**CSEE5590/490: Big Data Programming**

**Project Proposal (Increment 2)**

**Due Date: Friday, Mar 26, 2021**

**Project Title:** Analysis for determination of a relationship between energy demand and weather.

**Team Members:** Joe Goldsich, Anna Johnson, Kyle Son, and Bill Yerkes

**Introduction:** Information overload can make it difficult to understand issues and solve problems. The appearance of excessive quantity of information prevents the understanding of the problem and the ability to construct a solution to the problem. Big Data technologies takes this paradigm and turns it inside out. The students working on this project are going to attempt to construct a model which will be able to take weather and energy related data to be able to forecast energy demand.

**Goals and Objectives:** Utilize the tools and technologies learned from CSEE 5590 to be able to analyze collected data so that it will be possible to determine if there is a relationship between weather and energy consumption and if a relationship exists determine the possibilities of using that relationship to predict future energy needs.

**Motivation:** The global population continues to increase, and the weather patterns seem to be getting more extreme, from extending periods of both above and below normal temperatures in various parts of the world and in the United States. The demand and consumption of energy increase with the population and with the extreme weather, the need for air conditioning in the summer and for heating in the winter. The recent crisis in Texas has demonstrated what can happen if the energy providers are not able to meet the demands of the consumers. Being able to forecast accurately future demand and plan accordingly can help prevent or mitigate such crises in the future.

**Significance:** Better planning of resources for Utility Companies can result in reduced cost to the consumers and more reliable service. This also dips into the area of public safety, as loss of power during extreme weather with no warning can be dangerous for vulnerable groups.

**Objectives:**

Visualize the load and marginal supply curves.

What weather measurements and cities influence most the electrical demand, prices, generation capacity?

Can we forecast 24 hours in advance better than the TSO?

Can we predict electrical prices by time of day better than TSO?

Forecast intraday price or electrical demand hour-by-hour.

Potentially identify the abilities of different energy sources/ systems of regulation to keep up with fluctuating demand.

What is the next generation source to be activated on the load curve?

**Features:**

Hive (Map Reduce):

Use Hive and Map Reduce to store and gather metrics on the data to be able to feed the Spark, GraphX and MLLib portions of the project.

Sqoop (MySQL):

Sqoop and MySQL can be used to query our assembled database to identify patterns between extreme weather and fluctuations in energy consumption. MySQL can also be used to identify notable power outages in the past, or times when utility companies were able to keep up with rising demands.

Solr & Lucene (Search Engines):

We have not covered these technologies at this point in time. We need to determine how search engines can be applied to the project.

Cassandra (No SQL):

We will determine if we need to store data in a NoSQL repository, versus storing the data in Hive and MySQL, to be able to perform the required work to make a predictive model. (We have not covered Cassandra as of yet, do not have enough information as to how it can benefit in solving the problem.)

Spark / GraphX / MLLib (Programming and Analytics):

Use Spark, GraphX, and MLLib to analyze the weather and electrical data to be able to create a model which can predict the demand/cost of electricity based upon supplied weather information

**Approaches/Methods:**

The team’s initial form of communication was via Email. The discussion over email resulted in the team deciding to communicate via Discord, which has allowed for the team member to communicate on a regular basis. The team also meets periodically via Zoom to review items and discuss plans.

The team uses Google Docs and Google Slides to collaborate on documentation. The team also uses GitHub as the repository for source code and final version of documentation.

The team met to discuss several ideas on what topic to do the project on and came to a consensus to do the project on Energy prediction based on Weather. The team reviewed the information on the Kaggle site. The team decided to leverage the knowledge we had gained over the first half of the course to investigate both the technologies, Hadoop, MapReduce, Hive, Sqoop, Cassandra, Solr, that were covered and the data, Weather and Energy datasets.

Each team member worked on their own. They utilized the various tools to do analysis on the data for the project. This allowed the team members to get more practice with the various tools and to get familiar with the data. This approach led to discussions of common issues found and exchanges of ideas between team members. The team shared with each other their findings on various tools and data. The team also started discussion on what approach the team should take for the next iteration.

**Story Telling:**

**Life**

1. **Who** are the people or communities in need of help?

