Paper 3185-2015

Big-Game Hunting for Customer Behavior Trends Armed with SAS Visual Statistics with Hadoop and Hive

Mark Konya, Ameren Missouri; Kathy Ball, Devon Energy;
Paul Dick, SAS Institute

ABSTRACT

Your electricity usage reveals a lot about your family and routines. Information collected from electrical smart meters captures data that can be converted into patterns of behavior that can be utilized to optimize power systems.

Demand response programs represent an effective way to cope with rising energy needs and increasing electricity costs. Utilities receive customer data from smart meters, which track and store customer energy usage. The data collected is sent to the energy companies every fifteen minutes or hourly. With millions of meters deployed, this quantity of information creates a data deluge for utilities, as each customer generates about three thousand data points monthly and more than thirty-six billion reads are collected annually for a million customers. Large amounts of data must be processed to effectively control demand response, including transport, storage and access to the data.

Complex problems such as demand response rely on managing and digesting enormous volumes of machine-generated data to detect customers who are flexible enough in their electrical usage to curtail load at a given point in time. Traditional data warehouse systems can be costly and are not designed to handle data deluges of such enormity. The data scientist is the hunter and finding demand response candidate patterns are the prey in this cat-and-mouse game of finding the right customers willing to curtail electrical usage for a program benefit. The data scientist must connect large siloed data sources, external data, and even unstructured data to detect common customer electrical usage patterns, build dependency models, and score them against their customer population.

Taking advantage of Hadoop's ability to store and process data on commodity hardware with distributed parallel processing is a game changer. With Hadoop, no dataset is too large and SAS Visual Statistics leverages machine learning, artificial intelligence, and clustering techniques to build descriptive and predictive models. All data can be usable from disparate systems, including structured, unstructured, and log files. The data scientist can utilize Hadoop to ingest all available data at rest, analyze customer usage patterns, system electrical flow data, and external data such as weather.

This paper will be using Cloudera Hadoop with Apache Hive queries for analysis in SAS Visual Analytics and SAS Visual Statistics. This paper will showcase optionality within Hadoop for querying large datasets with open-source tools and importing these data into SAS for robust customer analytics, clustering customers by usage profiles, propensity to respond to a demand response event, and an electrical system analysis for demand response events.

INTRODUCTION

With the implementation of smart meters utilities are receiving oceans of customer data with typical sampling periods ranging from 5 to 60 minutes. Despite this deluge, or perhaps as a result of it, many have questions about how to leverage the power of meter data. In this paper we will focus on how to apply SAS analytics and business intelligence capabilities to big data via Hadoop using SAS, Cloudera Impala, and Apache Hive to identify demand response opportunities with customers.

Figure 1: Hadoop Symbols to be used in Analysis, Source: SAS and Cloudera.com

This paper will also demonstrate how to manipulate and analyze customer usage data collected from smart meters sets using SQL and familiar scripting languages along with using Apache Hive and Impala in the Hadoop cluster. In doing so, utilities can progress toward meeting demand response goals and optimizing customer-facing demand response programs. Additionally, this paper will show how to leverage real-time interactive analysis of data stored in Hadoop to respond to dynamic commodity markets for demand response optimization.

TACKLING THE BIG DATA ISSUE WITH HADOOP

Utilities will have to harness the power of data to truly understand what drives each customer's power consumption behavior. In doing so, utilities can design effective demand response programs and enroll the best customers to achieve stated goals. Embracing new technology like Hadoop and high-speed integration will allow utilities to create advantage in demand side management to react in real-time and near real-time to market, customer and environmental conditions and meet the goals of the new electric utility market.

For this paper we used Cloudera 5.3.2 for all queries in Hadoop. Hadoop is an open-source software framework for storing and processing big data in a distributed fashion on inexpensive hardware; its purpose is to store massive amounts of data and provide access to the data for fast processing. Apache Hive project within Hadoop provides a data warehouse view of data for the Hadoop Data File system (HDFS). Programming in Hive lets you summarize data, create ad-hoc queries, and analyze large datasets in the Hadoop cluster. HDFS allows you to scale across all data nodes. Performing queries within Cloudera Hadoop is accomplished via Hue which is a web application that interacts with Hadoop clusters. Hue applications let you work with Hive and Cloudera Impala queries, MapReduce jobs and Oozie workflows. Figure 2 shows the architecture for Cloudera Hadoop and how Hue fits within the Hadoop programs. Figure 3 is the typical way to operationally access the Hive and Impala programs within the Cloudera Hadoop framework.

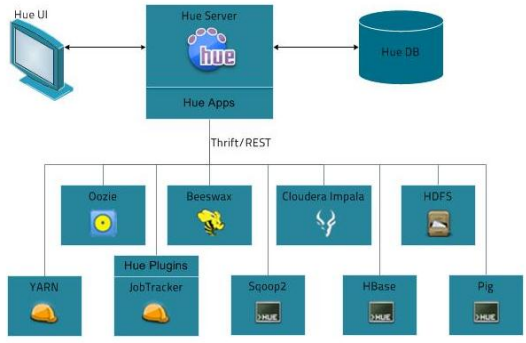


Figure 2: Source: Cloudera.com

Figure 3: Source: Cloudera.com

Hourly smart meter data for millions of customers can be stored and accessed through Hadoop as well as hourly weather data from state climatology offices. These data sources are analyzed to tackle the demand response problem: find the right customers who can curtail load when needed to reduce peak electrical demand. As seen below, the data is loaded into Hadoop from source files.

