Energy Consumption

Analytics and Prediction

with

Apache Spark

By: Laksh Lumba & Kshitij Kaushik

Table of Contents

Abstract

Energy fluctuation is one of the biggest problems in developing countries like India. To be exact, demand for energy changes with the time and various other factors such as weather conditions like in summer we need to switch on the AC, similarly in winter we require heater. Thus an energy requirement varies over the period of time. Also we know that storing the energy for future demands is very inefficient. So as the demand increases energy companies are required to meet up these requirements even during the peak time. One of the solutions to meet the requirement is generating enough amount of electricity to meet the demands. But this leads to wastage to energy. Another solution is analyzing the previous year data and predicts the future consumption and demands. So In this project our main target is to predicts the next year daily energy consumption by analyzing the last four year per minute given energy consumption.

Code Repository

https://github.com/lakshlumba/Team_8_Final_Project_Energy_Perdiction_Spark.git

Available Data Source

Previous four year data (December-2008 to November 2012) has been given to us:
https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption

Dataset Information:

- **Date:** Date in format dd/mm/yyyy
- **Time:** Time in format hh:mm:ss
- **Global_Active_Power:** household global minute-averaged active power (in kilowatt)
- **Global_Reactive_Power:** household global minute-averaged reactive power (in kilowatt)
- **Voltage:** minute-averaged voltage (in volt)
- **Global_Intensity:** household global minute-averaged current intensity (in ampere)
- **Sub_Metering_1:** energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered).
- **Sub_Metering_2:** energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.
- **Sub_Metering_3:** energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.

Problems we are solving

- What will be next day energy consumption i.e. on 27th December 2012?
- What will be Average Revenue Loss for the particular day next year, if there will be full day outage and tariff plans are given below:

Tariff plan

Time Period	Tariff (Rupees per KWh)
12 AM to 5 AM	4
5 AM to 7 AM	6
7 AM to 10 AM	12
10 AM to 4 PM	4
4 PM to 8 PM	6
8 PM to 10 PM	10
10 PM to 12 AM	6

- What will be next year week wise energy consumption starting from 27th December 2012 to 26th December 2013?
- Device Usage Pattern from the previous year data?
- Peak Time Load for the next one month during weekend and Weekdays? Assuming Peak time is during (7am to 10 am)

Aggregate the minutely given data into

- Hourly data
- Daily data
- Monthly data
- Yearly data.

