Location based Clustering of Cloud Datasets for Forensic Analysis

Asmita Ballal
TGPCET
Nagpur

Sulabha Patil
TGPCET
Nagpur

R.V.Dharaskar, Ph.D
MPGI
Nanded

ABSTRACT

Cloud environment is offering different services to the users and more and more companies are working to tap the benefits being provided by this environment. Data mining algorithms are proven algorithms to find hidden useful information from large database. K-Means clustering algorithm is one of the very popular and high performance clustering algorithms. The main aim of this work is to implement and deploy K-Means algorithm in Google Cloud using Google App Engine with Cloud SQL.

Key words

Cloud computing, cloud SQL, cluster, dataset, google app engine

1. INTRODUCTION

Cloud Computing has become one of the most talked about technologies in recent times and has got lots of attention from media as well as analysts because of the opportunities it is offering.[1] These services can be hosted and managed by the user organization (on a reduced hardware base), or by a third-party provider. Consequently, the software and data comprising the organization’s application may be physically stored across many different locations, potentially with a wide geographic distribution.Telecommunication companies routinely generate and store enormous amounts of high-quality data, have a very large customer base, and operate in a rapidly changing and highly competitive environment.[2]Mobile phone proliferation is on the increase with the worldwide cellular subscriber base reaching 4 billion by the year end of 008.[10]Clustering is one of the well-known Data mining techniques to find useful pattern from a data in a large Database[3].K-Means clustering [4] is one of the most famous clustering algorithms applied in different types of domains.[5] The need for mobile forensics is because of-

• Use of mobile phones to store and transmit personal and corporate information
• Use of mobile phones in online transactions
• Law enforcement, criminals and mobile phone devices .

2. CLOUD COMPUTING

A cloud has several uses, offering a variety of services and can be deployed in more than one way. Consequently, several definitions of cloud computing have been proposed (Mell & Grance, 2011; Schubert, et al., 2010; Wyld, 2009). [1] There are three main levels of service for users of cloud computing (Mell & Grance, 2011):

in the Software as a Service (SaaS) model, a client can make use of software applications made available from the cloud provider.

the Platform as a Service (PaaS) model provides an application programming interface (API) for clients to create and host custom-built applications.

the Infrastructure as a Service (IaaS) model is the leasing of virtualized computing resources such as processing power, volatile memory and persistent storage space to host virtual machines.

In addition to the different levels of deployment, a cloud can be categorised by its organizational deployment, with consequent impact on the geographical location and storage architecture of data held:

- in a private cloud, the infrastructure is operated solely by the organization who owns the cloud.
- a community cloud is shared between several organizations, either because of a common organizational goal, or in order to pool IT resources.
- public clouds will usually be owned by a provider organization, which will maintain the cloud facilities in one or more corporate data centres.
- a hybrid cloud is a composition of two or more of the above deployment options. Hybrid clouds can be used to provide load balancing to multiple clouds.

3. CLUSTERING ALGORITHM

Today’s business world is fast and dynamic in nature. It involves lot of data gathered from different sources. The most challenging task of the business people is to transform these data into useful information called knowledge. Data mining techniques are used to achieve this task. Clustering is a type of unsupervised learning technique that can be used to explore data sets in order to discover the natural structure and unknown but valuable behavioral patterns of customers hidden in it [6].Clustering techniques group data items based on their similarities. Euclidean distance is a common choice for measuring the similarities. A data item is assigned to a cluster whose center is the most similar to the data item. K-Means algorithm [7] follows the partitional or non-hierarchical clustering [8]. K-means uses the mean of the data items in a cluster as the center of that cluster. The main drawback of K-Means is the number of clusters must be known in advance, which is defined by K.

