CSEE5590/490: Big Data Programming

Increment 2

Project Title: Energy Demand Analysis in Spain

Team Members: Claire Ndofor, Wes McNall, Scott Howard, Shelby Mohar

Introduction:

Forecasting in energy markets is one exceedingly helpful tool in making the transition to a renewable-based electrical infrastructure (Rolnick et al, 2019). By improving forecasting, we can also increase the efficiency of a power grid and help reduce the usage of peak demand on power plants, which are generally less efficient than their counterparts. While the short-term results have the potential to improve 24-hour and hour-by-hour predictions, this work also has the potential predict energy prices for consumers.

Background:

The data is collected from the five largest cities in Spain: Madrid, Barcelona, Valencia, Seville, and Bilbao between the years of 2015 and 2019. This data has the ability to impact every community that uses an electrical grid. Not only is it advantageous at the individual level to be able to predict the cost of an electric bill, but it is also extremely helpful to be able to predict energy usage at a macro level as communities across the globe begin to make the transition to renewable energies in response to climate change. As stated in the introduction, forecasting in energy markets is an exceedingly helpful tool in making the transition to a renewable-based electrical infrastructure (Rolnick et al, 2019).

Goals and Objectives:

- Motivation:
    - Forecasting in energy markets is one exceedingly helpful tool in making the transition to a renewable-based electrical infrastructure, as stated in “Tackling Climate Change with Machine Learning” (see resources for link to paper). Our goal is to demonstrate this by leveraging Big Data analysis tools on a dataset that consists of energy usage and weather data for five large cities in Spain.
- Significance:
    - Predict energy usage to increase efficiency of electrical production
    - Predict energy price
    - Locate areas that would benefit from renewable energies
- Objectives:
    - Predict energy usage based on the weather
    - Predict energy prices by:
        - Time of day
        - Day of the week
        - Time of year
    - Analyze the factors that affect the fluctuations in energy usage, as well as the sources of energy
- Features:

- dt_iso (datetime index localized to CET)
- generation biomass (in MW)
- generation fossil brown coal/lignite (in MW)
- generation fossil coal-derived gas (in MW)
- generation fossil gas (in MW)
- generation fossil hard coal (in MW)
- generation fossil oil (in MW)
- generation fossil oil shale (in MW)
- generation fossil peat (in MW)
- generation geothermal (in MW)
- city_name
- temp (in kelvin)
- temp_min (in kelvin)
- temp_max (in kelvin)
- pressure (in hPa)
- humidity (in %)
- wind_speed (in m/s)
- wind_deg (wind direction)
- rain_1h (rain in last hour in mm)

Dataset

Our dataset is comprised of two .csv files:

- weather_features.csv – contains information about the weather
- energy_dataset.csv – contains information about the production, price, and variation of energy resources

The two files can be joined by a timestamp. The dataset can be found on Kaggle with the heading “Hourly energy demand generation and weather”. See resources for link.

Features Developed:

This section is dedicated to the features developed in this increment, and a guide to the files within the team repo.

HiveQL: (Wes)

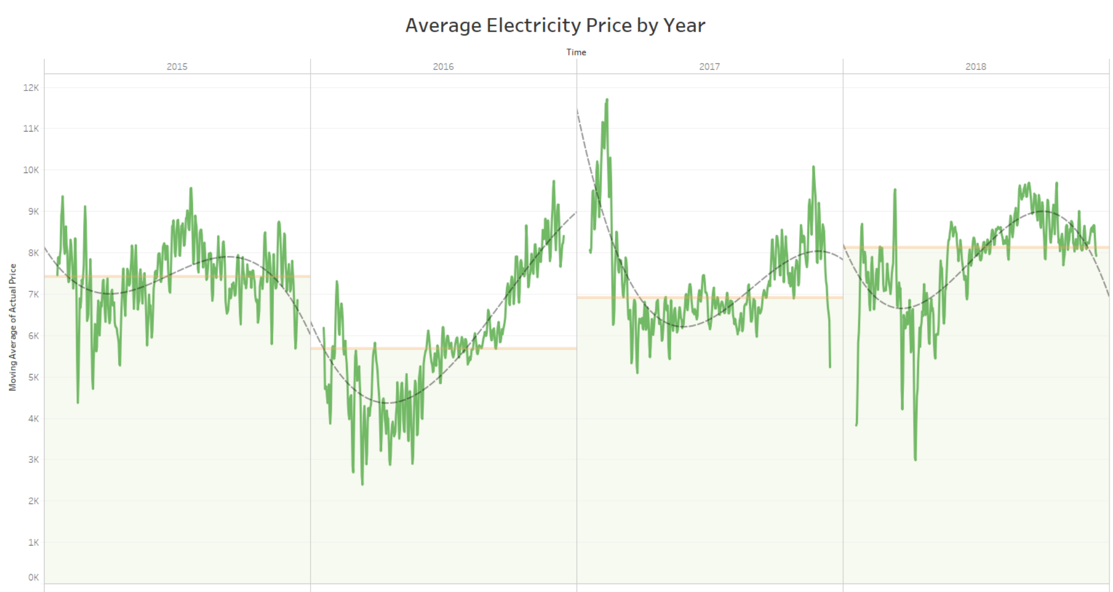
During a past class our professor mentioned using Graphs in this project and as soon as I heard that I knew I wanted to use Tableau to visualize some key aspects of the data, not only to learn more about it but to show key findings.

With the data already loaded from the previous Increment I took to asking some questions about the data and then visualizing the data to see what was interesting about it. Because we are dealing with trends of prices over time, that was a key aspect that I wanted to be able to visualize. Using Tableau I could add some extra visuals without having to calculate, such as trend lines for each particular year. Looking at the Average Electricity Price per Year Graph we can see there are clear lows and highs between the years, which will require some further investigation as to why those trends exist

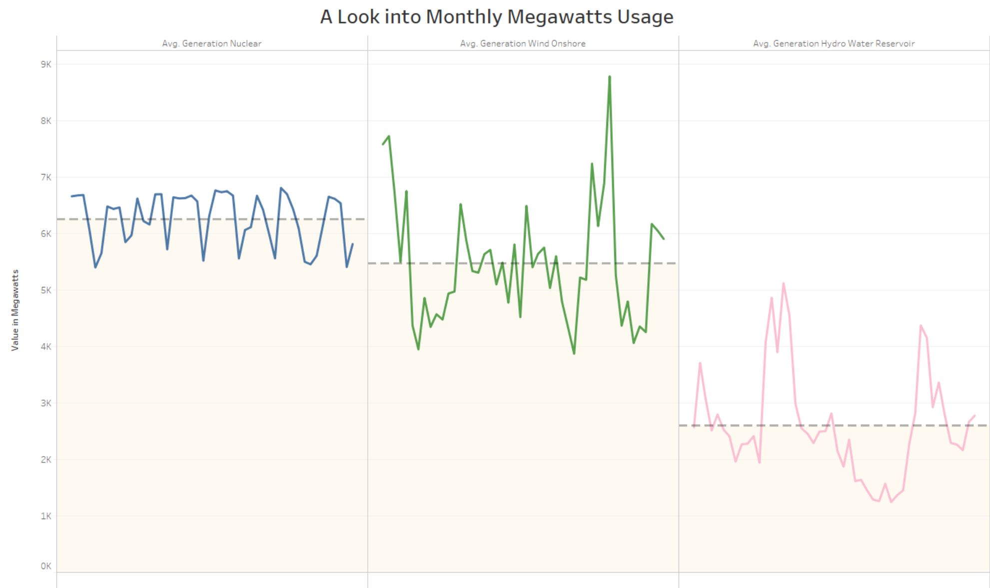
With a wide dataset, part of what I wanted to accomplish this Increment was to determine what columns had interesting and worthwhile data and what columns could be more or less ignored. By writing a large HiveQL query that included summary statistics over time, it would allow us to look for trends and determine which were worth investigating further. All separate Megawatts Usage graphs were trends I found interesting enough to highlight, and all others within the dataset I left ignored from the graphs and queries

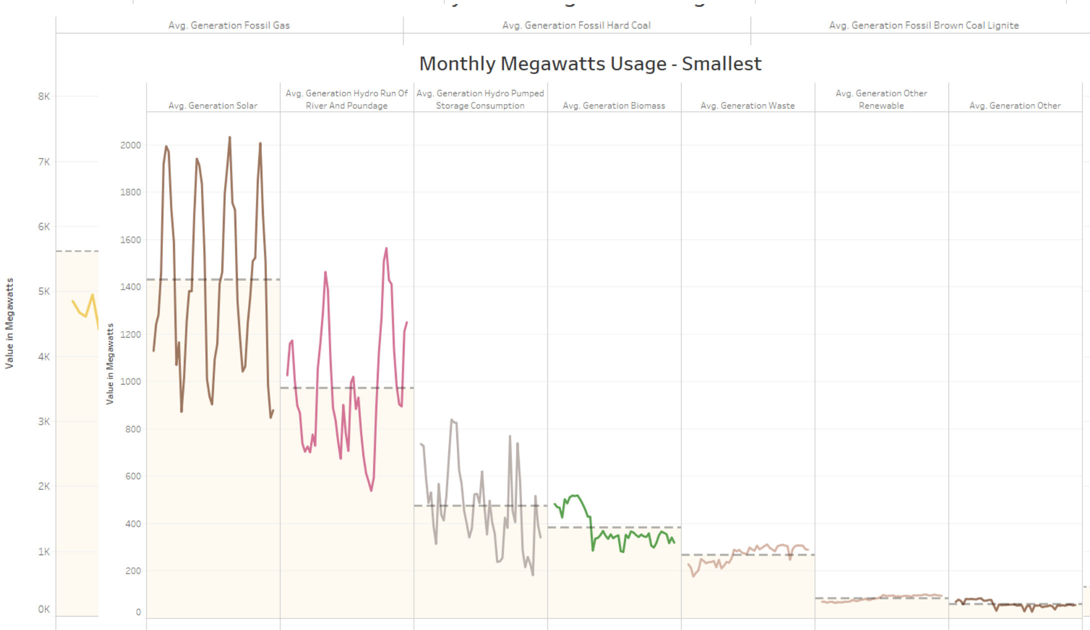
Something I wanted to see was not only trends over months and years, but just over the course of a day. Specifically the average prices over different times of day. The Prices by Hour of Day Graph shows that there is a fluctuation of the cost throughout the course of the day. This is to be expected and the highs and lows also match times that make sense for what time most people will be working and most people will be sleeping

With the 5 different cities in the dataset I wanted to explore the quantitative difference between the locations and see if there was any interesting information that varied between them. Well, the answer was that there isn't, but this wasn't an unfortunate discovery it was a happy one! This



means that these prices were being fairly priced between all of the different locations within the region, meaning that the pricing is independent of location which was a good thing to learn.