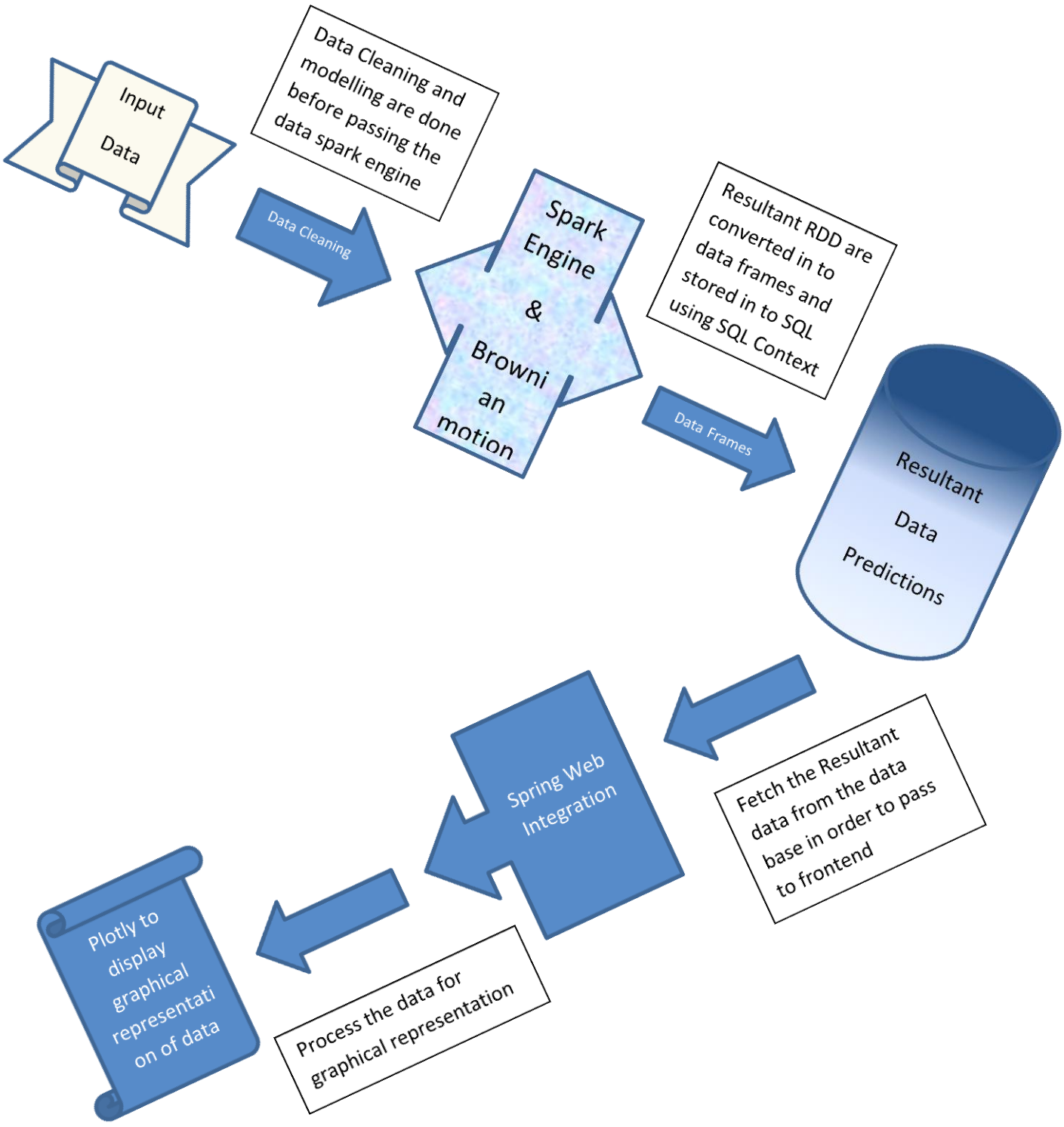
Technology and Implementation

- Apache Spark 1.4.1
- Scala
- Maven
- MySQL
- Spring Integration
- Plotly.js
- Tableau

Why Spark and Scala

- We use Spark machine learning Algorithms in order to analyze the given data because of its lightning fast cluster computing facilities.
- Another main reason of Apache Spark is ease of coupling with the My SQL. We can easily covert the RDDs in to data frames, these data frames can easily be stored in the form of tables.
- Another main reason for using Spark is its MLlib algorithms in order to predict the result.
- We use Scala because Spark is scala based platform which embeds the JVM libraries, As a java developer it is very easy to comprehend with Scala as it is JVM based functional language and is well known for its concurrent capabilities.
- To write map reduce machine learning algorithms, we need to very few lines of code.

Flow Diagram



Flow Implementation

Data Cleaning: We cleaned the input data by the removing the missing data ("?") and then we remove the duplicate data from the file. Once the data is cleaned we did the data sampling, means we confirmed the result with small amount of data and followed the same procedure throughout the data.

Data Prediction: We use the Brownian motion and Spark Linear Regression Algorithms in order to predict the various results. As per Brownian motion change of amount in one unit of time is normally distributed with μ and σ, where μ is the drift and σ denotes the volatility. Brownian motion suggests the following equation:

$$S_{i+1} = S_i \mu \Delta t + S_i \sigma \varepsilon \sqrt{\Delta t}$$

- S_{i+1}: Predicted value
- S_i: Present known value
- μ: Drift
- σ: Volatility
- ε: Random Number

Now we will discuss the approach to calculate the next month peak load during weekday and weekend.

