

Abstract

A new concept has been tested for the past two years in New York City called CitiBike. The service is essentially a bike share program for the city. The unique thing about CitiBike is that bikes are removed and docked into bike stations with the capability to record data about where the bike has traveled. Valuable information can be gleaned from this data, but at millions of data entries per month, storage and software demands grow strained very quickly. Hence, the use of Hadoop, an open source framework for storing and processing large data sets, should be considered to make use of this data. Hadoop is a Big Data solution that has four key distinctions: accessibility, scalability, simplicity, and robustness. Each allows for a Big Data solution that can be customized to CitiBike's needs, experience, and required data storage size. Along with the storage capacity needs, CitiBike needs a way to analyze the data. We can use Hadoop compatible programming such as Hive and Pig to extract deep insights from the data. By implementing a Hadoop Big Data system that includes Hive and Pig programming with CitiBike data, we can demonstrate how CitiBike can improve its operational procedures, expand its marketing capabilities, increase customer satisfaction, and improve its profit margin as well. This project, while minimized for the purpose of this assignment, has been developed to demonstrate these capabilities and how CitiBike can leverage them to achieve their business goals.

Introduction

Big Data is the future and the future is here. Big Data is different than traditional data that has been used in the past. Traditional data is something of the sort where we can make rows and columns in a matrix and place the individual values of data into cells that will be used for processing and analyzing at a later time. This type of data works really well when the data is based on transactions, such as those found in retail and financial industries. For years companies in these industries have used traditional data effectively in relational databases to understand their customers and improve their operating margins. However, Big Data is different, requires a different approach, and can create a lot more opportunities for improving business models, marketing, profitability, and fraud/security prevention to name a few. With CitiBike generating millions rows of data regularly, traditional data systems will not be effective long-term so using a new Big Data approach could serve CitiBike well.

More about Big Data and Why it Helps

Big Data is often described as data that varies in complexity, is generated at different velocities, has varying degrees of ambiguity, and cannot be processed using traditional technologies, processing methods, algorithms, or any commercial off-the-shelf solutions

(Krishnan, 2013). It is data that is collected from almost any source, and those sources include everything from basic new articles and documents, to video, to audio, sensor data from machines, social media forums, and much more.

Big Data has three main dimensions: volume, velocity, and variety. Data volume is characterized by the amount of data that is generated continuously (Krishnan 2013). The volume is generated by the company, its customers, competitors, the operating environment, and more. Some examples of where the volume of data is expanding rapidly are with machine data, **sensor data**, application logs, clickstream logs, emails, contracts, and **geo-spatial data**. Velocity is another dimension of Big Data and is the constant stream of data from sensors, mobile networks, social media, and large websites such as Google, Facebook, Amazon, and Yahoo. The key to handling velocity is a processing engine that can work at extremely scalable speeds on extremely volatile data. Finally, data variety is the final dimensionality of Big Data that must be addressed. Since data comes in so many different forms, this increases the processing complexity and processing requirements include graphs, video, audio, image, distributed processing capabilities, and scalability. Other dimensions of Big Data are ambiguity, viscosity, and the virality of data.

Big Data has been around for a long time, but the lack of automated processes and ability to efficiently store the large amounts of data collected has always kept the process slow and exclusive only to those with extensive statistical skills. All of that has changed nowadays with the improvements in software, database architecture, and database storage. This makes handling and processing Big Data much faster, it can be scaled to different sizes much easier than before, and the processing of the data is much more flexible than before.

In addition to the improvements in automated technology and data, the amount of information that can be gathered these days is staggering. The use of sensors in the environment and on machines, smart phones, text data, video data, audio data, news articles, journal documents, etc. has made gathering data almost overwhelming. Yet the new Big Data infrastructure is designed to handle such data, which makes gathering such Big Data much easier. The most exciting part is then extracting information from the data which can be used, ultimately, to improve profitability and to predict future trends.

CitiBike has a need to handle the volume of data produced by its system for three main reasons. First, it is regularly receiving and recording sensor data from each bike. Second, there is a geo-spatial element to the data because the latitude and longitude coordinates are recorded for each beginning and end location of the trip. Third and lastly, the company has tens of thousands of users each day so the data arrives and accumulates quickly. Having the capacity to handle such a quickly growing amount of data is another reason why CitiBike needs a Big Data solution. This analysis will demonstrate some of the extra insights that can be gained by leveraging this geo-spatial data. This project analysis will also demonstrate how to use the Big Data solution to extract insights to improve operations, marketing decisions, and customer satisfaction – all of which ultimately improve the vaunted profit margin.

Today we live in an insight economy where information is capital now. What that means is information is king in a competitive market because a company must first gather the data or information about virtually every area related to their business in order to gain an advantage over their competitors.

Businesses are using Big Data in almost every way today. Debates and decision making processes are often decided by large amounts of data analytics applied to the data. Tracking and monitoring customers is becoming a greater importance to gain insights to their behaviors and decision making patterns. Big Data is being used to automate traditional processes or develop new products and services based on the insights found in the data. Improvements in pricing, profitability, supply chain optimization, effective financial forecasting, workforce performance management, fraud detection, and security advancements are all possible through the use of Big Data.

Furthermore, three fundamental trends in the business world have moved the way we engage in business, provide services, and measure the value and profitability of our efforts. First, the business model has been transformed by globalization and connectivity so business can evaluate data from many more customer touch points. This allows businesses to improve their understanding of their services compared to the competition and how to adjust their process for better customer service. Second, globalization has changed the commerce of the world and the variety of data that is available for consumption and analysis. Third, the valuation of personalized services, and the value perceived in that by customers, has played a large role in the shift to creating a customized experience through the use of Big Data.