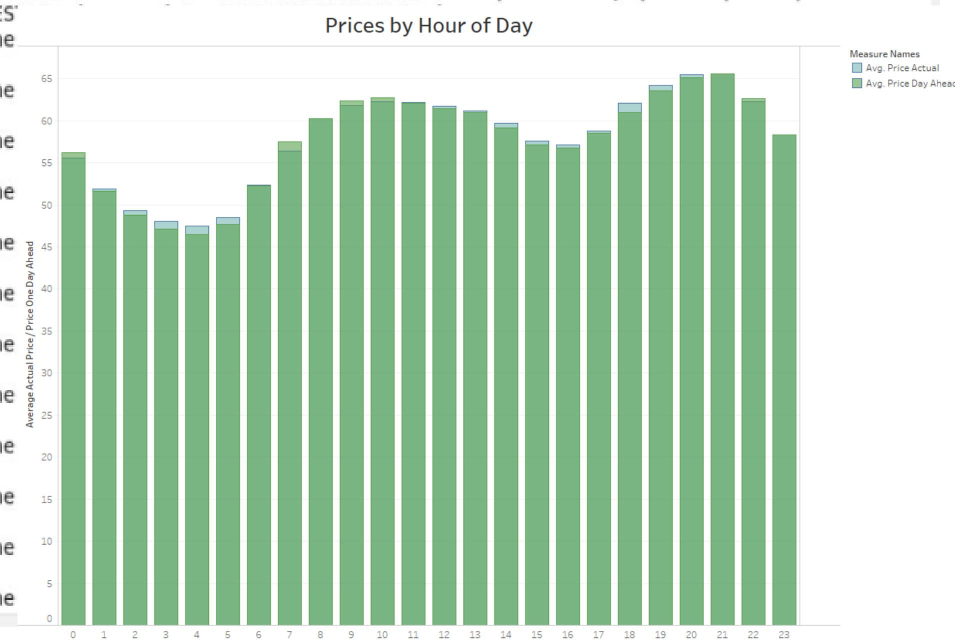




Sqoop: (Shelby)

Within Cloudera, we used Sqoop to transfer the merged dataset from Hive to mySQL.

```
mysql> SHOW TABLES;
+------------------------+
| Tables_in_group_project |
+------------------------+
| energy                 |
+------------------------+
1 row in set (0.00 sec)

mysql> DESCRIBE energy;
+------------------------------------------+-----------+------+-----+--------
----------+--------------------------+
| Field                                    | Type      | Null | Key | Default
          | Extra                    |
+------------------------------------------+-----------+------+-----+--------
----------+--------------------------+
| time                                     | timestamp | NO   |     | CURRENT
_TIMES
| gene
| gene
| gene
| gene
| gene
| gene
| gene
| gene
| gene
| gene
| gene
| gene
```

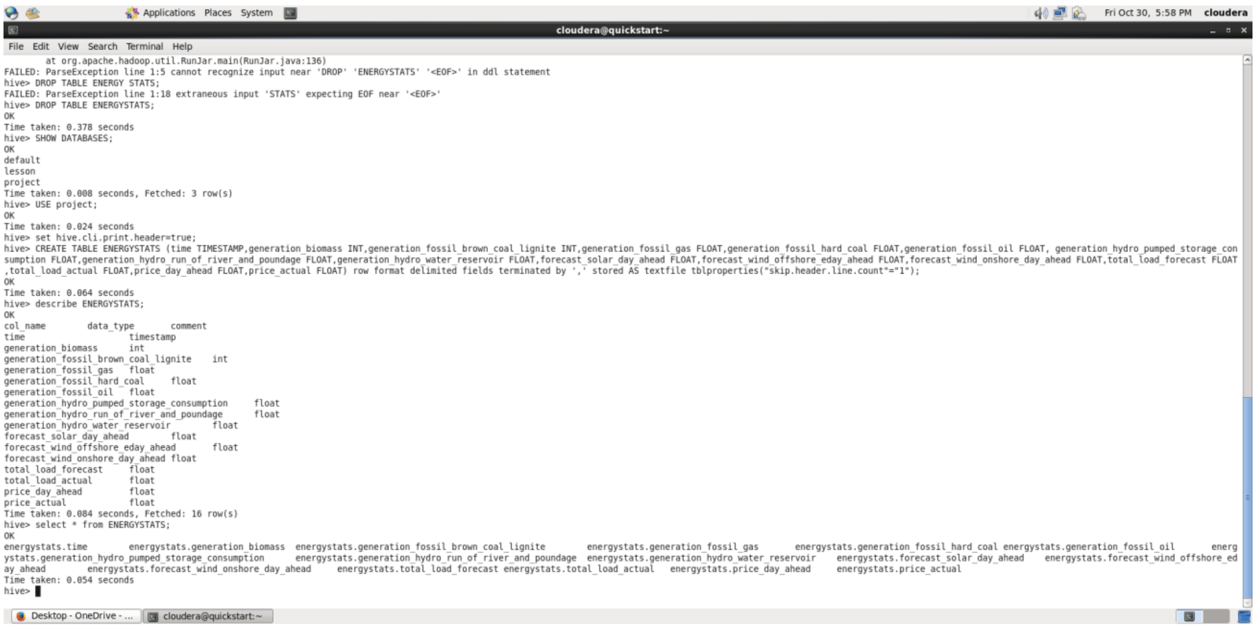


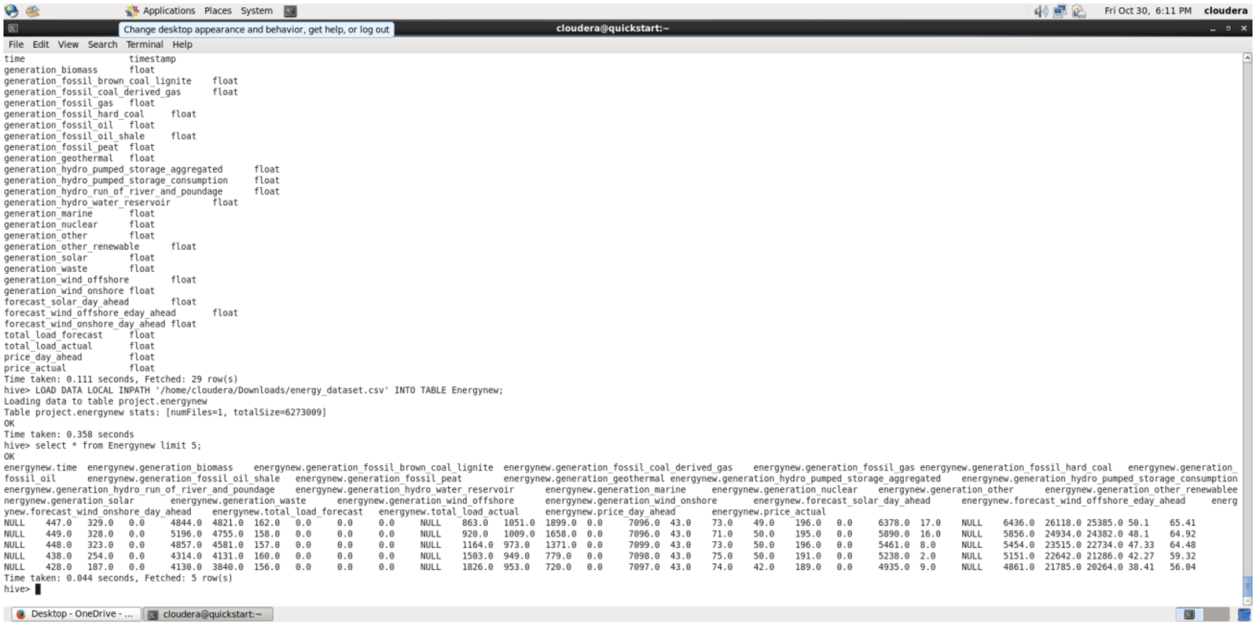
mySQL: (Claire)

So Basically, my aim was to import our dataset into mysql and see what queries I can run to get information from this dataset. While working on this, I did realize most of the columns were hard for me to interpret what their values mean and how they were related to each other. That is one of the aspects I will have to focus on so it's easier for me to decide on what to get out of this data.