FINDING JUST THE DATA NEEDED IN HADOOP CLUSTERS

If not already stored in Hadoop clusters, data needs to be uploaded into Hadoop and joined with other datasets as needed. Figure 4 shows the datasets used for this project. We loaded hourly electrical usage data from 75,000 residential customers and corresponding weather data for that time period into Hadoop. This project reviewed data from 2011 for customers in Columbia, South Carolina.

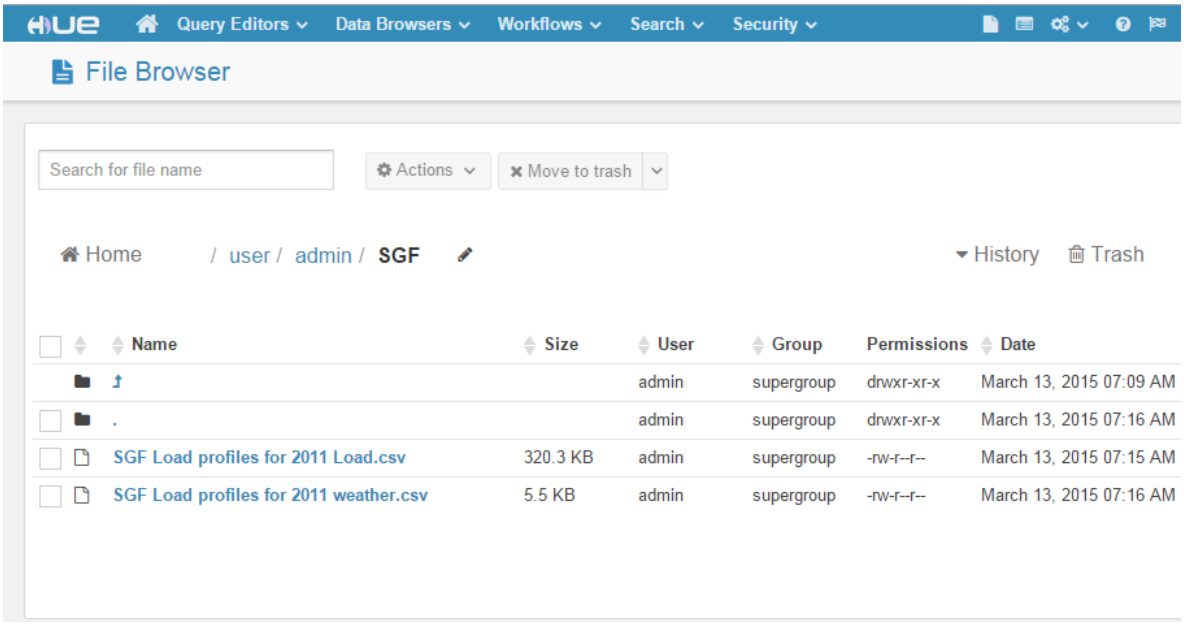


Figure 4: Viewing Datasets in Hadoop via File Browser.

Figures 5 and 6 show variables involved in this paper to determine the best demand response candidates for analyses. This data was then queried to extract the data for specific days when demand response events may have occurred for the utility company, i.e., days when it is necessary to curtail load during high demand time intervals.

