Clustering Of Electricity Consumption Behavior Dynamics Toward Bigdata Applications

Lakshmi Narayana Reddy .P[1], G.Shobana[1]
[1]*Assistant Professor , Final year Information Technology,*

Abstract-The government agencies and the large multinational companies across the world focus on energy conservation and efficient usage of energy. The need of using energy in an efficient way is the need of developing countries like India and China .The emergence of smart grid meters gave us access to huge amount of energy consumption data. This data provided by smart meters can be used efficiently to provide insights into energy conservation measures and initiatives. Various energy distribution companies harness this data and get unpredictable results about customer's usage pattern; they then after performing analysis predict the demand and consumption of users. This analysis helps them to decide the tariff at different point of time. The companies are trying to overcome the bottleneck in capital investment cost of data. Further, processing Big Data for chart generation and analytics is a slow process and is not fast enough to support real time decision making. Our paper showcases a Business Intelligence tool which uses Apache Hadoop to efficiently handle the existing problems. Taking the advantage of this tool, energy distribution companies can reduce the investment by using community hardware that runs Hadoop. The usage of distributed computing tools also reduces the processing time significantly to enable real-time monitoring and decision making. This tool will also reduce carbon footprint and other related problems in energy distribution including loses and theft. In future this same analysis can be done on other utility resources such as gas and water.

Keywords – Big Data, Power Consumption,smart meter data, clustering liferstyles, hadoop.

I. INTRODUCTION

There are many smart meters that are deployed across to monitor the electricity or power consumption. This initiates a challenge to manage these data that are obtained through hourly or quater-hourly intervals. This information needs to be decoded and displayed that is comprehensible to the users. There are many more ubiquitous devices, smart grid technologies etc. The information provided should be understood by all the customers. Since the customer needs varies the information given needs to be according to the needs of the customer pricing and packaging.The data which is obtained provides a significance to understanding for the customer. By this data we can define and design a customer segment to consuming high energy levels. It also equips to use the resource correctly and tracking their necessities with the resource consumed and their outcome. It can find potential in smart grid management through finding energy-saving techniques.

The focus of this paper is to cluster the power consumption using load charts and to understand the needs of the customers. Based on this findings we can handle the customers power consumption more effectively. It provides segmentation of lifestyle based on the power consumption.

II. LITERATURE SURVEY

This paper[1] tells us about the load profiling applications in a deregulated market of power consumption. The procedure for this application is described in this paper. The paper explains about how the new participators are appearing and how significant the viable patterns. The strengths and weaknesses of each models are described. Eventually the results are discussed based on the algorithm defined. I am picking the fact that the total territory consumed power can be considered.

This paper [2] tells about each growing field in accordance with power consumption and how it can be used in day to day life. 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 viewpoint. 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 shows how temporal resolution of power demand profile saffects the quality of the clustering process, the consistency of cluster membership (profiles exhibiting similar behavior), and the efficiency of the clustering process. This work uses both raw data from household consumption data and synthetic profiles. The motivation for this work is to improve the clustering of electricity load profiles to help distinguish user types for tariff design and switching, fault and fraud detection, demand-side management, and energy efficiency measures. The key criterion for mining very large data sets is how little information needs to be used to get a reliable result, while maintaining privacy and security.

III. ARCHITECTURE

The proposed method performs well in the general population as well as in sub-populations. 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.

The proposed framework or architecture provides the following techniques which is shown in fig.1. This figure depicts how the data is obtained, how it has been moved from traditional MySql database to sqoop. Sqoop is a tool used to import MySql data to hadoop framework. Sqoop is a tool that is used to transfer relational database applications to hadoop framework. Relational databases include MySql,oracle to hadoop HDFS. It is also used vice versa to export from Hadoop file systems to relational databases. The database that we are going to take is an electricity consumption database. This database uses the data of past 4 years of data of electricity consumed. The data is preprocessed so that while performing hadoop functions and getting the output it will show us exactly what we require. The data is obtained from the government through open source websites and also from github. This data is now sent to the sqoop as an interface to transfer the relational data to hadoop framework. Once it has been transferred from traditional relational data to hadoop framework, we can use any one of the following to produce the required result.