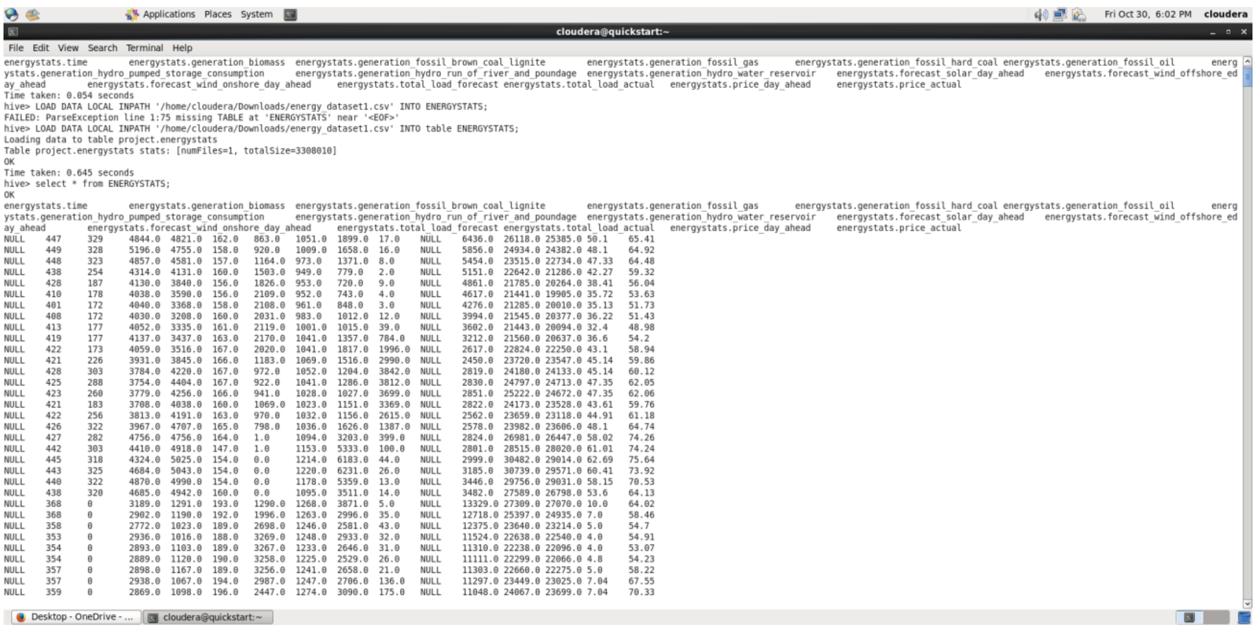
Importing dataset to mysql

First I had to create a table in HIVE , import the dataset from hdfs to hive

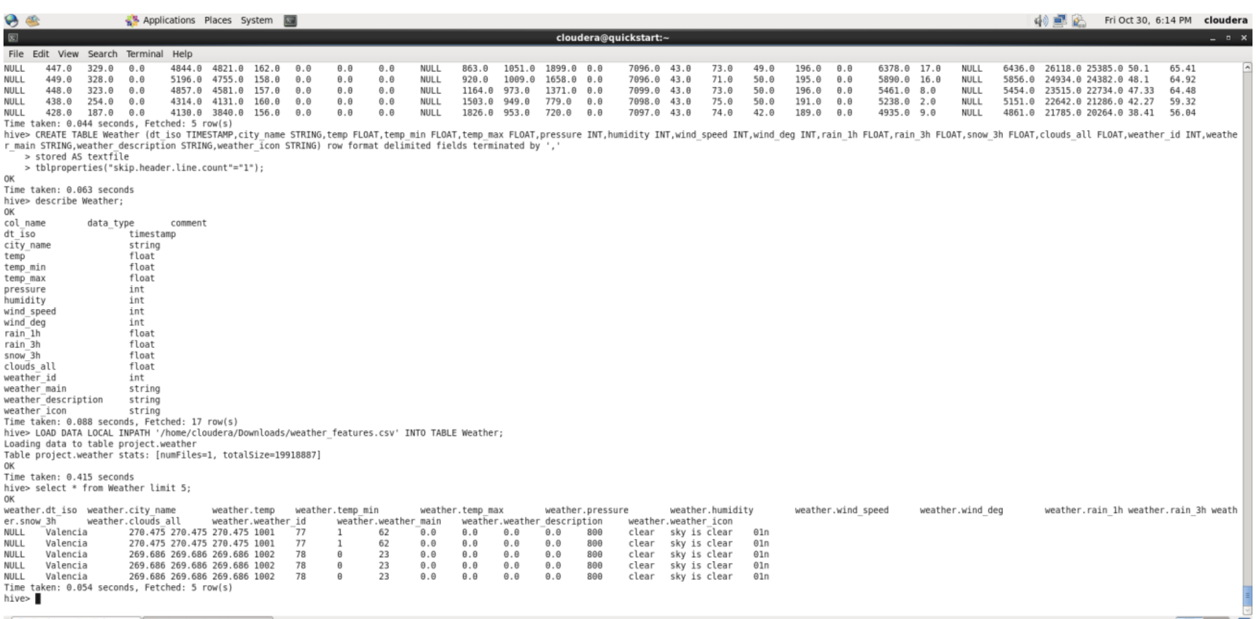




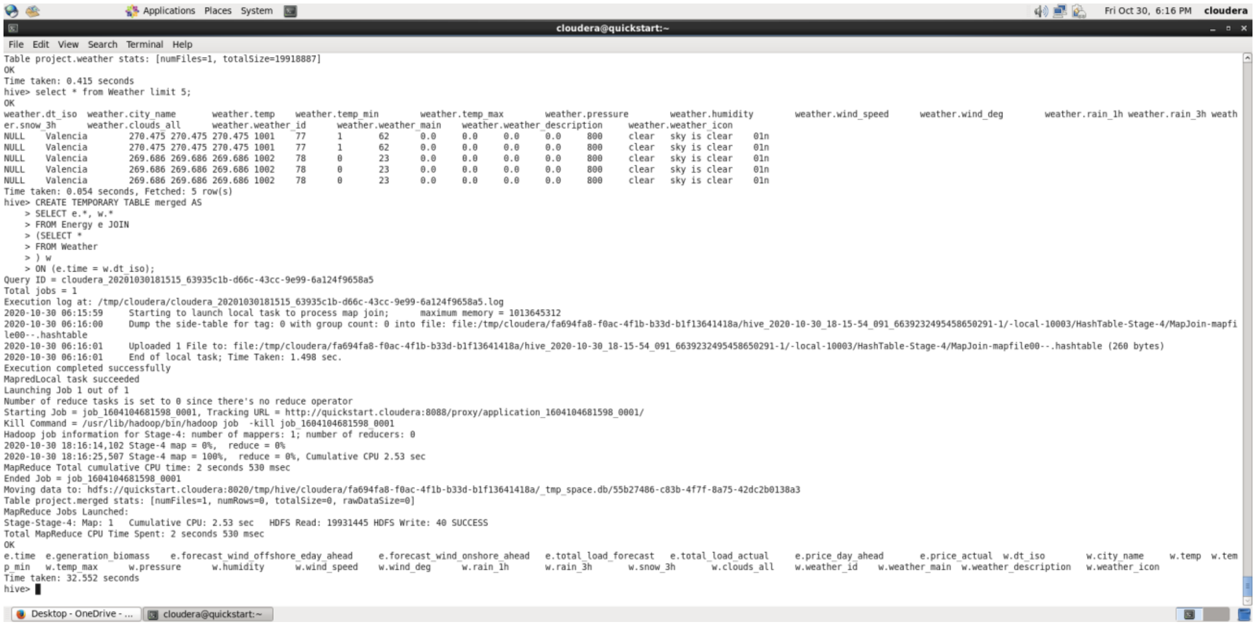

```
time                    timestamp
generation_biomass      float
generation_fossil_brown_coal_lignite    float
generation_fossil_coal_derived_gas      float
generation_fossil_gas   float
generation_fossil_hard_coal     float
generation_fossil_oil   float
generation_fossil_oil_shale     float
generation_fossil_peat  float
generation_geothermal   float
generation_hydro_pumped_storage_aggregated      float
generation_hydro_pumped_storage_consumption     float
generation_hydro_run_of_river_and_poundage      float
generation_hydro_water_reservoir        float
generation_marine       float
generation_nuclear      float
generation_other        float
generation_other_renewable      float
generation_solar        float
generation_waste        float
generation_wind_offshore        float
generation_wind_onshore float
forecast_solar_day_ahead        float
forecast_wind_offshore_eday_ahead       float
forecast_wind_onshore_day_ahead float
total_load_forecast     float
total_load_actual       float
price_day_ahead         float
price_actual            float
Time taken: 0.111 seconds, Fetched: 29 row(s)
hive> LOAD DATA LOCAL INPATH '/home/cloudera/Downloads/energy_dataset.csv' INTO TABLE Energynew;
Loading data to table project.energynew
Table project.energynew stats: [numFiles=1, totalSize=6273009]
OK
Time taken: 0.358 seconds
hive> select * from Energynew limit 5;
OK
energynew.time  energynew.generation_biomass    energynew.generation_fossil_brown_coal_lignite  energynew.generation_fossil_coal_derived_gas    energynew.generation_fossil_gas energynew.generation_fossil_hard_coal   energynew.generation_
fossil_oil      energynew.generation_fossil_oil_shale   energynew.generation_fossil_peat        energynew.generation_geothermal energynew.generation_hydro_pumped_storage_aggregated    energynew.generation_hydro_pumped_storage_consumption
energynew.generation_hydro_run_of_river_and_poundage    energynew.generation_hydro_water_reservoir      energynew.generation_marine     energynew.generation_nuclear    energynew.generation_other      energynew.generation_other_renewablee
nergynew.generation_solar       energynew.generation_waste      energynew.generation_wind_offshore      energynew.generation_wind_onshore       energynew.forecast_solar_day_ahead      energynew.forecast_wind_offshore_eday_ahead     energ
ynew.forecast_wind_onshore_day_ahead    energynew.total_load_forecast   energynew.total_load_actual     energynew.price_day_ahead       energynew.price_actual
NULL    447.0   329.0   0.0     4844.0  4821.0  162.0   0.0     0.0     0.0     NULL    863.0   1051.0  1899.0  0.0     7096.0  43.0    73.0    49.0    196.0   0.0     6378.0  17.0    NULL    6436.0  26118.0 25385.0 50.1    65.41
NULL    449.0   328.0   0.0     5196.0  4755.0  158.0   0.0     0.0     0.0     NULL    920.0   1009.0  1658.0  0.0     7096.0  43.0    71.0    50.0    195.0   0.0     5890.0  16.0    NULL    5856.0  24934.0 24382.0 48.1    64.92
NULL    448.0   323.0   0.0     4857.0  4581.0  157.0   0.0     0.0     0.0     NULL    1164.0  973.0   1371.0  0.0     7099.0  43.0    73.0    50.0    196.0   0.0     5461.0  8.0     NULL    5454.0  23515.0 22734.0 47.33   64.48
NULL    438.0   254.0   0.0     4314.0  4131.0  160.0   0.0     0.0     0.0     NULL    1503.0  949.0   779.0   0.0     7098.0  43.0    75.0    50.0    191.0   0.0     5238.0  2.0     NULL    5151.0  22642.0 21286.0 42.27   59.32
NULL    428.0   187.0   0.0     4130.0  3840.0  156.0   0.0     0.0     0.0     NULL    1826.0  953.0   720.0   0.0     7097.0  43.0    74.0    42.0    189.0   0.0     4935.0  9.0     NULL    4861.0  21785.0 20264.0 38.41   56.04
Time taken: 0.044 seconds, Fetched: 5 row(s)
hive>
```



