HOW BIG DATA TECHNOLOGY CAN HELP ANALYZE VEHICLE RECALLS

Mohammed-Usman Ali, Northwestern University

***Abstract*— this project is based on data from the Government of Canada – Vehicle Recalls Database. It is primarily used by the Defect Investigations and Recalls Division to record and monitor recall completion rate pertaining to vehicles, tires and child restraints. The database consist of recalls covering the years 1975 to 2016 in Canada. It brings forward one of the largest vehicle recall database which includes the defect year, the manufacturer, the model make, units affected, recall date and general comments. The data points help record all the type of defects with associated manufacturer over the past 36 years.**

The purpose of this study is to take a simplistic approach in leveraging and showcasing the simplistic yet efficient functionality of the Cloudera's Hadoop ecosystem to bring forward basic inferences from the Vehicle Recalls Database. Moreover, the aim is to bring forward interesting yet meaningful patterns that can provide ideas as to which car manufacturer has been incurring recall infractions and the associated business impact in regards to the safety concerns.

It is important to note that the Hadoop Ecosystem is merely a tool that has provided the analyst to conform and analyze the vast array of data. At the end it is the analyst who must take ownership and make key decisions about the data.

I. Introduction

BIG data is the new oil. It is the next step in the evolution of innovation and productivity. Statements like these have made their way to the forefront of many organizations. In addition, there's hardly an organization that is not aware of the opportunities big data affords [1]. Over the past several years, there also has been significant technology innovation in the automotive industry which has enabled automakers to vastly improve overall safety in their vehicles. While the effort has improved overall safety, recalls have significantly increased, with 2014 being the most expensive year for automotive recalls [2]. Both the evolution of innovation in the big data and automotive industry has created two dynamics forefront which are beginning to significantly complement each other.

It's important to understand that vehicle recalls have basically turned into a public nightmare for numerous automotive companies over the past decade. From a sheer numbers perspective, General Motors filed the highest number of recalls in 2014 in the United States. Problems ranging from unsecured floors to ignition-switch defects were linked to at least 13 death. Moreover, there were about 54 separate recall campaigned that affected as many as 26 million vehicles [3]. These 26 million vehicles were equivalent to 40% of General Motor's fleets and almost 11% of all vehicles in operation. Another use case that has also blown the confidence of auto buyers has been the German auto maker Volkswagen that has announced that it will recall 8.5 million diesel vehicles due to cars that were fitted with software that allowed them to cheat emissions. From a business and cost perspective, this Volkswagen, diesel recall is expected to cost millions and damage the brand's reputation.

As we can clearly see that auto recalls over the past decade have clearly been vast and shocking when just looking at it from a numbers perspective. It's also interesting to see that majority of recalls have been recent and in the same growth period of automotive innovation and big data growth. From a strictly value proposition perspective, Big Data certainly has the ability and power to assess a vehicle error captured through various sensors in a vehicle and possibly provide the driver with an update. Other Big Data application in the car industry have been in the form of warranties and vehicle management. Recalls could have been prevented or managed better as numerous automotive organization could have used Big Data to predict that a failure was happening or used it for a much better management of recalls [4]. It's also interesting to note that the automotive industry can learn from the aviation industries which use remote diagnostics to remotely manage and predict airplane maintenance.

We can clearly see that Big Data Technology & the automotive industry have direct complementary strengths. Based on various industry leading articles we will take a simplistic and a minor step in show casing big data technologies by leveraging a recalls database. Using the Hadoop Cloudera stack, some questions I will explore are

- Which Automotive Manufacturer has incurred the most recalls?
- Are there any specific trends in regards to automotive recalls since 1975?

M. U. Ali is a graduate student in the Master's of Predictive Analytics (MSPA) program at the Northwestern University's School of Professional Studies, Chicago, IL 60611 USA (e-mail: usmanali@u.northwestern.edu)

Again, I will aim to take a simplistic approach in analyzing the large dataset and showcasing the power of Hadoop.

II. Big Data Overview

Big Data consist of three characteristics that make the overall realm of Big Data unmatched. More specifically Variety, Velocity and Volume are three of the industry leading characteristics that define Big Data. Big Data comes forward as an enabler concept & technology that helps Big Organizations overcome the challenges and complexities of data as it has grown exponentially over the past few decades [5]. Variety, refers to the ability that data can be stored in multiple formats. It can range from emails, to tweets, to sensor data. There's no control over the input data format or the structure of the data. Velocity refers to the ability that data can stream in a high speed and continuous fashion for data processing. Rather than having batch process in which data was analyzed overtime, the big data processing engine can work at highly scalable and extreme speeds in a relatively small amount of time [6]. Lastly, Volume refers, to the amount of data that can be generated continuously without problem. These data points include, log data, SCOM server data and clickstream data.

