ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

Analysis and Data Mining of Call Detail Records using Big Data Technology

Nirmal Ghotekar[1]

Computer Engineering, Indira College of Engineering and Management, Savitribai Phule Pune University, Pune, India[1]

Abstract: Call Detail Record (CDR) is a very valuable source of information in telecom industry; it opens new opportunities and option for telecom industry and maximize its revenues as well as it helps the community to raise its standard of living in different ways. However, I need to analyse Call Detail Record in order to extract its big value which helps to find new business opportunities. Real time streaming data processing is became new trends in Call Detail Record processing. It helps to analyse Call Detail Record in real time and helps in finding real time location of any customer and also behaviour of network in real time. But these Call Detail Records has a huge volume, variety of data and different data rate, while current telecom systems are designed without considering these issues in mind. Call Detail Records can be seen as Big Data source, and hence, it is applicable to use Big Data technologies (for storage, processing and analysis) such as Hadoop in Call Detail Record analytics. There are considerable research efforts to address the Call Detail Records analysis challenges face in telecom industry. This project presents the use of Big Data technology in Call Detail Records analysis by giving some Call Detail Record analytics based application examples, highlighting their architecture, the utilized Big Data and Data Mining tools and techniques, and the Call Detail Record use case scenarios. In this project I am processing Call details record and calculating traffic in Erlang and using this I am creating cluster of base stations and checking its performance for different new pricing and bundling can be seen and management decision can be made.

Keywords: Convergent billing, K-means clustering algorithm, machine learning, normalisation, bundling, Hadoop.

I. INTRODUCTION

Telecom as a center for communication, telephony, video and internet, it has extreme verity of data such as, CDR data i.e user data, network data, and subscribers' personal and billing data. CDR is a record contains detailed information about a telecom transaction, such as call start time, end time, duration in seconds, call parties, cell ID, requested websites, type of data if calling or internet . And it also gives each event details that occur in the network. CDR Life cycle generally begins with CDR generation of a call, it is completed according to the events occurs in the call (call end, call join, etc.), then it is collected by different network elements. After that it goes through the mediation system; mediation systems format the raw CDRs to predefined standard format which common and compatible to others telecom system modules. Mediation system collect all billable as well as non billable events. Non billable event for checking system performance and other log purpose. Finally, it is written in the file system for later process use.

Call Detail Record (CDR) is telecom most valuable source of customer data, it is used in telecom fundamental processes (i.e. charging, settlement, billing, network efficiency, fraud detection, revenue assurance, churn detection [5], value added services, business intelligence, etc.) [3]. It consist of calling number, called number, duration and different flags as show in TABLE I. Moreover, CDR may help to improve many existed processes and services in areas such as business intelligence, marketing, transportations and networking etc.

A. Big data technology: Hadoop

Apache Hadoop is an open-source software framework which work on commodity hardware. It is used for distributed storage and distributed processing of very large data sets. It consists of computer clusters or node built from commodity hardware. All the modules in Hadoop are designed such a way that all hardware failures should be automatically handled and restore by the framework.

The core of Apache Hadoop frameworks are (a) Hadoop Distributed File System (HDFS): storage part, and (b) MapReduce: processing purpose. Hadoop splits files into large blocks and distributes them across nodes or cluster. It then transfers packaged code into different nodes to process the data in parallel. This approach takes advantage of data locality– nodes manipulating the data they have access to – to allow the dataset to be processed faster and very efficiently. The supercomputer architecture that relies on a parallel file system where computation and data are provided in distributed manner via high-speed networking. [8]

B. Data Pre-processing:

Normalization is scaling technique. It is used mainly in pre-processing stage. In this, we can find new normalise range from an existing or given data range. It can be more helpful a lot for the prediction or forecasting purpose. As

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

we know there are so many ways to predict or forecast but all can vary with each other. So to maintain the large variation of prediction and forecasting the Normalization technique is required to make them closer. But there is some existing techniques such as Min-Max, Z-score & Decimal scaling. I am using Min-Max technique. [9]

Advantages:
1. The technique which provides linear transformation on original range of data.
2. The technique which keeps relationship among original data.
3. Min-Mix Normalization. Min-Max normalization is a simple technique where data can be fit in a pre-defined boundary.