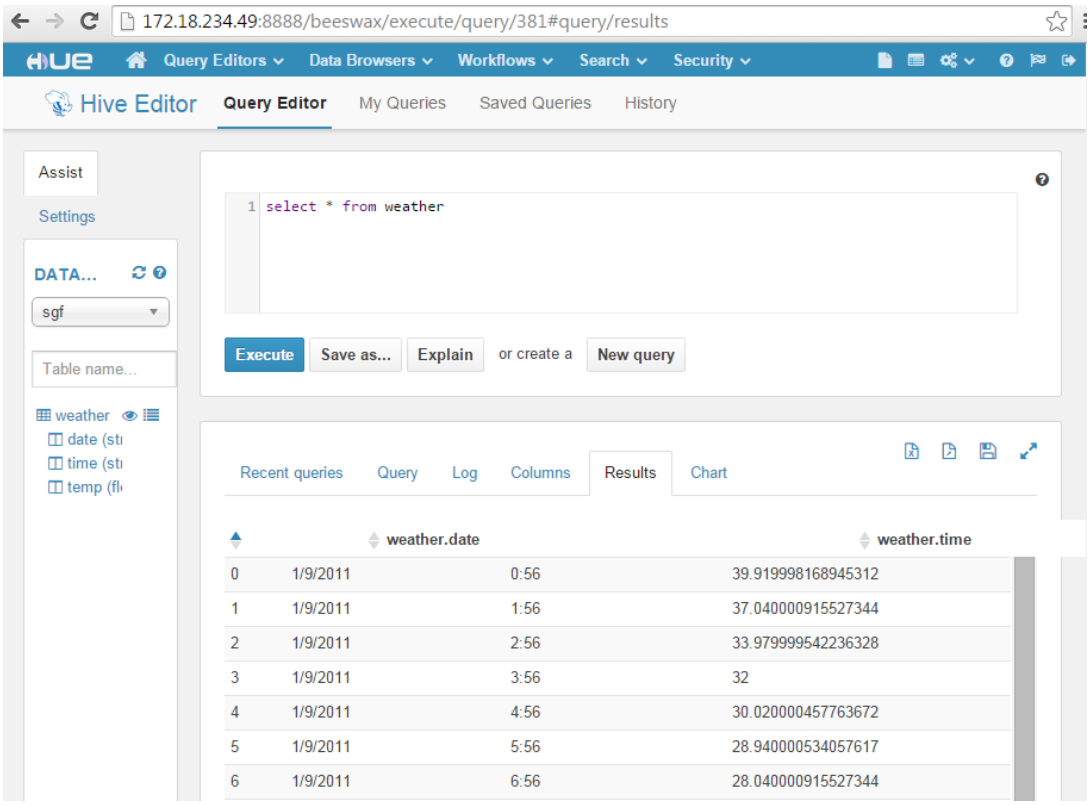
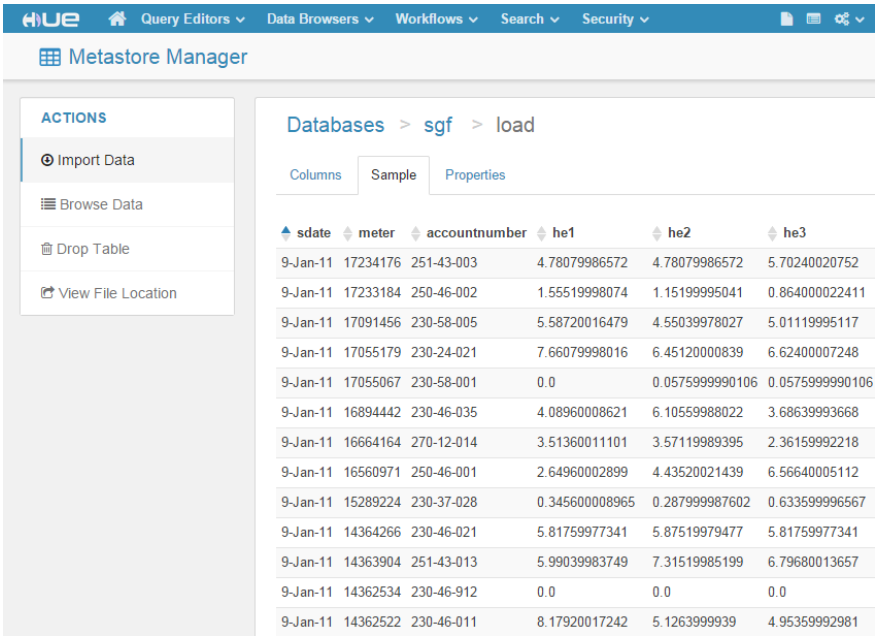


Figure 5: Weather Date Used, Source – SC State Climatology Office



sdate	meter	accountnumber	he1	he2	he3
9-Jan-11	17234176	251-43-003	4.78079986572	4.78079986572	5.70240020752
9-Jan-11	17233184	250-46-002	1.55519998074	1.15199995041	0.864000022411
9-Jan-11	17091456	230-58-005	5.58720016479	4.55039978027	5.01119995117
9-Jan-11	17055179	230-24-021	7.66079998016	6.45120000839	6.62400007248
9-Jan-11	17055067	230-58-001	0.0	0.0575999990106	0.0575999990106
9-Jan-11	16894442	230-46-035	4.08960008621	6.10559988022	3.68639993668
9-Jan-11	16664164	270-12-014	3.51360011101	3.57119989395	2.36159992218
9-Jan-11	16560971	250-46-001	2.64960002899	4.43520021439	6.56640005112
9-Jan-11	15289224	230-37-028	0.345600008965	0.287999987602	0.633599996567
9-Jan-11	14364266	230-46-021	5.81759977341	5.87519979477	5.81759977341
9-Jan-11	14363904	251-43-013	5.99039983749	7.31519985199	6.79680013657
9-Jan-11	14362534	230-46-912	0.0	0.0	0.0
9-Jan-11	14362522	230-46-011	8.17920017242	5.1263999939	4.95359992981

Figure 6: Hourly Load Data Used, Source – Anonymous

Using information from the South Carolina State Climatological Office, the three hottest day were selected from 2011 as this was an abnormally hot year placing extreme demands on the utility's electrical system. High temperature days are typical demand response events when the utility needs to determine which customers can curtail their load to reduce overall electrical demand.

Figure 7 shows the query used to extract weather information for hourly temperature readings to determine when a demand response event is likely to occur. Note that column names may need to be changed so as not to conflict with programming commands. Note that date column names needed to be changed to run in Impala and was changed to sdate and for this reason.