To start with firstly we aggregate the minutely based data in to hourly basis data and then filtering the data based on peak time frames and splitting it into weekday and weekend. Further we convert the data in to hourly based, finally monthly based. Aggregated monthly based data is used to predict the peak time load of week days and weekends using Spark Linear Regression and Brownian motion prediction.

Please check the below for full implementation.

https://github.com/lakshlumba/Team_8_Final_Project_Energy_Perdiction_Spark.git

Next we will discuss the approach to calculate the Average Revenue Loss for particular if there is power outage and with the given tariff plan.

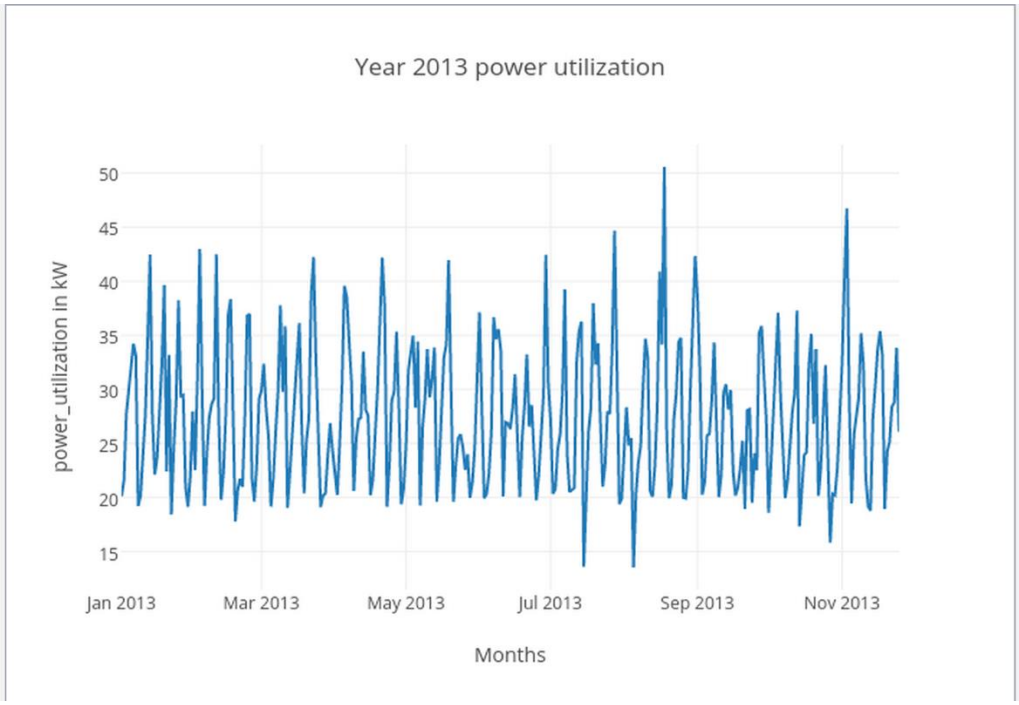
In this approach we aggregate the minutely data into hourly data and then we calculate the total cost per day by multiplying the power consumption and tariff plan for that particular hour. We did this for four years daily data and find the average value of the total cost.

Results and Discussions:

