Big Data a Big Win for Political Campaigns

Chanah Stern, Northwestern University

***Abstract*— This paper explores the meaning of Big Data and examines the differences between Big Data and traditional data. It explores the components of a Big Data ecosystem and, more specifically, multiple components that comprise the Hadoop ecosystem. Additionally, it demonstrates an applied Big Data use case and reviews the resulting analyses generated from Hadoop that are visualized in Tableau. The use case leverages data from the 2016 US presidential primary election results augmented with county-by-county census data. This paper highlights the benefits and value that Big Data provides for political campaigns.**

***Index Terms*— Big Data, Hadoop, HDFS, Hive, MapReduce, Tableau, political campaigns**

I. INTRODUCTION

POLITICAL campaigns in the modern era must rely on Big Data and advanced analytical techniques in order to win. Barack Obama's 2012 campaign developed a Big Data platform and utilized the data to mobilize the vote and turn the election in the favor of Mr. Obama. The campaign developed sophisticated models within their Big Data platform to predict most likely voters and tailored their campaign to target those individuals. (CSC, n.d.) The investment paid off and subsequently the Republicans and other smaller scale campaigns have recognized that in order to be competitive, they must factor Big Data and analytics into their campaign strategy.

This paper explores the possibility of gaining strategic insight from Big Data solutions that political parties can leverage during the course of their campaigns. The data utilized was downloaded from the data science website, Kaggle, and is based on US 2016 presidential primary election results augmented with county specific information. The paper demonstrates how a non-partisan campaign consulting agency could leverage Big Data to guide a campaign strategy for increasing voter turnout and converting votes in their favor.

II. BIG DATA OVERVIEW

Big Data is a term used to describe the very large volumes of data generated by our increasingly digital universe. Big Data offers organizations great potential to mine for intelligence that cannot be analyzed using traditional computing techniques. The amount of data that is being generated has grown exponentially in recent years, with industries capturing more data than they did in the past from an assortment of channels.

The size, variety, structure, and speed at which Big Data is generated and streamed into systems distinguishes it from traditional data. Big Data allows for more predictive analytics as compared to the reactive analytics that traditional data provides (SAS, n.d.). Traditional data includes structured data such as documents, financial transactions, stock records, and personnel files. Big Data includes unstructured data such as photos, audio and video, three dimensional models, simulations, and location data, all which can be analyzed. Traditional data is often delivered in batches while Big Data arrives in endless streams that need to be processed immediately when it is most valuable.

Big Data allows for companies to learn more about their customers, partners, and businesses as they gain insight from an increasing number of sources. Big Data has exploded in recent years with our ever increasing interconnected world. Retailers track every customer click on websites giving them the ability to provide customized shopping experiences and tailored promotions. Radio-frequency identification (RFID) tags on products and even groceries provide a wealth of information about the status and location of goods enabling manufacturers to anticipate shortages, defects, and problems in supply chain. Medical devices can report health status back to physicians in real time without a patient having to step into a doctor's office. Social media sites collect large amounts of data that can reveal detailed facts about customers, communities, and trending movements that all types of enterprises, including governments, can capitalize on. There has been a proliferation of streaming and still images from which computers can process and extract meaningful information. Location information is reported from millions of smart phones that can be used to detect traffic patterns and suggest optimal driving routes to drivers. The 'Internet of Things', a network of millions of sensors that are attached to everyday objects, generates streams of data that businesses and individuals can utilize for smart and automated decision making such as lawn sprinklers determining when to water the grass based on rainfall amount, temperature, weather forecast, sunlight, and other data.

The introduction of cloud computing and cluster architecture which provide companies affordable solutions for some of the challenges that Big Data introduces such as storage, complex processing, scalability, and reliability also have contributed to the expansion of Big Data by making the vast riches contained

C. Stern is a graduate student in the Masters of Predictive Analytics (MSPA) program at Northwestern University. chanahstern2017@northwestern.edu

This paper was prepared for an assignment due on August 28, 2016 for CIS 436, Section 55.

within the data more easily available. Even companies that have not yet begun to invest in a more comprehensive Big Data solution such as Hadoop, are able to harness some of the power of Big Data through the cloud (Florissi, 2012).

III. Big Data Technologies Differ from their Traditional Predecessors

Traditional relational databases cannot handle the quantities of data that are being generated and received, aren't able to process and classify unstructured data, and cannot apply analytics quickly enough before storing data. Many organizations simply cannot capture much of the data that is available and have to ignore precious information (Zikopoulos, 2012). Furthermore, traditional database related technologies mandate that the data they will process must be well described prior to loading data. This requirement slows organizations down as they cannot ingest rapidly changing data. Furthermore, traditional technologies tend to scale vertically which is very costly and necessitates significant downtime when platforms are expanded. Big Data solutions, on the other hand, tend to scale horizontally allowing for additional hardware to be added to the platform at a lower cost and without requiring the existing platform to be offline as the additional hardware is added to the platform.