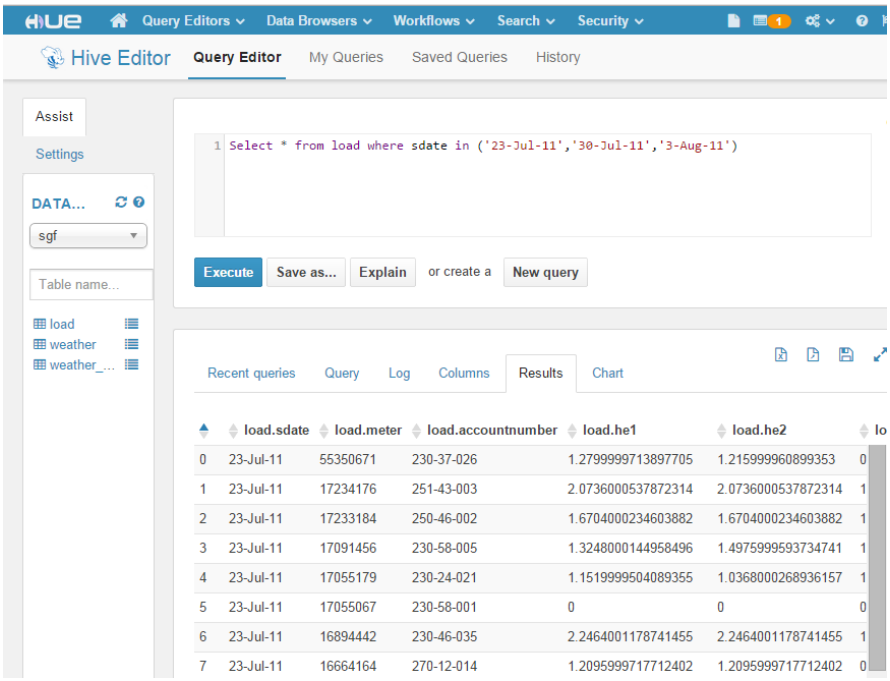


Figure 7: Hive Query to collect weather data for demand response events

Programming in Impala allows users to speed up the time it takes to run a query. In this example, the query took several minutes to complete in Hive versus seconds in Impala. Figure 8 shows the query executed in Impala to collect weather data for this project.

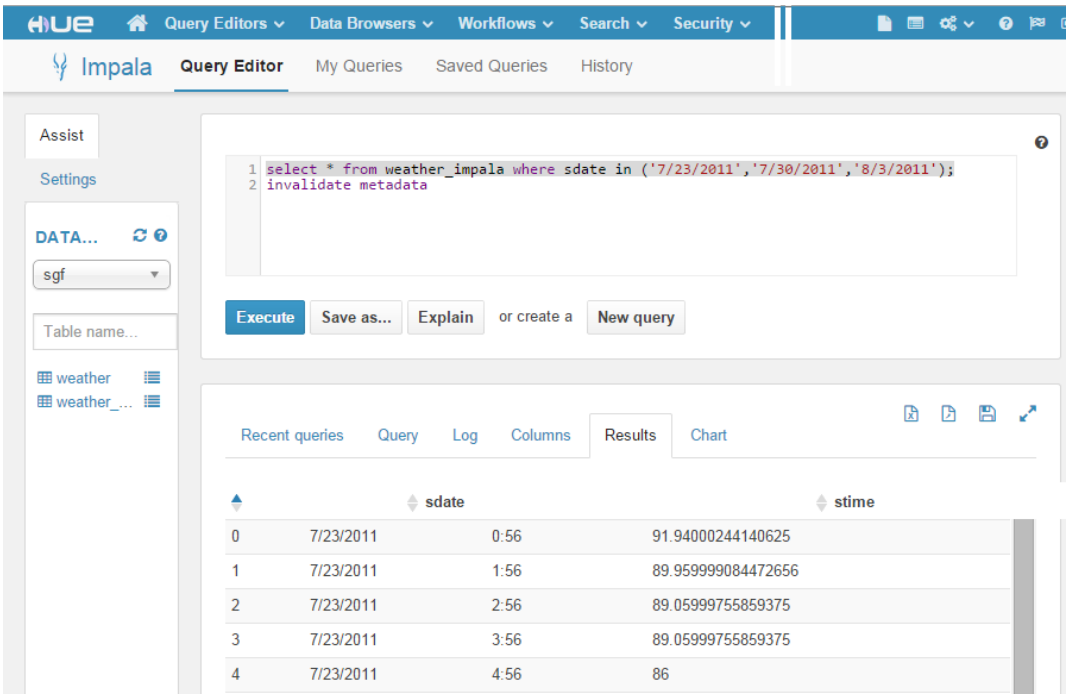


Figure 8: Impala Query to collect weather data for demand response events

Lastly, Hadoop has built in functionality which allows users to view and chart data prior to starting any analytical exercise. Figures 9 and 10 show results from a Hive query plotting electrical load for 5 pm versus metered load and an Impala query illustrating temperature over time.