Clearly, to take advantage of Big Data, for CitiBike the first step is about gathering the data. But with the volume, velocity, and variety of data that is available, how would CitiBike do that?

How Big Data Works and Technology Available for CitiBike

Big Data is regularly classified by five main characteristics: volume, velocity, variety, ambiguity, and complexity. Big Data technology must be able to handle each of these characteristics effectively in order to provide sufficient value to a business. While there are a number of important Big Data technologies, the first important Big Data platform is Hadoop.

Hadoop provides the architecture to solve Big Data processing on a cheaper commodity platform with faster scalability and parallel processing (Krishnan, 2013). Hadoop uses a file distribution system that is highly fault-tolerant, scalable, and can run on common hardware called HDFS. It can process extremely large files, process high-throughput rates from streaming data, and doesn't require special hardware requirements. This is a key component for compatibility and cost effectiveness for CitiBike's needs, but CitiBike also needs programming ability to extract information from the data.

MapReduce is a programming model for processing large data sets and is often used on top of HDFS for handling job scheduling and execution. Its ability to process distributed computations on large amounts of data on clusters of commodity servers makes it very flexible and an important interface on Hadoop. MapReduce can be combined with SQL to allow exploration with normal SQL functions and regular business intelligence tools.

HBase is another piece of database architecture that fills in some of the gaps left by Hadoop HDFS. It is used for operations other than MapReduce execution, operations not easy to work with HDFS, and when random access to data is needed. It satisfies two types of uses: 1) enables developers to deploy programs that can quickly read/write to specific subsets of large amounts of data without having to process the entire data set; 2) provides a transactional platform for running high-scale, real-time applications as an ACID –compliant database while still handling the large volume of data on the Hadoop platform. It is the database of choice for all Hadoop deployments due to its scalable and flexible architecture and support of key-value outputs from MapReduce.

Hive is another open-source data warehousing solution that has been built on top of Hadoop. It has the advantage of being able to support SQL functions with MapReduce. Hadoop cannot answer low-latency queries and deep analytical functions need to harness some SQL like language with MapReduce, so this is where Hive is so effective; it fills in the gaps left by Hadoop.

Pig is another Big Data technology. It is primarily a scripting language for exploring large data sets. It can be used in two modes for executing states: local mode or MapReduce mode, which makes it useful when working with Hadoop. It is mainly designed to analyze large data sets. Combined with that trait and its MapReduce mode, it is ideal for working with the Hadoop platform.

Pig Latin, the programming language used with Pig, is a simple yet powerful high-level data flow language similar to SQL that executes MapReduce jobs. It is a Hadoop extension that simplifies Hadoop programming by giving you a high-level data processing language while keeping Hadoop's simple scalability and reliability.

Zookeeper is essentially the coordinator of tasks between nodes on Hadoop. It is an open-source distributed NoSQL database that handles resource and application coordination. It is primarily designed to store coordination data and not large data volumes.

While there are more Big Data technologies, these are some of the most commonly used technologies, and the technologies we will be most concerned with during the CitiBike dataset analysis. However, it isn't all roses with Big Data and some of the difference should be recognized and planned for.

Big Data has four major differences from traditional technologies. Data volume is much, much larger and includes data both internally and externally. Data variety is much greater including different formats from different sources (video, audio, text, machine, etc.). Data ambiguity is more because of the complexity of the data and ambiguous metadata

associated with the data. Finally, the speed or velocity of the data is much higher (can be real-time) than traditional data. Thus, traditional technologies cannot handle the scalability, throughput, and flexibility required by Big Data.

In terms of processing data, traditional data processes can be described as collecting, processing, managing, and finally generating. In a traditional environment, data is first analyzed and a set of requirements is created. Discovery and model creation followed by the database structure finalize the process. With Big Data, the data is first collected and loaded onto a target platform. A metadata layer is then added and data structure is created. Finally, the data is transformed and analyzed. These volume and complexity differences require Big Data to operate with a file-driven architecture with a programming language interface as opposed to a database-driven architecture.

Furthermore, regular database issues, while improvements have been made, still provide challenges to Big Data. Storage is the first problem as we do not consume all the storage that is available on a disk, the cost per byte for fast-performing storage has been expensive, and the cost drives up the overall cost of ownership. Transporting or moving data between different systems is another challenge. Processing data has evolved over time but remains a challenge of Big Data due to the volume, variety, and velocity of the data. Finally, speed or throughput is a challenge and this can certainly be an issue from a value and financial perspective.

By understanding the technology, knowing how to apply it properly to our CitiBike business environment, and taking advantage of the relatively small learning curve to learn the programming languages developed to run MapReduce tasks to extract insights out of the system, we can create a data platform capable of meeting the requirements of accessibility, scalability, usability, and robustness for CitiBike.

Dataset Description and Specifications

The dataset used for this demonstration is provided directly from CitiBike and includes all data recorded each time a bike is used on a trip longer than sixty seconds in duration. Each row of data is one single trip taken by a CitiBike customer. Within each row a number of data values are recorded and stored. Trip duration is stored by number of seconds the trip took; the trip start date and time and the trip finish date and time are recorded; start station id and name, end station id and name, and their latitude and longitude coordinates are recorded; the bike id number is collected; the user type is listed as either customer (which means this person only purchased a 24-hour or 7-day pass) or subscriber (which means this person is an annual member); gender (0 equals unknown, 1 equals male, 2 equals female); and year of birth.

Preprocessing of the data involved decisions to improve the representative qualities of the data. First, trips taken by staff members to service and inspect the CitiBike system were omitted. Trips taken from the designated “test stations” which were used check the system were removed. Finally, any trip under sixty seconds was removed because it

could represent potentially false starts or from users trying to re-dock a bike to ensure it was secure.

