Project Report

on

Data Analysis of deviation in power generation

Submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY

DEGREE

Session 2018-19
in

Computer Science Engineering

By
ISHANK BANSAL
1503210102
ABHISHEK BANSAL
1503210006

Under the guidance of

MR. ANAND KUMAR SRIVASTAVA

ABES ENGINEERING COLLEGE, GHAZIABAD

AFFILIATED TO
DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, U.P., LUCKNOW
(Formerly UPTU)

STUDENT'S DECLARATION

I ISHANK BANSAL and ABHISHEK BANSAL hereby declare that the work being presented in this report entitled "Data Analysis of deviation in power generation" is an authentic record of our own work carried out under the supervision of Mr."ANAND KUMAR SHRIVASTAV"

The matter embodied in this report has not been submitted by us for the award of any other degree.

Dated:

Signature of students(s)

Ishank Bansal (CSE)

Abhishek Bansal (CSE)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Signature of HOD
(Prof. (Dr.) ShaileshTiwari)
(Computer Science & Engineering Department)
Date.....

Signature of Supervisor
(MR. Anand Kumar Srivastava)
(Sr. Assistant Professor)
(Computer Science & Engineering Department)

CERTIFICATE

This is to certify that Project Report entitled “Data Analysis of deviation in power generation” which is submitted by ABHISHEK BANSAL AND ISHANK BANSAL in partial fulfillment of the requirement for the award of degree B.Tech in Department of Computer Science Engineering of Dr. A.P.J. Abdul Kalam Technical University, is a record of the candidate own work carried out by him under my supervision. The matter embodied in this thesis is original and has not been submitted for the award of any other degree.

Supervisor: Mr. Anand Kumar Srivastava

Date:

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Professor Anand Kumar Srivastava, Department of Computer Science & Engineering, ABES Engineering College, Ghaziabad for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Professor (Dr.) ShaileshTiwari, Head, Department of Computer Science & Engineering, ABES Engineering College for his full support and assistance during the development of the project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Signature:

Name : ISHANK BANSAL

Roll No. : 1503210102

Date :

Signature:

Name : ABHISHEK BANSAL

Roll No. : 1503210006

Date :

ABSTRACT

The optimal usages of Electricity is a major concern for the utilities in our country. With the advent of smart meters, the frequency of collecting household energy consumption data has increased, making it possible for advanced data analysis by which we can show the mapping with energy generation, which was not possible earlier [5]. These days with emerging developments in all sectors and growing demands, electricity has become priority for every individual and every organization. The basic procedure for power supply includes power generation, power transmission and power distribution to the destinations. Naturally owing to few technical faults, losses may occur due to power dissipation by some devices.

In developing countries like India, power cut / failure is one of the most prevalent issues which not only cause economic losses but also affects the industry development. It hampers functioning of industries and factories, due to shortage of power supplied to them and it also causes the shortage of power supply to homes. It leads to loss of revenue of Government, ultimately it is the country's economy and development.

Our project's objective is to analyze the data of the power generated across India and particularly of the state Uttar Pradesh where electricity failure is one of the major problem due to which agricultural, industrial and various industries gets adversely affected. The analysis will be done on the deviation of the real-time data of site meritindia.in and upsldc.org.

In this project we are analyzing Deviation parameter which shows the power generation and power consumption in state and country level.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN: Artificial Neural Network

KDD: Knowledge-discovery in data base

CHAPTER 1

INTRODUCTION

1.1 Problem Introduction

The optimal usages of Electricity is a major concern for the utilities in our country. With the advent of smart meters, the frequency of collecting household energy consumption data has increased, making it possible for advanced data analysis by which we can show the mapping with energy generation, which was not possible earlier [1].These days with emerging developments in all sectors and growing demands, electricity has become priority for every individual and every organization. The basic procedure for power supply includes power generation, power transmission and power distribution to the destinations. Naturally owing to few technical faults, losses may occur due to power dissipation by some devices. In this project we are analyzing Deviation parameter which shows the Power generation and power consumption in state and country level.