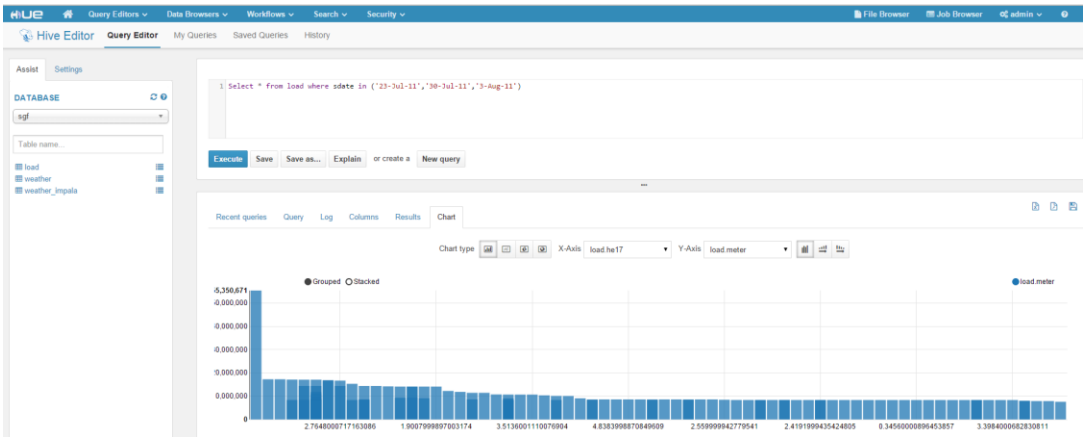


Figure 9: Hive Chart comparing High electrical usage across all customers at that time

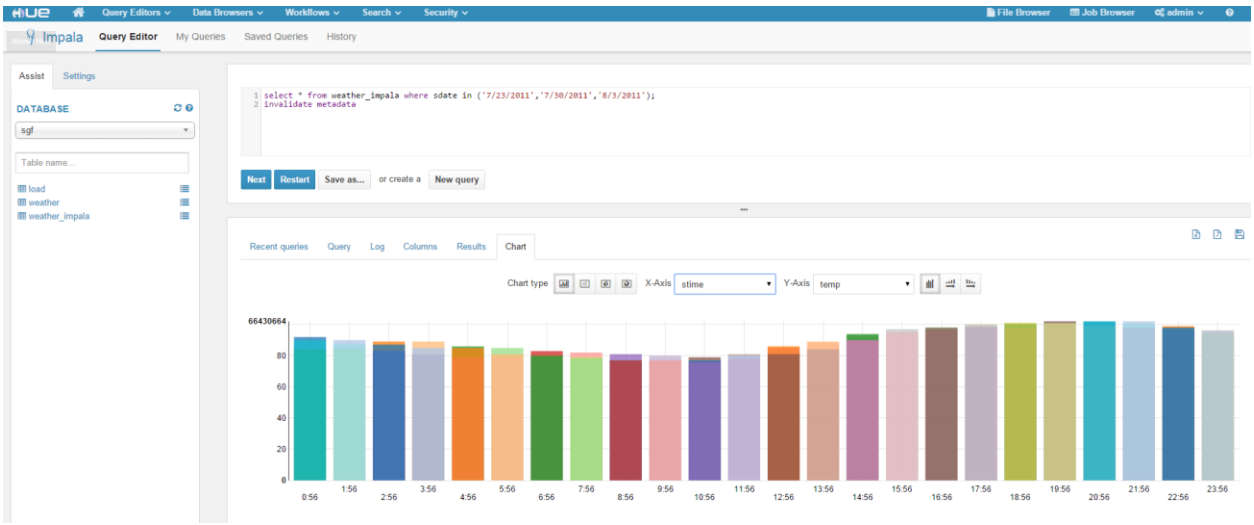


Figure 10: Impala Chart comparing Temperature distributions over times of a day

ANALYZING THE DATA

The problem to be solved for demand response is to identify customers who are likely to participate in a demand response event based on their electrical usage profile. Some utilities simply enroll customers willing to participate in electrical load curtailment into a demand response program. This paper provides one analytical approach to help identify customers likely to help the utility meet electrical curtailment and demand response goals.

Typical smart meter data consists of customer electrical usage (Kilowatt-Hours, or KWH) collected hourly in the time domain. There are some time series, however, which are better viewed and understood in the frequency domain. For example, in order to monitor the "health" of rotating machinery accelerometers can be placed on machine components; resulting acceleration, velocity and displacement vectors can then be monitored in the frequency domain to detect incipient failures.

Similarly, electrical load profiles can contain deterministic periodic components characteristic of cycling home equipment such as refrigerators, air conditioners and heating systems. Load profiles containing these cycling loads can be broken down into periodic components by transforming the load profile from the time domain to the frequency domain using an integral transform such as the discrete Fourier transform.

In this paper we performed a proof of concept to demonstrate that PROC SPECTRA could be used to transform an analytically-generated time series to identify periodicities or cyclical patterns whose Fourier components were known. Clusters of time series with similar frequency content were then grouped using two different methods to illustrate how customers with similar frequency content in their load profiles could be grouped.

PROC SPECTRA, which uses the finite Fourier transform to decompose data series into a sum of sine and cosine functions with different amplitudes and wavelengths, was employed to transform the analytically-generated time series. Shown below is standard SAS code to use PROC SPECTRA.

```
proc spectra data=a out=b coef;
var x;
run;
```

DEFINING STANDARD PROFILES AND PERFORMING THE TRANSFORMATION

In this proof-of-concept nine time series were used as proxies for customer load profiles. Each profile, defined by the parameters indicated below, was then transformed to the frequency domain using PROC SPECTRA. Results yielded power spectral densities as a function of frequency for each of the standard load profiles.