Taking a step back, it is critically important to note that every organization will have a different business model to capture each Big Data characteristics accordingly. The level of conformation to each of the 3 V's will uniquely be adapted within each organization by first understanding their long term data value proposition. The opportunities created by leveraging big data is clearly dependent on the long term business objective that each organization must establish.

III. Big Data Technologies & Ecosystem

Big data technologies such as Hadoop's Distributed File System and MapReduce technologies offer significant advantages from traditional technologies. It's important to note that traditional Relational Database Management Systems (RDBMS) and BI technologies, remain centralized with a single point of failure. There's always a fixed schema and a relational database which can contain up to a Terabyte of data. In stark contrast, Big Data technologies, are distributed on commodity servers, are highly scalable and have massive parallel processing power. More importantly, there's no single point failure and it can handle both structured and unstructured data using NoSQL platforms. Lastly, Big data technologies have the ability to store up to Petabytes and Exabyte's of data. Using our Vehicle Recall database initially in excel caused my system to slow down extremely. When it was uploaded into Hadoop it was very easy to query and transform.

The Big Data ecosystem consist of 4 key components that bring together the power to ingest and collect data from various sources, data storage and computing infrastructure, data transformation and analytics and finally data visualization and delivery. The first component brings forward the ability to pull data from various sources and collect various types of data. In the Big data world coupled with the Internet of Things, nowadays every aspects of our lives generate data. Data can now come from Social media interactions, Car sensor data and basic public data domains that have been tracked for years. Holistically, the first component helps to ingest and collect data.

The second component is data storage and computing infrastructure. This components accommodates the ability to store and process large volumes of data by using Hadoop proprietary storage cluster infrastructure, the distributed file system and Hadoop's massively parallel processing engine. There's also a NoSQL component that helps store unstructured data. For this project we used a single node cluster by leveraging Cloudera's Hadoop stack.

The third component is data transformation and analytics which encompasses various database like tools and languages such as MapReduce, Pig and Hive. This key Hadoop components brings forward the ability to run queries to analyze and transform the data accordingly. For our Vehicle Recall database we will aim to showcase Hive.

The final component is the visualization and delivery. There is a vast selection of different tools available on the market to help deliver and visualize data. This can be contained through various Business Intelligence Software and/or visualization software's such as Tableau, QlikView and R.

IV. Hadoop Ecosystem

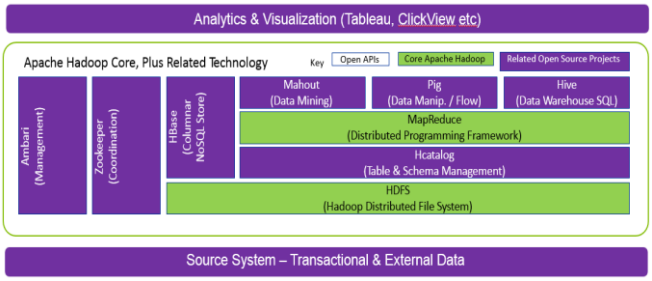
The Hadoop ecosystem is diverse and grows by the day. It's impossible to keep track of all of the various projects that interact with Hadoop in some form Hadoop. There are modules contained within the Hadoop project — Hadoop Common, Hadoop Distributed File System, Hadoop YARN and Hadoop MapReduce. The main strength is to use these components together which give users the tools to support additional Hadoop projects that we'll mention in below (Figure 1.0). The other key strength is Hadoops ability to process large data sets in real time while automatically scheduling jobs and managing cluster resources through the use of each component.