IV. BIG DATA ECOSYSTEM

As implied in the very name of 'Big Data', data is at the heart of the Big Data ecosystem. Any Big Data ecosystem will offer multiple components available to the consumer for storing, processing, accessing, and finally presenting the data. In the next section we will take a deep look at Hadoop, one of the most commonly used Big Data ecosystems in use today, and examine the Hadoop solutions available for each phase of the Big Data lifecycle.

V. HADOOP ECOSYSTEM

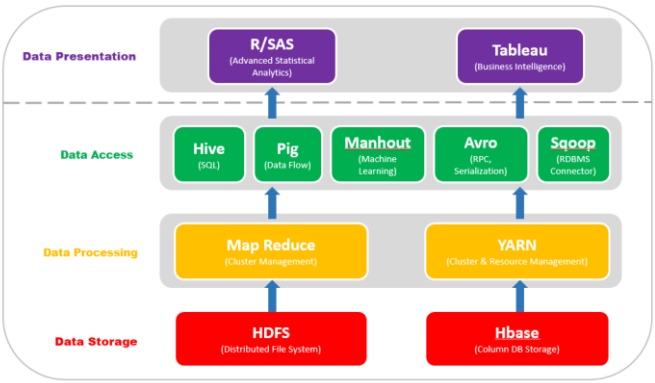
Hadoop is an open source affordable technology that utilizes cluster architecture to tackle the problems of massive amounts of data and the need for quick processing. Hadoop provides the ability to easily scale as a company's data needs expand and is very robust with built in recovery from several possible failures.

Hadoop is comprised of two key components: the Hadoop Distributed File System (HDFS) and the MapReduce mechanism for rapid processing of large amounts of data. Hadoop splits, scatters, and replicates files across hundreds to thousands of low cost nodes in a cluster. Data is replicated at least three times throughout the cluster and all files are equal in size so that when there is a failure, replacing hardware and restoring data from one of the copies is fast and inexpensive.

The MapReduce mechanism ensures application performance by taking advantage of data locality and parallel processing. Each task that is executed in Hadoop has two sequential phases: the mapper phase and the reduce phase. The mapper phase assigns a computation that should be performed to all relevant nodes. Each node that was assigned a computation task will execute the job in parallel on the data that is co-located with it, greatly decreasing the amount of information that needs to be passed across a network and the time it takes to process an entire computation. Each node that performs computations in the mapper phase produces an intermediate result list with key/value pairs and sometimes sorts or aggregates results before distribution to the reduce phase. The reduce phase will aggregate and sort output from the mapper phase for a final result.

A job tracker manages and coordinates the MapReduce process. The job tracker is a node that assigns tasks, keeps record of which nodes participate in a job, and reschedules jobs when a failure is detected. Programmers invoke MapReduce by providing a location of the data to be included, a map function that will process data and generate intermediate key/value pairs, and a reduce function that merges the output from the map function. The entire MapReduce process will return a list of key/value pairs as a result.

Figure 1: *Hadoop Ecosystem*



There are many components within the Hadoop ecosystem that add capabilities above and beyond HDFS and MapReduce as seen in the **figure 1** above. We will examine several of these components in detail.

On the **data storage** level, **HBase** is a database management system that runs on top of HDFS and integrates with MapReduce. The standard Hadoop HDFS is optimized for streaming access of large files, doesn't support a variable data schema very well, and its files are write once only files. HBase, however, can do quick lookups on small amounts of data that reside in a larger data set, supports a flexible data model, and allows users to modify data.

On the **data processing** level, **YARN** is essentially version 2 of MapReduce. YARN enables Hadoop to add flexibility in its processing approach since it decoupled MapReduce's resource management and scheduling from Hadoop data processing. As a result, it enables interactive querying in Hadoop simultaneous with MapReduce jobs which it was unable to do in the earlier implementations of MapReduce.

On the **data access** level, **Hive** is a SQL like querying language that enables users who are familiar with SQL to more easily query data stored in HDFS. A Hive query is translated to MapReduce jobs under the covers. Hive does have some limitations, most notably latency. If consumers need real time analytics of streaming data, Hive would not be an optimal solution. Furthermore, if an organization has an HBase

implementation, using Hive might negate one of the advantages of HBase which is its ability to perform quick lookups! Like Hive, **Pig** is a language that is intended to decrease the time it takes for programmers to develop MapReduce routines by providing an intuitive and concise language. However, it is not as similar to SQL as Hive so there is a steeper learning curve for experienced SQL programmers. **Sqoop** which is short for 'SQL plus Hadoop' is a tool that enables easy transfer of data between relational databases and Hadoop and can be used to load data into Hive or Hbase. Sqoop can be implemented through saved jobs that are run repeatedly to load incremental updates to data or it can be implemented as a one-time load based on a free form SQL query. **Mahout** provides scalable machine learning algorithms which is heavily used in the area of recommendation engines.