The major three methods that can be used are Hive, Mapreduce and Pig. Applying the algorithms to these will improve the processing speed. Hive, Pig and Mapreduce are used in different scenarios. It is based on the team and differently skilled people that also affect the method that is going to be used. Pig and hive are alternative sources to mapreduce. Pig and hive are used when only a scripting language is required. Mapreduce is when java is used to code the program to the hadoop framework. Most of the works can be done using java programming that is through hadoop. This project is performed on all three methods. This project is a scalable project that can be used on different datasets. Keeping all the techniques available for future developers is necessary. Only a few command lines are required to perform operations on hadoop hive and pig. These few lines are included as a command line interface through linux. By using Pig and hive code need not be reviewed every time. The efficiency is handled by the previous developers.

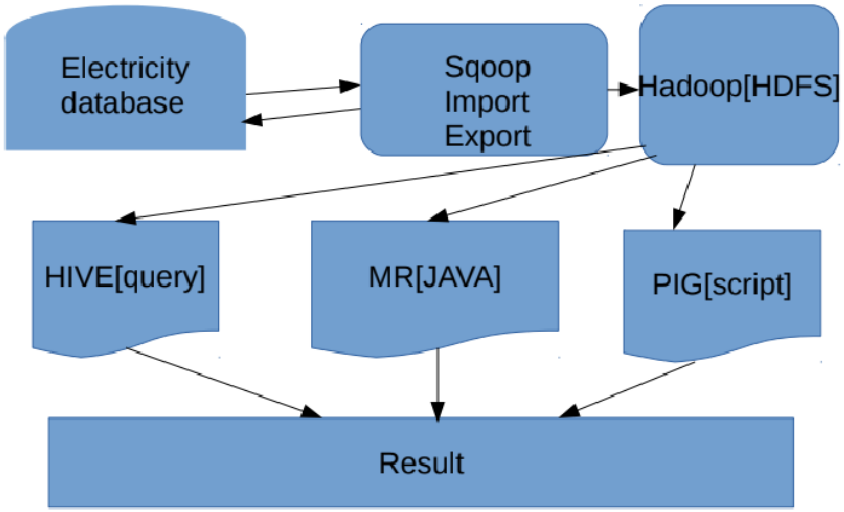

Fig. 1 Archtecture of database to hadoop

3.1. Implementation of modules

There are three modules in this project needs to be implemented. The modules are as follows:

- Data Preprocessing Module
- Data Migration Module With Sqoop
- Data Analytic Module With Hive
- Data Analytic Module With Pig
- Data Analytic Module With MapReduce

These modules are arranged as a sequence and can be executed only with the data preprocessing, Data migration and any one of he analytic module that is needed. The analytical modules are chosen based on the data that is used. In this project all the data analytical modules are used. This is a scalable project that can be used in other datasets. First

the data has to be collected and stored in databases such as MySql,NoSql in one of the traditional databases. After loading into the traditional database system the other modules are to be loaded into the hadoop framework using sqoop. The data of electricity is obtained from open source websites such as github and mygov.in. These websites provide genuine data for opensource project and other related purposes. The data consists of electricity payment, billing and customer details. All the data is used to show that whether the data is obtained from household or industry or providing energy for industry or household. The billing details tells us how much they paid each month. The payment details shows us how the person has paid each month. The payment and customer details show whether high volume of electricity is used, whether the customer requires it for household or industry purposes. The data preprocessing is done when the data needs to be consistent, not applicable to real-time standards etc. But this is genuine data from the website ensures that data need not be preprocessed.

The next module is on how the data is transferred from traditional databases to hadoop framework. Sqoop is the tool that is used to transfer the data from database to hadoop. There are basically two functions that are performed in sqoop. One is sqoop import and other is sqoop export. Sqoop import is used to transfer from relational databases to hadoop file system. All the data and records are stored as text. These are stores in text files or binary data. The sqoop export tool is used to take the data from hadoop framework back to relational database. This is done because the representation is easier in relational database system. The hadoop files are given as input to the system and then it is converted into records as in a relational database. There are few steps involved in setting up a hadoop file system. Installing java in the system. We need to configure hadoop with bringing changes to certain files like core-site.xml, hdfs-site.xml,mapred-site.xml and yarn-site.xml. Download and install sqoop. The extension is usually with .tar. The latest version of sqoop is preferred. Configure sqoop using your system that is associated with the local user. Import the tables that are needed for hadoop framework. For importing the jdbc needs to be connected to MySql. The table is imported as a text file or binary file. Considering the following table 1 that is imported in the hdfs.

ID	Name	Age
100	Ramu	29
101	Shyam	28

The file is inserted as follows
100 ,Ramu , 29,
101 ,Shyam ,28.