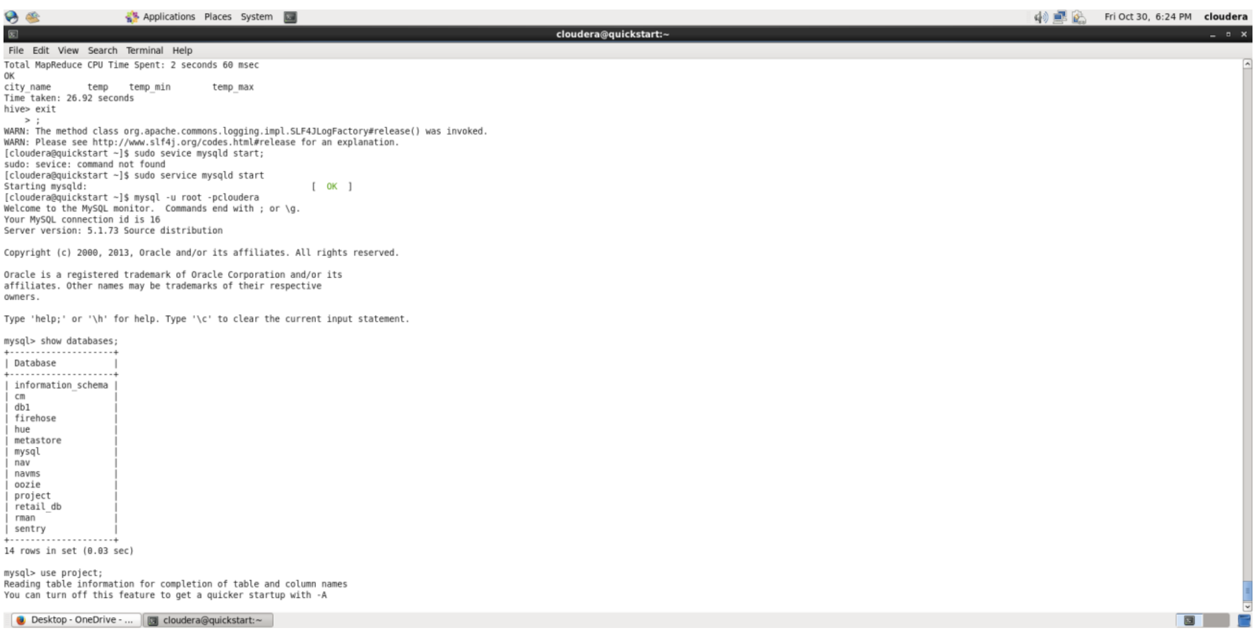

```
energystats.time        energystats.generation_biomass  energystats.generation_fossil_brown_coal_lignite        energystats.generation_fossil_gas       energystats.generation_fossil_hard_coal energystats.generation_fossil_oil       energ
ystats.generation_hydro_pumped_storage_consumption      energystats.generation_hydro_run_of_river_and_poundage  energystats.generation_hydro_water_reservoir    energystats.forecast_solar_day_ahead    energystats.forecast_wind_offshore_ed
ay_ahead        energystats.forecast_wind_onshore_day_ahead     energystats.total_load_forecast energystats.total_load_actual   energystats.price_day_ahead     energystats.price_actual
Time taken: 0.054 seconds
hive> LOAD DATA LOCAL INPATH '/home/cloudera/Downloads/energy_dataset1.csv' INTO ENERGYSTATS;
FAILED: ParseException line 1:75 missing TABLE at 'ENERGYSTATS' near '<EOF>'
hive> LOAD DATA LOCAL INPATH '/home/cloudera/Downloads/energy_dataset1.csv' INTO table ENERGYSTATS;
Loading data to table project.energystats
Table project.energystats stats: [numFiles=1, totalSize=3308010]
OK
Time taken: 0.645 seconds
hive> select * from ENERGYSTATS;
OK
energystats.time        energystats.generation_biomass  energystats.generation_fossil_brown_coal_lignite        energystats.generation_fossil_gas       energystats.generation_fossil_hard_coal energystats.generation_fossil_oil       energ
ystats.generation_hydro_pumped_storage_consumption      energystats.generation_hydro_run_of_river_and_poundage  energystats.generation_hydro_water_reservoir    energystats.forecast_solar_day_ahead    energystats.forecast_wind_offshore_ed
ay_ahead        energystats.forecast_wind_onshore_day_ahead     energystats.total_load_forecast energystats.total_load_actual   energystats.price_day_ahead     energystats.price_actual
NULL    447     329     4844.0  4821.0  162.0   863.0   1051.0  1899.0  17.0    NULL    6436.0  26118.0 25385.0 50.1    65.41
NULL    449     328     5196.0  4755.0  158.0   920.0   1009.0  1658.0  16.0    NULL    5856.0  24934.0 24382.0 48.1    64.92
NULL    448     323     4857.0  4581.0  157.0   1164.0  973.0   1371.0  8.0     NULL    5454.0  23515.0 22734.0 47.33   64.48
NULL    438     254     4314.0  4131.0  160.0   1503.0  949.0   779.0   2.0     NULL    5151.0  22642.0 21286.0 42.27   59.32
NULL    428     187     4130.0  3840.0  156.0   1826.0  953.0   720.0   9.0     NULL    4861.0  21785.0 20264.0 38.41   56.04
NULL    410     178     4038.0  3590.0  156.0   2109.0  952.0   743.0   4.0     NULL    4617.0  21441.0 19905.0 35.72   53.63
NULL    401     172     4040.0  3368.0  158.0   2108.0  961.0   848.0   3.0     NULL    4276.0  21285.0 20010.0 35.13   51.73
NULL    408     172     4030.0  3208.0  160.0   2031.0  983.0   1012.0  12.0    NULL    3994.0  21545.0 20377.0 36.22   51.43
NULL    413     177     4052.0  3335.0  161.0   2119.0  1001.0  1015.0  39.0    NULL    3602.0  21443.0 20094.0 32.4    48.98
NULL    419     177     4137.0  3437.0  163.0   2170.0  1041.0  1357.0  784.0   NULL    3212.0  21560.0 20637.0 36.6    54.2
NULL    422     173     4059.0  3516.0  167.0   2020.0  1041.0  1817.0  1996.0  NULL    2617.0  22824.0 22250.0 43.1    58.94
NULL    421     226     3931.0  3845.0  166.0   1183.0  1069.0  1516.0  2990.0  NULL    2450.0  23720.0 23547.0 45.14   59.86
NULL    428     303     3784.0  4220.0  167.0   972.0   1052.0  1204.0  3842.0  NULL    2819.0  24180.0 24133.0 45.14   60.12
NULL    425     288     3754.0  4404.0  167.0   922.0   1041.0  1286.0  3812.0  NULL    2830.0  24797.0 24713.0 47.35   62.05
NULL    423     260     3779.0  4256.0  166.0   941.0   1028.0  1027.0  3699.0  NULL    2851.0  25222.0 24672.0 47.35   62.06
NULL    421     183     3708.0  4038.0  160.0   1069.0  1023.0  1151.0  3369.0  NULL    2822.0  24173.0 23528.0 43.61   59.76
NULL    422     256     3813.0  4191.0  163.0   970.0   1032.0  1156.0  2615.0  NULL    2562.0  23659.0 23118.0 44.91   61.18
NULL    426     322     3967.0  4707.0  165.0   798.0   1036.0  1626.0  1387.0  NULL    2578.0  23982.0 23606.0 48.1    64.74
NULL    427     282     4756.0  4756.0  164.0   1.0     1094.0  3203.0  399.0   NULL    2824.0  26981.0 26447.0 58.02   74.26
NULL    442     303     4410.0  4918.0  147.0   1.0     1153.0  5333.0  100.0   NULL    2801.0  28515.0 28020.0 61.01   74.24
NULL    445     318     4324.0  5025.0  154.0   0.0     1214.0  6183.0  44.0    NULL    2999.0  30482.0 29014.0 62.69   75.64
NULL    443     325     4684.0  5043.0  154.0   0.0     1220.0  6231.0  26.0    NULL    3185.0  30739.0 29571.0 60.41   73.92
NULL    440     322     4870.0  4990.0  154.0   0.0     1178.0  5359.0  13.0    NULL    3446.0  29756.0 29031.0 58.15   70.53
NULL    438     320     4685.0  4942.0  160.0   0.0     1095.0  3511.0  14.0    NULL    3482.0  27589.0 26798.0 53.6    64.13
NULL    368     0       3189.0  1291.0  193.0   1290.0  1268.0  3871.0  5.0     NULL    13329.0 27309.0 27070.0 10.0    64.02
NULL    368     0       2902.0  1190.0  192.0   1996.0  1263.0  2996.0  35.0    NULL    12718.0 25397.0 24935.0 7.0     58.46
NULL    358     0       2772.0  1023.0  189.0   2698.0  1246.0  2581.0  43.0    NULL    12375.0 23640.0 23214.0 5.0     54.7
NULL    353     0       2936.0  1016.0  188.0   3269.0  1248.0  2933.0  32.0    NULL    11524.0 22638.0 22540.0 4.0     54.91
NULL    354     0       2893.0  1103.0  189.0   3267.0  1233.0  2646.0  31.0    NULL    11310.0 22238.0 22096.0 4.0     53.07
NULL    354     0       2889.0  1120.0  190.0   3258.0  1225.0  2529.0  26.0    NULL    11111.0 22299.0 22066.0 4.8     54.23
NULL    357     0       2898.0  1167.0  189.0   3256.0  1241.0  2658.0  21.0    NULL    11303.0 22660.0 22275.0 5.0     58.22
NULL    357     0       2938.0  1067.0  194.0   2987.0  1247.0  2706.0  136.0   NULL    11297.0 23449.0 23025.0 7.04    67.55
NULL    359     0       2869.0  1098.0  196.0   2447.0  1274.0  3090.0  175.0   NULL    11048.0 24067.0 23699.0 7.04    70.33
```