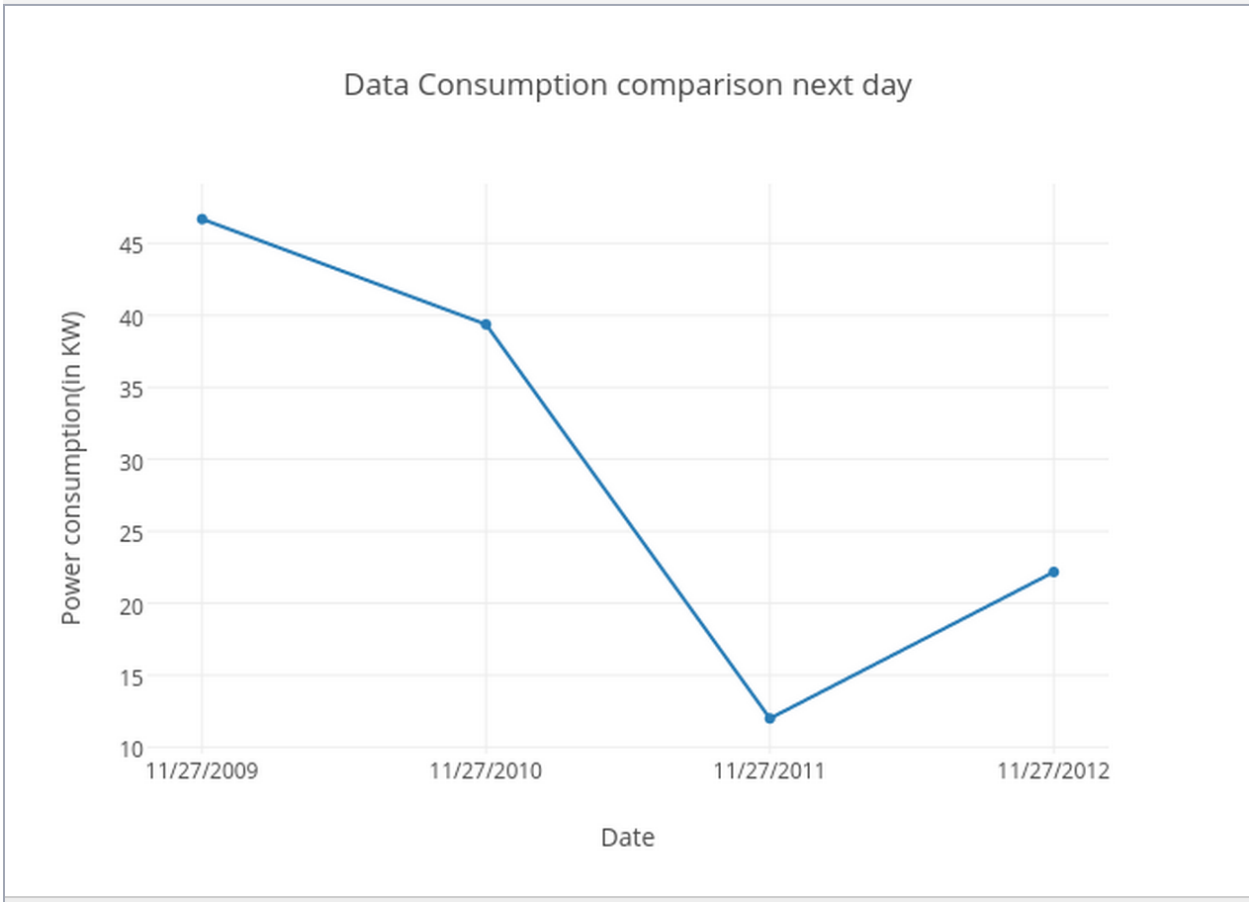
To capture the result, we convert the result RDDs in to Spark Data Frames and we use the Spark SQL context in order to save the resultant data frames in to tables. In order to present the graphical representation, we use the spring web and plotly.js libraries to display the data in the form of bar charts and histograms.