1.1.1 Motivation

In many poor countries economic growth is hampered by inadequate and irregular supplies of electricity. The scarcity and unpredictable supply of electricity in different parts is due to improper distribution and generation of electricity as well as lack of inadequate generating capacity. As a result, it is widespread across much of the developing world. Power cuts leads to lost government revenues, reducing the ability of the public sector to pay for the maintenance of existing facilities or to invest in new power generation; it places unexpected strains on already taxed and often inadequate infrastructure, increasing the risk and frequency of power shortages; and it reduces the availability of electricity to paying businesses and consumers [2].

1.1.2. Project Objective

Our project's objective is to analyze the data of the power generated across India and particularly of the state Uttar Pradesh where electricity failure is one of the major problem due to which agricultural, industrial and various industries gets adversely affected. These days with emerging developments in all sectors and growing demands, electricity has become priority for every individual and every organization. The analysis will be done on the deviation of the real time data of site meritindia.in[2] and upsldc.org[3].

1.1.3. Scope of the Project

In developing countries like India, power cut / failure is one of the most prevalent issues which not only cause economic losses but also affects theindustry development. It hampers functioning of industries and factories, due to shortage of power supplied to them and it also causesthe shortage of power supply to homes. It leads to loss of revenue of Government, ultimately it is the country's economy and development. Our project's objective is to analyze the deviation between the powers generated and consumption data of electric data state as well as country.

Chapter 2

LITERATURE SURVEY

2.1. Fraud Detection in electric power distribution networks

This methodology is based on knowledge-discovery in databases (KDD) process using datamining, commonly used to extract previously unknown, non-trivial, and useful patterns from different data sources. The steps are presented in the Figure 1.[6]

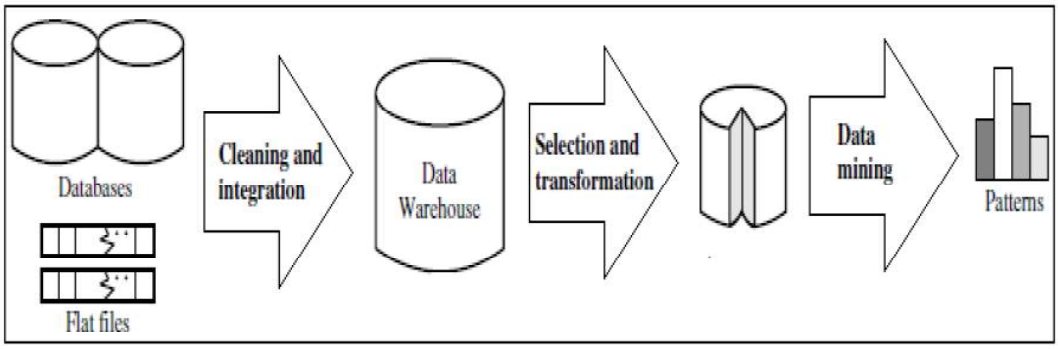

Figure 1: KDD process [6]

One of the usual ways to decrease non-technical losses rates is to perform local inspections to check if there are any thefts or hoaxes being committed by the consumers. For that purpose, the field staff is responsible for creating a methodology to generate an inspection schedule with suspected consumers. However, this task can be a too complex on due to the large amount of existing consumers in an electric power distribution network. Whenever a field staff performs an inspection, it returns only two possible outcomes: consumer is fraudster (there is irregularity) or non-fraudster (there is no irregularity). Therefore, whenever a fraud is found, appropriate measures are applied and the situation returns to normal. Nevertheless, if a fraud is not found some costs such the inspector's hours and vehicles costs are unnecessarily spent, instead of being applied to inspect others consumers that are committing energy frauds. Increasing the hit rate of identifying irregularities is needed to reduce costs in the fraud

detection, saving operational costs, even more importantly, identifying the fraudsters of the distribution network[7].