Consumers of energy utility companies (gas and electricity) are one portion of people who need help, as the recent crisis in Texas has demonstrated. The other people who need help are the producer of the energy being consumed (Utility Companies) and their suppliers. Being able to accurately forecast demand can help them better plan on how much to produce, how much raw materials to keep in stock, and how to better plan to use their resources, including labor, to meet their client’s demands. In the case where it is not possible for utility companies to keep up with demand, adequate warning could be given to energy consumers so they can plan for loss of power.

2. **What** problem happened to them?

Recently in Texas with the severe cold weather, the demand for electricity far exceeded the capacity the Utility Companies were able to provide. In addition, the cold weather caused some producers of gas to halt production. If the Utility Companies could have foreseen the spike in demand, it may have been possible for them to take actions to mitigate the negative impacts which were caused by the shortages in energy as compared to the demand that was required to keep people warm.

3. **When** did the problem take place?

This recent issue occurred in February 2021. There were also rolling blackouts in Kansas City Missouri in February 2021 because of the cold. There have been rolling blackouts in California because of the heat over numerous years.

4. **Where** means two things:

a. The environment and settings that the people or the community is living in, and

Due to a rapidly changing climate, these problems can affect a wide range of environments. Communities that are not accustomed to extreme cold or hot weather are particularly vulnerable to energy shortfalls when hit with unexpected demand. Notably in the case of Texas, homes there were designed to shed heat, so when they were hit with uncharacteristically cold weather, the infrastructure was especially unequipped to deal with power outages.

b. The place/location where the problem takes place.

The problem of energy shortages can occur anywhere, in any country. The recent cases have been in Kansas City Missouri and in Texas. They have occurred in California. They can occur anywhere.

5. **Why** means the possible causes and/origin of the problem.

The inability of the Utility Companies to accurately predict demand and a lack of excess capacity to handle a reduction in production capacity by outside forces results in a failure to meet the demands of the consumer. IT companies have HA and redundancy built into their server farms to help prevent outages of their services. Utility companies need to have the same HA and redundancy built into their systems, and they need to understand and be able to reasonably predict how much demand there will be from their consumers.

6. **How**: If you would like, you can add a dimension of how. How did it happen? Sometimes, the answer to how can be covered by what, when, and where.

In the case of Texas, lawmakers were warned years in advance of the possibility of this situation happening. They ignored recommendations that they winterize their energy infrastructure, and that contributed greatly to the energy shortages as some systems failed. Analysis of energy consumption vs. weather patterns could provide the necessary information to utility companies and lawmakers in advance of disastrous situations and could help hold them accountable for their lack of preparedness.

**Workflow:**

Gain knowledge and experience with Big Data Programming tools.

Select the problem the team wishes to solve.

Research aspects of the problem, data, tools, why, benefits, issues.

Select the tools the team wishes to use to solve the problem.

[Repeat following steps as necessary]

Refine the data associated with the problem.

Design Model for solving the problem.

Construct Model for solving the problem.

Test/validate Model for solving the problem.

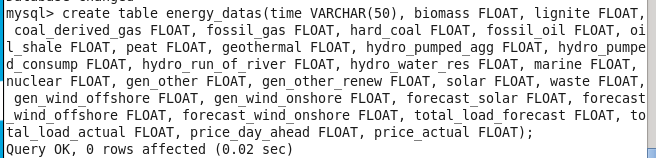
Create a presentation of the solution the team has created.

**Working screens from project:**

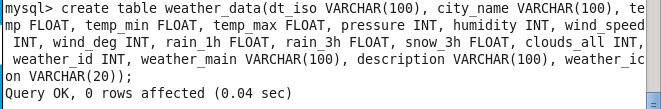
**Anna Johnson:**

Create tables in MySQL:

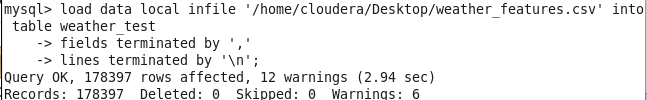
Energy dataset:



Weather dataset:

