Hadoop Operations Management for Big Data Clusters in Telecommunication Industry

N. Kamalraj
Asst. Prof., Department of Computer Technology
Dr. SNS Rajalakshmi College of Arts and Science
Coimbatore-49

A. Malathi, Ph.D.
Asst. Prof., PG and Research Department of Computer Science, Govt. Arts College
Coimbaore-18

ABSTRACT

This paper describes how data mining techniques are used in Hadoop for cloud data where it is an open source implementation. Extraction of useful information from raw data is always referred by the term DM. The techniques of DM are integrated into the normal day-to-day life has become very popular. Data mining are useful to improve the efficiency for the reduction of cost in the businesses field. In the cloud computing paradigm, the applications and techniques of data mining are most wanted. The users can retrieve meaningful information from virtually integrated data warehouse and it is allowed by implementing the data mining in cloud computing for reducing the cost of storage and infrastructure. This paper aims at predicts the churn customer in telecommunication industry where the dataset is stored in cloud and implemented using data mining techniques in Hadoop. In this paper, classification is used to analyze the dataset of telecommunication industry and classify the numerical and text data and predict the churners who are likely to switch from current network, and the clustering is used to group the result of the classification from the given data set for the best prediction of numerical and text data together in Hadoop. Hadoop is an environment easy to implement the classification; clustering techniques and cloud are used to store the data set for the industry.

Keywords

Datamining, Hadoop, HDFS, Map/Reduce.

1. INTRODUCTION

Data Mining is an effective method to analyze data from various angles and retrieve useful information from the data. From the data set, it is easy to find out the correlation of data patterns, categorize the data and classify the data by using data mining. On the other hand, storing and transferring of large amount of data is the challenging one. The main bottleneck is to maintain the data flow and data resource. In the scientific discovery after the experimental, theoretical and computation science, the computing of intensive data is to be assumed as the fourth paradigm, and the major challenge is the handling of large data sets [1].

In everybody's life, an internet plays the vital role for both use of personal and profession. In the recent years, the cloud computing becomes the revolutionary concept. Due to the huge availability, low cost and mobility, the cloud computing usage becomes more popular than past years. On the other side, for the security of the company's information and data it becomes more threat. In the various fields like spatial data, science, Engineering, medicine and business the knowledge discovery from database becomes more vital and used and the techniques of data mining have been involved.

2. THE CLOUD COMPUTING

The resource of computer and the storage space of network can be attained by means of cloud computing where it is the subscription-based service [7]. Cloud computing facilitates end-users or small companies to use computational resources such as software, storage, and processing capacities belonging to other companies (cloud service providers). Cloud services include SaaS (Software as a Service), IaaS (Infrastructure as a Service) and PaaS (Platform as a Service) [1] [4]. Big corporates like Amazon, Google and Microsoft are providing cloud services in various forms. Amazon Web Services (AWS) provides cloud services that include Amazon Elastic Compute Cloud (EC2), Simple Queue Service (SQS) and Simple Storage Service (S3). Google provides Platform as a Service (PaaS) known as Google App Engine (GAE) and facilitates hosting web applications. Microsoft also provides cloud services in the form of SQL Azure, Windows Azure, Windows Intune etc. By using these services, users can exploit the benefit of mass storage and processing capacity at a low cost. Developers can use these services to avoid the mass overhead cost of buying resources, e.g., processors and storage devices.

2.1 Cloud types in Cloud Computing

The various types of clouds can be supported by cloud computing based on the needs. The services of public cloud can be mostly used by the small business owner or the home user [4].

1. With an internet connection, the cloud space can be accessed by anyone which is called Public Cloud.

2. For an organization or specific group the cloud space is obtained and only that group can access the space called Private Cloud.

3. Where the requirements of cloud are the same for one or more organizations and the cloud space is shared by those organizations it is called Community Cloud [7].

4. The cloud space is the mixture of community, private or public cloud and at least two clouds are combined is known as Hybrid Cloud.

3. DATAMINING TASKS

At times, the data can be considered to be a gold mine for strategic planning for research and development in this area which is often referred to as Data Mining (DM) and Knowledge Discovery in Databases (KDD).

There are six set of activities that are used in the Data mining [5].

- Classification
- Estimation
- Prediction
- Association rules

- Clustering
- Description and Visualization

Classification, Prediction and Estimation are the examples for supervised learning or directed data mining whereas clustering, association rules, description and visualization are the examples of unsupervised learning or undirected data mining [3].