Finally, for the **presentation** of data which is key to delivering value from the data that is stored and processed in Hadoop, there are many components that don't necessarily reside within the Hadoop ecosystem but can integrate with Hadoop to provide a seamless presentation layer. For example, there are R and SAS connectors that data analysts can take advantage of for advanced statistical analytics. Business Intelligence experts can utilize Tableau for visualizing and presentation of data in Hadoop.

VI. DESCRIPTION OF DATA LEVERAGED IN POLITICAL CAMPAIGN USE CASE

There are two main data sources used for this exercise. The first, primary_results, is a county by county listing of the 2016 US presidential election primary results. The entire set of attributes for all 24,612 records is loaded to Hadoop, but only the elements from the primary_results data source that are leveraged in this paper are listed in **Table 1**.

Table 1 - *Elements in **primary_results** utilized in this paper*

Element	Description	Possible Values
state	State where the primary or caucus was held	Any one of the 50 US states
state_abbr	Two letter state abbreviation	Any one of the two letter abbreviations of the 50 US states
fips	Federal Information Processing Standard (FIPS) county code for the county	Valid FIPS code. The reader may go to the US Census Bureau website to search for FIPS codes.
party	Party for which the results are being reported	Democrat or Republican
votes	Number of votes the candidate received in the county being reported on	Whole number, 0 or greater

The second data source, county_facts, relates to primary_results via state abbreviations and fips codes and provides valuable county by county data that will augment the primary results data. The county data will enable the campaign consulting agency to assist the parties in tailoring their campaigns by analyzing the primary results as compared to demographics of the counties. Once again, the entire set of attributes for all 3,196 records is loaded to Hadoop and available for use. However, only the elements from the county_facts data source that are leveraged in this paper are listed in **Table 2.**

Table 2 - *Elements in **county_facts** utilized in this paper*

Element	Description	Possible Values
state	State where the primary or caucus was held	Any one of the 50 US states
state_abbr	Two letter state abbreviation	Any one of the two letter abbreviations of the 50 US states
area_name	County where the results originated	Valid US county. The reader may go to the US Census Bureau website to search for counties.
fips	Federal Information Processing Standard (FIPS) county code for the county	Valid FIPS code. The reader may go to the US Census Bureau website to search for FIPS codes.
PST045214	2014 population estimate for the county	Whole number, 0 or greater
AGE775214	Percent of 2014 population estimate age 65 or older	Decimal value from 0.0 to 100.0
RHI725214	Percent of 2014 population estimate that his Hispanic or Latino	Decimal value from 0.0 to 100.0

VII. DATA PREPARATION

The data and accompanying data dictionary were downloaded from Kaggle to a local drive. The two data sources files, primary_results.csv and county_facts.csv, were then transferred to a folder on the Cloudera File System (**figure 2**). From there, Cloudera's Hue user friendly interface was used to load the data into Cloudera so that it would be available in the Hadoop HDFS. The 'Upload File' utility was used to upload both sources. **Figure 3** shows both files uploaded to Cloudera. Once the files were uploaded to Cloudera, Hue's Metastore Manager was used as it provides a very intuitive interface for creating tables from a file. A similar process was followed for both data sources. **Figure 4** shows the first step where the file is selected and a table name is provided. **Figures 5 and 6** demonstrate selecting a

delimiter for the input file and specifying column names, respectively. For these data sources the defaults were selected (comma delimited, column titles, and data types from input files). Finally, in **Figures 7,8, and 9**, we can see that the tables were created and that the expected columns exist within the tables.