Right now, the data is provided on a monthly basis. It comes as a .CSV file which makes it easy to transfer to tab delimited files if need be. Each file contains a month's worth of CitiBike trips and can easily reach more than one million rows of data. The actual file size, once extracted, is usually a few hundred megabytes in size. It's easy to see that by combining a few datasets or creating one large cumulative history of trips that CitiBike would easily be dealing with gigabytes of data. Furthermore, that dataset will continue to grow each month. Scalability and ability to process large volumes of data are two important requirements for CitiBike's situation. To gather the absolute latest data, CitiBike could monitor the data real-time as it streams in from the city system to monitor. All of these requirements are possible with a Hadoop system, and that makes it a perfect fit.

It should be noted that for the purpose of this project, I was using a personal virtual machine on one computer. Thus, I did not have the computing power that would be required to process the millions of trips supplied by the datasets from CitiBike. Instead, I took a sample of three thousand trips from one single day to reduce the processing requirements to match those of my single computer. In a real application, this data would be processed in parallel over multiple nodes. To determine the amount of storage needed, we determine the amount to be stored and multiply it by four – three times for backup storage space and once more for extra space. Thus, if we needed to store 1TB per year, we'd need 4TB of storage space in our Hadoop cluster system.

Results and Insights

With CitiBike, simply learning about the data using Exploratory Data Analysis methods can produce valuable insights on the benefits of using Big Data. As mentioned, the emphasis on this study is to discover information that can be used to improve regular operations of CitiBike and/or to enhance the marketing campaigns designed to take advantage of opportunities provided by CitiBike.

In terms of return on investment opportunities, reducing operational costs by understanding the customer and how the customer uses CitiBike is an important technique to utilize. For example, we've learned that almost ninety percent of the CitiBike users are men. This type of information could be used when ordering different size bikes. Since men are typically physically taller and larger than females, making more purchases of larger sized bikes rather than smaller ones would be one way to maximize the likelihood that a bike is being used in the system. It also improves the chances that there are enough of the correctly sized bikes for the number of customers.

For the future, the same information is extremely important to understand if CitiBike were to, say, decide to rent safety equipment, such as helmets and bike gloves, at their stations. Since they would know that approximately ninety percent of the helmets and

bike gloves would need to accommodate men, they could save money in their approach to the project. By reducing frivolous expenditures, it can help reduce operational costs, improve customer satisfaction, and increase CitiBike's return on investment by simply understanding this one insight – and there are many more too!

For another example, insights about type of customer can be useful as well. Are CitiBike users usually annual membership subscribers? What percent of customers only have the 7-day or 24 hour pass? Percentages are great to know but on an average day, approximately how many customers are using CitiBike?

While knowing what type of membership a CitiBike customer is using may sound trivial to some people, for marketers this can be extremely valuable information. If the majority of customers are using the annual membership subscription then, most likely, they are a resident of the city. City residents will be much more concerned and interested in local events, activities, and local news. Marketers can and should understand this and can tailor their messages accordingly.

An advertisement offering coupons to local mom and pop stores and restaurants might be more appealing to this group compared to customers who purchases 7-day or 24-hour subscriptions. These shorter subscriptions generally indicate shorter stays in the city and probably include a much higher percentage of tourists. Here tourist related advertisements and marketing campaigns would probably be more effective. By drilling down even deeper into the data, we can determine which CitiBike stations are used the most often by these customers (tourists) to start or end their CitiBike trips. In this case we could continue to maximize our marketing value by focusing more tourist related advertisements at those stations.

Another important insight from this data is how long do customers ride a CitiBike on average and what is the longest trip taken thus far? The longest time on a CitiBike is just over three hours, but this is uncommon as the average trip time is between eleven and twelve minutes. For marketers, this is important to know because it provides them with the amount of time they have to capture the customer's attention, especially if they are actually advertising on the bikes as well. Ads can be tailor to these time limits and advertising media can be creatively adjusted to take advantage of advertising opportunities on the actual bicycles. For the CitiBike operators, this information is important because it allows them to understand how much time is being put on each bike each day. Again, once we find an insight, the value of Big Data is that we can usually drill down even further and gain an even deeper understanding of it. In this case we can determine the average of how many times a bike is used per day and how long it's used on each trip. The operator can then create a predictive model to determine when the bike will most likely need routine service or need new parts like new tires or a chain. This can help predict operation maintenance expenses.

Furthermore, a Big Data system could be created at any of CitiBike's stations that include a maintenance facility as well. The total use time for each bicycle could be tracked, stored, and updated regularly at the end of the day when the current day's trip activity

data is downloaded. Threshold limits or maintenance “digital flags” could be established to trigger notices to conduct certain maintenance tasks. As a bike enters a CitiBike station with a maintenance facility, if any of its usage limits are met, then it is removed from service until the maintenance task is performed. This type of well-organized maintenance program based on Big Data capabilities can improve the longevity of bicycles, improve customer satisfaction rates, improve service, and save costs in the long-term.

Deeper insights could be obtained by combining the Big Data system for trip data with a CitiBike maintenance log system. These insights could include how long to expect certain bike parts to last, when will most bikes in the system need new tires, and how to take advantage of these predicted events by adjusting inventory, financial, and staffing levels accordingly. When we can see the changes or stresses on the operational system beforehand, we can make the appropriate adjustments to compensate accordingly or make changes to take advantage of them.

If they haven’t created maintenance facilities yet, by knowing which stations are the most popular end points by customers, it can be used to determine where to strategically locate a maintenance facility. If the system of performing maintenance is based on storing and following an individual bike’s usage time is implemented, then it’s best to have a maintenance facility located where it has the best chance of “catching” each bike before the bike spends too much time over its usage limit.