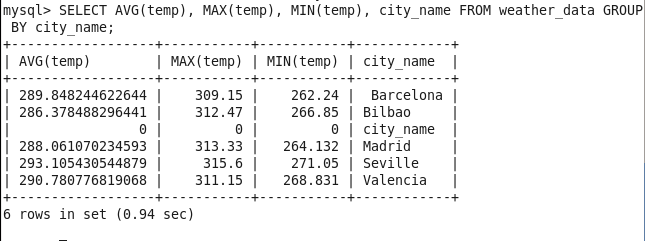


Load Data into MySQL tables from csv files:



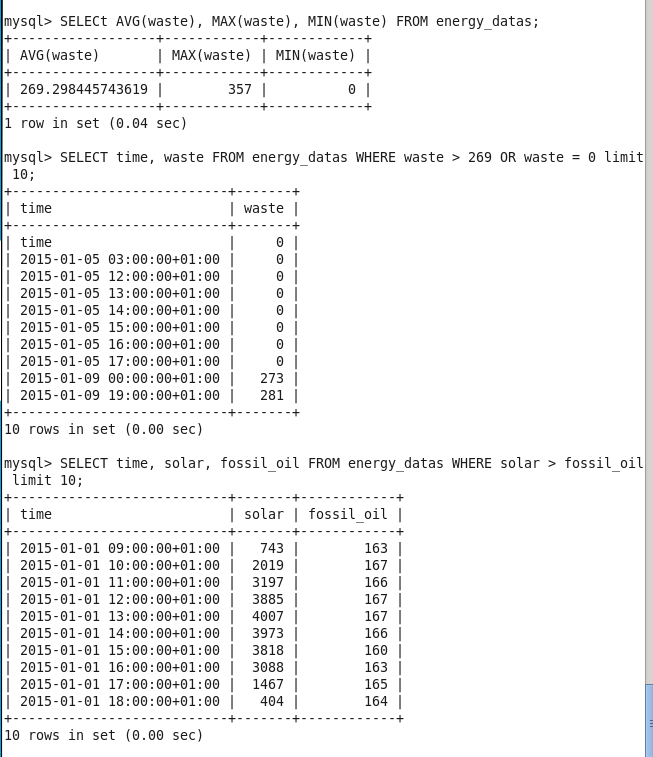
Run Queries on data in MySQL:

Query on weather dataset to get statistics on weather for each city:



Queries on energy dataset:

1. Find the average, min, and max for the ‘waste’ column
2. Find times where there was above average waste, or no waste at all
3. Find times where more energy was generated from solar sources than from fossil oil sources

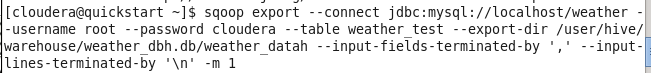


These queries helped us to gain a better insight into our datasets. We can apply this knowledge in the next steps of our project to identify important features to compare between the energy and weather datasets to determine the relationship between them.

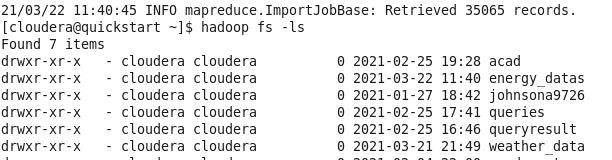
Lastly, it’s important that we are able to move these datasets between Hadoop and MySQL so that we can utilize the different functionality of both the Hive and MySQL databases. This can be done using Sqoop.

Exporting and importing tables between Hadoop, Hive, and MySQL:



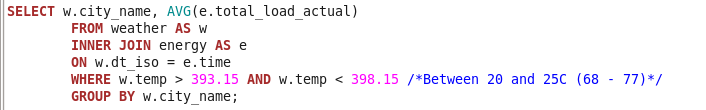


The MySQL tables in HDFS after importing using sqoop:

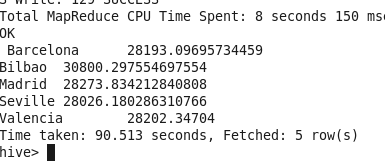


**Joe Goldsich:**

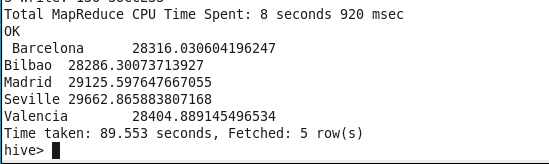
An example of a simple query on a join of the two data sets using HiveQL:



