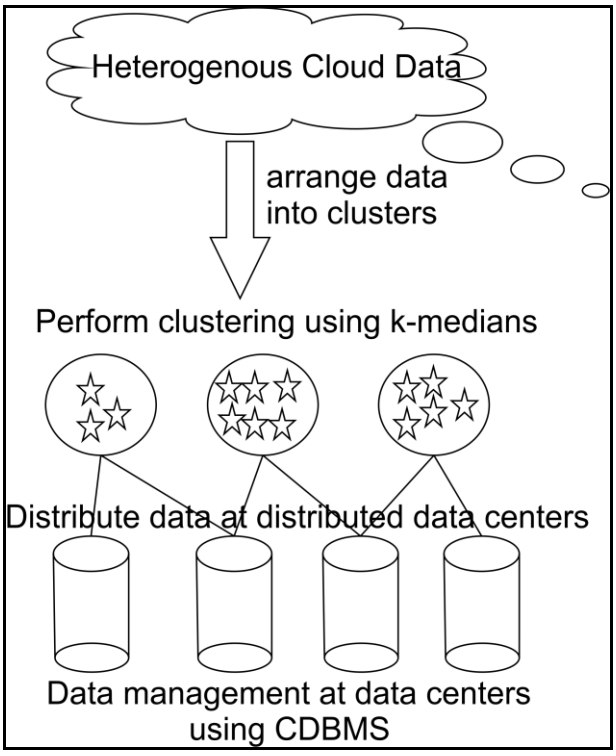

Fig 2: Managing Data through clustering in cloud

Cloud environment consist of data coming from different sources, in varied forms and of different types. The traditional data management techniques are meant for handling traditional data, this data was of a particular type and of a bounded size. Now the data which is available across networks is growing at a very high rate, it's of varied forms, has huge velocity and needs to have authentic veracity. The data also has different requirements in terms of security because the data will be stored at a third party untrusted host. Thus, the traditional systems fail to handle the requirements of data management in cloud. Since the data which is available is heterogeneous and of huge size thus, a different method apart from the traditional ones to handle this kind of data are required. The proposed approach uses clustering techniques to organize cloud data into clusters and this data will then be distributed across varied data centers at different geographical locations. Figure 2 provides a brief overview of the proposed approach

The following are the steps of the proposed method.

5.1 Organize data into clusters

In cloud computing several end users use the computing resources that are provided to them as services. Therefore data which is generated by these cloud users is heterogeneous in nature and its size is usually big. Therefore in order to deal with the heterogeneous nature of cloud data clusters of these data are formed based on the degree of similarity between the various data's. Thus data's which have less cohesion will be kept in same cluster and those that have high cohesion will be kept in another cluster. For the proposed approach a k-medians clustering method will be used.

Where, k=number of clusters=number of available data centers

k-medians clustering is highly efficient in cloud computing as it's a variant of k means clustering and thus is one of the fastest clustering mechanism. Since in cloud environment speed is a critical issue as data is distributed across various geographical locations and to be in competition speed provides an edge for organizations. The input parameter 'k' can also be modified according to the number of available data centers. So the steps in the proposed method are:

5.1.1 Choose a set of k points {c1, c2.., ck} where k is number of data centers and then form clusters {C1, C2..., Ck}

5.1.2 Recalculate the centroids of clusters again. Keep on repeating this till no more data moves between the clusters.

5.2 Assign the clusters to geographically distributed datacenters based on certain parameters

After the clusters have been formed using k-medians clustering the next step in the proposed method is to distribute these clusters amongst data centers which are located at geographically dispersed locations. This distribution is based on several factors such as channel bandwidth, channel capacity, distance between the cluster and the location of data center and the compute capacity of a particular data center i.e. whether the data center can handle the data traffic of the given channel or not. The channel capacity is defined as the rate at which information can be transferred in a reliable manner across a network and has a significant impact on the performance in cloud based environment. The distribution of clusters is also based on content delivery networks and distance of the edge nodes.