2.2. Knowledge-Discovery process

Table 1: Key attributes of the model

#	Name	Type	Description
1	Location	nominal	Consumer location (city + neighborhood)
2	Business class	nominal	Business class (e.g. residential, industrial, commercial, among others)
3	Activity type	nominal	Activity type (e.g. residence, drugstore, bakery, public administration, among others)
4	Voltage	nominal	Consumer voltage (110v, 220v)
5	Number of phases	nominal	Number of phases (1, 2, 3)
6	Situation	nominal	Is consumer connected? (yes, no)
7	Direct debit	nominal	Type of direct debit
8	Metering type	nominal	Type of electricity metering
9	Mean consumption	numeric	Mean consumption during the previous 12 months
10	Service notes	nominal	Are there service notes requested by consumers during the previous 12 months? (Yes or no)
11	Ownership exchange	nominal	Are there ownership exchanges during the previous 12 months? (Yes or no)
12	Query debits	nominal	Are there query debits during the previous 12 months? (Yes or no)
13	Meter reading	nominal	Are there meter reading notes during the previous 12 months? (Yes or no)
14	Inspection (output)	nominal	The inspection result (Fraudster or Non-fraudster)

The field staff is responsible for creating a methodology to generate an inspection schedule with suspected consumers. However, this task can be a too complex on due to the large amount of existing consumers in an electric power distribution network. Whenever a field staff performs an inspection, it returns only two possible outcomes: consumer is fraudster,there is irregularity or non-fraudster ,there is no irregularity.

CHAPTER 3

SYSTEM DESIGN AND METHODOLOGY

3.1. System Design

The proposal is to collect the real-time data from meritindia.in [2] andupsldc.org[3] using web-scraping in 2 separate files.The web-scraping code is written using python. Setting up of master-slave architecture is done on HDFS. Cleaning using Hadoop(Java Framework) in which the map-reduce programs will be written using Java and Pig Script. Now, when we have a bulk of cleaned data, we can do analysis over the data. All this will be done on a Linux platform.