K-Means Algorithm :-

- Select K points as the initial centroids
- Repeat
- Form K Clusters by assigning all points to closest centroid
- Recompute the centroid of each cluster
- Until the centroids don’t change

The Initial centroids will be chosen randomly. K-Means generates different clusters in different runs.[9]

4.EXPERIMENTAL METHODOLOGY

The K-Means algorithm was to be implemented in java so the software chosen was Eclipse IDE for design and development of the application. To deploy the application in Google, we download the Google App Engine Plug-In from Google’s

official site. To create Database and table we use Google Cloud SQL. Figure 1 shows the components of the application.
There are four important components:
Observation. In: Le

- Client user interface
- Google App Engine
- Cloud SQL
- Client browser window

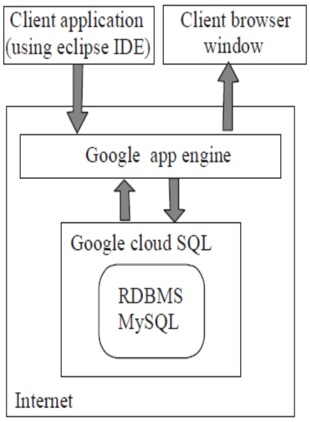

Fig. 1: Architecture of cloud data mining

Client Application is the user interface which contains the Java implementation of K-means algorithm created using Eclipse IDE.
To deloy the application Google App Engine is chosen.The real time data is taken from Reality Mining Dataset.Cloud Sql was used for storing the data where database and tables are created. Google Cloud SQL is a MySQL database that lives in Google's cloud. It has all the capabilities and functionality of MySQL, with a few additional features and a few unsupported features as listed below. Google Cloud SQL is easy to use, doesn't require any software installation or maintenance and is ideal for small to to medium-sized applications. Google Cloud SQL is currently available for Google App Engine applications that are written in Java or Python. It can also be accessed from a command-line tool, and other admin and reporting tools are available.The Client browser is used to view the output of clusters.
Steps for implementation and deployment of K-means in cloud:
Step-1: Create Database and table in Cloud SQL
Step-2: Take the real-time Dataset from Reality Mining dataset and store it in to respective table
Step-3: Write and execute a sample SELECT query and check whether it works well or not in Cloud SQL of Google API's console.
Step-4: Design the User Interface and write code in Java for K-Means
Step-5: Debug and deploy the application in Google App Engine Cloud
Step-6: Go to the browser window and view the output clusters.

5. CONCLUSION

It is widely accepted that K-Means algorithm is very popular clustering algorithm to analyse any real world problems. K-Means algorithm is more efficient algorithm for mining large Databases and Cloud computing provides solution for storing large database with less cost. So, in this paper, we focused the implementation of K-Means algorithm in the Cloud environment.
With increased connectivity options and higher storage capacities and processing power, abuse of mobile phones can become more main stream. Mobile phones outsell personal computers and with digital crime rates rising, the mobile phone may be the next avenue for abuse for digital crime. Mobile phones with their increased connectivity options may become a source of viruses that infect computers and spread on the internet.

6. REFERENCES

[1] George Grispos,Tim Storer, William Glisson."Calm Before the Storm: The Challenges of Cloud Computing", *Digital Forensics.*

[2] Gary M Weiss,"Data Mining in the Telecommunications Industry",*IGI Global* 2009 (486-491)

[3] A. Mahendiran, N. Saravanan, N. Venkata Subramanian and N. Sairam."Implementation of K-Means Clustering in Cloud Computing Environment", Research Journal of Applied Sciences, Engineering and Technology 4(10): 1391-1394, 2012

[4] MacQueen, J.B., 1967. Some Methods for Classication and Analysis of Multivariate Cam, L.M. and J. Neyman, (Eds.), 5 Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, USA.

[5] Julie, S., 1982. A survery of the literature of cluster analysis. Comput. J., 25(1): 130-134.

[6] Indranil Bose and Xi Chen ," Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn",In Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong.

[7] Pang-Ning, T., S. Michael and V. Kumar, 2006,"Introduction to Data Mining." Pearson AddisonWesley.

[8] Jain, A.K. and R.C. Dubes, 1988. "Algorithms for Clustering Data", Prentice Hall, New Jersey.

[9] Murat, E., C. Nazif and S. Sadullah, 2011," A new algorithm for initial cluster centers in k-means algorithm", Elsevier B.V., Pattern Recognition Let., 32: 1701-1705.

[10] Rizwan Ahmed, R. V. Dharaskar, "Mobile Forensics: An Introduction, Standard Practices, and Tools", National Conference on Wireless Communication and Networking (WINCON), 9-10 Jan 2009, L&T Institute of Technology, Powai Mumbai, pp.7-23.