• Graphical Representation & Plotting:

- Total Year Power Consumption 2013

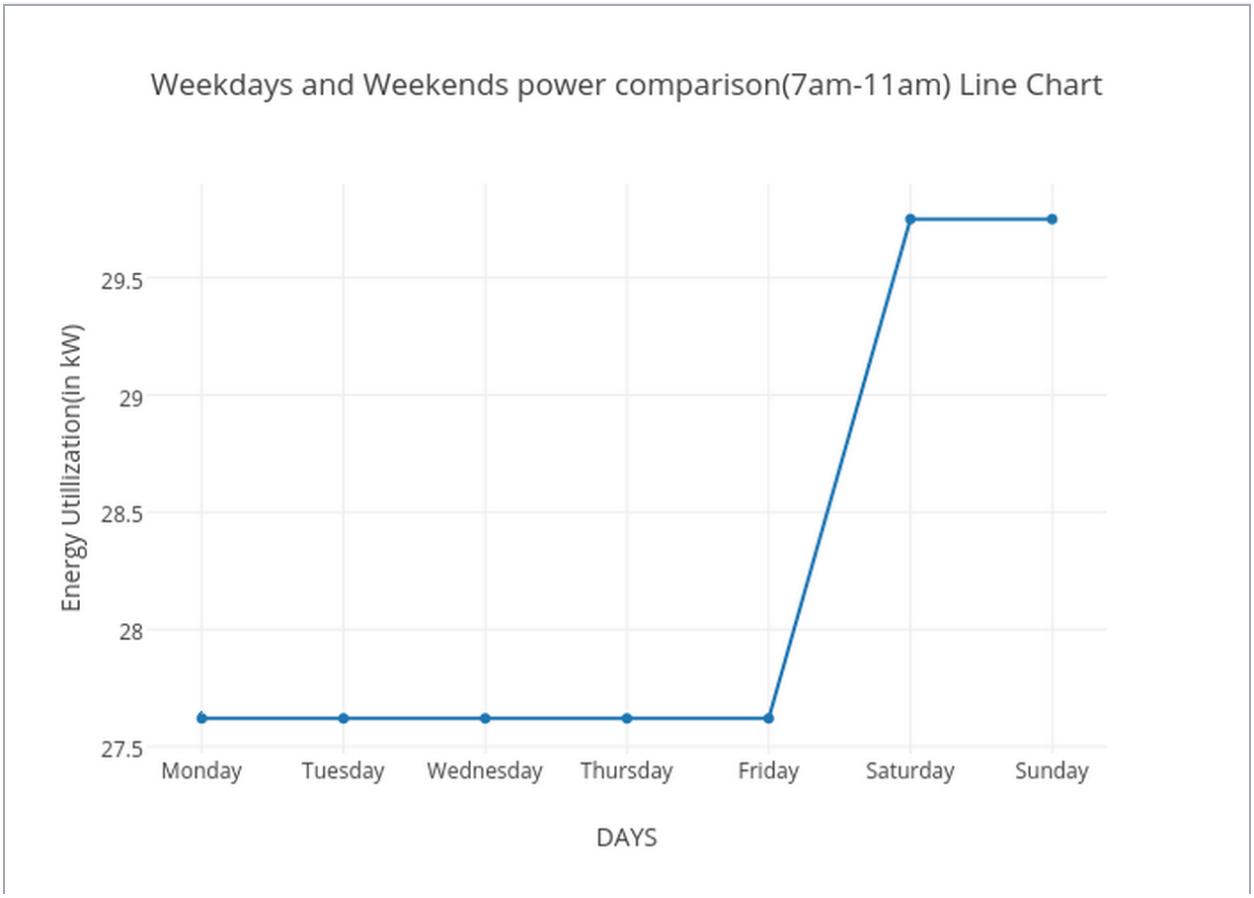


- Next Day Power Consumption Comparison (4 years)