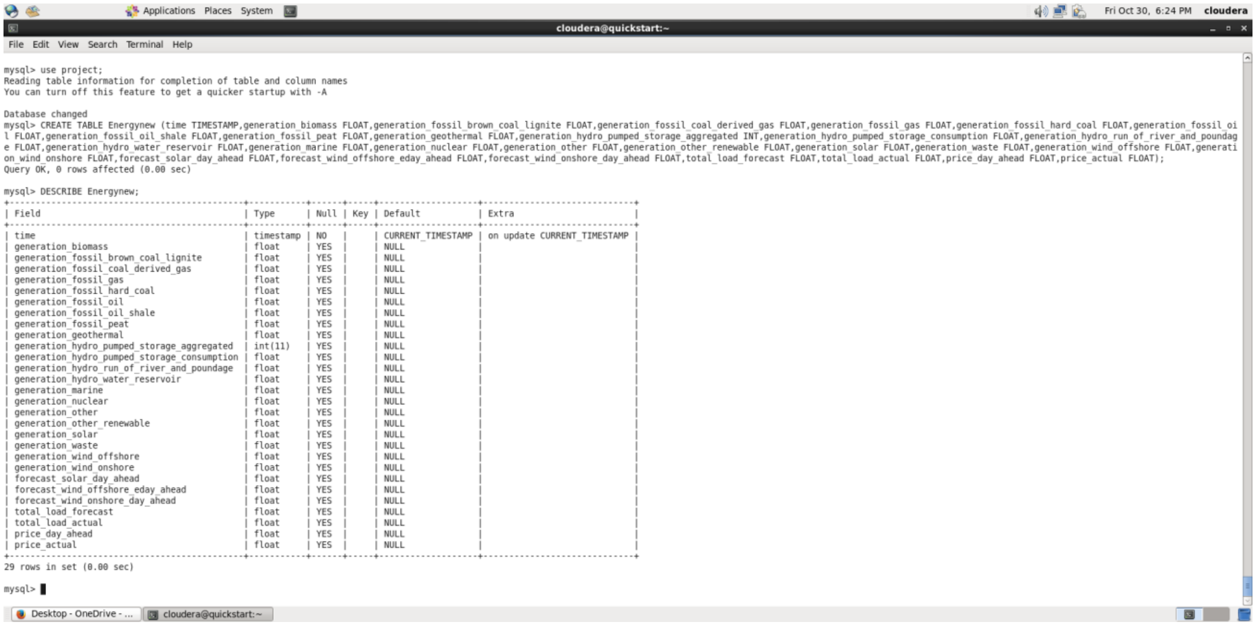



```
NULL    447.0   329.0   0.0     4844.0  4821.0  162.0   0.0     0.0     0.0     NULL    863.0   1051.0  1899.0  0.0     7096.0  43.0    73.0    49.0    196.0   0.0     6378.0  17.0    NULL    6436.0  26118.0 25385.0 50.1    65.41
NULL    449.0   328.0   0.0     5196.0  4755.0  158.0   0.0     0.0     0.0     NULL    920.0   1009.0  1658.0  0.0     7096.0  43.0    71.0    50.0    195.0   0.0     5890.0  16.0    NULL    5856.0  24934.0 24382.0 48.1    64.92
NULL    448.0   323.0   0.0     4857.0  4581.0  157.0   0.0     0.0     0.0     NULL    1164.0  973.0   1371.0  0.0     7099.0  43.0    73.0    50.0    196.0   0.0     5461.0  8.0     NULL    5454.0  23515.0 22734.0 47.33   64.48
NULL    438.0   254.0   0.0     4314.0  4131.0  160.0   0.0     0.0     0.0     NULL    1503.0  949.0   779.0   0.0     7098.0  43.0    75.0    50.0    191.0   0.0     5238.0  2.0     NULL    5151.0  22642.0 21286.0 42.27   59.32
NULL    428.0   187.0   0.0     4130.0  3840.0  156.0   0.0     0.0     0.0     NULL    1826.0  953.0   720.0   0.0     7097.0  43.0    74.0    42.0    189.0   0.0     4935.0  9.0     NULL    4861.0  21785.0 20264.0 38.41   56.04
Time taken: 0.044 seconds, Fetched: 5 row(s)
hive> CREATE TABLE Weather (dt_iso TIMESTAMP,city_name STRING,temp FLOAT,temp_min FLOAT,temp_max FLOAT,pressure INT,humidity INT,wind_speed INT,wind_deg INT,rain_1h FLOAT,rain_3h FLOAT,snow_3h FLOAT,clouds_all FLOAT,weather_id INT,weathe
r_main STRING,weather_description STRING,weather_icon STRING) row format delimited fields terminated by ','
    > stored AS textfile
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.063 seconds
hive> describe Weather;
OK
col_name        data_type       comment
dt_iso                  timestamp
city_name               string
temp                    float
temp_min                float
temp_max                float
pressure                int
humidity                int
wind_speed              int
wind_deg                int
rain_1h                 float
rain_3h                 float
snow_3h                 float
clouds_all              float
weather_id              int
weather_main            string
weather_description     string
weather_icon            string
Time taken: 0.088 seconds, Fetched: 17 row(s)
hive> LOAD DATA LOCAL INPATH '/home/cloudera/Downloads/weather_features.csv' INTO TABLE Weather;
Loading data to table project.weather
Table project.weather stats: [numFiles=1, totalSize=19918887]
OK
Time taken: 0.415 seconds
hive> select * from Weather limit 5;
OK
weather.dt_iso  weather.city_name       weather.temp    weather.temp_min        weather.temp_max        weather.pressure        weather.humidity        weather.wind_speed      weather.wind_deg        weather.rain_1h weather.rain_3h weath
er.snow_3h      weather.clouds_all      weather.weather_id      weather.weather_main    weather.weather_description     weather.weather_icon
NULL    Valencia        270.475 270.475 270.475 1001    77      1       62      0.0     0.0     0.0     0.0     800     clear   sky is clear    01n
NULL    Valencia        270.475 270.475 270.475 1001    77      1       62      0.0     0.0     0.0     0.0     800     clear   sky is clear    01n
NULL    Valencia        269.686 269.686 269.686 1002    78      0       23      0.0     0.0     0.0     0.0     800     clear   sky is clear    01n
NULL    Valencia        269.686 269.686 269.686 1002    78      0       23      0.0     0.0     0.0     0.0     800     clear   sky is clear    01n
NULL    Valencia        269.686 269.686 269.686 1002    78      0       23      0.0     0.0     0.0     0.0     800     clear   sky is clear    01n
Time taken: 0.054 seconds, Fetched: 5 row(s)
hive>
```

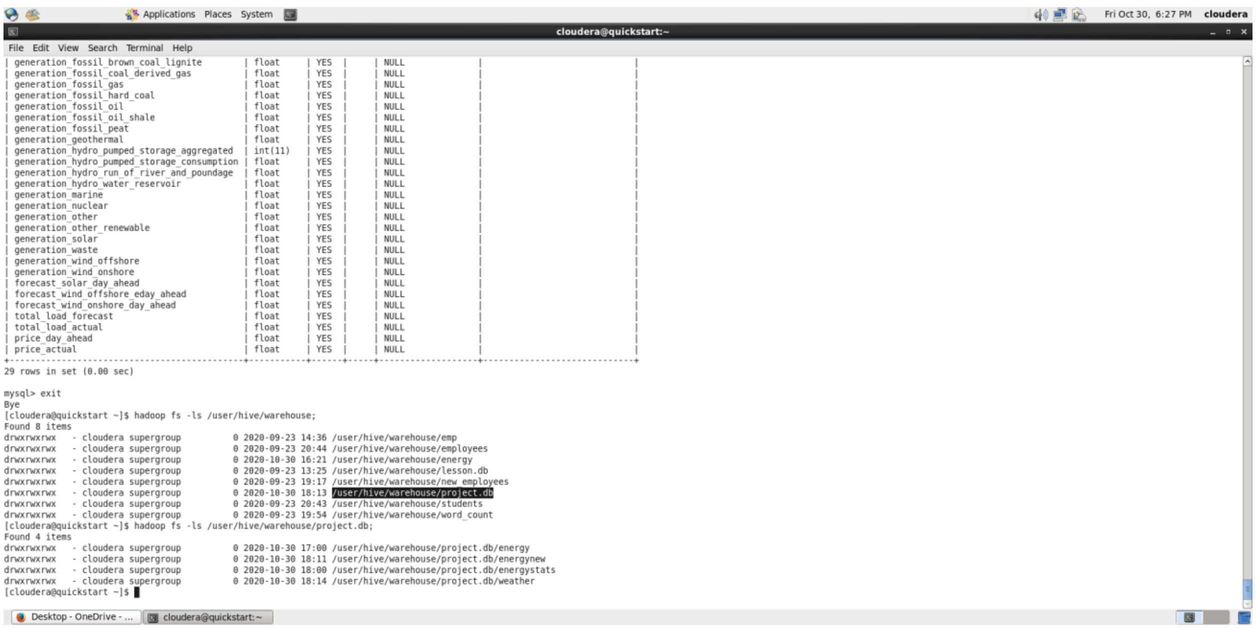


One thing I noticed was some of my columns returned NULL values and I tried fixing this by making sure ROW DELIMITERS was set but this didn't help at all. Secondly I had to create a table with similar columns in mysql so I can export data from hive to mysql via sqoop