However, more than just knowing which station is a popular end point, it is important to know about the area as well. CitiBike can use the latitude and longitude coordinates for each trip to understand the area where its customers operate. By taking the maximum and minimum latitude and longitude coordinates, we can create a four quadrant grid that represents the entire area in use by CitiBikes. Each quadrant can be examined in more detail to understand the activities in each quadrant compared to other quadrants and to understand the flow traffic between quadrants.

When analyzing the travel area quadrants of the first eight hours of August 1st, 2014 we can obtain a larger number of insights. First, we see that it is almost evenly distributed as to which quadrant has more CitiBike users who start their trip in that quadrant. Approximately twenty-four percent leave from quadrants one (northwest city quadrant) and another twenty-four percent from quadrant four (southeast city quadrant). It’s about twenty-six percent respectively for quadrant two (northeast city quadrant) and quadrant three (southwest city quadrant). With all things being equal, we should see an even distribution between quadrants as we do.

Interestingly though, when we run the same analysis on the end stations, we see a significant percentage difference between quadrants. For example, almost thirty-seven percent of users stop their trip within quadrant two (northeast sector), twenty-eight percent stop their trip in quadrant three (southwest sector), while only 20% and 15% stop their trips in either quadrant four (southeast) or quadrant one (northwest) respectively. These figures indicate a migration from an evenly distributed amount of bikers from each quadrant to the southwest and northeast sections of town.

Why is this and can we take advantage of this insight? We would have to do more investigating or collecting more data to understand what might explain this bike migration. Are there more commercial, entertainment, or residential operations in one quadrant compared to the other? Does time of day, month, or year make any difference when we see this travel migration?

If this pattern was over an entire day of trips, this insight suggests that more bikes leave but not returned to quadrants one and four each day. The issue here is that there always needs to be a minimum number of bikes in each quadrant/station to begin the day. If one or more quadrants are losing bikes to migration trends (as previously discussed), then eventually the minimum number of bikes to begin the day in that quadrant will be reached or lower. This is important because now CitiBike can react accordingly by learning which quadrants finish with a surplus of bikes compared to where they started the day. Then they can implement a CitiBike repopulation or repositioning plan to mitigate these inevitable migration patterns identified through Big Data systems and technologies. We can even learn exactly how many bikes will be available for repositioning and, therefore, determine the logistics and costs to reposition the bikes before implementing the idea.

Continuing with the quadrants, we can identify the most often used start stations and end stations within each quadrant. This allows the marketers to get even more detail on where to advertise to maximize their return. For example, if you are advertising drinks for sale, you may find better results advertising or locating in stations where more users complete their trip and are thirsty. Or maybe it would be better to advertise at the stations where most of the long duration rides finish, and therefore has customers who have expended more energy and may be thirstier. The same is true for products often needed to begin a bike ride. Those products may find better profits by advertising or offering their products at a station where more users begin trips. Small insights like these are what allow the company to begin to save 10% here or there, and then even more savings in operational efficiency as it gets better at using insights.

Marketers love to understand the makeup of their customers in terms of age or age group. Big Data allows CitiBike to determine age brackets of their users divided by ages 10-19, 20-29, 30-39, 40-49, 50-59, and 60-69.

As expected, almost half of the CitiBike users are in two age groups, the 20-29 (20%) and 30-39 (27%) year old group. Of the customers aged 40-49, they made up 17% of the total users and 11% of customers were aged 50-59. Of those 20-29 years old, 80% are men. 84% are men in the 30 – 39 year old age group. Men also make up 87% and 86% of the population respectively in the 40-49 and 50-59 year old age ranges. Furthermore, we can see that virtually two thirds of the CitiBike subscribers are between 20 and 50 years old with over 80% being men. Advertising strategies, based on these numbers, should have more focus on products and services targeted at men aged 20 to 50 years old. What about the females? Well, if any advertising was directed towards them it should be focused on

goods and services aimed at females aged 20-39 only, as there are too few females in other brackets to earn a good return on the advertising expenditure.

Furthermore, knowing that the majority of users are men, if the company wanted to expand to improve profits or for any other reason, one great area for improvement would be to understand why so few women, percentagewise, don't use CitiBike service compared to men. This could be accomplished through the use of surveys and other means for getting customer feedback. If the reason is identified and can be solved easily, a huge opportunity could be available to get more women bike riders which means more subscriptions and more demographic variety to offer to potential advertisers.

Finally, with Big Data, we can break down our questions even further into what happened on specific days, days of the week, and time of day. For example, we asked how many trips began in each of the first 8 hours of the day. We learned that almost 400 trips started at midnight and that dropped to only 53 trips starting between 3 and 4 am. At 5 AM the number jumps to 208 and then increases very quickly to 823 at 6 AM to almost 1200 at 7 AM. We can compare data from any time of the day, to any particular day of the month, or weekdays compared to weekends, etc. All of which could be used by marketing to build advertising campaigns around or to offer campaigns to potential advertisers.

Summary of Results

Some of the main results from our findings are very helpful for understanding the CitiBike customer base. We saw that almost nine out of ten customers are men. Virtually two thirds of all users are between 20 and 49 years old. There is a migration to the stations in the northeast and southwest quadrants throughout the day. Most trips are about twelve minutes long. And we learned the station names where most tourists finish their trips, amongst other interesting insights.

We learned there is some very clear geo-spatial information combined with customer information that can be used to answer important operational and marketing questions. By identifying common customer traits/actions associated with station locations or general areas of operation, CitiBike can really maximize how they use their resources and plan their marketing strategies. This improves operational efficiency, marketing strategy, customer service, and ultimately improves profit margin.