Variable	Y11	Y12	Y21	Y22	Y23	Y31	Y32	Y33	Y34
TS	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
PI	3.14159	3.14159	3.14159	3.14159	3.14159	3.14159	3.14159	3.14159	3.14159
N_OBS	96	96	96	96	96	96	96	96	96
T_COS	24	24	12	12	12	8	8	8	8
T_SIN	1	1	1	1	1	2	2	2	2
F_COS	0.041667	0.041667	0.083333	0.083333	0.083333	0.125	0.125	0.125	0.125
F_SIN	1	1	1	1	1	0.5	0.5	0.5	0.5
W_COS	0.261799	0.261799	0.523598	0.523598	0.523598	0.785398	0.785398	0.785398	0.785398
W_SIN	0.261799	6.28318	6.28318	6.28318	6.28318	3.14159	3.14159	3.14159	3.14159
A	1	1	1	1	1	1	1	1	1
B	0	0	0.25	0.25	0.25	0.1	0.1	0.1	0.1
PH_COS	3.1416	2.3562	0	0.7854	1.5708	0	0.7854	1.5708	0
PH_SIN	0	0	0	0.7854	0.3927	0	0	0	0
ERR_WT	0.01	0.005	0.01	0.01	0.005	0.005	0.005	0.005	0.005
OFFSET	2	2	2	2	1.5	2	2.5	2	1.5

Y(T) = OFFSET +A*COS(W_COS*T+PH_COS) + B*SIN(W_SIN*T+PH_SIN) +ERR_WT*RAND(-50,50)

Figure 11: Standard time series definitions

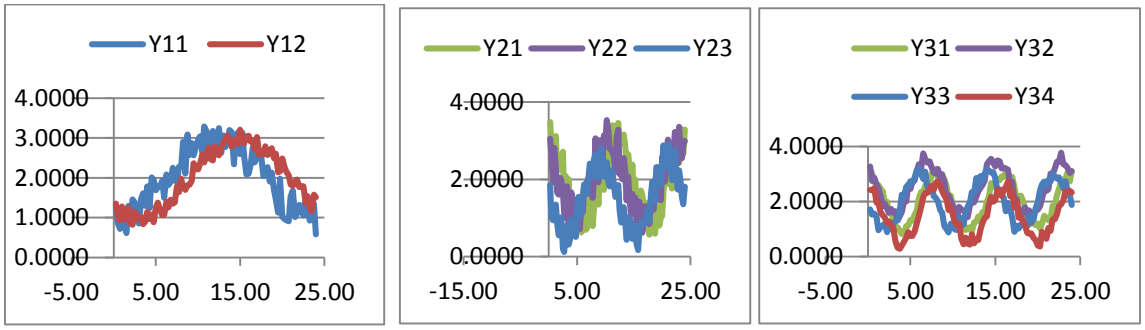
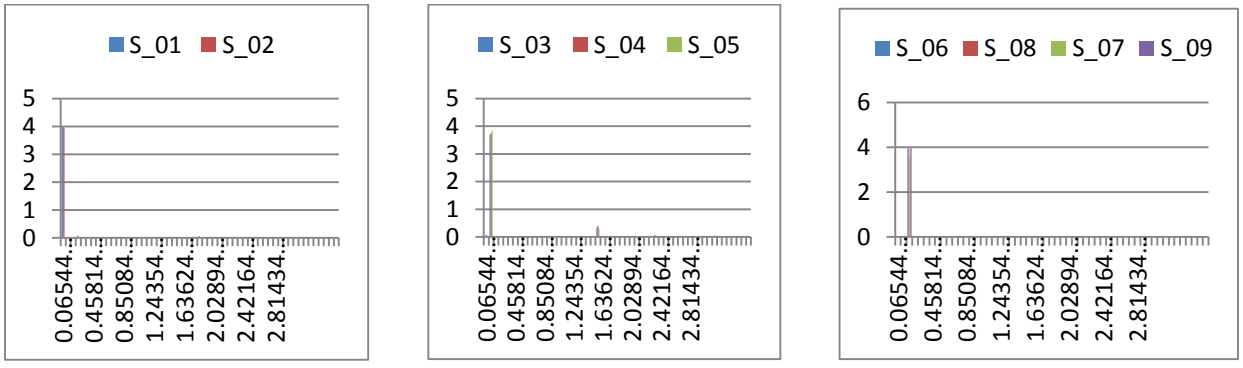


Figure 12: Standard time series plots in the time domain



Spectral Densities for Y11 and12 Spectral Densities for Y21, Y22, Y23 Spectral Densities for Y31–Y34

Figure 13: Transformed standard time series plotted in the frequency domain

CLUSTERING IN THE FREQUENCY DOMAIN

Power spectral densities for all profile pairs were next correlated using PROC CORR to determine which pairs contained similar frequency content. Clusters were easily extracted by sorting correlation coefficients for profile pairs in descending order and then grouping profiles with significant ($p<0.05$) and large correlation coefficients. Referring to Figure 14 for an example, column S_09 is clustered with rows S_06, S_07, S_08, and S_09 which corresponds to Y31, Y32, Y33, and Y34 of the standard time series, respectively. This result is expected because Y31 through Y34 have the same frequency content as shown in Figure 11.

