IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Data Mining, Understanding and Modelling of Telecommunication Data-Set using Big Data Technology

Nirmal Ghotekar[1], Prof Ashish Manwatkar[2]

[1,2] *Department of Computer Engineering Indira College of Engineering and Management, Pune (India)*

ABSTRACT

Today the most competitive industries is considered as a Telecommunication industry. Due to competition and advancement of distributed computing every company tries to gain retrieve knowledge for their own benefits. By knowing customer's behaviour they can have more benefit in their business. This call detail record dataset size is huge in size. In this paper we have proposed work to analyse telecom data get insight into customer's behaviour. By this we will get user's daily behaviour and accordingly we get network traffic pattern. Call detail record is such large set of data that contain user's behaviour.

We conduct detailed analysis of network usage and subscriber behaviour using a largescale data set collected from a nationwide cellular data network. The data set records consist of millions of subscribers and large quantity event record and over thousands of base stations. We analyse individual subscriber behaviours patterns and observe a significant variation in network usage among different subscribers. We characterize subscriber mobility, subscriber behaviour and temporal activity patterns and identify their relation to traffic volume. We then investigate how efficiently radio resources are used by different subscribers or can be used by different subscribers. Traffic usage of different application also used for analysis. We also analyse the network traffic and optimised it from the point of view of the base stations and find significant temporal and spatial variations in different parts of the network, while the aggregated behaviour appears predictable. Broadly, our observations deliver important insights into network-wide resource usage. We describe usage in pricing, network protocol design and resource and spectrum management.

Keywords: - K-means, enhance K-means, Distributed processing, Machine learning, Normalisation, Spark.

I. INTRODUCTION

Call Detail Record (CDR) is telecom most important customer data which give insight for the customer. It can be used in telecom for its fundamental processes such as generating bills, finding any fraud in usage, finding location of the customer. Also used to find on road traffic. CDR is consist of duration, calling and called number and different flags as show in Call Detail Record table. Moreover, CDR may help to improve many existed processes and services in areas such as business intelligence, marketing, transportations and networking etc. There are many distributed tools can be used for analysing this data. Hadoop and Spark are most widely used tools in distributed computing. Spark performance is far better than Hadoop. Sometimes it can be 100 times faster than Hadoop.

II. DISTRIBUTED PROCESSING

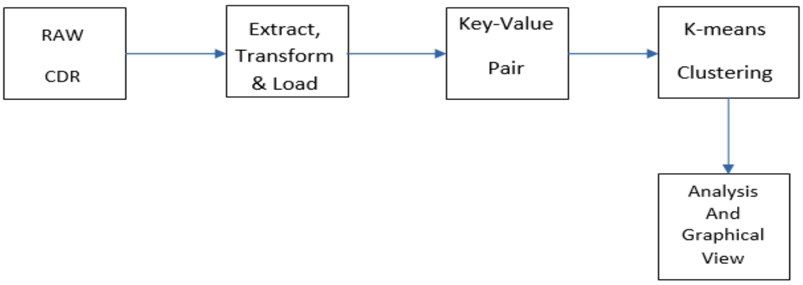
Apache HADOOP is apache open source framework used to develop data processing applications which are executed in a distributed computing environment such that time required to process data is to be reduce dynamically. It is open-source software framework which also work on commodity hardware. Similar to other local file system of personal computer system, in Hadoop, also files and data resides in a distributed file system which is called as a Hadoop Distributed File system (HDFS). Data Processing model is based on concept of ' Data Locality' wherein computational logic is sent to cluster nodes containing data. This computational logic is nothing but a compiled version of a high level programming model written in a high level language such as Java. Such a program or code, processes data stored in Hadoop HDFS in distributed way. Applications built using HADOOP are work on large size data sets distributed across clusters of commodity computers. Commodity computers hardware are widely available and cheap. These are mainly useful for achieving greater computational power at low cost.

There are two and a half types of machine in a HDFS cluster. First is Data-nodes, where HDFS actually stores the data, there are usually quite a few of these. Second is Name-node - the ‘master’ machine. It controls all the meta-data for the cluster. Eg - what blocks are used to make up a file, and what data-nodes those blocks are stored. Secondary Name-node, is another name node and this is NOT a backup name-node, but is a separate service that keeps a copy of both the all edit logs, and filesystem image, merging them periodically to keep the size reasonable. This is soon being deprecated in favour of the backup node and the checkpoint node, but the functionality remains similar (if not the same)). Spark is another tool to perform distributed computing.

II. SYSTEM OVERVIEW

System consist of different module. It consist of CDR collector, ETL operator, key-value pair creator. The value generated from it is used for performing clustering. KMeans applied on this data. Graph is shown as line graph to view performance of the system. Spark framework is used perform clustering and data filtration. Scala or java language can be used to write done algorithm

Fig. 1 System Diagram



III. SYSTEM ALGORITHM

Part A: Data filtering and pre-processing

Step 1) To view network traffic performance, to check Base stations performance and to check customer behaviour call data such as call details record and network traffic.

Step 2) Do following steps for behaviour and performance analysis check

Step 3) Place CDR data files on Hadoop Distributed File System such that data can be computed distributed manner.

Step 4) Filter required field in the CDR record for analysis.

Step 5) Perform Extract, transform and load operation.

Step 6) Calculate and measure traffic in Erlang from extracted data. Erlang traffic is per hour basis.

Step 7) Normalized monthly traffic of each base station to its maximum value.

Step 8) Put filter and formatted record for analysis.