Understanding that CitiBike records approximately a million bike usage transactions a month and growing, storage space and processing requirements continue to grow and that's why Hadoop, with its scalable nodes and MapRequest processing, is a fitting selection.

Implementing Hadoop

Certainly there are infrastructure costs associated with transitioning from traditional h to Big Data technology, and open-source technology with commodity hardware (Hadoop)

helps alleviate some of those costs; however, most of the costs will be external of infrastructure.

First, there needs to be cultural shift in the company to optimizing business decisions through quantitative measures (Big Data analytics). Finding the right people who can think creatively to maximize the use of Big Data, operate, manage, and interpret the Big Data results within the existing ecosystem is where most of the real costs are incurred. Also, while Hadoop is a low-costs open-source Big Data platform that can be used on commodity hardware, it does have two major ease-of-adoption and cost challenges: how to leverage existing SQL and existing BI tools with data in Hadoop, and the ability to compress data at the most granular level which will reduce storage requirements and drive down the number of nodes and simplify the infrastructure (Savitz, 2012).

The benefits of leveraging Big Data will outweigh any IT investments into infrastructure and employee training that is associated with its implementation. The main question is by how much will the investments costs outweigh those for training, maintaining, analyzing, and understanding the new source of Big Data.

For CitiBike, we will be using a large variety of data, the possibility of having data that is constantly streaming, and data that can use geo-spatial coordinates to extract data. Basically, it will be sensor and geo-spatial data that is a large volume considering the company has tens of thousands of trips each day.

As a platform, I would recommend the Hadoop platform for a couple of reasons. First, it has scalability and the ability to handle the amount of volume and variety of data the CitiBike will be utilizing. Second, its ability to use commodity hardware greatly decreases the implementation costs. MapReduce will also be used on top of Hadoop's HDFS because it integrates so well with Hadoop and an effective job scheduling and execution program is absolutely necessary when handling multiple tasks simultaneously.

I'd also recommend Zookeeper to keep things organized. It's an open-source data warehousing solutions and therefore keeps the implementation costs to a minimum. Zookeeper is essentially a coordinator of tasks between nodes. Thus, it's an important component to handle the larger data volumes being collected from CitiBike.

Finally, Hive is important because it allows us to use SQL language with MapReduce to process low-latency queries and deep analytical processes that are not easily handled by Hadoop. PIG can be used as well because it is another way to access the HDFS without having to learn to write tedious and complicated MapReduce programs in Java. With either Hive or PIG, we use a much easier language to learn which converts our statements into MapReduce statements that are usable by the Hadoop system. This reduces the learning and training curve for staffers when implementing the system. Combined, each of these Big Data technologies should be extremely helpful for the CitiBike to transition from traditional databases to reap the benefits of Big Data technology.

Once we have the system in place, we have the programming tools to extract data, and we can collect, store, process, analyze, and find meaningful relationships and insights in the data, we need to understand how to deploy this information into the system so it can be used properly. The information we have demonstrated being extracted in this project is mainly useful for operation and marketing divisions. Thus, decisions need to be made based on how frequently and/or how much data needs to be analyzed to make timely decisions.

For operations, they may only be interested in long term analysis so it's important to have a solid, clean historical archive system set up for any non-current data. Does this system need to be updated daily, monthly, or weekly? What is considered archival data? Another interesting phenomena with Big Data is that answers often lead to more questions, which means more data, or the need to gather data in a different way. Is the system designed to handle different formats of data? If so, will it be compatible with current formats and accessible so we may use each at once?

Marketing has similar needs but the time scales may be different or vary from campaign to campaign. One campaign may only run a week and therefore we may need almost instant real-time data to analyze the campaign results and make appropriate adjustments to the campaign. Quick, agile adjustments, by having and recognizing the performance deficiencies of the current campaign can make or break a campaign. The same is true when exploring the data to create new marketing campaigns; we want to use the latest and most accurate data possible to make decisions. Thus, real-time data processing may be helpful for the marketing department at CitiBike.

Hadoop with its HDFS, ability to handle structured and unstructured data, MapReduce, scalable nodes, and easy learning programming languages will allow CitiBike to handle such requirements.

Conclusion

It's clear that CitiBike can benefit a lot from Big Data and the implementation of Hadoop with data technology. For CitiBike's situation, it is relatively straightforward as to how the system would be implemented and, while even greater insights could be found using other analytics such as Mahout and machine learning, for now and for the foreseeable future CitiBike should be able to extract plenty of valuable insights from the effective Hadoop related programming languages of Hive and PIG. It is also important to note how CitiBike can improve its system in the future by incorporating the ability to process data real time. This too can be accommodated by Hadoop.

We have seen how Hadoop, by using Hue, can produce some visuals as well which can augment or emphasize certain data points or insights found by CitiBike. By implementing a system that gathers data regularly, processes it on a consistent basis, and is updated frequently, we can produce a data system with Hadoop that is designed to produce specific visualizations and the hard data to help upper management in operational and marketing divisions at CitiBike to make more effective decisions. This accomplishes

CitiBike's goal to improve its operational procedures, expand its marketing capabilities, increase customer satisfaction, and improve its profit margin. Finally, this is all accomplished with open-source technology using commodity hardware. Which means it is accomplished with minimal expense, plenty of support, and unlimited scalability – all are things that almost every company cherishes.

References:

Janssen, C. (n.d.). *NoSQL*. Retrieved October 2, 2014, from Techopedia website:

<http://www.techopedia.com/definition/27689/nosql-database>

Krishnan. (2013). *Data Warehousing in the Age of Big Data*. Waltham, MA: Morgan Kaufmann.