3.1 Classification and Prediction

Assigning a newly presented object to an already existing predefined class by means of examining the features of the object comes under the classification task. The result from the classification is called prediction. The following are the techniques that are present in the classification. Classification by decision tree induction, Bayesian classification, Rule-based classification, back propagation classification, support vector machines, Genetic algorithms, rough set approach, fuzzy set approach. The predictions are as follows: Linear regression and Non-linear regression.

3.2 Estimation

The continuous value outcome of the process is known as estimation. The following methods are used for evaluating the estimation of the process. Holdout method, Random Sub-sampling, Cross validation and bootstrap.

3.3 Association Rules

In the databases, the existence of the relationships between the set of objects is implied by means of the association rules.

3.4 Clustering

Clustering is a process of grouping the objects that are having the same or similar characteristic into one group or cluster. The following are the categorization of the major clustering methods: Outlier analysis, density-based methods, Grid-based methods, hierarchical methods, model-based clustering methods and Partition methods.

4. HADOOP INTRODUCTION

An open source implementation of the MapReduce parallel processing framework is called Hadoop. Hiding the details of distributing data to processing nodes, collecting the results of the computation, restarting subtasks after a failure and the parallel processing is known as Hadoop [7]. This framework allows developers to write relatively simple programs that focus on their computation problems, rather than on the nuts and bolts of parallelization.

4.1 Hadoop Architecture

The following architecture explains the concepts behind the Hadoop.

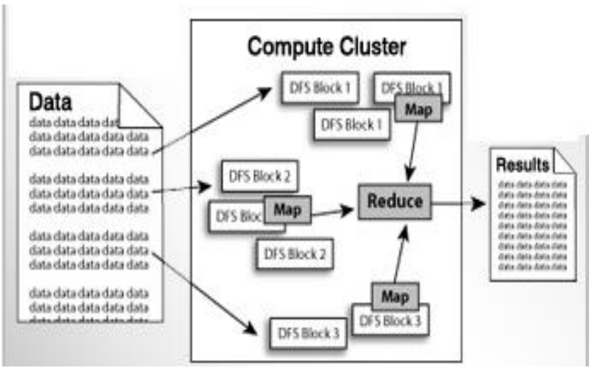


Figure 1: Architecture of Hadoop

4.2 Components of Hadoop

The following are the Hadoop components: Hadoop Distributed File System (HDFS) and Map Reduce.

1. Distributed file system (HDFS) :
In a large cluster, across the machines, the system that is designed to store the very large files is known as Hadoop's Distributed File System (HDFS). HDFS is attracted by the file system of Google [6]. Each file in Hadoop DFS is stored as the sequence of blocks and each block is of the same size except the last block in the file. For fault tolerance, the blocks in the files are replicated. Each file in HDFS is configured by means of the replication factor and the size of the block. At any time, the HDFS files should have only one writer, and are “Write Once” manner.

- For the entire cluster, there should be a single namespace
- Replicates data 3x for fault-tolerance
- Store large data sets
- Cope with hardware failure
- Emphasize streaming data access

2. MapReduce Framework
In a cluster across the nodes, the user defined Reduce/Mapping jobs are executed and a cluster of machines are tackled by the Hadoop Map/Reduce framework. There are two levels in this computation, i.e., a reduce phase and a map phase. The key/value pairs of the dataset are the input for the computation.

There will be the re-distribution of the tasks to the remaining nodes if the nodes go wrong in the middle of the computation. The above execution of tasks takes place in the fault-tolerant manner. Re-run of the failed tasks with small run-time and good load balancing is enabled by means of many map and reduce tasks. Executes user jobs specified as “map” and “reduce” functions [6]

- Manages work distribution & fault-tolerance
- Process large data sets
- Cope with hardware failure
- High throughput

Google introduced and patented MapReduce- a software framework to support distributed computing on large data sets on clusters of computers. The name “MapReduce” and the inspiration came from map and reduce functions in functional programming.

In MapReduce, the Map function processes the input in the form of key/value pairs to generate intermediate key/value pairs, and the Reduce function processes all intermediate values associated with the same intermediate key generated by the Map function, as shown in the figure below.