End to end visual demonstration of loading the data to Hadoop

Figure 2: *Files transferred to Cloudera file system*

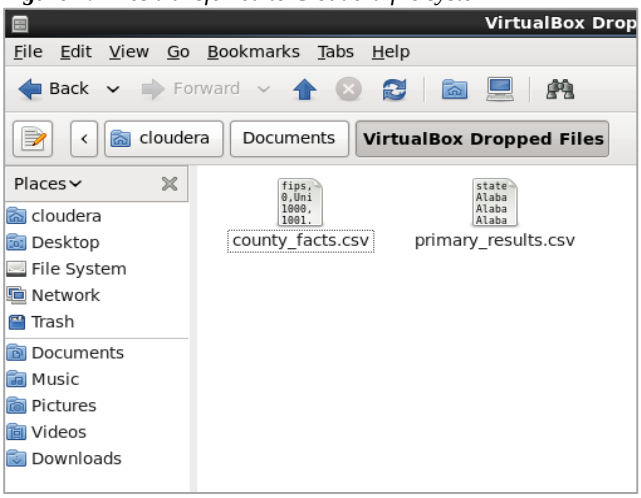


Figure 3: *Data source files uploaded to Cloudera HDFS*

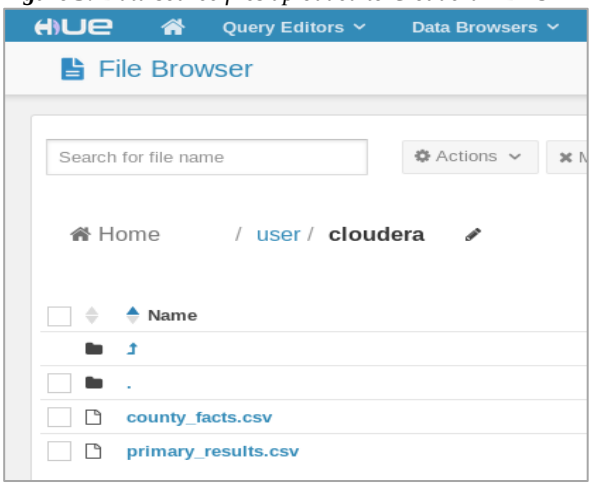


Figure 4: *Select a file to create a table*

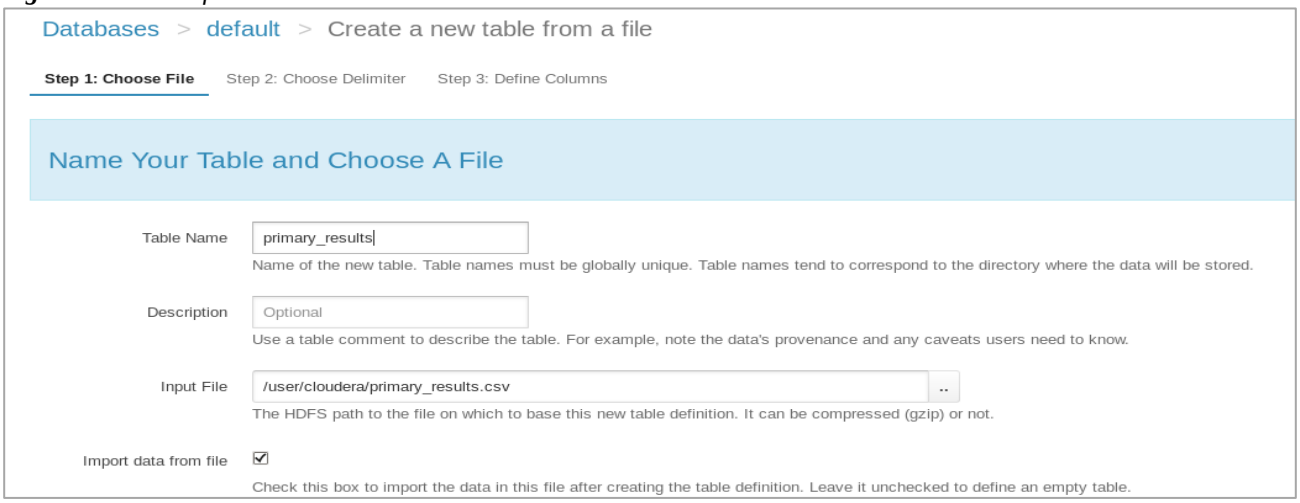


Figure 5: *Select a delimiter for the input file*

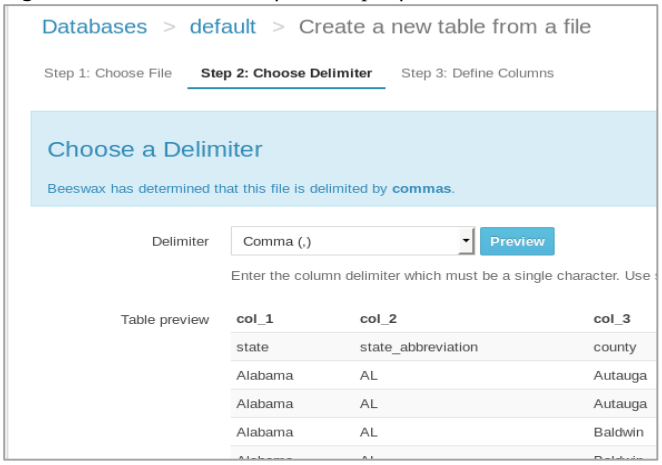


Figure 6: *Select column titles and data types*

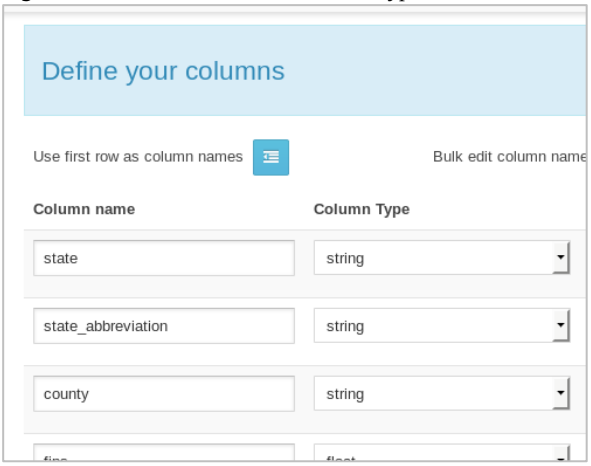


Figure 7: *Tables for each data source have been created*

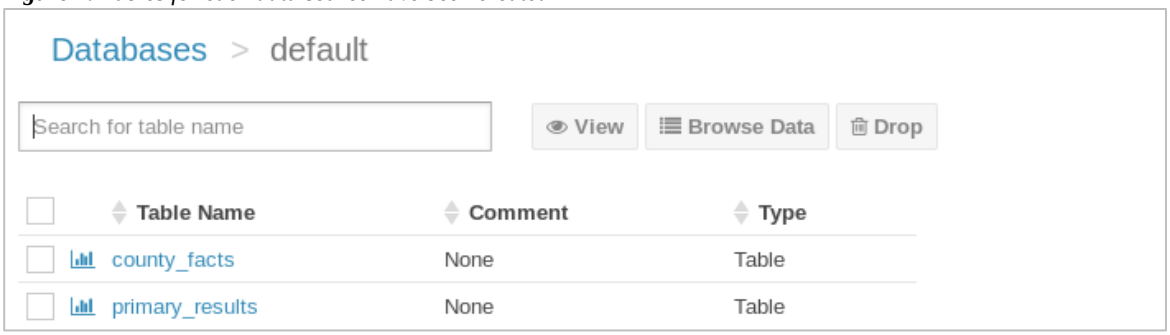


Figure 8: *Sampling of county_facts columns data*

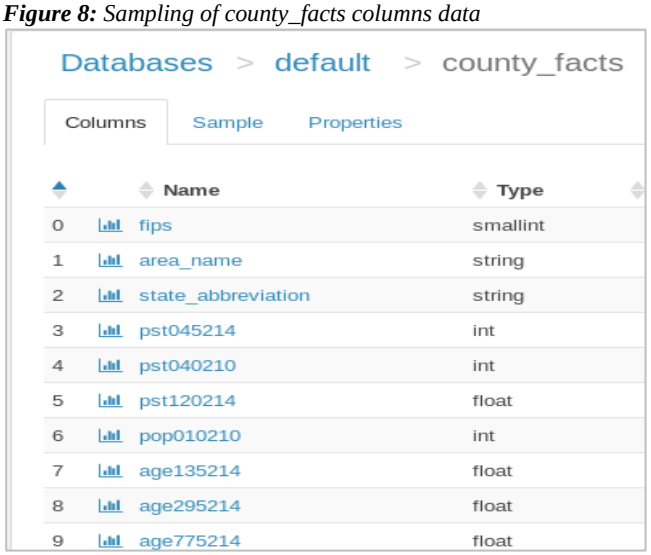