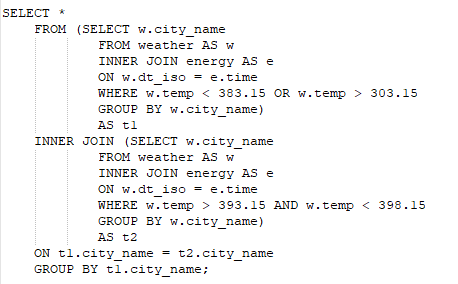
And the results of this simple query:



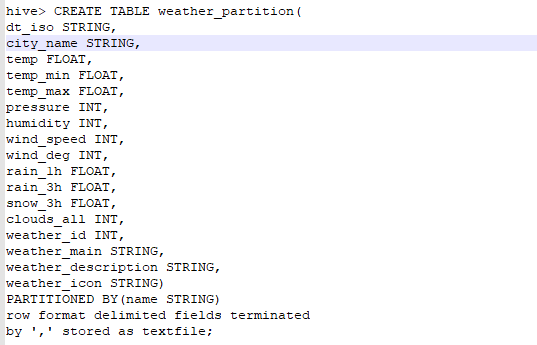
With the exception of Bilbao (which has a very temperate climate) most of the cities experienced an increase in their energy usage at more extreme temperatures (below 10C and over 30C):

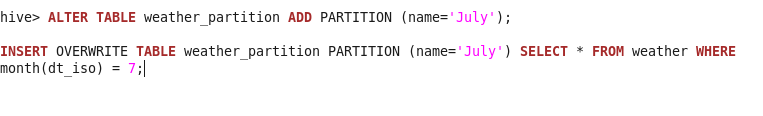


A slightly more complicated Select query was very slow to run (10 minutes slow).

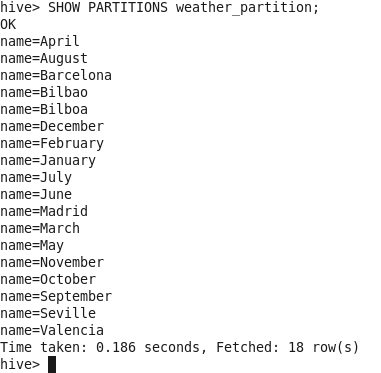


Creating and loading partitions into a new table for the weather dataset to look for performance gains:

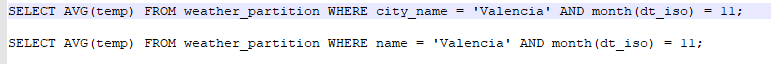




The result of all the partitions:



And some queries using the partitions and not using them. This did result in significantly quicker query times (~58seconds down from ~99 seconds).



**Kyle Son**:

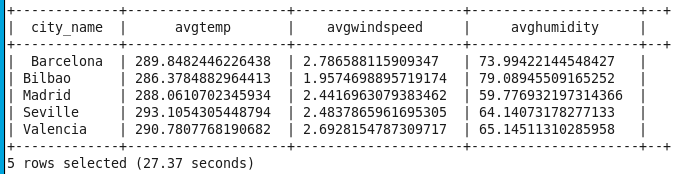
**Hive(Beeline)**

In hive, I used beeline command for the better visualization of the table then, I merged two tables to named merged