Savitz, E. (2012). *The Big Cost of Big Data*. Retrieved October 2, 2014, from

Forbes website: <http://www.forbes.com/sites/ciocentral/2012/04/16/the-big-cost-of-big-data/>

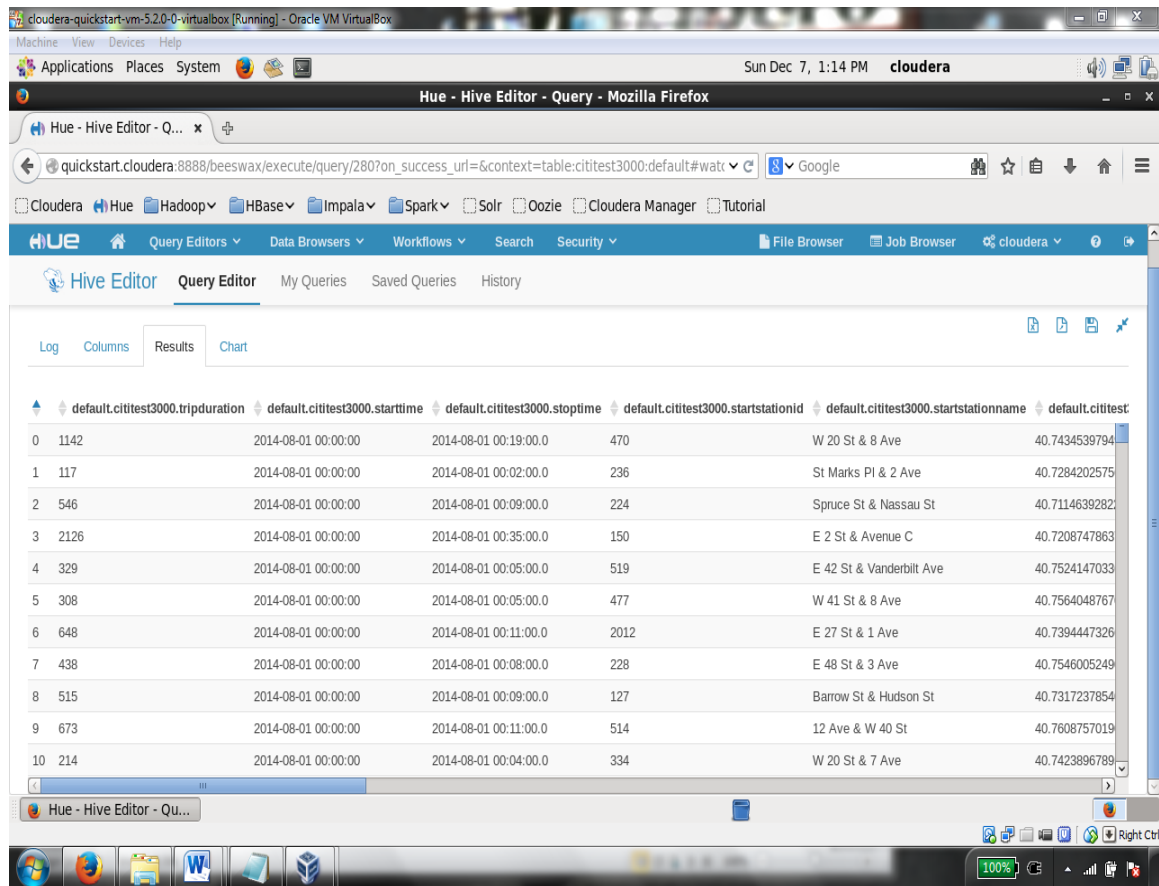
System Data (n.d.). Retrieved December 3, 2014, from CitiBike NYC website:

<http://www.CitiBikenyc.com/system-data>

Appendix:

Displaying every query and visual graph discussed in this analysis would create a very large number of additional pages. Thus, for brevity, only a select few graphs, codes, and tables have been selected for display to keep the results close to the assignments recommended length (15 pages).

Hive Table:



The screenshot shows the Hue Hive Editor interface within a Mozilla Firefox browser. The interface includes a top navigation bar with various tools like Query Editors, Data Browsers, Workflows, Search, Security, File Browser, and Job Browser. The main content area displays a table of test results with columns for trip duration, start time, stop time, start station ID, start station name, and test ID. The table contains 11 rows of data, each representing a different test run.

	default.cititest3000.tripduration	default.cititest3000.starttime	default.cititest3000.stoptime	default.cititest3000.startstationid	default.cititest3000.startstationname	default.cititest
0	1142	2014-08-01 00:00:00	2014-08-01 00:19:00.0	470	W 20 St & 8 Ave	40.7434539794
1	117	2014-08-01 00:00:00	2014-08-01 00:02:00.0	236	St Marks Pl & 2 Ave	40.7284202575
2	546	2014-08-01 00:00:00	2014-08-01 00:09:00.0	224	Spruce St & Nassau St	40.7114639282
3	2126	2014-08-01 00:00:00	2014-08-01 00:35:00.0	150	E 2 St & Avenue C	40.7208747863
4	329	2014-08-01 00:00:00	2014-08-01 00:05:00.0	519	E 42 St & Vanderbilt Ave	40.7524147033
5	308	2014-08-01 00:00:00	2014-08-01 00:05:00.0	477	W 41 St & 8 Ave	40.7564048767
6	648	2014-08-01 00:00:00	2014-08-01 00:11:00.0	2012	E 27 St & 1 Ave	40.7394447326
7	438	2014-08-01 00:00:00	2014-08-01 00:08:00.0	228	E 48 St & 3 Ave	40.7546005249
8	515	2014-08-01 00:00:00	2014-08-01 00:09:00.0	127	Barrow St & Hudson St	40.7317237854
9	673	2014-08-01 00:00:00	2014-08-01 00:11:00.0	514	12 Ave & W 40 St	40.7608757019
10	214	2014-08-01 00:00:00	2014-08-01 00:04:00.0	334	W 20 St & 7 Ave	40.7423896789