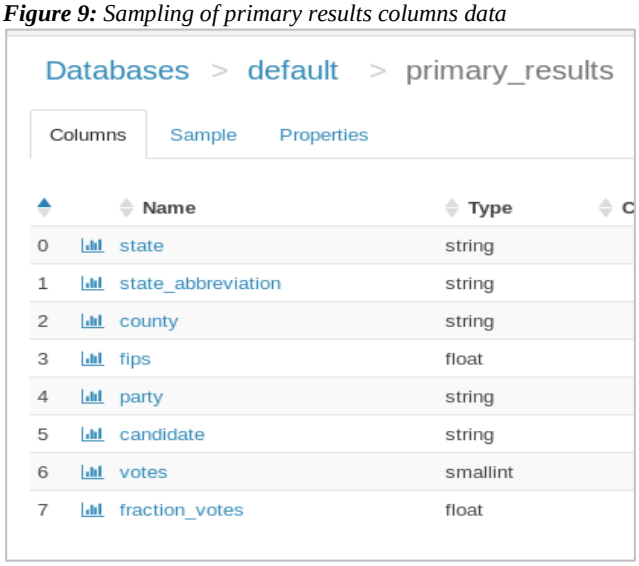


Figure 9: *Sampling of primary results columns data*



VIII. DATA ANALYSIS

Once the data were loaded and available in Hadoop, I decided to use Impala for querying for two reasons: its similarity to SQL which I have a lot of experience with and its speed so that I could efficiently run multiple queries as needed. An overarching objective of the analysis is to focus on states with the highest population since the general election in the US is determined by the electoral college which is indirectly determined by population. As such, this consulting agency has concluded that the most efficient use of campaign resources is to focus on states with top populations.