C. Clustering
It is method for grouping similar type of data object together. Intra cluster objects are similar to each other in characteristics. Inter cluster objects are different from each other. It may be depends on characteristics of object we are trying to group.

Methods for clustering
1. Hierarchical (Agglomerative)
- Initially each point in cluster by itself.
- Repeatedly combine the two nearest clusters.
- E.g. Hierarchical Clustering algorithm.

2. Point Assignment
- Maintain a set of clusters.
- Place points or data object into their nearest cluster.
- It is based on Euclidean distance.
- E.g. K-means Clustering algorithm.

D. Extract-Transform-Load (ETL):
It is a process in data warehousing responsible for pulling data out of the source systems and placing it into a data warehouse(a) Data extraction: In this data is extracted from homogeneous or heterogeneous data sources; (b) Data transformation: In this the data is transformed for storing in the proper format or structure for the purposes of querying and analysis; (c) Data loading: In this the data is loaded into the final target database, more specifically, an operational data-store, data-mart, data-warehouse or any storage system.

E. Convergent Billing:
Convergent billing enables to issue a single bill for a customer taking any type of service. A customer having presence only in a particular zone, spanning across the country can have a single bill for all the services he takes, whether the bill for the particular service is prepared or not from this system. [3]

Main components of the system:
- Customer Relationship Management (CRM)
- Billing
- Accounting
- Mediation
- Provisioning

II. RELATED WORK

In paper [1] using K-means behaviour of base station is checked. Tempo-spacial analysis is done to find nightburst phenomenon. By using Moran's I in spatial dimension abnormal station is find out. Also call arrival is model as Poisson process. There is no pricing scheme is describe. K-means performance also not improved in this.

In paper [2] bundling is telecom is describe. How different services are bundled together and provided to customer so that it can add new customer to the operator. BSNL tender document [3] gives insight into convergent billing system.It mainly consist of CRM, billing, mediation and accounting. Different preprocessing is explain in paper [9]
GSM network planning is explained in paper [7]. Traffic channel availability is well within limits which implies that the network is well planned.SDCCH congestion is beyond threshold limits which implies operator is losing big revenue due to this.
Major study in measurement analysis of subscriber and network behavior in a large scale 3G data network done in paper [4]. They have made several important observations related to traffic load, mobility and resource efficiency. they have indicated the implications of these observations in pricing, protocol design and resource management.

III. SYSTEM OVERVIEW

In this project I will present Call Detail Record processing and analysing using big data analysis. Such data generated in telecom industry is very large. Monthly more than on Terabyte of data generated. Due to this I have proposed distributed clustered based analytics in Hadoop system. Hadoop is open source, commodities hardware based architecture to perform operation on big data. In this project I am presenting what is present mean by CDR based billing system, what are operation performed on CDR data, how CDR helps to telecom industry to perform various operation such as churn prediction, fraud detection, subscriber pricing and bundling scheme and to find out behaviour of networks traffic.

This paper is for research on convergent billing system and data mining on CDR data. In this my system is consist of different stages. First store CDR flat file (raw) to Hadoop System. Raw data is encrypted data which is send over network. This raw data is send in .dat or .cdr file format. Raw CDR's are collected from different switches. That raw file decrypted and ETL operation perform on them. Formatted data extracted from ETL operation are used for analysis. Formatted data contain customer number, location, call start time, call duration.