- Hadoop Distributed File System: A distributed file system that provided high-performance access across Hadoop Clusters.
- MapReduce: A distributed processing frame-work named MapReduce. Brings forward two distinct task that Hadoop Performs. Map job which converts data into individuals key and value pairs. The reduce job takes output from Map and combines into a smaller set of data.
- Hbase: The non-relational Hadoop database is a distributed, column oriented database, with HDFS for the underlying storage.
- Hcatalog: A metadata abstraction layer for referencing data without using the underlying filenames or formats, insulates users and scripts from how and where the data is physically stored
- PIG: A Scripting platform that are used to analyze large data sets, since their structure is amenable to substantial parallelization.
- Hive: The Hadoop Data Warehouse, a tool developed at

Facebook, facilitates easy data summarization, ad-hoc queries, and the analysis of large datasets stored in Hadoop compatible file systems

- Zookeeper: A centralized Cluster coordination service that maintains configuration information and Naming, and provides distributed synchronization and group services useful for a variety of distributed systems. HBase is not operational without ZooKeeper
- Sqoop: A Data integration tool, designed to efficiently transfer bulk data between Hadoop and structured datastores, such as Relational Databases.
- Flume: Provides distributed log management services, and can function as an Event Queue Manager or a Log Aggregator, collecting log data from many diverse sources and moving them to a centralized data store.
- Mahout: A scalable machine learning library that implements many different approaches to machine learning.
- Ambari: An open source operational framework for provisioning, managing and monitoring Apache Hadoop clusters that includes an intuitive collection of operator tools and a robust set of APIs that hide the complexity of Hadoop, simplifying the operation of clusters.

Figure 1.0 – Hadoop Ecosystem.



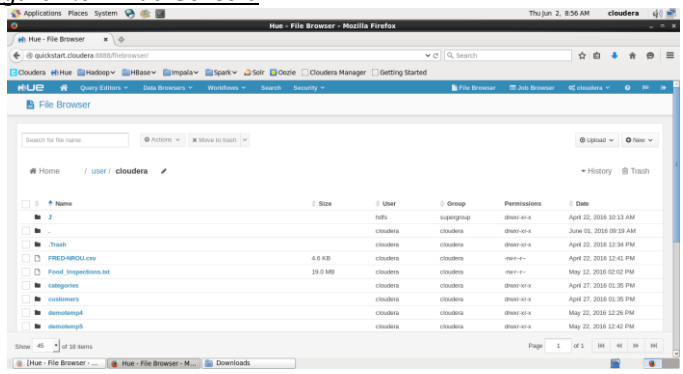
V. Data Set & Data Preparation

This project is based on the data from the Government of Canada – Vehicle recall database. The database contains records from 1975 to 2016. It's important to note that the data contained both French and English attributes as per the requirement of the Canadian Government. Moreover there are 15 attributes and a total of 81,296 records. The total size of the file stands around 100 megabytes and significantly slows down the overall pc performance when opening it in excel.

From a data preparation perspective, we removed the French labelled attributed for ease of use and understanding. This helped truncate the size of our database, although truncating would be indifferent when using Big Data tools as the sheer processing power still would processes in the same amount of time. Our final data set contained 10 attributes with a total of 81,296 records.

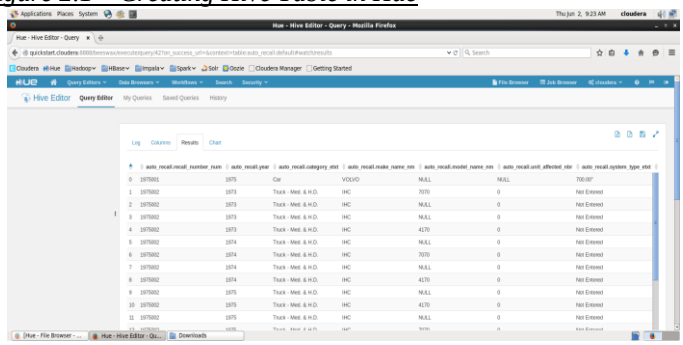
For our analysis I will leverage Cloudera's Hue console and showcase the simplicity brought forward when uploading a large data set into a Big Data tool. The Hue Console is specifically created by Cloudera and leverages a simplistic web browser functionality to access various Hadoop Components described earlier. For the Vehicle Recall database, I first need to upload our dataset into Hadoops HDFS using the Hue console. The workflow is extremely simple, I uploaded the whole dataset by clicking "Upload" and selecting our dataset. It took about 10 seconds to load about 41 megabytes. The Recall database was in a csv format. Uploading into HDFS is usually the first step taken before any analysis or transformation can be performed.

Figure 2.0 – Hue Console



In our analysis I will solely leverage Hive due to its ease of use, SQL like similarities and basic visualization functionalities. In order to leverage Hive, I created a Hive table through the metastore functionality in Cloudera. This basically pulls the data from HDFS and automatically defines the schema of the data. The beauty surrounding this functionality is that there is no manual effort in defining the schema. The Hive editor automatically recognizes the type of data and defines it accordingly.