Several queries to help drive campaign strategy by identifying areas of untapped potential for each campaign in the top 15 states by population are run. Four analyses are conducted, two of which are intended to assist both parties, and one for each of the individual parties. The queries and the rationale behind the analyses conducted are provided in **Table 2**. The results of each of the queries are downloaded to .csv files and exported out of Cloudera for visualization in Tableau. Enterprise solutions would be able leverage seamless integration between Tableau and Hadoop, eliminating the extra step of the .csv download and import to the BI tool. Samplings of the results are provided in **Table 3**.

Table 2 - *Queries and analyses conducted*

Name	Rationale	Query
Percent Voting	*Examine voter turnout in the primary elections in the top most populated states. This is intended to be assistive to both parties in targeting areas with low turnout for a get out the vote effort for the general election.*	`SELECT state_abbreviation,2014_population, round((total_votes/2014_population)*100) as pct_vote FROM (SELECT SUM(votes) as total_votes,state_abbreviation FROM primary_results GROUP BY state_abbreviation) A INNER JOIN (SELECT state_abbreviation as state_abbr, SUM(PST045214) as 2014_population FROM county_facts`

Name	Rationale	Query
		WHERE length(state_abbreviation)>0 GROUP BY state_abbreviation ORDER BY 2014_population DESC LIMIT 15) B ON A.state_abbreviation = B.state_abbr order by pct_vote ASC
Percent Voters by Party	*Examine what percentage of voter turnout in each state in the primary elections was attributable to each party for the top most populated states. This is intended to be assistive to both parties in targeting areas where their party is not doing well.*	SELECT sa, party, round(((votes_per_party_per_state/votes_per_state)*100))as party_pct FROM (SELECT state_abbreviation as sa, party, SUM(votes) as votes_per_party_per_state FROM primary_results GROUP BY sa, party) A INNER JOIN (SELECT state_abbreviation, SUM(votes) as votes_per_state FROM primary_results GROUP BY state_abbreviation) B ON A.sa = B.state_abbreviation ORDER BY A.SA
Percentage of population that is Hispanic or Latino	*Examine what percentage of the population in the top most populated states are Hispanic or Latino. This is intended to be assistive to the Republican party which has historically not done well with this demographic. The Republicans should utilize the results of this analysis to reach out to these heavily populated Hispanic and Latino areas.*	SELECT state_abbreviation, area_name, fips, hispanic_latino, 2014_population FROM (SELECT state_abbreviation as state_abbr, SUM(PST045214) as 2014_pop FROM county_facts WHERE length(state_abbreviation)>0 GROUP BY state_abbreviation ORDER BY 2014_pop DESC LIMIT 15) A INNER JOIN (SELECT * FROM (SELECT state_abbreviation, area_name, fips, round(RHI725214) AS hispanic_latino, PST045214 as 2014_population, rank() OVER (PARTITION BY state_abbreviation ORDER BY RHI725214 DESC) AS hispanic_rank FROM county_facts WHERE length(state_abbreviation)>0) inner_hispanic WHERE hispanic_rank < 4)B ON A.state_abbr = B.state_abbreviation
Percentage of population that is age 65 or older	*Examine what percentage of the population in the top most populated states are age 65 or over. This is intended to be assistive to the Democrat party which has historically done better with younger voters. The Democrats should utilize the results of this analysis to reach out to these areas that have higher concentrations of senior citizens.*	SELECT state_abbreviation, area_name, fips, seniors, 2014_population FROM (SELECT state_abbreviation as state_abbr, SUM(PST045214) as 2014_pop FROM county_facts WHERE length(state_abbreviation)>0 GROUP BY state_abbreviation ORDER BY 2014_pop DESC LIMIT 15) A INNER JOIN (SELECT * FROM (SELECT state_abbreviation, area_name, fips, round(AGE775214) AS seniors, PST045214 as 2014_population, rank() OVER (PARTITION BY state_abbreviation ORDER BY AGE775214 DESC) AS seniors_rank FROM county_facts WHERE length(state_abbreviation)>0) inner_seniors WHERE seniors_rank < 4) B ON A.state_abbr = B.state_abbreviation