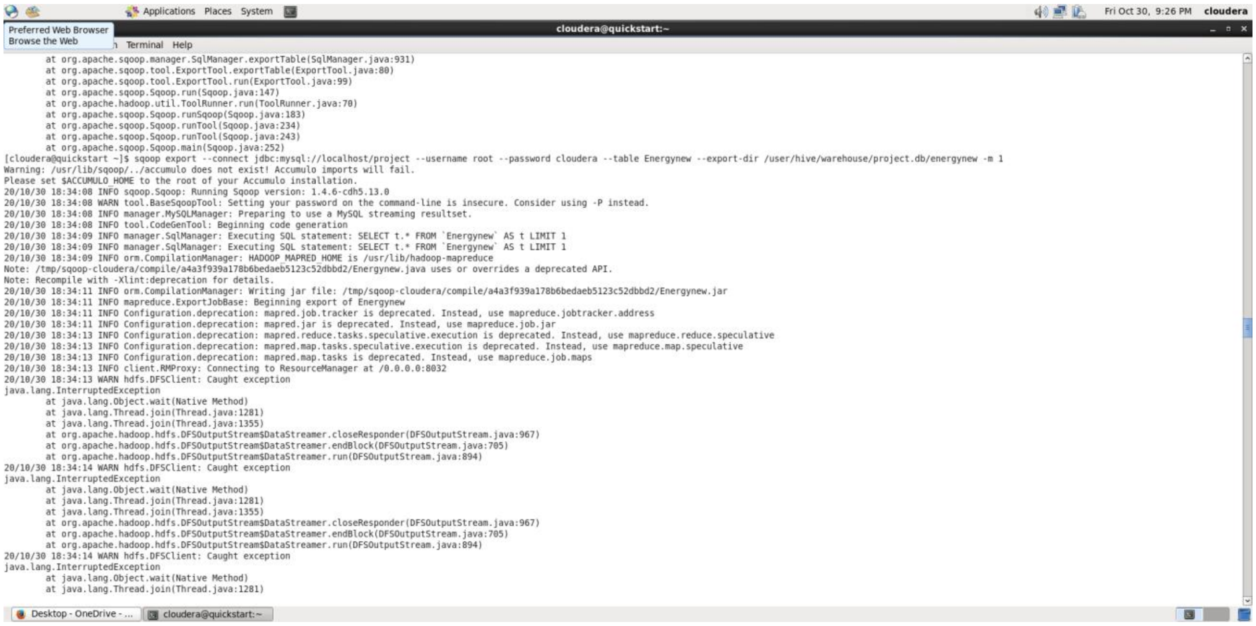


```
mysql> use project;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> CREATE TABLE Energynew (time TIMESTAMP,generation_biomass FLOAT,generation_fossil_brown_coal_lignite FLOAT,generation_fossil_coal_derived_gas FLOAT,generation_fossil_gas FLOAT,generation_fossil_hard_coal FLOAT,generation_fossil_oi
l FLOAT,generation_fossil_oil_shale FLOAT,generation_fossil_peat FLOAT,generation_geothermal FLOAT,generation_hydro_pumped_storage_aggregated INT,generation_hydro_pumped_storage_consumption FLOAT,generation_hydro_run_of_river_and_poundag
e FLOAT,generation_hydro_water_reservoir FLOAT,generation_marine FLOAT,generation_nuclear FLOAT,generation_other FLOAT,generation_other_renewable FLOAT,generation_solar FLOAT,generation_waste FLOAT,generation_wind_offshore FLOAT,generati
on_wind_onshore FLOAT,forecast_solar_day_ahead FLOAT,forecast_wind_offshore_eday_ahead FLOAT,forecast_wind_onshore_day_ahead FLOAT,total_load_forecast FLOAT,total_load_actual FLOAT,price_day_ahead FLOAT,price_actual FLOAT);
Query OK, 0 rows affected (0.00 sec)

mysql> DESCRIBE Energynew;
+---------------------------------------------+-----------+------+-----+-------------------+-----------------------------+
| Field                                       | Type      | Null | Key | Default           | Extra                       |
+---------------------------------------------+-----------+------+-----+-------------------+-----------------------------+
| time                                        | timestamp | NO   |     | CURRENT_TIMESTAMP | on update CURRENT_TIMESTAMP |
| generation_biomass                          | float     | YES  |     | NULL              |                             |
| generation_fossil_brown_coal_lignite        | float     | YES  |     | NULL              |                             |
| generation_fossil_coal_derived_gas          | float     | YES  |     | NULL              |                             |
| generation_fossil_gas                       | float     | YES  |     | NULL              |                             |
| generation_fossil_hard_coal                 | float     | YES  |     | NULL              |                             |
| generation_fossil_oil                       | float     | YES  |     | NULL              |                             |
| generation_fossil_oil_shale                 | float     | YES  |     | NULL              |                             |
| generation_fossil_peat                      | float     | YES  |     | NULL              |                             |
| generation_geothermal                       | float     | YES  |     | NULL              |                             |
| generation_hydro_pumped_storage_aggregated  | int(11)   | YES  |     | NULL              |                             |
| generation_hydro_pumped_storage_consumption | float     | YES  |     | NULL              |                             |
| generation_hydro_run_of_river_and_poundage  | float     | YES  |     | NULL              |                             |
| generation_hydro_water_reservoir            | float     | YES  |     | NULL              |                             |
| generation_marine                           | float     | YES  |     | NULL              |                             |
| generation_nuclear                          | float     | YES  |     | NULL              |                             |
| generation_other                            | float     | YES  |     | NULL              |                             |
| generation_other_renewable                  | float     | YES  |     | NULL              |                             |
| generation_solar                            | float     | YES  |     | NULL              |                             |
| generation_waste                            | float     | YES  |     | NULL              |                             |
| generation_wind_offshore                    | float     | YES  |     | NULL              |                             |
| generation_wind_onshore                     | float     | YES  |     | NULL              |                             |
| forecast_solar_day_ahead                    | float     | YES  |     | NULL              |                             |
| forecast_wind_offshore_eday_ahead           | float     | YES  |     | NULL              |                             |
| forecast_wind_onshore_day_ahead             | float     | YES  |     | NULL              |                             |
| total_load_forecast                         | float     | YES  |     | NULL              |                             |
| total_load_actual                           | float     | YES  |     | NULL              |                             |
| price_day_ahead                             | float     | YES  |     | NULL              |                             |
| price_actual                                | float     | YES  |     | NULL              |                             |
+---------------------------------------------+-----------+------+-----+-------------------+-----------------------------+
29 rows in set (0.00 sec)

mysql>
```