Figure 2.1 – Creating Hive Table in Hue



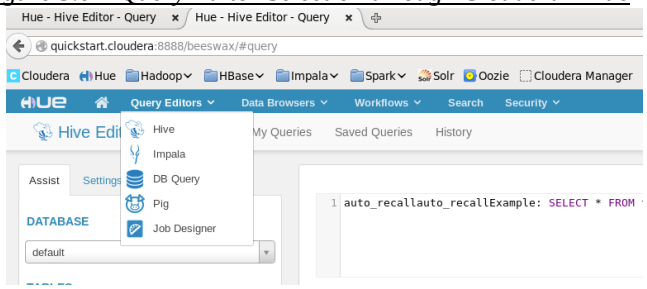
From an overall perspective, this whole process took about 5 minutes, just the sheer simplicity of these big data tools really brings forward a new realm of efficiency. Now think about how long this would take when inputting a large set in a traditional relational database.

VI. Data Analysis

As highlighted earlier, I aim to showcase the Hive component within Cloudera's Hadoop stack. In addition, it's interesting to note that within the Hue console, users also have

the ability to easily select any other Hadoop components that will help with transformation and data analysis. This is highlighted in Figure 3.0, where Impala and Pig can be easily leveraged.

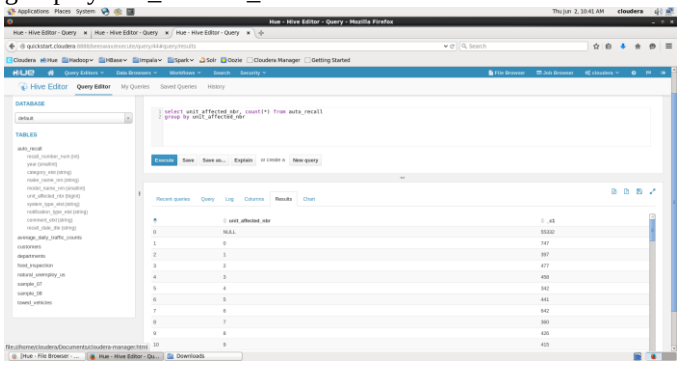
Figure 3.0 – Query Editor Selection through Cloudera - Hue



The Hive Query Editor is an SQL like tool to query the data and that provides basic visualizations in an efficient manner. The Hive language is known as HiveQL. The first effort from a data perspective, was to look through our recall database to bring forward any data quality issues and missing values. For this reason I put together a basic query to bring forward any discrepancies. The query was run in under 5 seconds and the results have been highlighted in Figure 3.1. When looking at automotive recalls, the number of unit affected is one of the most critical attributes. The results of the query brought forward around 55,332 NULL values which is significant. It's important to note that the vehicle recall database is from 1975, hence the missing values maybe due to the lack of centralized technology to record the recalls.

Figure 3.1 – Null values in Units Affected

HiveQL: select unit_affected_nbr, count(*) from auto_recall group by unit_affected_nbr

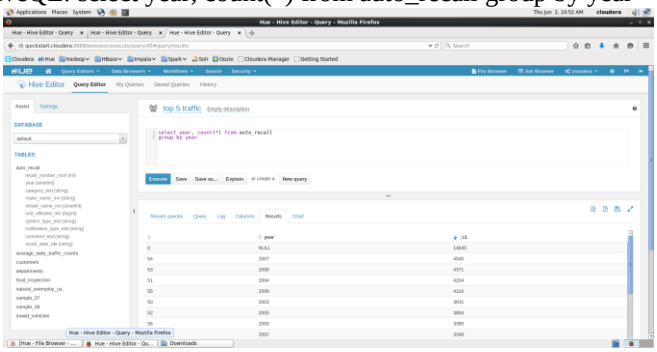


Next I aim to analyze which year had the highest number of automotive recalls. In figure 3.2, we can clearly see that the year 2007 had the highest number of recalls in Canada with abour 4,500 units affected. Again, in figure 3.2 we see that there's also a significant number of NULL values in the years.