Correlation Matrix

		S_01	S_02	S_03	S_04	S_05	S_06	S_07	S_08	S_09
S_01	Spectral Density of Y11	1.0000	0.9996	-.0279	-.0108	-.0242	-.0238	-.0251	-.0246	-.0246
S_02	Spectral Density of Y12	0.9996	1.0000	-.0267	-.0095	-.0230	-.0217	-.0231	-.0226	-.0225
S_03	Spectral Density of Y21	-.0279	-.0267	1.0000	0.9994	0.9995	-.0270	-.0253	-.0262	-.0258
S_04	Spectral Density of Y22	-.0108	-.0095	0.9994	1.0000	0.9993	-.0266	-.0250	-.0258	-.0254
S_05	Spectral Density of Y23	-.0242	-.0230	0.9995	0.9993	1.0000	-.0244	-.0228	-.0237	-.0233
S_06	Spectral Density of Y31	-.0238	-.0217	-.0270	-.0266	-.0244	1.0000	0.9999	0.9999	0.9999
S_07	Spectral Density of Y32	-.0251	-.0231	-.0253	-.0250	-.0228	0.9999	1.0000	1.0000	1.0000
S_08	Spectral Density of Y33	-.0246	-.0226	-.0262	-.0258	-.0237	0.9999	1.0000	1.0000	1.0000
S_09	Spectral Density of Y34	-.0246	-.0225	-.0258	-.0254	-.0233	0.9999	1.0000	1.0000	1.0000

Figure 14: Spectral Density clustering methodology, Source SAS Enterprise Guide

PRINCIPAL COMPONENT ANALYSIS

An alternative to correlation analysis of profile pairs is to perform a principal component analysis (PCA) of power spectral densities using PROC PRINCOMP. PCAs are helpful when numerous measures are collected on a number of observed variables which can be reduced to unboserved principal components. PCA is a variable reduction procedure where some variables used are highly correlated with one another. PCA creates linear combinations of observations called principal components. Each of the cascading components has a descending explanatory contribution for variance that is orthogonal (uncorrelated with) the preceding components. Principal components are orthogonal because they are the eigenvectors of the covariance matrix, which is symmetric. However the observable contributions to each component may not be interpretable if the component loadings are not clearly separated into distinct groups.

EIGENVECTOR TIME SERIES CLUSTERING RESULTS

Figure 15 shows the result of the PCA for the spectral densities calculated for the nine standard time series analyzed. Note that PRIN1 has constituent components Y31, Y32, Y33, and Y34. These series all have similar frequency components as defined analytically in Figure 11, which is the expected result. Similar groupings occur for PRIN2 and PRIN3 as expected.

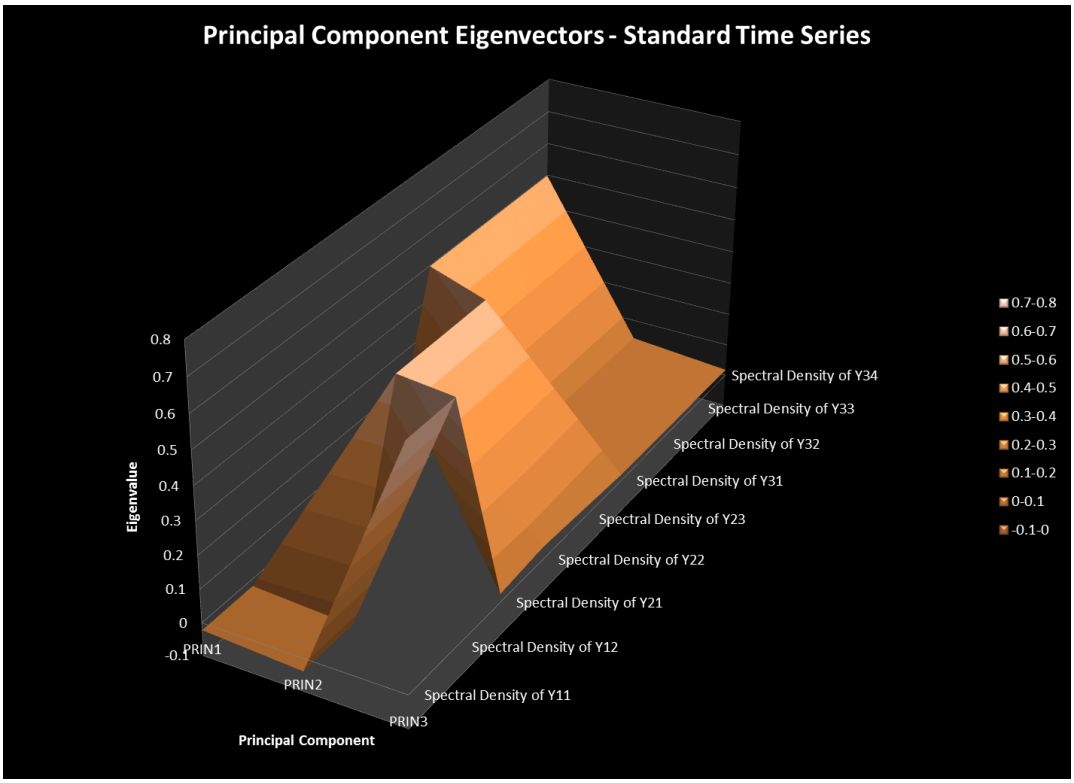


Figure 15: Eigenvector clustering for time series data

By extending this transformation-clustering method to actual load profiles it should be possible to identify large groups of customers with similar frequency content in their load usage patterns in near-real-time. Comparing these patterns to known equipment cycling patterns should also make it possible to identify customers utilizing different equipment types, the vintage of equipment, or possibly other extrinsic factors which influence equipment cycling.

