Implementation of K-Means Clustering in Cloud Computing Environment

A. Mahendiran, N. Saravanan, N. Venkata Subramanian and N. Sairam
School of Computing, SASTRA University, Tanjore, Tamilnadu, India

Abstract: Data mining algorithms are proven algorithms to find hidden useful information from large database. Cloud Computing is a web-based technology whereby the resources are provided as shared services. The large volume of business data can be stored in Cloud Data centers with low cost. Both Data Mining techniques and Cloud Computing helps the business organizations to achieve maximized profit and cut costs in different possible ways. 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, database, dataset, google app engine

INTRODUCTION

Clustering is one of the well-known Data mining techniques to find useful pattern from a data in a large database (Fayyad, 1996). These patterns are very useful for the knowledge workers such as financial analyst, Manager etc… to take right managerial decisions.

K-Means clustering (MacQueen, 1967) is one of the most famous clustering algorithms applied in different types of domains such as Biology and Zoology, Medicine and Psychiatry, Sociology and Criminology, Geology, Geography and Remote sensing, Pattern recognition and Market research, and Education etc… (Julie, 1982) to find the useful patterns.

Today’s business world is fast and dynamic in nature. It involves lot of data gathered from different sources. These data are stored in Data Warehouses. 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.

Cloud Computing offers several benefits to the business organization to cut the initial investments to establish infrastructure for storage and compute. Many business organizations have already started migration of their business data into cloud data centers. Most often they need to mine useful information from the data stored in the cloud data centers with regards to business decisions. So the main objective of this work is to incorporate and implement K-Means Data mining technique into Cloud environment.

K-means algorithm: K-Means algorithm (Pang-Ning et al., 2006) follows the partitional or nonhierarchical clustering approach (Jain and Dubes, 1988). It involves partitioning the given data set into specific number groups called Clusters (Klosgen and Zytkow, 1996). Each cluster is associated with a enter point called centroid. Each point is assigned to a cluster with the closest centroid. The main drawback of K-Means is the number of clusters must be known in advance, which is defined by K.

- 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. The centroid is nothing but the mean of the points in the cluster. Euclidean distance is used to measure the closeness. K-Means generates different clusters in different runs (Murat et al., 2011).

EXPERIMENTAL METHODOLOGY

We implement the K-Means algorithm in java so we choose 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 our application. There are four important components:

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

Corresponding Author: A. Mahendiran, School of Computing, SASTRA University, Tanjore, Tamilnadu, India
Client Application (using eclipse IDE) is the user interface and the K-Means implementation in Java that has been created using Eclipse IDE. The second major component is Google App Engine, which we chose to deploy our application. The real-time dataset has been taken from the Machine Learning Repository, 2012 (archive.ics.uci.edu/ml/datasets.html) for analysis. The third component, Cloud SQL, was used for storing the data set. We created databases and tables in Cloud SQL.

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 Machine Learning Repository, 2012 and store it in respective tables.

**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.

**EXPERIMENTAL RESULTS**

As we stated in the introduction, the main objective of this work is to implement the K-Means algorithm in cloud.
Cloud environment. For the experiment we took two data sets from the well known real world data base “Machine learning repository, 2012”.

The first data set we used was “Iris Dataset” (Fisher, 1936). It consists of 5 attributes and 150 instances. The attributes are sepal width, sepal length, petal width, petal length and class label. It has three classes Iris flowers namely:

- Iris setosa
- Iris versicolor
- Iris virginica

We have created a database called “KMean” and two tables to store the data sets of iris and Blood Transfusion Service Center Data Set in Cloud SQL using MySQL. As one of the properties of K-Means is that it assumes the number of clusters K is known in advance. For the first experiment, we assume that k = 3. Figure 2 shows the output clusters of Iris dataset obtained from the experiment. Since the value of K is 3, the number of output clusters was three.

The second data set we used was “Blood Transfusion Service Center Data Set”. It consists of 5 attributes and 748 instances. We have taken first 200 data for analysis. As one of the properties of K-Means is that it assumes the number of clusters K is known in advance. Here we assume that k = 2. Figure 3 shows the output clusters of Blood Transfusion Service Center Data Set obtained from the experiment. Since the value of K is 2, the number of output clusters was two.

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 and the experimental results shows that it works well in the Cloud.

REFERENCES


Pang-Ning, T., S. Michael and V. Kumar, 2006. Introduction to Data Mining. Pearson Addison Wesley.