Table 3 – *Sampling of query results*

Percent Voting

Recent queries | Query | Log | Columns | Results | Chart

	state_abbreviation	pct_vote
0	CA	7
1	WA	7
2	NY	9
3	TX	13
4	NJ	13

Percent Voters by Party

Recent queries | Query | Log | Columns | Results | Chart

	sa	party	party_pct
0	AK	Republican	98
1	AK	Democrat	2
2	AL	Democrat	30
3	AL	Republican	70
4	AR	Republican	66

Percent Hispanic or Latino

Recent queries | Query | Log | Columns | Results | Chart

	state_abbreviation	area_name	fips	hispanic_latino	2014_popu
0	CA	San Benito County	6069	58	58267
1	CA	Tulare County	6107	63	458198
2	CA	Imperial County	6025	82	179091
3	TX	Maverick County	32767	95	57023
4	TX	Webb County	32767	95	266673

Percent 65 or older

Recent queries | Query | Log | Columns | Results | Chart

	state_abbreviation	area_name	fips	seniors	2014_population
0	CA	Calaveras County	6009	25	44624
1	CA	Amador County	6005	25	36742
2	CA	Sierra County	6091	27	3003
3	TX	Jeff Davis County	32767	30	2204
4	TX	Menard County	32767	30	2147

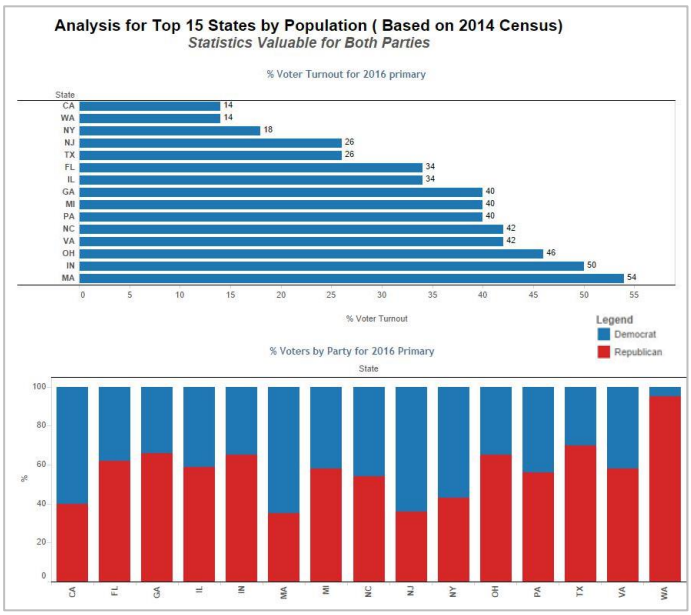
IX. DATA RESULTS & OBSERVATIONS

As described previously, the results of the queries that were run in Hadoop were downloaded and imported into Tableau so that the analyses could be visualized and easily interpreted by the political campaigns. The first dashboard that is developed, **Dashboard 1**, is intended to provide analyses that is valuable to both the Republican and Democrat parties. In the **top** chart on the dashboard, we can see that **Percent Voting** analysis for the top 15 states will inform each party that California, Washington State, New York, New Jersey, and Texas all have very large untapped potential with voter turnouts in the primary under 30% in the primary elections. These states should be targeted for a 'get out the vote' campaign. The campaigns can also consider targeting the other states that had less than 50% turnout if they have sufficient resources.

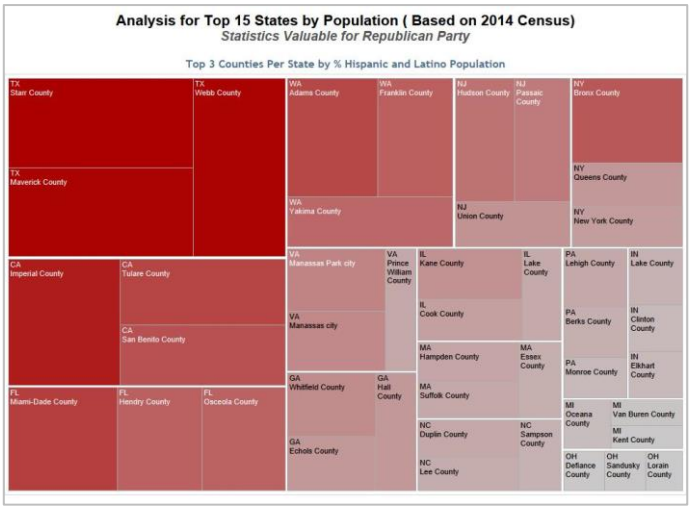
The campaigns can use the visualization on the **bottom** part of the dashboard to understand which states their parties are not doing well in. For example, the Democrat party has a lot of work to do in Washington state with almost 100% of the turnout in the primaries going toward the Republican race. While not as significant, the Republicans have some similar catching up to do in Massachusetts and New Jersey with less than half of primary voters casting their vote in the Republican race.