3.2 System Architecture /Diagrammatical View

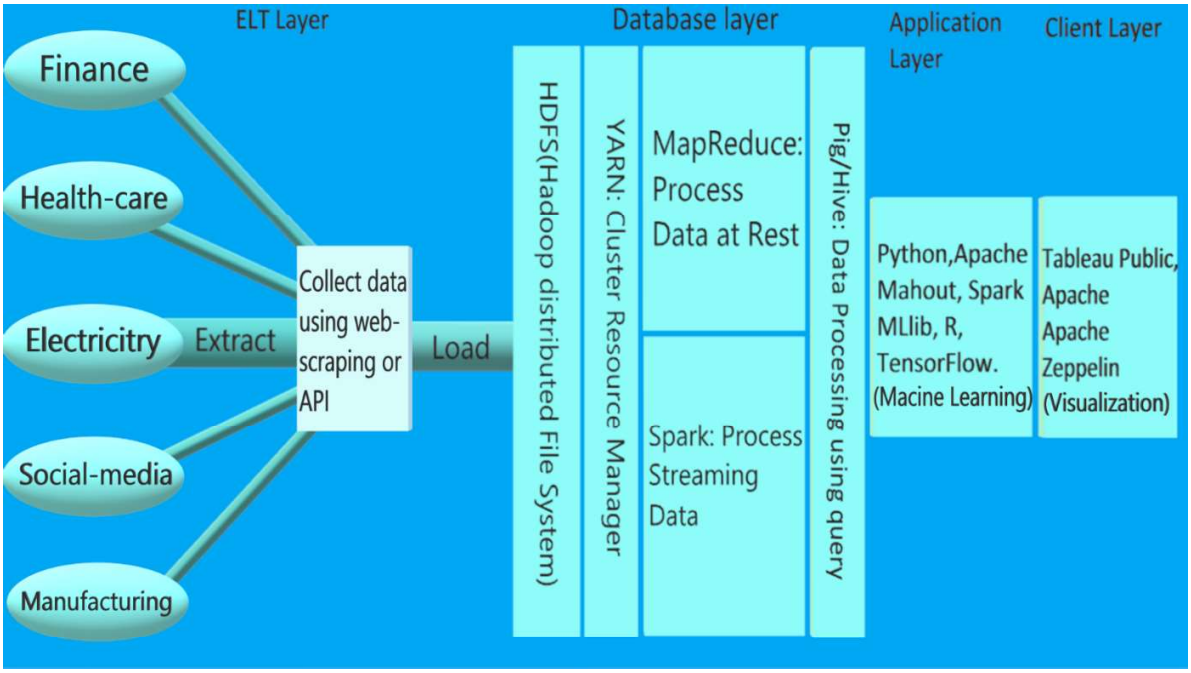

Figure3.2: Architecture Diagram

3.2.1 ETL LAYER

ETL stands for Extract, Transform and Load. It includes the processes required to manage raw data which is mostly homogeneous and enormous. Let me explain each of these processes in detail:

Extract: Extraction of data is the most important step of ETL which involves accessing the data from all the Storage Systems. The storage systems can be the RDBMS, Excel files, XML files, flat files, ISAM (Indexed Sequential Access Method), hierarchical databases (IMS), visual information etc. Being the most vital step, it needs to be designed in such a way that it doesn't affect the source systems negatively. Extraction process also makes sure that every item's parameters are distinctively identified irrespective of its source system.

Transform: Transformation is the next process in the pipeline. In this step, entire data is analyzed and various functions are applied on it to transform that into the required format. Generally, processes used for the transformation of the data are conversion, filtering, sorting, standardizing, clearing the duplicates, translating and verifying the consistency of various data sources.[4]

Load: Loading is the final stage of the ETL process. In this step, the processed data, i.e. the extracted and transformed data, is then loaded to a target data repository which is usually the databases. While performing this step, it should be ensured that the load function is performed accurately, but by utilizing minimal resources. Also, while loading you have to maintain the referential integrity so that you don't lose the consistency of the data. Once the data is loaded, you can pick up any chunk of data and compare it with other chunks easily.

3.2.2 DATABASE LAYER

3.2.2.1 HDFS(Hadoop Distributed file System)

Apache HDFS or **Hadoop Distributed File System** is a block-structured file system where each file is divided into blocks of a pre-determined size. These blocks are stored across a cluster of one or several machines. Apache Hadoop HDFS Architecture follows a Master/Slave Architecture, where a cluster comprises of a single NameNode (Master node) and all the other nodes are DataNodes (Slave nodes).HDFS can be deployed on a broad spectrum of machines that support Java. Though one can run several DataNodes on a single machine, but in the practical world, these DataNodes are spread across various machines.

3.2.2.2 NameNode

NameNode is the master node in the Apache Hadoop HDFS Architecture that maintains and manages the blocks present on the DataNodes (slave nodes). NameNode is a very highly available server that manages the File System Namespace and controls access to files by clients. I will be discussing this High Availability feature of Apache Hadoop HDFS in my next blog. The HDFS architecture is built in such a way that the user data never resides on the NameNode. The data resides on DataNodes only.

3.2.2.3 Functions of NameNode

It is the master daemon that maintains and manages the DataNodes (slave nodes).It records the metadata of all the files stored in the cluster, e.g. The location of blocks stored,the size of the files, permissions, hierarchy, etc. There are two files associated with the metadata:

FsImage: It contains the complete state of the file system namespace since the start of the NameNode.

EditLogs: It contains all the recent modifications made to the file system with respect to the most recent FsImage.