From this telecom traffic is measured per hour. Unit for this traffic is Erlang. In many ways Erlang can be

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

measured. For my system traffic is measured in Erlang A. Monthly traffic is average hourly basis. Max value of every base station is considered for clustering. And on this data clustering algorithm is performed. This Erlang traffic is normalized for min-max value of traffic. Accordingly graph is plot for to show centroid of cluster variation in hourly basis as shown in Fig.2. This graph is used to show daily traffic patterns on hourly basis. Also see cluster after new plan launch.

By using different month's data we can analyse effect of new plan on customer and its behaviour can be analyse. This algorithm's performance is checked on single Hadoop node and on double Hadoop node. We can do Time and Space correlation of data as well as churn prediction on CDR data. I thus proposed a solution to perform Hadoop based big telecom data analysis.

TABLE I CALL DETAIL RECORD

Units	Sample data	Field name
	940****111	Calling number
	202****102	Called number
Seconds	1-JUN-2016 01:20:53	Start time
Minute	2	duration
	2048	Sector of each cell

IV. SYSTEM ANALYSIS

System architecture of the system consist of different module. It consist of Hadoop distributed file system where CDR files are place. On that pre-processing is done and key-value pair is generated. And using Hadoop framework repetitive work perform on different node and parallel processing is done with auto repairing function of Hadoop. Scala language is used in this to implement K-means clustering. Graph can be generated as shown in Fig 2.

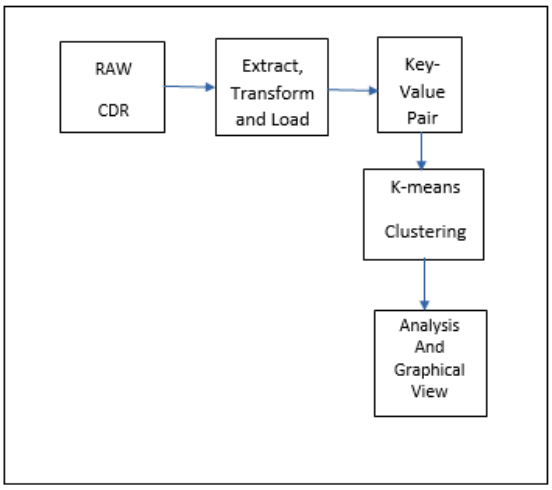

Fig.1 System Architecture

The general work of the system is as follows:
System Algorithm:-

Part A: Data pre-processing
Step 1) Place CDR flat files on Hadoop Distributed File System.
Step 2) Select required field for CDR analysis.
Step 3) Perform Extract, transform and load operation.
Step 4) Measure traffic in Erlang (type A) from extracted data. It is per hour basis.
Step 5) Normalized monthly traffic of each base station to its maximum value.
Step 6) Generate Key-Value pair to perform map-reduce operations.

Part B: Analysis and presentation
Step 7) Apply K-means algorithm and perform clustering.
Step 8) Graphical representation shows centroid of cluster variation in hourly basis.

Part C: Compare Performance
Step 9) As launch of new plan base station behaviour is checked. Base station behaviour and customer pattern checked before and after launch of new plan.

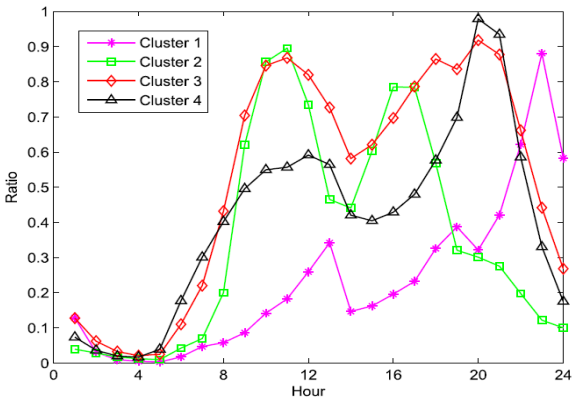

Fig. 2 Centroid pattern of each cluster. Cluster 1 to 4 are shown in above figure. Each consist of different base station