There are several kinds of imports that can take place from relational database to hdfs. Importing into a target directory takes place when a specified target is chosen in the hadoop file to be imported. We can import a subset of a table using the command where condition. Incremental import is one of the methods. In this method the records are inserted into the file one by one. Export syntax is as follows to export the file into the SQL table format.

```
sqoop export \
--connect jdbc:mysql          ://localhostdb \
--username root\
--table customer \
--export diremp/emp_data/
```

The next module is to choose which needs to be done based on the pros and cons given before. Firstly, in mapreduce the file requires at least ten libraries or more for a simple word count program that needs to be used. A tokenizer is created and all the words are taken as tokens at first. These tokens have data and their location. The data and location are separated and shown in the first mapping step. This mapping step is the third step of the actual mapreduceprorgam which requires a prerequisite of settling all the data properly. But since the data that we have used does not lack consistency, the data can be directly gone through mapping phase. The next phase is to group up all the same words in the file that is imported. Now these words and locations are grouped up based on the data similarity. The data which are same are grouped up. The output is displayed on a text file similar to the file that is imported.

The other two modules when chosen require scripting knowledge. They remove most of the overhead by not optimizing the queries that are created manually. These queries are executed by single command line interface. Hive uses a query language called the Hive query language (HQL) which is similar to SQL engine. Hive queries are compiled by the hive compiler. Hive queries are converted to mapreduce programs parallel across the hadoop cluster. This helps to focus more on the business problems than on the optimizing the queries.

The other way of executing the programs in hadoop is to use pig latin script. It uses only general SQL statements. Only filtering statements such as select are used mostly. The most preferable module that needs to be implemented is

mapreduce but since the effort required is much more when compared and it is not transferable to other programs, Hive and Pig are mostly used.

IV. RESULTS AND DISCUSSION

The result from this project is to use hadoop framework that is better than relational database format. There are many advantages that can be obtained from hadoop. They are used
instructured,semi structured and unstructured data. This can be used in large data storage. The average data that can be done using hadoop is in the size of petabytes and terabytes. The reading and writing data into hadoop are fast. Even though hadoop framework used pig or hive it is much faster in hadoop than in relational database. Hadoop usage is free, we can create databases but the efficiency is clearly seen in purchased products. Hadoop uses all types of data such as audio,video,analytics,data discovery etc. Relational database can be used mainly in OLTP. The hardware that is used in hadoop costs very less whereas high end servers are used in relational database.

V. CONCLUSION

The current system is designed mainly for power consumption. It is scalable to other datasets. The enhancement of this project is to implement apache spark on hadoop or on other standalone or it can run by itself. Spark is faster than hadoop running certain programs hundred times faster than hadoop and ten times faster on disk. The code that needs to be written on mapreduce is easier using spark. Certain command line interfaces can be used but since there is a wide range of languages that can be used, it is preferable to code for the necessary program. It also uses advanced analytics not only map and reduce functions are used here.

VI. REFERNCES

[1] I. P. Panapakidis, M. C. Alexiadis and G. K. Papagiannis, "Load profiling in the deregulated electricity markets: A review of the applications," in European Energy Market (EEM), 2012 9th International Conference on the, 2012, pp. 1-8.

[2] R. Granell, C. J. Axon and D. C. H. Wallom, "Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles," IEEE Trans. Power Systems, vol. 30, pp. 3217-3224, 2015.

[3] P. Zhang, X. Wu, X. Wang and S. Bi, "Shortterm load forecasting based on big data technologies," CSEE Journal of Power and Energy Systems, vol. 1, no. 3, pp. 59-67, 2015.

[4] Y. Wang, Q. Chen, C. Kang, M. Zhang, K. Wang, and Y. Zhao, "Load profiling and its application to demand response: A review," Tsinghua Science and Technology, vol. 20, pp. 117-129, 2015.

[5] N. Mahmoudi-Kohan, M. P. Moghaddam, M. K. Sheikh-El-Eslami, and S. M. Bidaki, "Improving WFA k-means technique for demand response programs applications," in Power & Energy Society General Meeting, 2009.PES '09. IEEE, 2009, pp. 1-5.

[6]420 IEEE transactions on smart grid, vol.5 “Household Energy Consumption Segmentation Using Hourly Data” JungsukKwac, Student Member, IEEE, June Flora, and Ram Rajagopal, Member,IEEE.

[7]IEEE transactions on power systems, “Variability and Trend-Based Generalized Rule Induction Model to NTL Detection in power companies”,CarlosLeon,Senior Member IEEE,FelixBiscarri,InigoMonedero Juan Igancio and RocioMillan.