It records each change that takes place to the file system metadata. For example, if a file is deleted in HDFS, the NameNode will immediately record this in the EditLog. It regularly receives a Heartbeat and a block report from all the DataNodes in the cluster to ensure that the DataNodes are live. It keeps a record of all the blocks in HDFS and in which nodes these blocks are located. The NameNode is also responsible to take care of the **replication factor** of all the blocks which we will discuss in detail later in this HDFS tutorial blog. In **case of the DataNode failure**, the NameNode chooses new DataNodes for new replicas,balance disk usage and manages the communication traffic to the DataNodes.

3.2.2.4 DataNode

DataNodes are the slave nodes in HDFS. Unlike NameNode, DataNode is a commodity hardware, that is, a non-expensive system which is not of high quality or high-availability. The DataNode is a block server that stores the data in the local file ext3 or ext4.

3.2.2.5 Functions of DataNode

These are slave daemons or process which runs on each slave machine.The actual data is stored on DataNodes. The DataNodes perform the low-level read and write requests from the file system's clients. They send heartbeats to the NameNode periodically to report the overall health of HDFS, by default, this frequency is set to 3 seconds.

3.2.3 APPLICATION LAYER

The programming language which is used for this project is Python. Python is an interpreter language which also is a high level language. It has been developed by Guido van Rossum and is maintained by the Python Software Foundation.

A very special characteristic compared to other programming language is the separation of code blocs. Separations are used for inner function or commands which contains other operation like the if operator or diverse kinds of loops[5]. Very common in other languages is to use curly brackets at begin and at end of a block or using for this separation task a specific key word. As a separation indentation are used in Python which are readable for humans and for the interpreter. Such indentation is composed of four spaces. In the year 2008 the version 3.0 was released which brought explicit changes in the syntax of Python. Because it is not easy to migrate every code to 3.X the version 2.7 got an extended maintenance [Pet08]. This leads to the fact that now both versions are in use. For Python exists numerous packages which extend the function of the original.

Apache Spark is an open source big data processing framework built around speed, ease of use, and sophisticated analytics. It was originally developed in 2009 in UC Berkeley's AMPLab, and open sourced in 2010 as an Apache project. Spark has several advantages compared to other big data and MapReduce technologies like Hadoop and Storm. First of all, Spark gives us a comprehensive, unified framework to manage big data processing requirements with a variety of data sets that are diverse in nature (text data, graph data etc) as well as the source of data (batch v. real-time streaming data). Spark enables applications in Hadoop clusters to run up to 100 times faster in memory and 10 times faster even when running on disk. Spark lets you quickly write applications in Java, Scala, or Python. It comes with a built-in set of over 80 high-level operators. And you can use it interactively to query data within the shell. In addition to Map and Reduce operations, it supports SQL queries, streaming data, machine learning and graph data processing. Developers can use these capabilities stand-alone or combine them to run in a single data pipeline use case. In this first installment of Apache Spark article series, we'll look at what Spark is, how it compares with a typical MapReduce solution and how it provides a complete suite of tools for big data processing.

3.2.4 Client Layer

Visualizing data is important regardless of the size of the data because it translates information into insight and action. The approach to visualizing Big Data is especially important because the cost of storing, preparing and querying data is much higher. Therefore, organizations must leverage well-architected datasources and rigorously apply best practices to allow knowledge workers to query Big Data directly. Big Data has been home to a great deal of innovation in recent years - thus there are many options available, each with their different strengths. Tableau's vision is to support any Big Data platform that becomes relevant to our users, and help them facilitate a real-time conversation with their data.[6]

CHAPTER 4
IMPLEMENTATION AND RESULTS

4.1 Software and Hardware Requirements

4.1.1 SOFTWARE REQUIREMENTS

4.1.1.1 Python for Web Scraping -

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser.

4.1.1.2. Apache Hadoop 2.X

Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model.

4.1.1.3. Apache Pig

Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called Pig Latin. Pig can execute its Hadoop jobs in MapReduce, Apache Tez, or Apache Spark

4.1.1.4 Core Java

Although the **Hadoop** framework is implemented **in Java**™, **MapReduce** applications need not be written **in Java**. **Hadoop** Streaming is a utility which allows users to create and run jobs **with** any executables (e.g. shell utilities) as the mapper and/or the reducer.

