IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-5, Issue-5 E-ISSN: 2347-2693

Data Analyzing using Big Data (Hadoop) in Billing System

R. Din[1], Prabadevi B.[2*]

[1]School of Information Technology and Engineering, VIT University, Vellore, India
[2]School of Information Technology and Engineering, VIT University, Vellore, Tamil Nadu, India

Corresponding Author: rajudintagala23@gmail.com , Mob.No: 9942373427

Online Available at: www.ijcseonline.org

Received: 20/Apr/2017, Revised: 28/Apr/2017, Accepted: 21/May/2017, Published: 30/May/2017

Abstract**—** Hadoop is an open source structure in java that grants differing kind of immense datasets transversely over different groups of PCs using many programing models on which tremens -dous data works. By and large we saw that on the off chance that we increment the measure of the datasets away media, then recovering of information sets aside longer opportunity to prepare. Significant explanation behind this is because of the heap forced on information. So to take care of this kind of issues we utilize Big Data developed to fill this need. In this paper Hadoop eco-frameworks like Sqoop, hive, pig latin and so forth are utilized. Likewise we investigate expansive volume of power charging framework information and increased more prominent exactness in results, too it figures quick and furthermore recuperates loss of information.

Keyword**—** Sqoop, Hive, Pig, Hadoop, Volume

I. INTRODUCTION

Information is vital piece of our life so to spare information and store information fastly and exact way is extremely important. In this paper, to investigate the power charging system we store a large number of information in MySQL database. For the most part, MySQL database has an issue that is MySQL database does not store billion measured datasets.

So to conquer that issue we utilize enormous information idea on which we change over all information base into Hadoop stockpiling. For this utilized many breaking down device, for example, sqoop,hive,pig.All these apparatus have distinctive diverse examining idea. All these device that are utilized here is a piece of Hadoop biological community, for example sqoop,hive, pig and mapreduce calculations. Mapreduce calculations is utilized proficiently. In guide lessen calculations as a me demonstrates right off the bat mapping then decreasing all dataset. Big information is a gathering of large dataset that can not be handled on conventional technique. It is not a solitary method or a device rather it Involves numerous territories of business and techno - logy. In this examination we store past year of information, for example, (2011, 2012, 2013.etc) and utilizing that biological community instrument break down on earlier year dataset by utilizing Hadoop that is open source Framework by utilizing that with respect to dataset does not make any issue, for example, no loss of information, No capacity impediment, cost is less and good to all stage on the grounds that Hadoop is open source Framework in light of java .

This work is identified with brilliant meter which carefully stores the previous years of information. On which perform diverse distinctive examination, for example, specific client most extreme unit at specific year , how much pay that year and how way he do the installment and so forth information are recovered.

II. RELATED WORK

Ramon GranellI et.al, has proposed impacts of Raw Data Temporal Resolution Using Selected Clustering Method on Residential Electricity Load Profiles[1].

There is growing interest in discerning behaviors of electricity users in both the residential and commercial sectors. With the advent of high-resolution time-series power demand data through advance metering mining this data could be costly from the computational view- point. One of the popular techniques is clustering, but depending on the algorithm there solution of the data can have an important influence on the resulting clusters. This paper demonstrates how worldly determination of energy request profile affects the nature of the bunching process,the consistency of group part deliver(profiles showing comparative conduct), and the efficiency of the bunching procedure.

Pei Zhang et.al, has proposed Short-Term Load Forecasting Based on Big Data Technolog -ies[2].With the construction of smart grid, lots of renew- able energy resources such as wind and solar are deployed in power system. It might make the power system load varied complex than before which will bring difficulties in short-term load forecasting area. To overcome this issue, this paper proposes a new short-term load forecasting framework based on big data technologies. First ,cluster an aliases per formed to classify daily load patterns for individual loads using smart meter data. Next, an association analysis is used to determine critical influential factors.

M.K. Sheikh-El-EslamiI et.al,has proposed improving WFA K-means Technique for Demand Response Programs Applications [3].There are several pattern-based clustering methods which are used for different applications such as pattern recognition ,data mining, etc. In recent years, some of these methods are implemented in power system studies ,especially for clustering load curves for designing suitable tariffs, demand response pro -grams selection, etc. Choice of the best clustering method for certain application is one of the most important issues which is case dependent and should be considered in using of clustering load curves

Carlos León et.al,has proposed Variability and Trend-Based Generalized Rule Induction Model to NTL Detection in Power Companies[4].This paper proposes a comp rehensive framework to detect non-technical losses (NTLs) and recover electrical energy (lost by abnormalities or fraud) by means of a data mining analysis, in the Spanish Power Electric Industry. It is divided into four sections: data selection, data preprocessing, descriptive, and pred- ictive data mining. The authors insist on the importance of the knowledge of the particular characteristics of the Power Company customer: the main features available in databases are described. The paper presents two innovative statistical estimators to attach importance to variability and trend analysis of electric consumption and offers a predictive model based on the Generalized Rule Induction (GRI) model.