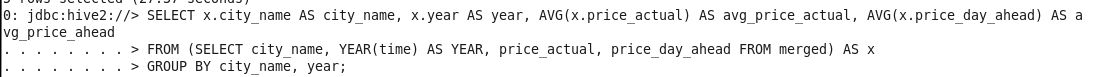
****

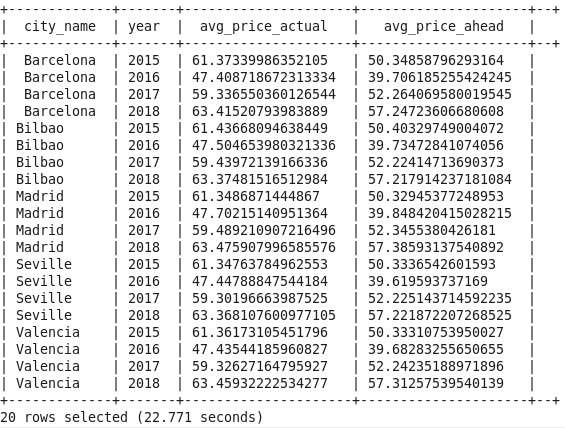
Displays avg temp, windspeed and humidity group by city name

****

****

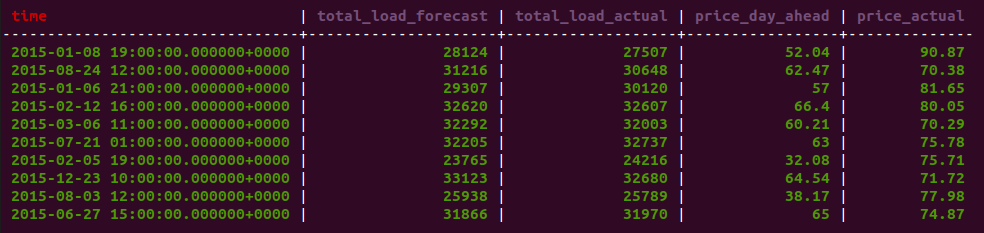
Show the chronological change of the price group by city name

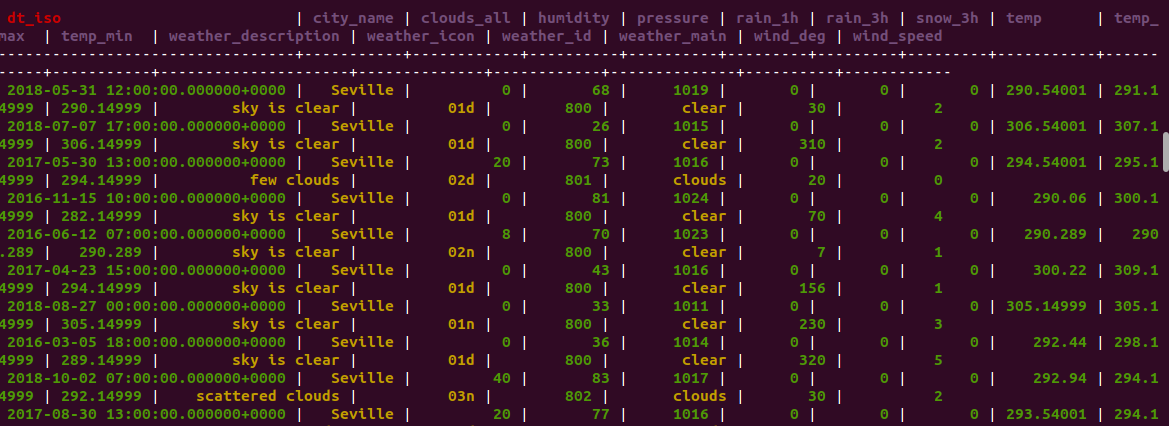
****

****

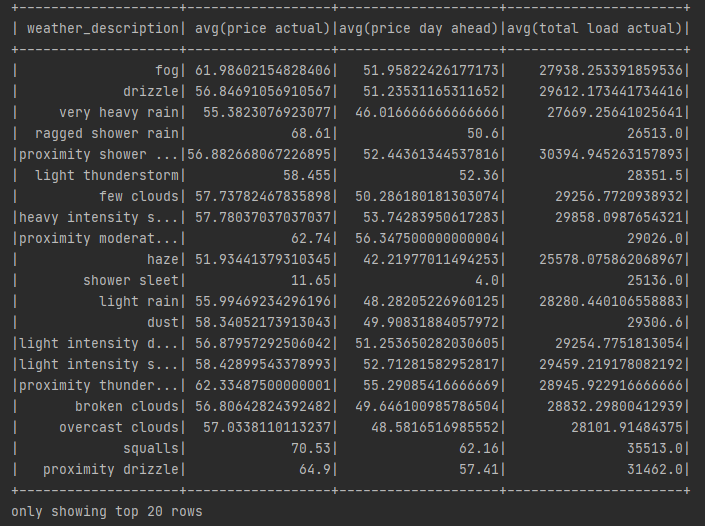
**Cassandra**

Perform some queries in cassandra, There is a limitation for our project because joining the table is not possible in cassandra