4.1.2 HARDWARE REQUIREMENT

4.1.2.1 Master Node (i5 Processor, RAM-8GB, 1TB Hard disk)

The **master nodes** in distributed **Hadoop** clusters host the various storage and processing management services, described in this list, for the entire **Hadoop** cluster. NameNode: Manages **HDFS** storage. To ensure high availability, you have both an active NameNode and a standby NameNode.

4.1.2.2 Slave Node (i3 Processor, RAM-8GB, 1TB Hard dis)k)

In a **Hadoop** universe, **slave nodes** are where **Hadoop** data is stored and where data processing takes place. The following services enable **slave nodes** to store and process data:**DataNode**: An **HDFS** service that enables the NameNode to store blocks on the **slave node**. RegionServer: Stores data for the HBase system.

4.1.3 OPERATING SYSTEM

4.1.3.1 Linux (Ubuntu 16.04/18.04)

The reason **why Linux** is the operating system of choice is pretty simple: because **Linux** is free and **Hadoop** distribution already includes native libraries that are built for **Linux**, the last makes it much easier to set up and maintain on **Linux** than on any other platform.

4.2 OUTPUT

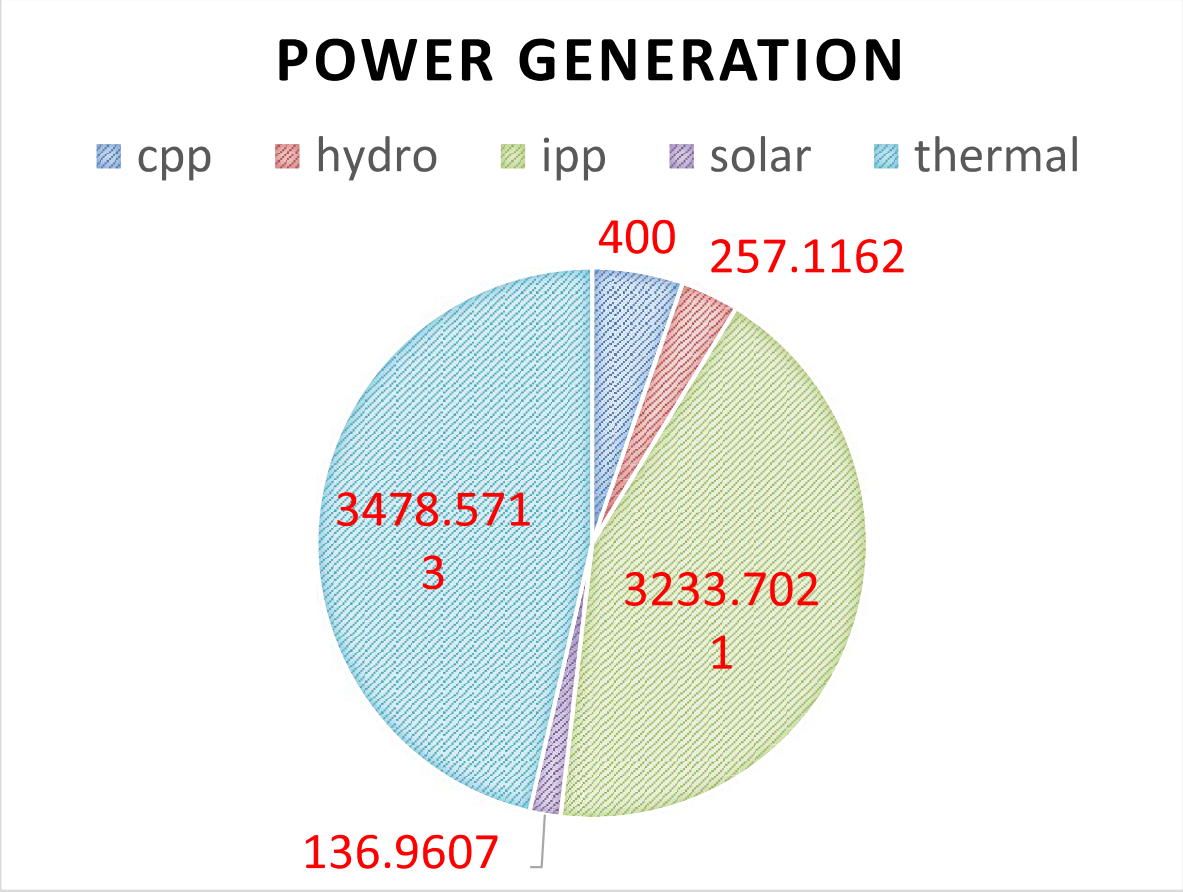

Figure4.2: Pie-chart showing deviation

The above pie-chart is the final output of our project. As we are calculating the deviation between the data of the two sites , meritindia.in and upsldc by collecting the data in the two csv files and uploading them on HDFS. The figures in the above pi-chart is showing the deviated result among the different power generated sectors i.e cpp,hydro,solar etc.

CHAPTER 5
CONCLUSION

5.1 Performance Evaluation

As we are using bigdata analytics therefore we can use our model for large data files as well hence our model is efficient in calculating the deviating efficiently.

5.2 Comparison with existing State-of-the-art Technologies

One of the usual ways to decrease non-technical losses rates is to perform local inspections to check if there are any thefts or hoaxes being committed by the consumers. For that purpose, the field staff is responsible for creating a methodology to generate an inspection schedule with suspected consumers. However, this task can be a too complex on due to the large amount of existing consumers in an electric power distribution network. Whenever a field staff performs an