Yi Yang et.al ,has proposed A Time Based Markov Model for Automatic Position-Dependent Services in Smart Home[5].

A smart home is likely in the near future. An important ingredient in an intelligent environment such as a home is automatic services ,which means home sys- tem itself could know or predict what the inhabitant want to do, and so provide inhabitant the services automatically. Many researches reveal that most of the services in smart home are location dependent so the automatic services must be built on the location awareness. In this paper we model inhabitant location pattern as a time based markov model (TMM).

Jungsuk Kwac et.al,has proposed Household Energy Consumption Segmentation Using Hourly Data [6] . The increasing US deployment of residential advanced metering infrastructure (AMI) has made hourly energy consumption data widely available. Using CA smart meter data, we investigate a house -hold electricity segmentation methodology that uses an encoding system with a pre-processed load shape dictionary. Structured approaches using features derived from the encoded data drive five sample program and policy relevant energy lifestyle segmentation strategies.

Stephen Haben et.al ,has proposed clustering and analysis of Residential Consumers Energy Behavioral Demand Using Meter Data [7] . Clustering methods are increasingly being applied to residential smart meter data, which provides a number of important opportunities for distribution network operators (DNOs) to manage and plan low-voltage networks. Clustering has a number of potential advantages for DNOs, including the identification of suitable candidates for demand response and the improvement of energy profile modeling. However, due to the high stochastic city and irregularity of household level demand, detailed analytics are required to define appropriate attributes to cluster. In this paper, we present in-depth analysis of customer smart meter data to better understand the peak demand and major sources of variability in their behavior.

III. Methodology

1.System Architecture –

The proposed method performs well in the general population as well as in subpopulations. Results indicate that the proposed model significantly improves predictions over established baseline methods analyzing electricity consumption.The goal of this study was to analyze how much of units consumed in last four years and how much amount they paid previous four year as the forecast for the following year.

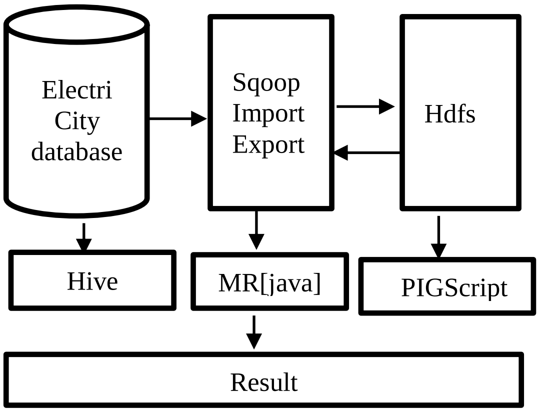


Fig 1.1 System Diagram

IV. Technique

In proposed technique 5 modules are used. Modules describe how the data are store and how way perform process on data. Modules are-

A. Data Preprocessing module-In this module we have to create Data set for Electricity Consumption it contain set of table such that customer details, billing details and payment details for last four years .and this data first provide in

MySQL database with help of this dataset we analysis this project. Data preprocessing Module diagram.

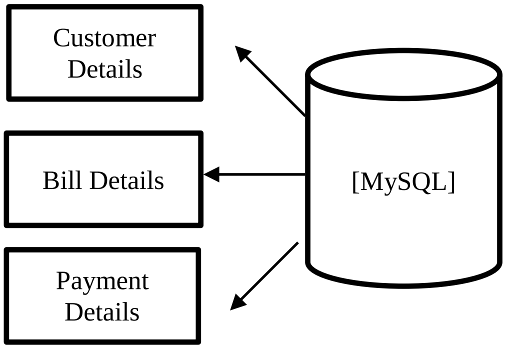


Fig 1.2 Data Preprocessing

B. Process with Sqoop module

Sqoop is a part of Hadoop eco system basically sqoop store the data structure type of data between relational databases and Hadoop.

In this module we fetch the dataset into Hadoop (HDFS) using Sqoop Tool. Using Sqoop we have to perform lot of the function, such that if we want to fetch the particular column or if we want to fetch the dataset with specific condition that will be support by Sqoop Tool and data will be stored in Hadoop (HDFS).