Part B: Analysis and presentation of telecom data

Step 9) Apply improved and efficient K-means algorithm and perform clustering.

Step 10) Show record such as number of calls and duration of call has been done for analysis.

Step 11) Graphical representation shows cluster variation for pattern visualisation.

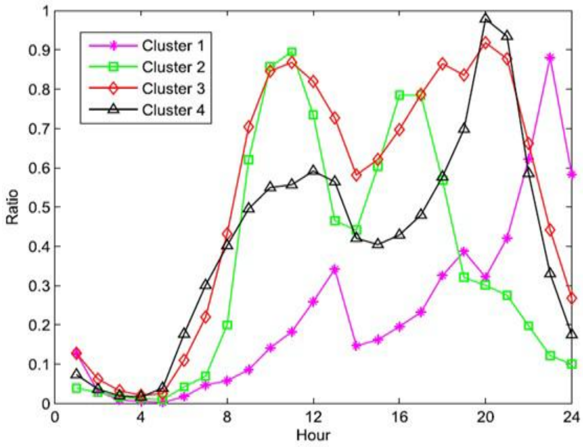
Part C: Compare Performance on basis of graph

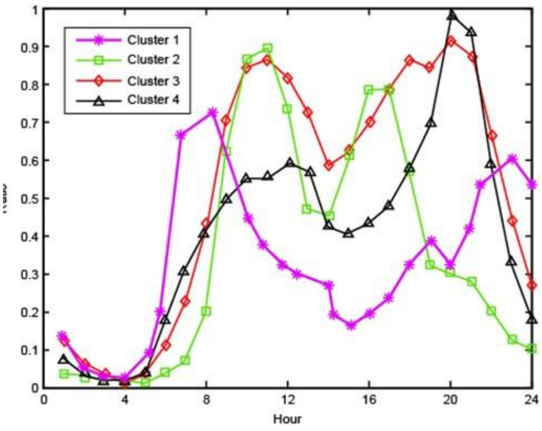
Step 12) compare Performance monthly basis.

IV. CLUSTER GRAPH

Graph 1 and 2 are line graphs to show variation in the performance of the network utilisation.

Fig. 2 and Fig. 3 Shows cluster graph





Above figure shows actual variation in users behaviour during the month if any new planning is or new plan is launched.

Call details Records

Column name	Sample data	Units
Calling party Number	98******12	
Called party Number	98******12	
Call Start time	1-Aug-2006 01:20:53	
Duration	20	Sec

Algorithm Performance Comparison

Algorithm	Centroid Selection	Accuracy	Performance
K-Means	Random Type	72	60
Improved KMeans	Selected Type	80	55

V. CONCLUSION

In this paper we have mainly focused on telecommunication problem of mining different types dataset which is very large in size, generally used data analysis techniques and case studies using CDR and Erlang measurement. By implementing K-means clustering algorithm we can understood daily traffic usage patterns. This helps for network planning and optimisation. Also helps to check performance of the system after new launch plan by seeing this cluster centroid graph in Fig.1 and Fig.2. Using clustering us can check different phenomenon's such as night burst phenomenon. We have also check performance of algorithm on different sizes of the system nodes. In previous paper [9] they used only K-means algorithm to check base station behaviour. Tempo-spatial analysis is done to find night burst phenomenon. Morans I in spatial dimension used to find abnormal station Also call arrival is model as Poisson process. There is no pricing scheme is describe. K-means performance also not improved in this. K-means performance can be increased on parallel or distributed processing. In our paper we proposed enhance K-means algorithm with Spark API. It improve KMeans algorithm efficiency and performance. As Spark do distributed processing, K-means on spark will help to improve both efficiency and performance. It is difficult to plan and configure telecom network. By understanding user behaviour a person can efficiently use available network and able to optimize the network. Our system will help managerial people to analyse network by using cluster graph as shown in Fig.2 and take fast decision. .

VI. ACKNOWLEDGEMENTS

I would like to express my gratitude to my Indira College Professor Ashish Manwatkar for providing me adequate facilities and knowledge to complete this paper. He give me more insight about the field. I express my gratitude for his support and suggestions regarding presenting paper in this international conference. I also thanks by head of department and Department of Computer Engineering for support and encouragement

REFERENCES

[1] M. Panda and S. P. Padhya, Traffic analysis and optimization of GSM network, International journal of computer science Issues, 2011.

[2] Hadoop basic. Available available on site: https://en.wikipedia.org/ wiki/ Apache Hadoop/

[3] S.Gopal Krishna Patro, Kishore Kumar Sahu Paper: Normalization: A Preprocessing Stage

[4] Holden Karau, Andy Konwinski, Patrick Wendell and Matei Zaharia Learning Spark, LIGHTNING-FAST DATA ANALYSIS, 2015

[5] Dean Wampler and Alex Payne, Programming Scala, SCALABILITY FUNCTIONAL PROGRAMMING + OBJECTS, 2014

[6] Satish and Rohan, Comparing Apache Spark and Map Reduce with Performance Analysis using K-Means algorithm, 2015

[7] K. Nazeer and Sebastian, Improving the Efficiency of the k-means Clustering Algorithm, 2009.

[8] Holden Karau, Andy Konwinski, Patrick Wendell & Matei Zaharia (Learning Spark LIGHTNING FAST DATA ANALYSIS-2015)

[9] D Yin, Y Zhang, W Zhou And S Zhang, (MEMBER of IEEE organisation) Computing on Base Station Behaviour Using Erlang Measurement and Call Detail Record, 2014