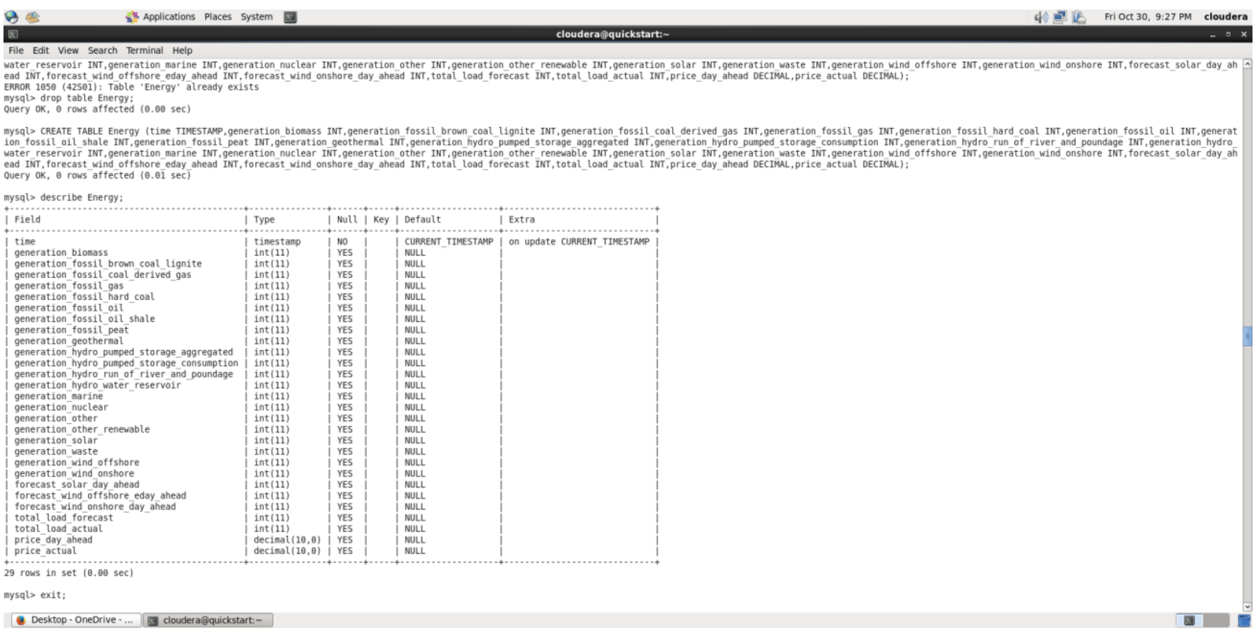


```
| generation_fossil_brown_coal_lignite        | float     | YES  |     | NULL              |                             |
| generation_fossil_coal_derived_gas          | float     | YES  |     | NULL              |                             |
| generation_fossil_gas                       | float     | YES  |     | NULL              |                             |
| generation_fossil_hard_coal                 | float     | YES  |     | NULL              |                             |
| generation_fossil_oil                       | float     | YES  |     | NULL              |                             |
| generation_fossil_oil_shale                 | float     | YES  |     | NULL              |                             |
| generation_fossil_peat                      | float     | YES  |     | NULL              |                             |
| generation_geothermal                       | float     | YES  |     | NULL              |                             |
| generation_hydro_pumped_storage_aggregated  | int(11)   | YES  |     | NULL              |                             |
| generation_hydro_pumped_storage_consumption | float     | YES  |     | NULL              |                             |
| generation_hydro_run_of_river_and_poundage  | float     | YES  |     | NULL              |                             |
| generation_hydro_water_reservoir            | float     | YES  |     | NULL              |                             |
| generation_marine                           | float     | YES  |     | NULL              |                             |
| generation_nuclear                          | float     | YES  |     | NULL              |                             |
| generation_other                            | float     | YES  |     | NULL              |                             |
| generation_other_renewable                  | float     | YES  |     | NULL              |                             |
| generation_solar                            | float     | YES  |     | NULL              |                             |
| generation_waste                            | float     | YES  |     | NULL              |                             |
| generation_wind_offshore                    | float     | YES  |     | NULL              |                             |
| generation_wind_onshore                     | float     | YES  |     | NULL              |                             |
| forecast_solar_day_ahead                    | float     | YES  |     | NULL              |                             |
| forecast_wind_offshore_eday_ahead           | float     | YES  |     | NULL              |                             |
| forecast_wind_onshore_day_ahead             | float     | YES  |     | NULL              |                             |
| total_load_forecast                         | float     | YES  |     | NULL              |                             |
| total_load_actual                           | float     | YES  |     | NULL              |                             |
| price_day_ahead                             | float     | YES  |     | NULL              |                             |
| price_actual                                | float     | YES  |     | NULL              |                             |
+---------------------------------------------+-----------+------+-----+-------------------+-----------------------------+
29 rows in set (0.00 sec)

mysql> exit
Bye
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse;
Found 8 items
drwxrwxrwx   - cloudera supergroup          0 2020-09-23 14:36 /user/hive/warehouse/emp
drwxrwxrwx   - cloudera supergroup          0 2020-09-23 20:44 /user/hive/warehouse/employees
drwxrwxrwx   - cloudera supergroup          0 2020-10-30 16:21 /user/hive/warehouse/energy
drwxrwxrwx   - cloudera supergroup          0 2020-09-23 13:25 /user/hive/warehouse/lesson.db
drwxrwxrwx   - cloudera supergroup          0 2020-09-23 19:17 /user/hive/warehouse/new_employees
drwxrwxrwx   - cloudera supergroup          0 2020-10-30 18:13 /user/hive/warehouse/project.db
drwxrwxrwx   - cloudera supergroup          0 2020-09-23 20:43 /user/hive/warehouse/students
drwxrwxrwx   - cloudera supergroup          0 2020-09-23 19:54 /user/hive/warehouse/word_count
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse/project.db;
Found 4 items
drwxrwxrwx   - cloudera supergroup          0 2020-10-30 17:00 /user/hive/warehouse/project.db/energy
drwxrwxrwx   - cloudera supergroup          0 2020-10-30 18:11 /user/hive/warehouse/project.db/energynew
drwxrwxrwx   - cloudera supergroup          0 2020-10-30 18:00 /user/hive/warehouse/project.db/energystats
drwxrwxrwx   - cloudera supergroup          0 2020-10-30 18:14 /user/hive/warehouse/project.db/weather
[cloudera@quickstart ~]$
```

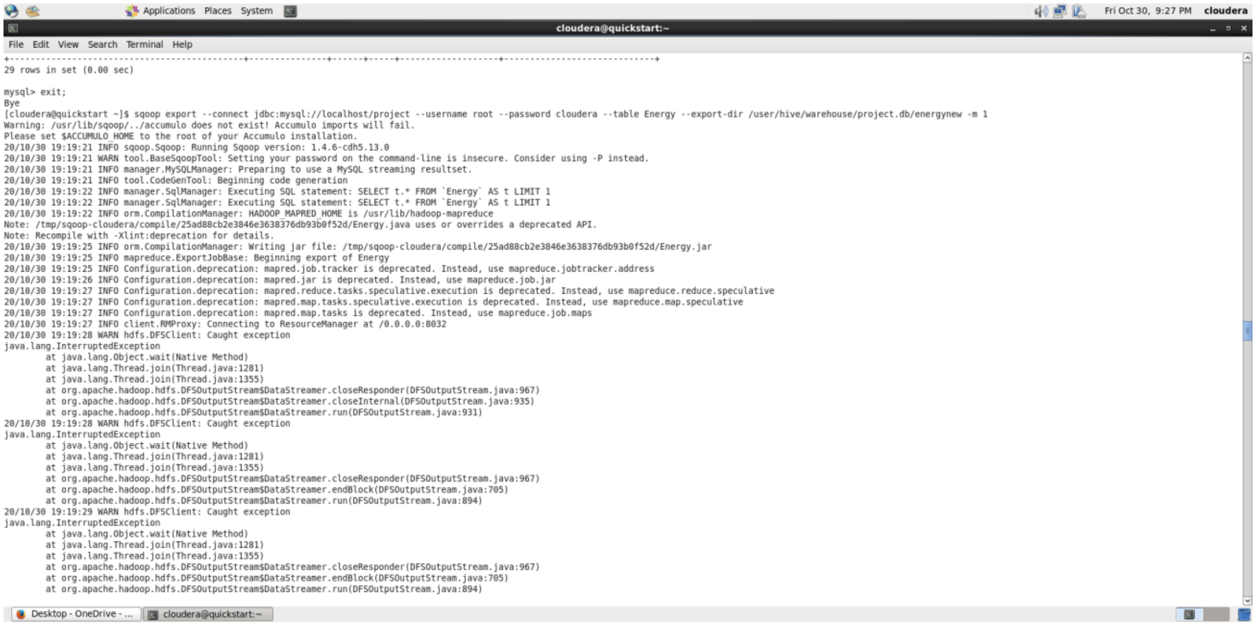
I succeeded to create a table in mysql which matched that in hive, Now the next step is to import the data from hive into mysql so I can run some queries



Unfortunately, this process kept failing, so I had to recreate my tables and made sure I used the correct datatypes.



That didn't work either.



For my future work with MySQL, I plan to:

- Find out why imported data in hive has NULL columns and fix
- Successfully move data into mysql and run queries to see relationships that exists amongst these columns

MapReduce: (Scott)

Starting with the basics in MapReduce I wanted to get some descriptive statistics for each column in our dataset. Fortunately, our dataset makes this easy since nearly all the columns are of the same type. Since I only have one datatype to worry about, I can get away with creating only one reducer to find the mean, min, and max. Each row is split up by column and written to the reducer with the field name as the key.

I wish to calculate a few more descriptive statistics, such as the median, quartiles, standard deviation, and variance but found them difficult to calculate due to the nature of MapReduce. I believe I can overcome a few of these limitations by using a secondary sort, changing the algorithm used to calculate the statistic or only calculating an approximation. In addition, I also want experiment with joining the weather dataset to start doing some complex grouping. A mapper-side join currently is not be possible without some preprocessing since the datasets are different lengths and some rows may be missing from the energy dataset.

cloudera@quickstart:~/project/Energy_Demand_Analysis/MapReduce/output

```
[cloudera@quickstart output]$ cat job_output.log
20/10/30 04:27:44 INFO mapreduce.Job:  map 0% reduce 0%
20/10/30 04:27:58 INFO mapreduce.Job:  map 100% reduce 0%
20/10/30 04:28:17 INFO mapreduce.Job:  map 100% reduce 100%
20/10/30 04:28:17 INFO mapreduce.Job: Job job_1604032870144_0008 completed successfully
20/10/30 04:28:17 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=28038648
                FILE: Number of bytes written=56365263
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=6062761
                HDFS: Number of bytes written=2170
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=10581
                Total time spent by all reduces in occupied slots (ms)=15436
                Total time spent by all map tasks (ms)=10581
                Total time spent by all reduce tasks (ms)=15436
                Total vcore-milliseconds taken by all map tasks=10581
                Total vcore-milliseconds taken by all reduce tasks=15436
                Total megabyte-milliseconds taken by all map tasks=10834944
                Total megabyte-milliseconds taken by all reduce tasks=15806464
        Map-Reduce Framework
                Map input records=35065
                Map output records=911263
                Map output bytes=26216116
                Map output materialized bytes=28038648
                Input split bytes=136
                Combine input records=0
                Combine output records=0
                Reduce input groups=26
                Reduce shuffle bytes=28038648
                Reduce input records=911263
                Reduce output records=26
                Spilled Records=1822526
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=221
                CPU time spent (ms)=18370
                Physical memory (bytes) snapshot=714465280
                Virtual memory (bytes) snapshot=3139694592
                Total committed heap usage (bytes)=643825664
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=6062625
        File Output Format Counters
                Bytes Written=2170

[cloudera@quickstart output]$
```