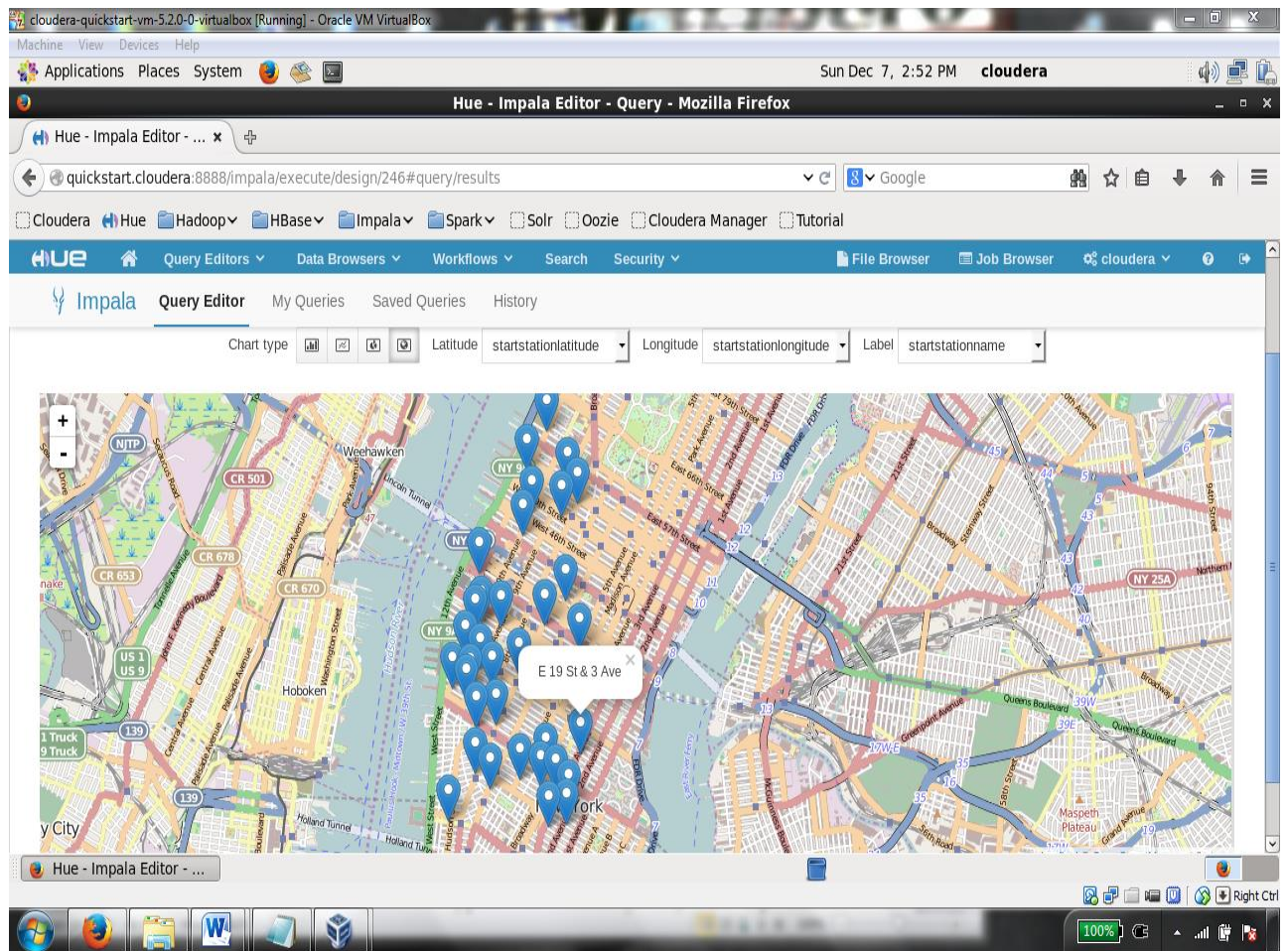
Quadrant breakdown- start of day = green, end of day = black:

Northwest quadrant (#1) <u>Start of Trip</u> Total – 720 users Total % = 24% Male- 620 – 96% Female – 25 – 4% Unknown – 75 <u>End of Trip</u> Total – 453 Total % - 15% (- 9%) Male – 353 – 95% Female – 19 – 5% Unknown - 61	Northeast quadrant (#2) <u>Start of Trip</u> Total – 778 users Total % = 26% Male- 648 – 94% Female – 44 – 6% Unknown – 86 <u>End of Trip</u> Total – 1104 Total % - 37% (+ 11%) Male – 932 – 95% Female – 48 – 5% Unknown - 124
Southwest quadrant (#3) <u>Start of Trip</u> Total – 770 users Total % = 26% Male- 583 – 95% Female – 30 – 5% Unknown – 157 <u>End of Trip</u> Total – 845 Total % - 28% (+ 2%) Male – 641 – 95% Female – 37 – 5% Unknown - 167	Southeast quadrant (#4) <u>Start of Trip</u> Total – 717 users Total % = 24% Male- 512 – 92% Female –46 – 8% Unknown – 159 <u>End of Trip</u> Total – 591 Total % - 20% (- 4%) Male – 423 – 91% Female – 44 – 9% Unknown - 124

Sample Hive query: To find how many trips begin in quadrant 1 using latitude and longitude coordinates.

```
SELECT startstationid, startstationlatitude, startstationlongitude, COUNT(*)  
FROM cititest3000  
WHERE startstationlatitude > 40.72593221 AND  
startstationlongitude < -73.98359122  
GROUP BY startstationid, startstationlatitude, startstationlongitude;
```

Sample display using geo-spatial features in Hadoop and Hive to create an actual map that shows the location where every trip started that began in this quadrant.