5.3 Manage data at each data centers using CDBMS

After the data has been distributed at different nodes this data is now managed using CDBMS. Cloud DBMS can easily handle the heterogeneous nature of data available in cloud. It can handle the query traffic by partitioning and allotting the queries across multiple, geographically dispersed and distributed data centers. It can easily cater to the needs of local data by splitting the nodes into multiple nodes in case of

heavy query traffic and also takes care of distributed queries by considering a distributed query as a union of several queries each handling individual nodes. The result is produced by joining the answers from different queries and thus presenting a unified result. Thus, the use of CDBMS to manage data at the data centers can help overcome the challenges of data management in cloud by providing high availability, data security up to a certain degree and also an optimized distributed query workload mechanism.

6. ADVANTAGES AND FEATURES OF THE PROPOSED METHOD

Mining of data helps in extracting useful information from raw data. K medians clustering algorithm is a variant of the popular k-means algorithm. The proposed approach is a combination of k-medians clustering technique along with CDBMS and therefore promises to fulfill several challenges imposed for data management in a cloud environment. Use of CDBMs in particular helps in solving varied issues such as scalability limits which can be encountered between and within datacenters in relational or column store databases.

Apart from this there is a lack of ability to handle mixed or heterogeneous workloads by existing environments such as Hadoop and Map Reduce [4] as these environments were not built for handling such kinds of workloads,the emphasis of such environments is more on the tolerance to failures .CDBMS used in this method can deal with these challenges up to a certain level.

Some of the advantages and Features of the proposed method over the existing ones are:

6.1 Efficient for mining of large datasets

The proposed method employs the use of k medians clustering technique and is thus an efficient technique in case of mining of very large datasets such as the ones that are prevalent in cloud environment k-medians is better than k- means algorithm since it does not use the square of distance as in k means. This paper focuses about how data can be managed in a cloud environment in an efficient manner using k-medians algorithm.

6.2 Useful for Data management in cloud based environment

This approach can help in solving the data management problem associated with cloud computing, though there are several data management tools available at present such as oracle database 12c [17], Googles Bigtable which deals with non relational data and Microsoft SQL Azure which is a fully relational data but none of them is fully capable in solving all the management problems associated with data in cloud. The proposed approach is useful for solving data management issues to a certain extent.

6.3 Independent of the Nature of Data

The proposed approach is irrespective of the nature of data i.e. it is not restricted upon the nature of data, whether the data is relational or non relational. Therefore it can handle data of heterogeneous nature as well with equal competence.

6.4 No performance bottlenecks

Since the clusters are assigned depending upon the content delivery networks and distance to edge nodes therefore, the chances of performance bottlenecks for distribution of data are no longer a phenomenon in cloud environments.

6.5 Ability to query data across distributed data centers

CDBMS has the ability to make a query across distributed database nodes which is envisioned as a future problem associated with growth of cloud computing.

6.6 High performance of a single node

Use of A2DB [7] as CDBMS, which can reuse previously executed queries' intermediate results yields high performance at a single node.

7. CONCLUSION AND FUTURE WORKS

Cloud computing is a new generation technology which boasts of being a very promising technology for future and is bound to change the entire scene of information technology industry. Cloud comes in four different types: public, private, community and hybrid. The various cloud service models are Infrastructure as a service, platform as a service and software as a service. It has several characteristics such as on demand self-service, agility and a metered service. Cloud data base management system is a data base management system for cloud and can manage heterogeneous data such as the ones available in cloud. Moving data to cloud requires us to identify the kind of data we are moving i.e. whether it is analytical data or transactional data. There is a growing need for data management in cloud as it allows several resources to be pooled together and can support data which is growing at a rapid rate with very high velocity. Our proposed approach provides management of cloud data through clustering and uses a k median which is a variant of k means as the clustering technique. It first organizes the cloud data to form clusters and then assigns these clusters to distributed data centers. At these distributed data centers CDBMS is used for effective management of data. The proposed method has several advantages such as scalability, mining of large data sets and the ability to handle heterogeneous data sets. Future works for the proposed method involves the study of distribution of these clusters into the data centers in a more detailed manner and also involves the research upon the factors that affect the distribution of data at these data centers.