In any case, grouping customers with similar frequency content in their load profiles should make it possible to target certain classes of equipment (and thereby customers) for load curtailment based on the cycling characteristics of equipment installed in customers' homes. Because there is a large amount of profile data available to utilities who have installed smart meters, this type of analysis can be realistically accomplished in the Hadoop environment in near-real-time using Fast Fourier Transforms and other high performance methods.

Note that good practices for frequency domain analysis must be followed to obtain usable results with these methods. There are many good references which document the theory of frequency domain analysis and its application. For example, refer to "Measurement and Analysis of Random Data" by Julius Bendat and Allan Piersol, John Wiley and Sons, Inc., New York, 1958.

PROPENSITIES FOR DEMAND RESPONSE

In some areas, utilities have the ability to control individual devices within a home such as HVAC temperature settings, Water Heaters, Interior lighting and exterior lighting. Figure 16 shows a kilowatt breakdown by components for devices in Columbia, SC. These metrics are available nationally and can be customized for electrical component profiles within a geographic area. From this graph, a utility can see which device has the highest electrical usage on an hourly basis and the amount of power that could be garnered for demand response curtailment.

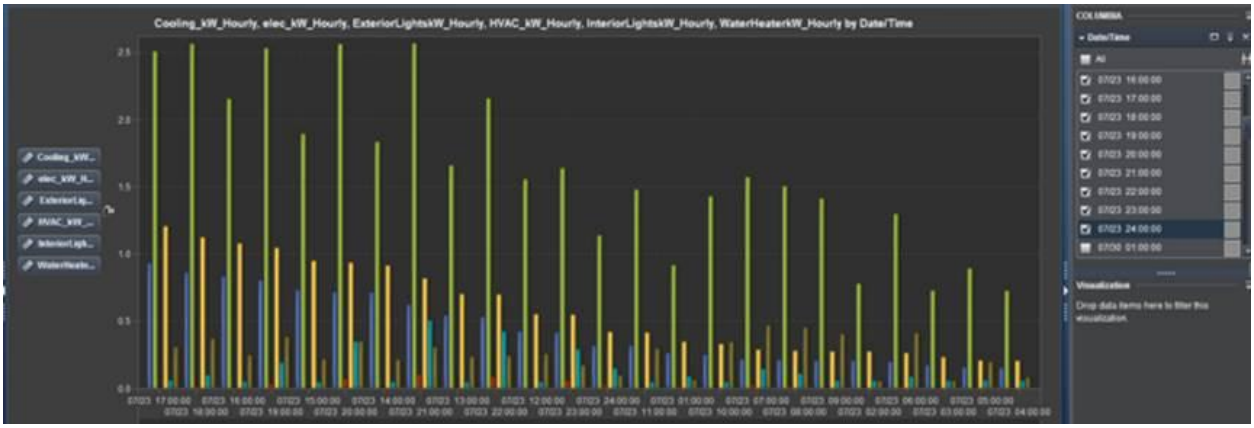


Figure 16: Electrical Device Usage by components for Columbia SC, Source: SAS Visual Analytics

CONCLUSION

The volume, variety, and velocity of data needed to analyze customer information and outside data for demand response event is complex. However, using SAS and Hadoop a utility can effectively sort through mountains of data in a short period of time using Hadoop functionality. This paper used Cloudera Hive and Cloudera Impala for querying data. However, newer techniques are being established that will allow utilities to query data in real time within Hadoop such as Apache Spark which runs programs up to one hundred times faster than Hadoop MapReduce in memory. Since utilities typically have millions of customers, hourly and sometimes sub hourly readings make demand response clustering quite difficult. Other programs within Hadoop such as ACID, which stands for four traits of database transactions: atomicity, consistency, isolation, and durability, add extra capabilities to Hive. ACID allows for streaming ingest of data versus dealing with Hive related partition issues, allows for slow changing dimensions allowing for insertion of individual records or updates to records, and data restatement. This new functionality allows Hive to INSERT, UPDATE, and DELETE.

The flexibility of being able to cluster customers on the fly based on their current electrical usage is a game-changer. Utilities will no longer have to rely on historical usage patterns which may or may not predict future behaviors and can instead target current behaviors of customers for demand response events. Near-real-time frequency domain clustering based on spectral density correlations or PCA will yield clusters of customers with similar frequency content in their load profiles. This intelligence can be leveraged to help identify groups of customers whose equipment usage profiles make them candidates for demand side management or other energy efficiency programs.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Mark Konya
Ameren Missouri
MKonya@ameren.com

Kathy Ball
Devon Energy
Kathy.Ball@dvn.com

Paul Dick
SAS Institute
Paul.Dick@sas.com