inspection, it returns only two possible outcomes: consumer is fraudster (there is irregularity) or non-fraudster (there is no irregularity). Therefore, whenever a fraud is found, appropriate measures are applied and the situation returns to normal. Nevertheless, if a fraud is not found some costs such the inspector's hours and vehicles costs are unnecessarily spent, instead of being applied to inspect others consumers that are committing energy frauds. Increasing the hit rate of identifying irregularities is needed to reduce costs in the fraud detection, saving operational costs, even more importantly, identifying the fraudsters of the distribution network[7]

5.3 Future Directives

In developing countries like India, power cut / failure is one of the most prevalent issues which not only cause economic losses but also affects theindustry development. It hampers functioning of industries and factories, due to shortage of power supplied to them and it also causesthe shortage of power supply to homes. It leads to loss of revenue of Government, ultimately it is the country's economy and development. Our

project's objective is to analyze the deviation between the powers generated and consumption data of electric data state as well as country.

With the help of Machine learning and predictive analysis we can also predict the demand of power needed in a particular state as in summers electricity needed is high while in winters that is quite low.

References

[1] “Theft and Loss of Electricity in an Indian State”,” Miriam Golden”,https://www.theigc.org/wp-content/uploads/Golden-Min-2012-Working-Paper.pdf, November 4, 2019

[2] “merit india website”, http://meritindia.in/, February 7,2019

[3] ”Upsldc website”, http://www.upsldc.org/real-time-data, February 7,2019

[4] ”Governing clean energy transition in India”,”Karoliina Isoaho”, https://www.wider.unu.edu/sites/default/files/wp2016-28.pdf, March 3,2019

[5] “Estimating the level and distribution of global electricity”, “James B. Davies”, https://www.wider.unu.edu/sites/default/files/wp2016-3.pdf

[6] “The role of electricity in the development of the global economy”,”James B.”, October 10,2019.

[7] ”Fraud Detection in Electric Power Distribution Networks using an Ann-Based Knowledge-DiscoveryProcess”,https://www.researchgate.net/publication.pdf, October 11,2019