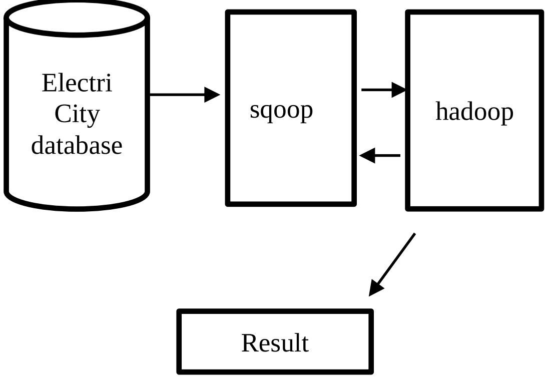


Fig.1.3 Sqoop Process

C. Data Analytic Module with hive-

In this module we have to analysis the dataset using HIVE tool which will be stored in Hadoop (hdfs). For analysis data set hive using HQL Language. Using hive we perform Tables creations, joins, Partition, Bucketing concept. Hive analysis the only Structure Language.

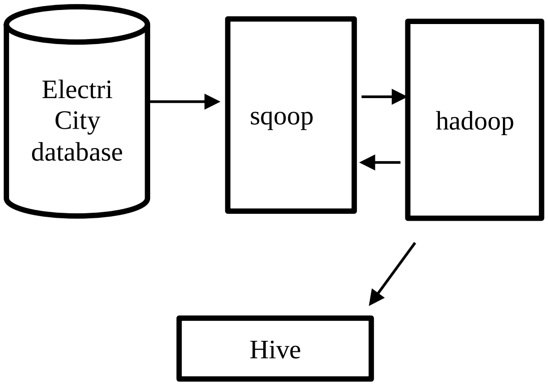
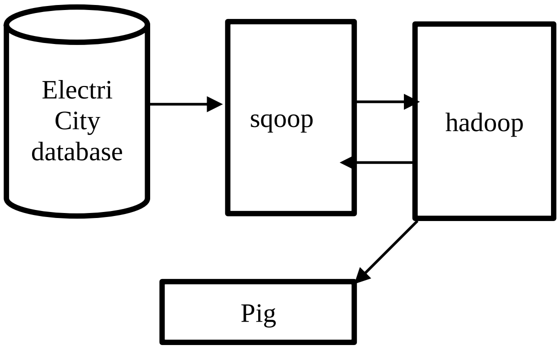


Fig.1.4 Hive Process

D. Data Analytic Module with pig-

Apache Pig is a high level data flow platform for execution MapReduce programs of Hadoop. The language for Pig is pig Latin. Pig handles both structure and unstructured language. In this module also used for analyzing the Data set through Pig using Latin Script data flow language.in this also we are doing all operators, functions and joins applying on the data see the result.



1.5 Pig Process

E. Data Analytic-Module with Map Reduce

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm include two tasks that is mapping and reducing.
In this module also used for analyzing the data set using MapReduce.MapReduce run by Java Program.

MapReduce is a mapping technique and a programming model for different-different platform based on java The MapReduce algorithm contains two important tasks, namely Map and Reduce.Map takes a set of data and divide it into another set of data, where individual elements are partition into tuples (key/value pairs) Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name this algorithms implies, the reduce task is always performed after the map job.

Map reduce algorithms basically basically contain three stage.

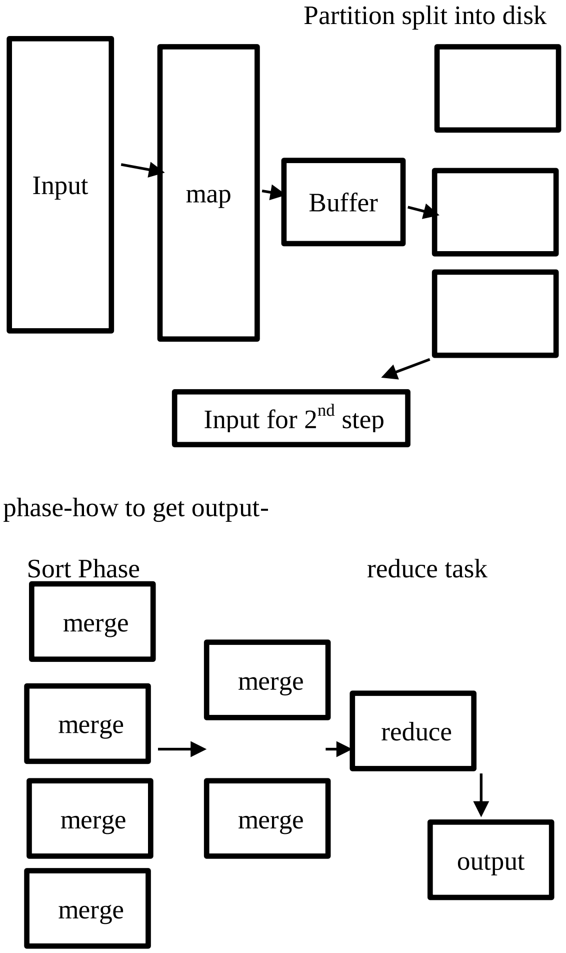
A..Map stage- The map or mapper's job is to process the input data. For the most part, the information is as document or

registry and is put away in the Hadoop record framework (HDFS). The info document is passed to the mapper function line by line.The mapper forms the information and makes a few little lumps of data.