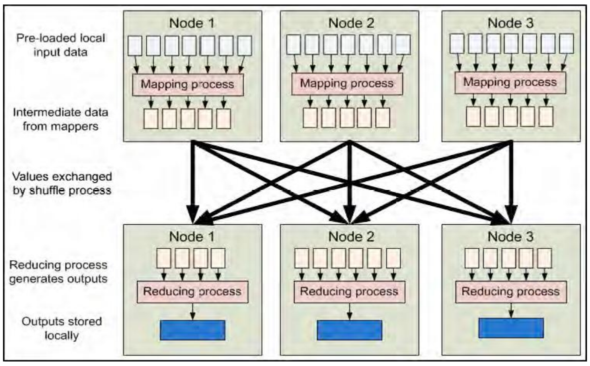


Figure 2: Map / Reduce Framework Illustration

The MapReduce framework automatically parallelizes and executes on a large cluster of machines. Partitioning the input data, scheduling the program's execution across a set of machines, handling faults, and managing the required inter-machine communication are all handled by the run-time system. This facilitates programmers with no experience with parallel and distributed systems to easily utilize the resources of a large distributed system.

5. DATA SET

In this paper, the following data sets for the Hadoop classification and clustering are taken into consideration. There are two data sets are used for the above function. The first data set is for the North Zone of Tamil Nadu and the second data set is for South Zone of Tamil Nadu. The following dataset are taken from Orange Dataset for the telecommunication industry [9].

5.1 North Zone Data set

The following attributes are in this data set:

- Account Number (acc.no)
- Area code
- Voice mail Service (vmail)
- Number of minutes per day (day.mins)
- Number of Calls per day (day.calls)
- Charge per day (day.charge)
- Number of International Minutes (intl.mins)
- Number of International Calls (intl.calls)
- Charges for International calls (intl.charge)
- Churn

5.2 South Zone Data Set

The following are attributes that are included in this data set for the classification and clustering of Tamil Nadu South Zone:

- Churn
- Area code
- Voicemail(vmail)
- Number of voicemail messages (vmail.msgs)
- Number of day minutes (day.mins)
- Number of day Calls (day.calls)
- Charge per day (day.charge)
- Number of Evening Minutes (eve.mins)
- Number of evening calls (eve.calls)
- Evening Charges (eve.charge)
- Minutes at Night (night.mins)
- Number of Night Calls (night.calls)
- Charges at Night (night.charge)
- Number of International Minutes (intl.mins)
- Number of International Calls (intl.calls)
- Charges for International calls (intl.charge)
- svc.calls

6. PROPOSED ARCHITECTURE

The following architecture explains our proposed system. In this paper, the telecommunication data set are taken from the cloud and put into the Hadoop framework for the classification and clustering to predict the churn customers.

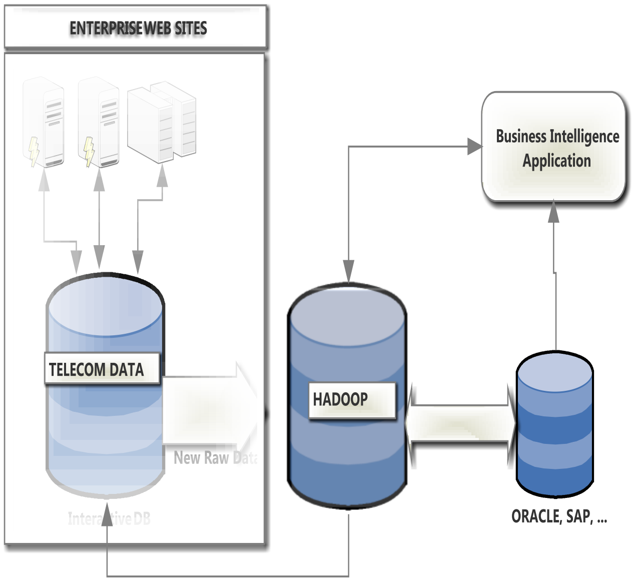


Figure 3: Proposed Method Architecture

7. RESEARCH METHODOLOGY

7.1 C4.5 Algorithm for classification

C4.5 is a program that creates a decision tree based on a set of labeled input data. This algorithm was developed by Ross Quinlan. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is commonly mentioned as a statistical classifier (”C4.5 (J48)” [2]. In this paper, C4.5 is used to classify the numerical and text attributes in Hadoop using HDFS (Hadoop Distributed File System) to predict the churn customers in telecommunication industry.

7.2 K-Means for Clustering

Estimation of the mean (vectors) of a set of K-groups is by means of a simple method called the K-means clustering algorithm [8].

Step #1: Randomly clustered seeds are selected initially. It is represented as the temporary.

Step #2: From each object to each cluster is computed by means of squared Euclidean distance and for the nearest (closest) cluster each object is allotted.

Step #3: The new centroid is calculated for each cluster and by the respective cluster centroid, every seed value is substituted.

Step #4: From an object to each cluster is calculated by means of the squared Euclidean distance and with the smallest squared Euclidean distance the object is allotted to the cluster.