****

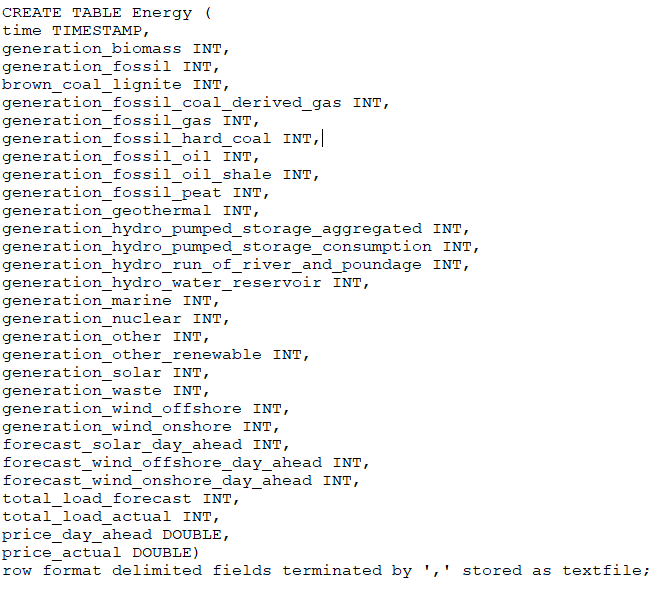
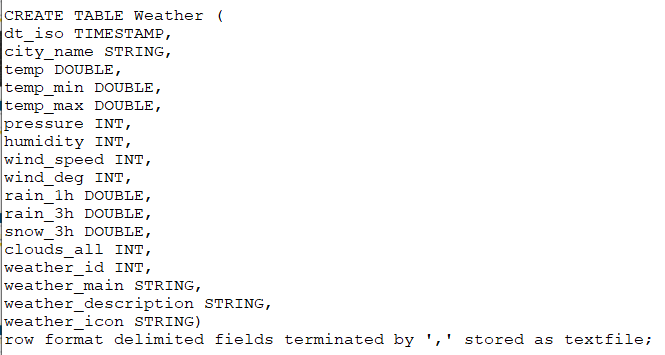
****

**Spark(Scala)**

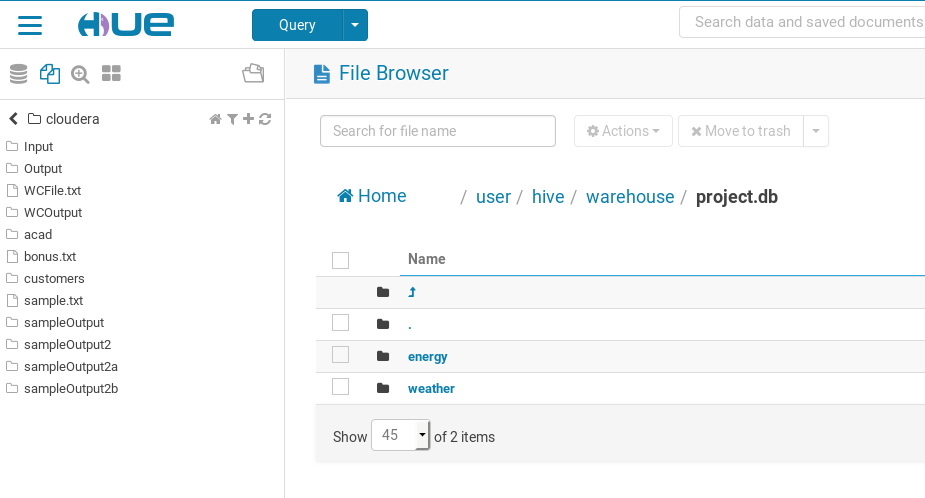
Load the csv file in the spark in the intellij, I merged two table by using spark sql join function and then do some queries based on joined table. **** ****