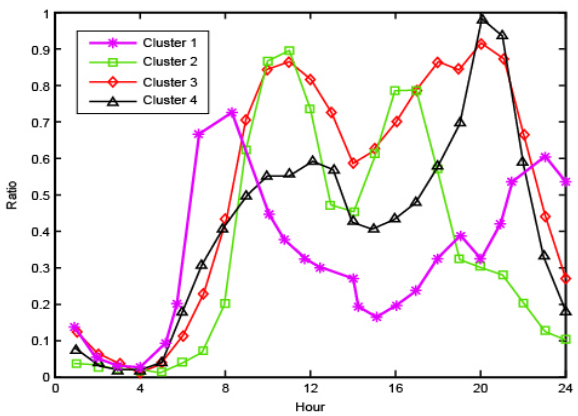

Fig. 3 Centroid pattern of each cluster after new plan launch. Check Cluster 1 in above figure.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

V. MATHEMATICAL MODEL

System's algorithm K-means clustering is defined as:
v = (v1, v2 .., v_j, . . . , vm)
x =(x1, x2 .., x_j, . . . , xm).
f (d)= minimum d(v, x)

f(d) = function to calculate and store minimum Euclidean distance
x_j= new mean or centroid of cluster

Where
v=the object vector
x=cluster vector
n=objects
m= dimension vector (centroid or mean for a cluster)
K =clusters
d(v, x)= function to calculate Euclidean distance between object and cluster vector

VI. CONCLUSION

In this project I have concentrates on telecommunication problem of large dataset analysis, generally used data analysis techniques and case studies using CDR and Erlang measurement. By implementing K-means clustering algorithm I can understood daily traffic patterns. This helps for fast and efficient way taking decision for network planning. Also helps to check performance of new launch plan by seeing this cluster centroid graph. Using clustering us can check hotspot and night burst phenomenon. I have also check performance of algorithm on multiple node.

In this project I studied Convergent billing system. Convergent charging platform combines different service networks, processing charging for both prepaid and postpaid subscribers. It enables operators to offer all types of services to all subscribers, providing convergent rating, charging and balance management.

ACKNOWLEDGMENT

I would like to express my gratitude to **Prof. Ashish Manwatkar** for providing me adequate facilities to complete this paper. I express my gratitude for her support and suggestions regarding dissertation. I also thank Department of Computer Engineering for support and encouragement.

REFERENCES

[1] Dandan Yin, Yanqin Zhang, Wuyang Zhou And Sihai Zhang, (MEMBER, IEEE) "Computing on Base Station Behavior Using Erlang Measurement and Call Detail Record", 2016.
[2] Bego˜na Garc´ıa-Mari˜noso, Comisi´on del Mercado de las Telecomunicaciones and Xavier Martinez-Giralt "Bundling in telecommunications", Universitat Aut`onoma de Barcelona Pau Olivella Universitat Aut`onoma de Barcelona, January 2008.
[3] BSNL TENDER DOCUMENT, Nov. 2009.
[4] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in Proc. IEEE INFOCOM, Apr. 2011, pp. 882-890. (Member, IEEE).
[5] Michael C. Mozer, Richard Wolniewicz, David B. Grimes, Eric Johnson, Howard Kaushansky "Churn Prediction in Wireless Industry".
[6] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
[7] M. Panda and S. P. Padhy, "Traffic analysis and optimization of GSM network," Int. J. Comput. Sci. Issues, 2011.
[8] Hadoop details website https://en.wikipedia.org/ wiki/Apache_Hadoop.
[9] S.Gopal Krishna Patro, Kishore Kumar Sahu. "Normalization: A Preprocessing Stage ".

BIOGRAPHY

Nirmal Ghotekar graduated with Bachelor of Electronic and telecommunication Engineering from Savitribai Phule Pune University, Pune, India. He has more than 7 years of industrial experiences with Government PSU. Currently he is pursuing his Master of Engineering study with Savitribai Phule Pune University, Pune, India in the Computer Engineering.