cloudera@quickstart:~

```
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hdfs dfs -cat energy_out8/*
forecast_solar_day_ahead        total: 50459320.00, min:  0.00, max: 5836.00, avg: 1439.06,
forecast_wind_onshore_day_ahead total: 191842272.00, min: 237.00, max: 17430.00, avg: 5471.20,
generation_biomass      total: 13440232.00, min:  0.00, max: 592.00, avg: 383.51,
generation_fossil_brown_coal_lignite    total: 15702683.00, min:  0.00, max: 999.00, avg: 448.06,
generation_fossil_coal_derived_gas      total:       0.00, min:  0.00, max:  0.00, avg:  0.00,
generation_fossil_gas   total: 197053216.00, min:  0.00, max: 20034.00, avg: 5622.70,
generation_fossil_hard_coal     total: 149158480.00, min:  0.00, max: 8359.00, avg: 4256.08,
generation_fossil_oil   total: 10454617.00, min:  0.00, max: 449.00, avg: 298.32,
generation_fossil_oil_shale     total:       0.00, min:  0.00, max:  0.00, avg:  0.00,
generation_fossil_peat  total:       0.00, min:  0.00, max:  0.00, avg:  0.00,
generation_geothermal   total:       0.00, min:  0.00, max:  0.00, avg:  0.00,
generation_hydro_pumped_storage_consumption     total: 16666608.00, min:  0.00, max: 4523.00, avg: 475.58,
generation_hydro_run_of_river_and_poundage      total: 34067880.00, min:  0.00, max: 2000.00, avg: 972.12,
generation_hydro_water_reservoir        total: 91298880.00, min:  0.00, max: 9728.00, avg: 2605.12,
generation_marine       total:       0.00, min:  0.00, max:  0.00, avg:  0.00,
generation_nuclear      total: 219530448.00, min:  0.00, max: 7117.00, avg: 6263.89,
generation_other        total: 2110771.00, min:  0.00, max: 106.00, avg: 60.23,
generation_other_renewable      total: 3001329.00, min:  0.00, max: 119.00, avg: 85.64,
generation_solar        total: 50209328.00, min:  0.00, max: 5792.00, avg: 1432.67,
generation_waste        total: 9442950.00, min:  0.00, max: 357.00, avg: 269.45,
generation_wind_offshore        total:       0.00, min:  0.00, max:  0.00, avg:  0.00,
generation_wind_onshore total: 191508528.00, min:  0.00, max: 17436.00, avg: 5464.49,
price_actual    total: 2029643.38, min:  9.33, max: 116.80, avg: 57.88,
price_day_ahead total: 1748825.63, min:  2.06, max: 101.99, avg: 49.88,
total_load_actual       total: 1005197248.00, min: 18041.00, max: 41015.00, avg: 28696.96,
total_load_forecast     total: 1006763520.00, min: 18105.00, max: 41390.00, avg: 28712.17,
[cloudera@quickstart ~]$
```

Cassandra: (Shelby)

Because joins aren't possible in Cassandra, it was necessary to keep the two tables separate. Furthermore, since Cassandra operates by a query-first approach, I created several tables within Cassandra such that each table was designed for a specific query. Though it did result it duplication of data, this design is good for high-load queries that usually happened in big data. The insights gleaned from these queries seemed rather unhelpful compared to the query capabilities of HQL and mySQL. Whereas HQL/mySQL can perform direct analysis on the data (such as calculating averages, join functions, etc.), it seems like there would have to be some secondary analysis step performed with any data returned from a Cassandra query.

'Cassandra Tables Creation.cql' – This file contains the script that was used to create and load data into five different Cassandra tables. Because the data is just text, the class used was SimpleStrategy. A replication factor of 3 was arbitrarily decided upon.

'Cassandra Queries.cql' – This file contains the queries used for each table. The result of the queries was stored into a unique txt file.

'Cassandra Results" – This folder contains the results of the five .cql queries used for each of the Cassandra tables, as well as screenshots of the successfully created tables.

```
cqlsh:group_project> DESCRIBE TABLES;

energy_by_price_actual  temp_by_time_and_city  energy_renewable_by_time
energy_fossil_by_time   weather_by_time
```

dt_iso	temp	city_name	clouds_all	humidity	pressure	rain_1h	rain_3h	snow_3h	temp_max	temp_min	weather_description	weather_icon	weather_id	weather_main	wind_deg	wind_speed
2018-05-31 12:00:00.000000+0000	292.04001	Bilbao	75	68	1018	0	0	0	293.14999	291.14999	broken clouds	04d	803	clouds	40	2
2018-05-31 12:00:00.000000+0000	295.32999	Madrid	40	43	1018	0	0	0	297.14999	293.14999	scattered clouds	03d	802	clouds	220	2
2018-05-31 12:00:00.000000+0000	296.14999	Barcelona	20	57	1017	0	0	0	297.14999	295.14999	few clouds	02d	801	clouds	130	5
2018-05-31 12:00:00.000000+0000	298.32999	Seville	0	34	1017	0	0	0	300.14999	297.14999	sky is clear	01d	800	clear	300	2
2018-05-31 12:00:00.000000+0000	299.14999	Valencia	20	39	1016	0	0	0	299.14999	299.14999	few clouds	02d	801	clouds	100	4
2016-12-20 20:00:00.000000+0000	276.26001	Madrid	0	70	1024	0	0	0	279.14999	274.14999	sky is clear	01n	800	clear	340	2
2016-12-20 20:00:00.000000+0000	280.51999	Bilbao	88	100	1026	0.3	0	0	282.14999	279.14999	light rain	10n	500	rain	0	1
2016-12-20 20:00:00.000000+0000	282.14999	Valencia	0	70	1021	0	0	0	282.14999	282.14999	sky is clear	01n	800	clear	300	3
2016-12-20 20:00:00.000000+0000	282.14999	Barcelona	75	87	1020	0.3	0	0	282.14999	282.14999	light intensity shower rain	09n	520	rain	0	0
2016-12-20 20:00:00.000000+0000	283.20999	Seville	0	93	1025	0	0	0	291.14999	278.14999	sky is clear	01n	800	clear	177	0
2015-01-08 19:00:00.000000+0000	269.29401	Madrid	0	65	978	0	0	0	269.29401	269.29401	sky is clear	01n	800	clear	353	1
2015-01-08 19:00:00.000000+0000	275.10599	Bilbao	58	88	1041	0	0	0	275.10599	275.10599	broken clouds	04	803	clouds	192	1
2015-01-08 19:00:00.000000+0000	276.95001	Valencia	0	83	1040	0	0	0	276.95001	276.95001	sky is clear	01n	800	clear	294	1
2015-01-08 19:00:00.000000+0000	278.944	Seville	0	90	1046	0	0	0	278.944	278.944	sky is clear	01n	800	clear	54	3
2015-01-08 19:00:00.000000+0000	283.45001	Barcelona	0	60	1036	0	0	0	283.45001	283.45001	sky is clear	01n	800	clear	315	2
2018-07-07 17:00:00.000000+0000	293.95001	Bilbao	40	88	1021	0.3	0	0	295.14999	293.14999	light rain	10n	500	rain	290	1
2018-07-07 17:00:00.000000+0000	298.64999	Barcelona	20	54	1018	0	0	0	299.14999	298.14999	few clouds	02n	801	clouds	0	1
2018-07-07 17:00:00.000000+0000	299.14999	Valencia	0	74	1018	0	0	0	299.14999	299.14999	sky is clear	01n	800	clear	120	1
2018-07-07 17:00:00.000000+0000	301.67001	Madrid	0	24	1017	0	0	0	305.14999	300.14999	sky is clear	01n	800	clear	270	2
2018-07-07 17:00:00.000000+0000	302.32999	Seville	0	31	1014	0	0	0	304.14999	301.14999	sky is clear	01n	800	clear	230	4

Project Management:

- Work completed:
  - Description: We have analyzed our dataset using Hive, mySQL, MapReduce, and Cassandra, as well as created some preliminary data visualizations.
  - Contributions:
    - Claire: mySQL queries (20%)
    - Wes: Hive table creation, HQL queries, visualizations (30%)
    - Scott: MapReduce queries (25%)
    - Shelby: Sqoop transfer from Hive to mySQL, Cassandra analysis, report composition (25%)
- Work to be completed:
  - Description: For our next increment, we will utilize Spark to gain more insights about our data.
  - Concerns: IntelliJ/Scala can be finicky, and some members of our group are using PySpark instead. It will be a challenge to coordinate our efforts when our setups are not the same.

Assignment 2 Questions:

- Who:
  - This dataset is about the people who use energy in Spain, whose energy production and grid was sampled for this dataset. There is no identifiable information on the individual level, meaning that there is little personal risk with this dataset.
- What:
  - The energy usage and sources of energy production of the people of Spain are what is being recorded by the data set. This addresses all of our questions in Assignment 1.
- When:
  - This data was collected between 2015 – 2019, meaning that the data is recent and therefore relevant. It is cross-sectional since the data was collected from several cities in Spain. This dataset contains real-time data.
- Where:
  - The data is collected from the five largest cities in Spain: Madrid, Barcelona, Valencia, Seville, and Bilbao. It could possible be extrapolated that the energy usage would be similar in the surrounding European countries with similar populations and weather as these five cities, and it is certainly possible that larger generalizations about predicting energy usage could be used for non-European locations.
- Why:
  - The data was collected by ENTSOE, a public portal for Transmission Service Operator (TSO) data and is publicly available.

References:

"Tackling Climate Change with Machine Learning"

https://arxiv.org/abs/1906.05433

"Hourly energy demand generation and weather – Electrical demand, generation by type, prices and weather in Space"

https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather?select=weather_features.csv

"Chapter 4. The Cassandra Query Language"

https://www.oreilly.com/library/view/cassandra-the-definitive/9781491933657/ch04.html

"Defining Application Queries"

https://cassandra.apache.org/doc/latest/data_modeling/data_modeling_queries.html

"LanguageManual Select"

https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Select

MapReduce:

https://nestedsoftware.com/2018/03/27/calculating-standard-deviation-on-streaming-data-253l.23919.html

https://hadoop.apache.org/docs/r2.6.0/api/org/apache/hadoop/mapred/lib/ChainMapper.html

https://hadoop.apache.org/docs/r2.6.0/api/org/apache/hadoop/mapred/lib/ChainReducer.html

https://hadoop.apache.org/docs/r2.6.0/api/org/apache/hadoop/mapred/Mapper.html

https://hadoop.apache.org/docs/r2.6.0/api/org/apache/hadoop/mapred/Reducer.html

https://hadoop.apache.org/docs/r2.6.0/api/org/apache/hadoop/mapreduce/Job.html

https://hadoop.apache.org/docs/r2.6.0/api/org/apache/hadoop/io/package-summary.html