**Bill Yerkes**:

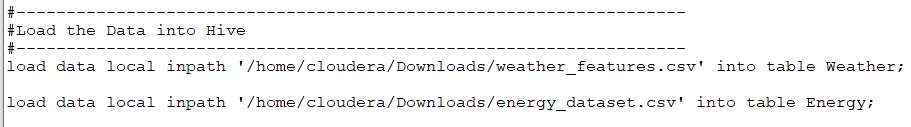
Create Tables in Hive:



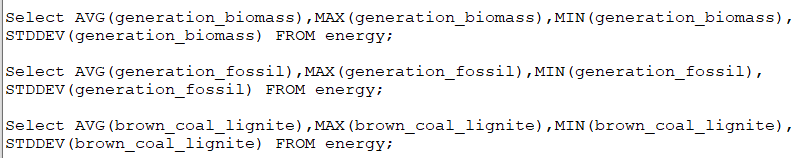
Hive Tables in Hadoop:

****

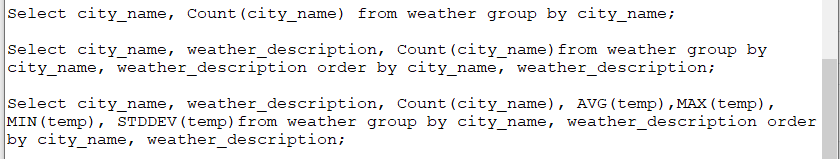
**Load the Data into Hive:**

****

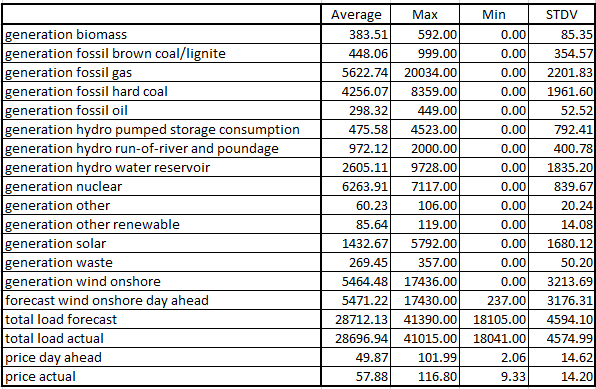
**Some of the queries on Energy Data Set:**

****

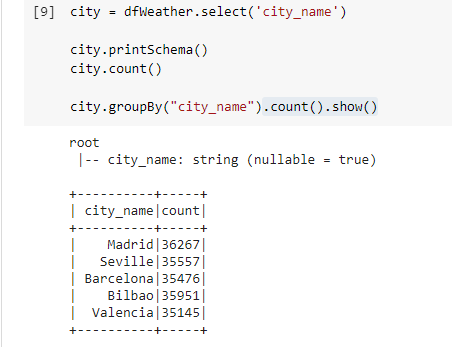
**Some of the queries on Weather Data Set:**

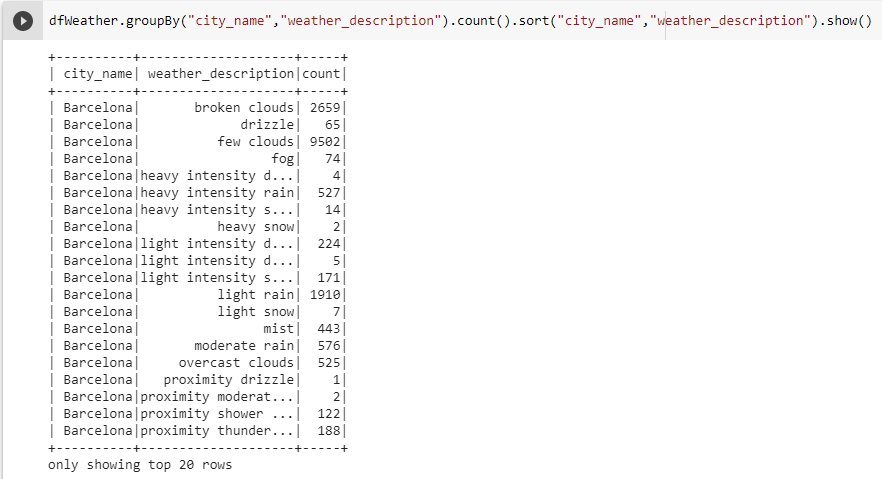
****

**Some of the metrics on the Data**

****

**Query and Results in Pyspark**

****

****

**Data and Parameters:**

[**Hourly energy demand generation and weather | Kaggle**](https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather)