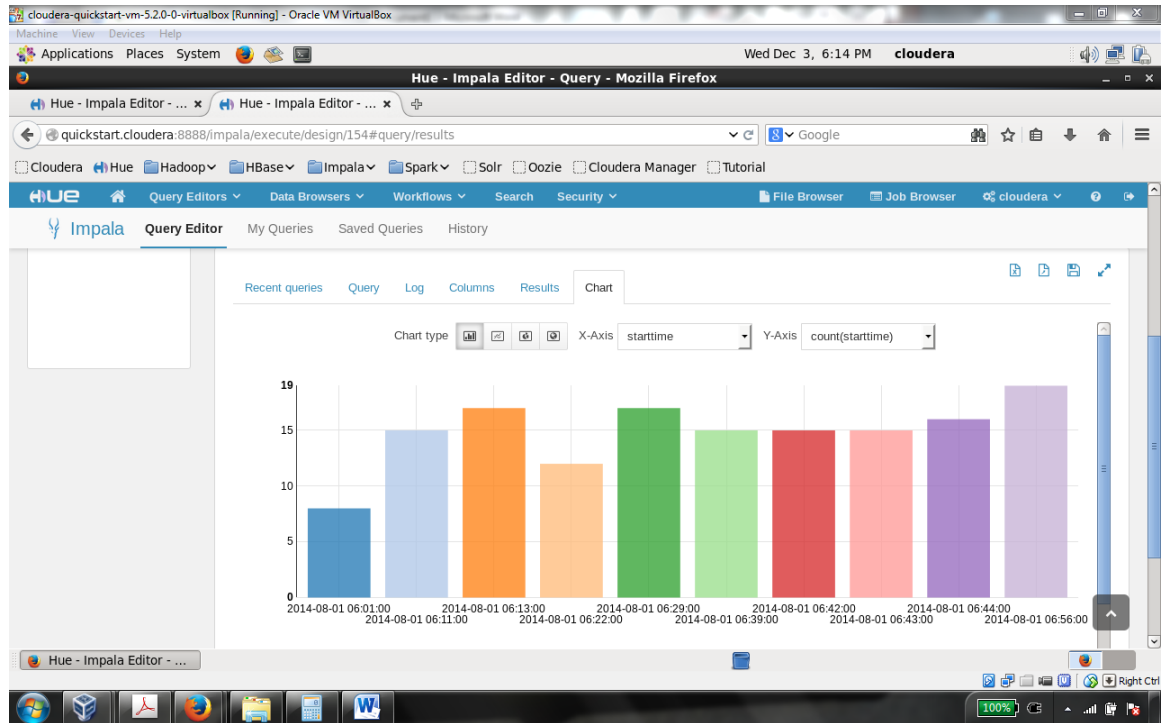
Sample demonstrating how a HIVE query can extract insights from the data.

Here we are trying to find the exact times when most users began a trip between the hours of 6am and 7am. Hive allows us to make the insights quite detailed and specific if necessary.

The screenshot shows the Hue Hive Editor interface in a Mozilla Firefox browser window. The browser address bar shows the URL: `quickstart.cloudera:8888/beeswax/execute/query/282#query/results`. The interface includes a top navigation bar with various tools like Query Editors, Data Browsers, Workflows, Search, Security, File Browser, and Job Browser. The main area is titled "Hive Editor" and contains a "Query Editor" tab. The query being executed is: `1 SELECT starttime, COUNT(starttime) FROM cititest3000 WHERE starttime LIKE '%2014-08-01 06%' GROUP BY starttime;`. Below the query editor, there are buttons for "Execute", "Save", "Save as...", "Explain", and "New query". The "Results" tab is active, displaying a table with the following data:

	starttime	_c1
47	2014-08-01 06:47:00	23
38	2014-08-01 06:38:00	21
46	2014-08-01 06:46:00	21
50	2014-08-01 06:50:00	21

Sample graph of when most trips begin between 6am and 7am – shown in chronological order over time:



Sample display of results using HIVE and IMPALA:

Results from the stations where most Customer users (those who only purchase 24 hour or 7 day access) end their trips:

Hive version:

The screenshot shows the Hue Hive Editor interface in a Mozilla Firefox browser. The query editor contains the following SQL query:

```
1: SELECT endstationname, COUNT(endstationid) FROM cititest3000 WHERE usertype LIKE 'Customer' GROUP BY endstationname;
```

The results tab displays the following data:

endstationname	count(endstationid)
Broadway & W 60 St	5
E 7 St & Avenue A	5
Monroe St & Classon Ave	5

Impala version:

The screenshot shows the Hue Impala Editor interface in a Mozilla Firefox browser. The query editor contains the following SQL query:

```
1: SELECT endstationname, COUNT(endstationid) FROM cititest3000 WHERE usertype LIKE 'Customer' GROUP BY endstationname;
```

The results tab displays the following data:

endstationname	count(endstationid)
E 7 St & Avenue A	5
Broadway & W 60 St	5
Monroe St & Classon Ave	5
Vesey Pl & River Terrace	4

Finally, another example of a visual aid using Hadoop....here we see the results of the amount and type of gender using CitiBike who ended their trip in quadrant 2 (the northeast city sector).