Step #5: The cluster centroids are recalculated based on the new membership assignment.

Step #6: Till no object moves clusters, iterate the Steps 4 and 5.

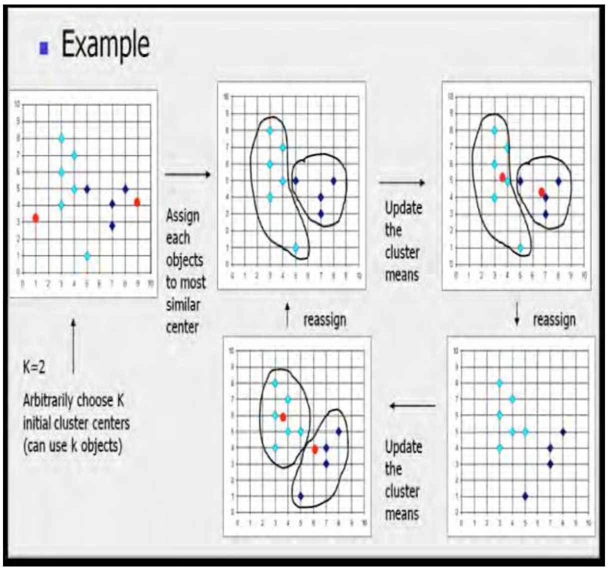

Example
Assign each objects to most similar center
Update the cluster means
reassign
reassign
Update the cluster means
K=2
Arbitrarily choose K initial cluster centers (can use k objects)

Figure 4: Original K-Means Clustering Algorithm Illustration

1. K-Means adaption to MapReduce framework

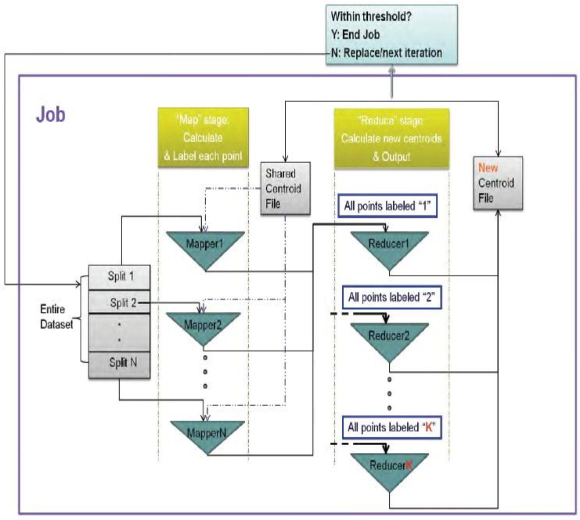


Figure 5: Map Reduce K-Means algorithm Illustration

As shown in Figure 3, the MapReduce adaption of the algorithms works as follows [3] [6]:

Step 0: To simplify the work, K points (K is the number of the clusters we wish to get out of the input dataset) are randomly selected from the input data file as initial cluster centers.

Step 1: Input data set (through a data file) is partitioned into N parts (controlled by the run-time system, not your program). Each part is sent to a mapper.

Step 2: In the map function, the distance between each point and each cluster center is calculated, and each point is labeled with the center index to which the distance is the smallest. Mapper outputs the key-value pairs of label assigned to each point, and the coordinates of the point.

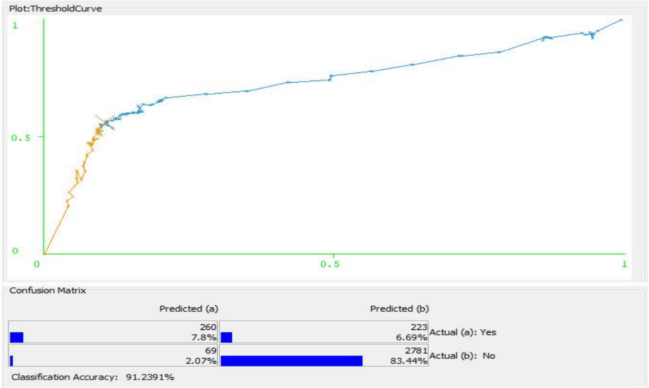
Step 3: All data points of the same current cluster (based on the current cluster centers) are sent to a single reducer. In the reduce function, new cluster center coordinates are easily computed. The output of the reducer is consequently the cluster index and its new coordinates.

Step 4: The new cluster coordinates are compared to the original ones. If the difference is within a preset threshold, then program terminates, and we have found the clusters. If not, use the newly generated cluster centers and repeat steps 2 to 4.