The second dashboard that is developed, **Dashboard 2**, is intended to provide analysis that is helpful to the Republican party. Since the Hispanic and Latino vote has not historically gone to the Republican party, the campaign should use the treemap to guide them where to focus campaign resources in the top 15 states. The treemap will show them which states have the counties with the largest Hispanic or Latino populations and which counties within these states they should target.

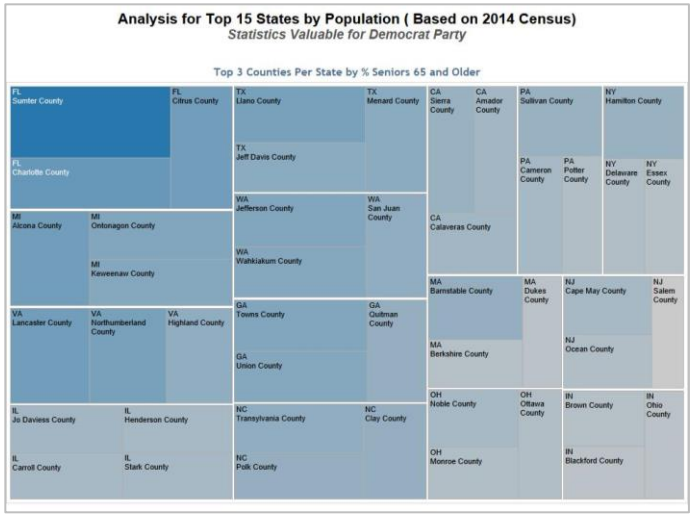
Dashboard 1



Dashboard 2



Finally, **Dashboard 3** is intended to provide analysis that is helpful to the Democrat party. The Democrat party does well with young voters but would gain advantage if they could win over some of the senior citizen votes. The Democrat party could use the treemap to guide them where to focus campaign resources within the top 15 states if they want to target seniors with their party's message. The treemap will show them which states have the counties with the largest senior population. Not surprisingly, we see that Florida which is a state where many seniors move to after retiring, is home to the county with the largest senior population in the country.



X. CONCLUSION AND NEXT STEPS

Big Data is a growing phenomenon in size and pervasiveness. Solutions to collect, process, and analyze big data are becoming more available and affordable. Organizations that capitalize on the opportunities of big data will improve the efficiency of their operations, increase profit margins, and enhance customer satisfaction. Organizations who do not recognize the emergence of Big Data as a fundamental piece of business in the modern day will find themselves left behind and unable to compete.

This paper demonstrated a sliver of capabilities afforded by a Big Data platform and proved that Big Data can provide a competitive edge for political campaigns. The next steps for the campaign consulting agency highlighted in this exercise would be to incorporate and correlate additional data sources including unstructured data sources such as social media. Sentiment analysis from Facebook and Twitter feeds would help campaigns understand what topics are trending and important to constituents. This would help them further tailor their campaign and messaging in key areas. Lastly, the consulting agency should also implement a platform to measure success and determine if the analyses and guidance resulted in additional votes for the campaigns.

REFERENCES

CSC. (n.d) *Big data meets presidential politics*. Retrieved August 7, 2016, from http://www.csc.com/big_data/publications/91710/94409-big_data_meets_presidential_politics

Florissi, Patricia. (2012, April 6). *Big Ideas: Demystifying Hadoop* [Video file]. Retrieved from https://www.youtube.com/watch?v=XtLXPLb6EXs&feature=youtu.be

Florissi, Patricia. (2012, July 17). *Big Ideas: Simplifying Cluster Architectures* [Video file]. Retrieved from https://www.youtube.com/watch?v=4M3cROio9vU&feature=youtu.be

Florissi, Patricia. (2012, February 27). *What is Big Ideas: How Big is Big Data?* [Video file]. Retrieved from https://www.youtube.com/watch?v=eEpxN0htRKI&feature=youtu.be

SAS. (n.d.). *Big data analytics: What it is and why it matters*. Retrieved April 19, 2015, from http://www.sas.com/en_us/insights/analytics/big-data-analytics.html

SAS. (n.d.). *SAS Solutions for Hadoop*. Retrieved April 19, 2015, from http://www.sas.com/en_us/software/sas-hadoop.html

Zikpolous, Paul. (2012, April 3). *What is IBM Big Data? Part 1* [Video file]. Retrieved from https://www.youtube.com/watch?v=B27SpLOOhWw&feature=youtu.be

Zikpolous, Paul. (2012, April 10). *What is IBM Big Data? Part 2* [Video file]. Retrieved from https://www.youtube.com/watch?v=W2Vnke8ryco&feature=youtu.be

.