B.**Reduce stage-**

This stage is the combination of the **Shuffle** stage and the **Reduce** stage. The Reducer's job is to process the data that comes from the mapper. After getting, all these stages it produces a new set of result, which will be stored in the HDFS. How shuffle and sorting work in partitioning

Ist phase-



2nd phase-how to get output-

Fig 1.6 MapReduce

V. Implementation

Data analyzing is performed using big data. It analyzes more and more data that studied on billions of data. It involves following steps:
Firstly,

- store thousands of data in MySQL database.
- Convert MySQL database to Hadoop platform via big data.

Secondly we convert in sqoop tool and analyze its performance

- Then convert that data to the hive and analyse its performance
- Then convert the dataset to pig and see the how partition and bucketing are perform using hive.

All these results in mapreduce analysis on the stored on local host.

VI. Result

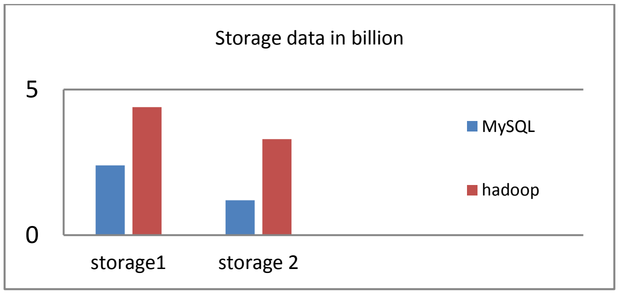


Fig 1.7 Result

The outcomes depict that proposed framework has more stockpiling limit than existing frameworks. Fig 1.8 clarifies how way storage , time and stack satisfy proposed framework objective.

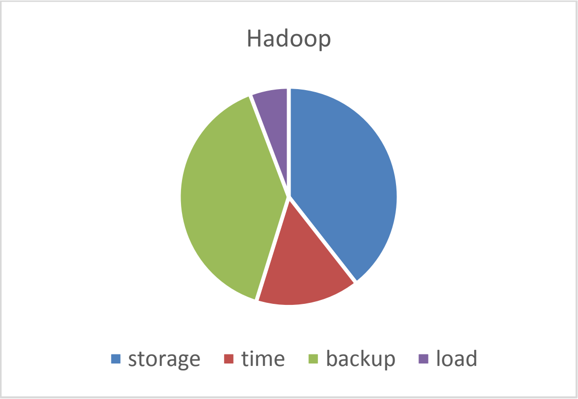


Fig 1.8 Graphical representation

When used Hadoop framework then all the results retrieved very fastly in very less time, load has also decreased than in MySQL database. It also give how way partitioning and bucketing are performed in MapReduce. So in this way analysis are perform in this paper.

VII. CONCLUSION

By using big data concept after converting all database to the Hadoop eco-system such as MySQL to sqoop and sqoop to hive, pig in turn applying MapReduce algorithms that tells how the data are analyzed and how to reduce time to access data and how to recover data. When accessed data using Hadoop, the accessing time is in milliseconds. Further for this project can be used with spark technology.

VIII. REFERENCES

[1] R Granell, CJ Axon, DCH Wallom, "*Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles*", IEEE Transactions on Power Systems, Vol.30, No.6, pp.3217-3224, 2015.

[2] Zhang, Pei, Xiaoyu Wu, Xiaojun Wang, Sheng Bi, "Short-term load forecasting based on big data technologies", CSEE Journal of Power and Energy Systems, Vol.1, No.3, pp.59-67, 2015.

[3] N. Mahmoudi-Kohan, M. P. Moghaddam, M. K. Sheikh-El-Eslami, S.M. Bidaki, "*Improving WFA k-means technique for demand response programs applications*", 2009 IEEE Power & Energy Society General Meeting, Calgary, pp.1-5, 2009.

[4] C León, F Biscarri, I Monedero, JI. Guerrero, J. Biscarri, R. Millán. "*Variability and trend-based generalized rule induction model to NTL detection in power companies*." IEEE Transactions on Power Systems, Vol.26, No.4, pp.1798-1807, 2011.

[5] Y. Yang, W. Zhiliang, Q. Zhang, Yang Yang, "*A time based markov model for automatic position-dependent services in smart home*", In Control and Decision Conference (CCDC), Chinaese, pp. 2771-2776, 2010.

[6] J Kwac, J Flora, R Rajagopal, "*Household energy consumption segmentation using hourly data*", IEEE Transactions on Smart Grid, Vol. 5, No.1, pp.420-430, 2014.

[7] S. Haben, C. Singleton, G. Peter, *"Analysis and clustering of residential customers energy behavioral demand using smart meter data*", IEEE Transactions on Smart Grid, Vol.7, No.1, pp.136-144, 2016.