8. IMPLEMENTATION RESULT

The above data sets (North and South zone data sets) are implemented in Hadoop environment. The C4.5 algorithm is used to classify the numerical and text data in data set and used to predict the churn customers and K-Means clustering is used to group the classified data to improve the prediction in telecommunication industry using Hadoop Environment for cloud data sets.

8.1 Classification Result of South Zone



```
Confusion Matrix
              Predicted (a)     Predicted (b)
                   260               223
                  7.8%             6.69%   Actual (a): Yes
                    69              2781
                 2.07%            83.44%   Actual (b): No
Classification Accuracy:  91.2391%
```

Figure 6: South Zone classification Result

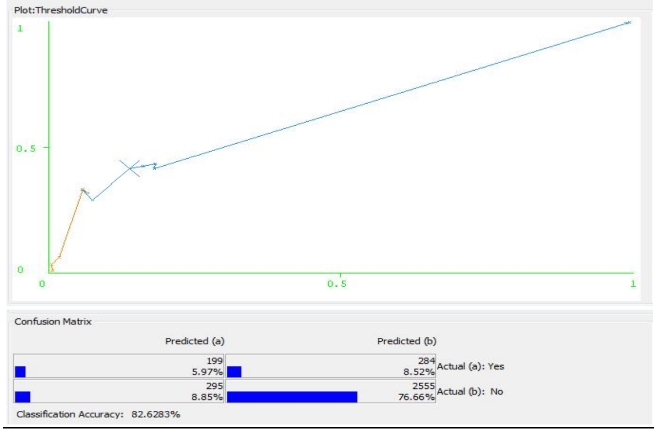
8.2 South Zone Clustering Result

Clustered Instances

Table 1: South Zone Clustering Result

0	922 (28%)
1	2411 (72%)

8.3 North Zone Classification Result



```
Confusion Matrix
              Predicted (a)     Predicted (b)
                   199               284
                 5.97%             8.52%   Actual (a): Yes
                   295              2555
                 8.85%            76.66%   Actual (b): No
Classification Accuracy:  82.6283%
```

Figure 7: North Zone classification Result

8.4 North Zone Clustering Result

Clustered Instances

Table 2: North Zone Clustering Result

0	1155(35%)
1	2178 (65%)

9. INTERPRETATION OF RESULT

In Hadoop environment, the cloud data set is classified to predict the churn customers in the telecommunication industry. For classification C4.5 decision tree algorithm is used and K-Means clustering algorithm is used for clustering in the Hadoop for the given cloud data set. The south zone classification result gives the 91.2391 % classification accuracy for the prediction and the confusion matrix helps to analyze the reasons for the positive churners and like this north zone data set gives 82.6283% classification accuracy and the confusion matrix for the churn customers. And the clustering in Hadoop is used to group the numerical and text data in the data set together. In south zone data set, there are 28% (922/3333) of text data and 72% (1144/3333) of numerical data in the cloud data set and there are 35% (1155) of text data and 65% (2178) of numerical data in the north zone cloud data set from out of 3333 instances.

10. CONCLUSION

This paper has presented the classification and clustering result for the given data set to predict the churn customers in telecommunication using Hadoop. It is hoped that this work will provide a better understanding of the performance of the classification and clustering techniques for the churn prediction. Hadoop was designed based on a new approach to storing and processing complex data. Therefore for large and complex data sets in the telecommunication industry, Hadoop can be used and cloud is used to store those data sets.

11. REFERENCES

[1] Ruxandra Stefania Petre., *Datamining In Cloud Computing*, Database system journals, pp.67-70.

[2] Jay Gholap, *Performance Tuning of J48 algorithm for Prediction of Soil Fertility*, pp.1-5.

[3] Himel Dev, Tanmoy Sen, Madhusudan Basak and Mohammed Eunus Ali, *An Approach to Protect the privacy of Cloud Data from Data Mining based Attacks*, pp.1-9

[4] Alexa Huth and James Cebula, *The Basics of Cloud Computing*, pp.1-4.

[5] Dr. Rajni Jain, *Introduction to Data Mining Techniques*, pp.1-11.

[6] B.Thirumala Rao, N.V.Sridevi, V.Krishna Reddy, L.S.S.Reddy, *Performance Issues of Heterogeneous Hadoop Clusters in Cloud Computing*, pp. 0-6.

[7] *Introduction to Cloud Computing and Hadoop*, pp.1-4.

[8] Zhexue Huang, *Extension to K-Means Algorithm for Clustering Large Data Sets with Categorical Values*, pp.283-304.

[9] Cellular Telecommunication Industry : 1990 – 2006.