8. ACKNOWLEDGMENTS

Kashish Ara Shakil would like to thank and acknowledge Dr. Mansaf Alam, Mrs. Seema Shakil and Mohammad Zaki for being a pillar of support in all her endeavors. She would also like to thank Shabih Shakeel for his contributions towards this work.

9. REFERENCES

[1] Meng-Ju Hsieh, Chao-Rui Chang; Li-Yung Ho; Jan-Jan Wu; Pangfeng Liu. 2011 SQLMR : A Scalable Database Management System for Cloud Computing, in Proc. 2011 International Conference on Parallel Processing (ICPP), pp. 315-324.

[2] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach,M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. 2008 Bigtable: A distributed storage system for structured data', ACM Trans. Comput. Syst., vol. 26, pp. 4:1–4:26.

[3] S. Ghemawat, H. Gobioff, and S.-T. Leung. 2003 The google file system, in Proceedings of the nineteenth ACM symposium on Operating systems principles, ser. SOSP '03. New York, NY, USA: ACM, 2003, pp.29–43.

[4] J. Dean and S. Ghemawat. 2004 MapReduce: simplified data processing on large clusters,' in Proceedings of the 6th conference on Symposium on Opearting Systems Design and Implementation, ser. OSDI 04, vol. 6. Berkeley, CA, USA: USENIX Association, 2004, pp. 10–10.

[5] Cassandra, http://cassandra.apache.org/

[6] Amazon, Amazon simple storage service, http://aws.amazon.com/s3/.

[7] Robin Bloor. 2011 What is Cloud DataBase

[8] Daniel Abadi. 2009 Data Management in the Cloud: Limitations and Opportunities', IEEE Data Engineering Bulletin, vol. 32 no. 1

[9] Amazon. 2011 Amazon Web Services – Running Databases in the Cloud.

[10] Oracle Data Center Best Practices: Managing Data with cloud computing http://www.oracle.com/us/products/database/cloud-computing-guide-1561727.pdf

[11] Kosinska, J., Kosinski, J., Zielinski, K. 2010 The Concept of Application Clustering in Cloud Computing Environments: The Need for Extending the Capabilities of Virtual Networks,' 2010 Fifth International Multi-Conference on Computing in the Global Information Technology (ICCGI),, pp.139-145

[12] Mansaf A. and Kishwar S. 2012 A Review on Clustering of Web Search Result,' ACITY 2012, Chennai, Advances in Computing & Information Technology, AISC 177, Springer-Verlag Berlin Heidelberg, pp. 153-159.

[13] Mansaf A. and Kashish A. S. 2013 Cloud database Management System Architecture International Journal of Computer Science and its Applications, Vol. 3, Issue 1, pp. 27-31.

[14] Tingting H., Haishan C., Lu H., Xiaodan Z. 2012 A survey of mass data mining based on cloud-computing,' In proc. Of 2012 International Conference on Anti-Counterfeiting, Security and Identification (ASID), vol. 1, no. 4, pp. 24-26.

[15] Cong W., Kui R., Shucheng Y., Urs, K.M.R. 2012 Achieving usable and privacy-assured similarity search over outsourced cloud data,' INFOCOM, 2012 Proceedings IEEE, pp.451-459.

[16] Jang-Jaccard, J., Manraj A., Nepal S. 2012 Portable key management service for cloud storage,' 2012 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), pp.147-156

[17] Oracle, "Plug into the Cloud with Oracle Database 12c," An Oracle White Paper , June 2013