This dataset contains 4 years of electrical consumption, generation, pricing, and weather data for Spain. Consumption and generation data were retrieved from ENTSOE a public portal for Transmission Service Operator (TSO) data. Settlement prices were obtained from the Spanish TSO Red Electric España. Weather data was purchased as part of a personal project from the Open Weather API for the 5 largest cities in Spain and made public here.

**GitHub Link:**

[BillYerkes/CSEE5590\_GroupProject (github.com)](https://github.com/BillYerkes/CSEE5590_GroupProject)

**Project Management:**

**Implementation Status Report**

**Work completed:**

Joe Goldsich:

Analyze Weather Data Set using Hive, Hadoop and MapReduce.

Anna Johnson

Analyze Data in MySQL, using Sqoop and Hadoop

Kyle Son

Analyze Merged data in using Hive, Hadoop and Cassandra

Bill Yerkes

Analyze Energy Data Set using Hive, Hadoop and MapReduce.

**Work to be completed:**

Determine tools to be able to make predictions on the relationship between weather and energy.

Clean the data removing or correcting items so as to be able to make a more accurate prediction with the constructed model.

Create a model for predicting energy demand based upon weather information.

Test and verify the model.

Review the results and generate a presentation.

**Issues or Concerns:**

The date / time fields for the data are stored currently as text / string. These values will need to be converted to data / time. Additionally the data is stored as Greenwich Meridian time with an off set.

The group also needs to get exposure to how to use GraphX / MLLib to be able to make the predictions based upon the data and to create visualizations to present the results.

The group will need to determine how to move forward using Spark. Determination of using Scala or Python as the programming language.

**Conclusion:** The team has familiarized themselves with the tools learned over the first half of the semester. The team has also investigated the two datasets to be able to better understand how to construct our solution. As the team gains greater knowledge about the tools associated with Big Data Programming, we feel confident that we should be able to successfully accomplish our goals by the end of the semester.

**The scientist:**

UMKC Students / CSEE 5590 Big Data Programming. .

Anna Johnson, Joe Goldsich, Jongkook Son, and Bill Yerkes

**Users**

There are two main users for our application. Consumers of energy utility companies and producers of the energy being consumed.

**The Society**

Thanks to our application, the overall total utility/energy cost for our society would decrease because each subject would be able to act appropriately according to the prediction of the application. Producers would be able to expand or decrease their production line based on the weather forecast. Consumers of energy would be able to avoid huge amounts of electricity bills because of a more efficient system.

**Video Presentation:**

[**https://youtu.be/DXZD-4CaSlI**](https://youtu.be/DXZD-4CaSlI)

**References:**

[**https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather**](https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather)

[**Thousands caught off guard by rolling blackouts in Kansas City metro Monday (fox4kc.com)**](https://fox4kc.com/news/thousands-caught-off-guard-by-rolling-blackouts-in-kansas-city-monday-afternoon/)

**https://www.texastribune.org/2021/02/17/texas-power-grid-failures/**

[**https://www.forbes.com/sites/arielcohen/2021/02/19/texas-energy-crisis-is-an-epic-resilience-and-leadership-failure/?sh=46d08806eee8**](https://www.forbes.com/sites/arielcohen/2021/02/19/texas-energy-crisis-is-an-epic-resilience-and-leadership-failure/?sh=46d08806eee8)

[**California rolling blackouts during summer heat wave caused by 3 main factors, report says | Fox Business**](https://www.foxbusiness.com/energy/california-rolling-blackout-power-outage-summer-heat-wave-water-climate-change)