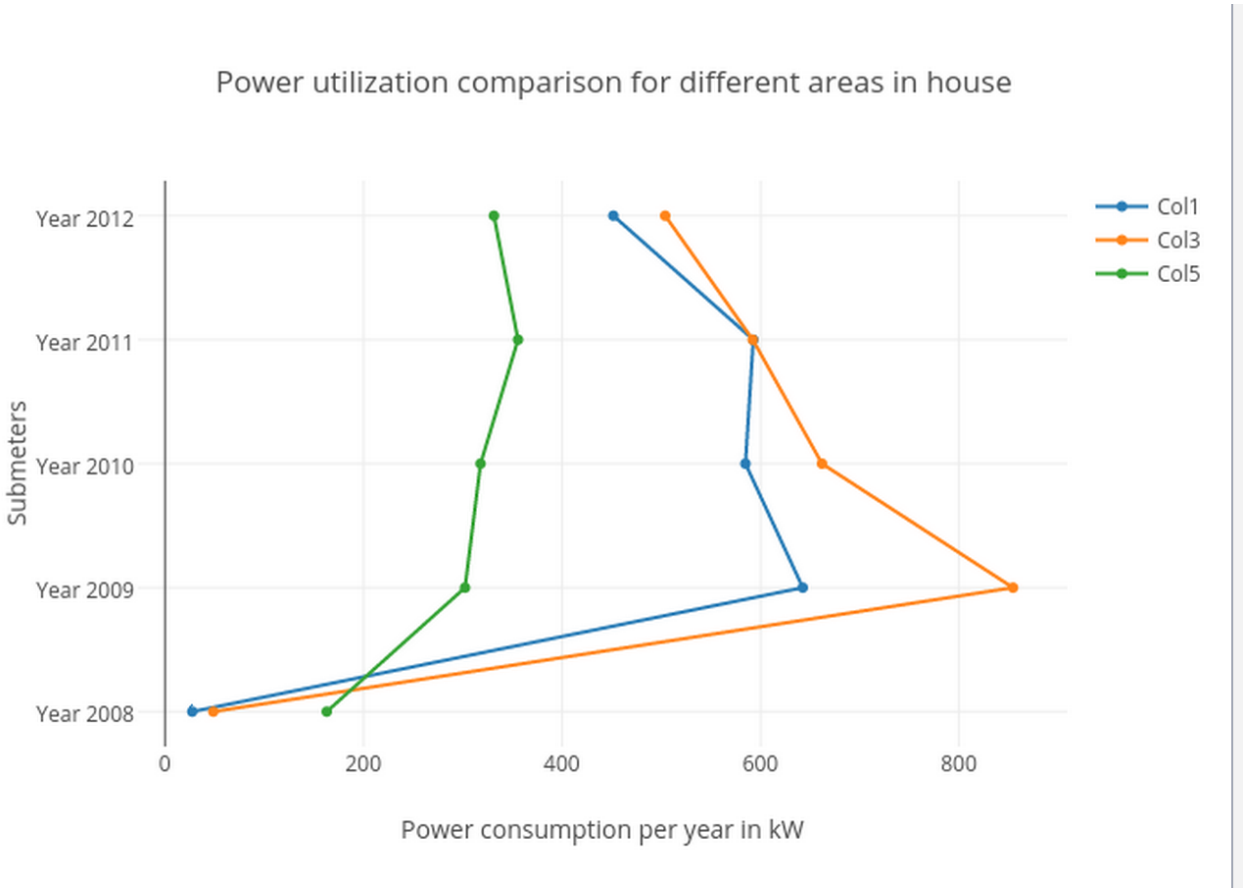


- Weekend and Weekday Power Comparison Peak Time (7-11)

Data calculated over a period of 1 month



- Sub metering of electrical devices in 3 categories and Comparison over the period of 4 years.



- Revenue Loss for a company for 1 house for 1 day

☐ Revenue Loss Prediction Data Actions

Number of Days	Revenue Loss for Company per house(cents/day)
1	86.77212988826814

References:

http://spark.apache.org/docs/latest/index.html

http://zinniasystems.com/blog/2013/12/12/predicting-global-energy-demand-using-spark-part-1/

https://www.kaggle.com/c/belkin-energy-disaggregation-competition

https://www.knime.org/files/knime_bigdata_energy_timeseries_whitepaper.pdf

http://www.greenbiz.com/article/big-data-energy-management-Siemens-MGM-Intel

http://spotfire.tibco.com/blog/?p=17868