Figure 3.2 – Year with the highest number of recalls

Answer: 2007

HiveQL: select year, count(*) from auto_recall group by year



Other data inferences from the vehicle recall database include the following:

Figure 3.3 - Which Manufacturer has incurred the most recalls since 1975? -

Answer: CHEVROLET

HiveQL: select make_name_nm, count(*) from auto_recall group by make_name_nm

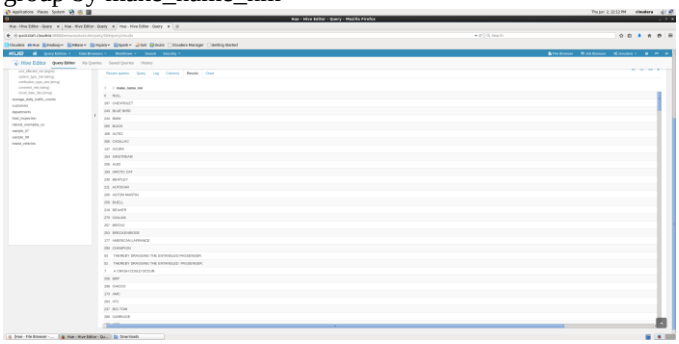


Figure 3.4 - Which Manufacturer has incurred the least amount of recalls since 1975?

Answer: ACCURIDE

HiveQL: select make_name_nm, count(*) from auto_recall group by make_name_nm

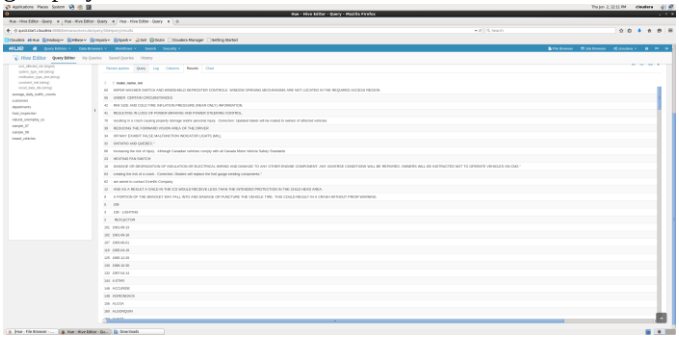


Figure 3.5 – Most Type of Recalls since 1975

Answer: SAFETY MANUFACTURER

HiveQL: select notification_type_etxt, count(*) from auto_recall group by notification_type_etxt

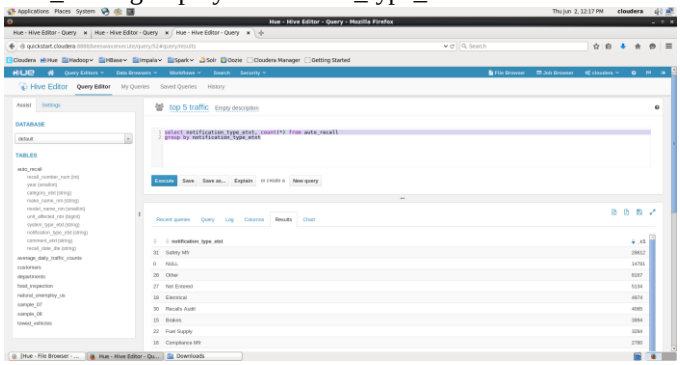
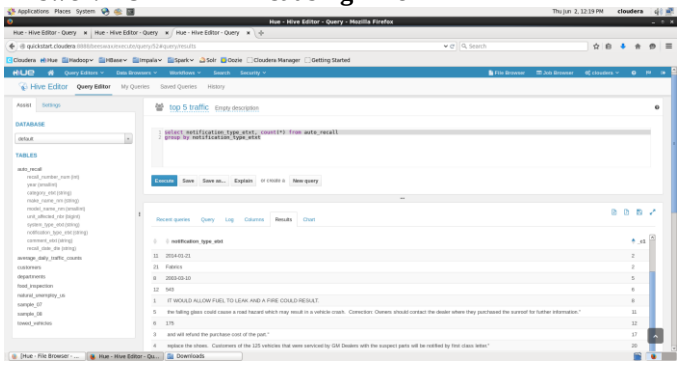


Figure 3.5 – Least Type of Recalls since 1975

HiveQL: select notification_type_etxt, count(*) from auto_recall group by notification_type_etxt

Answer: FUEL LEAK causing Fire



As we can clearly see, a simple HiveQL can bring forward many interesting inferences from a large dataset. The amount of time to pull each of these interesting inferences was minimal and the output was really simple to interpret. Again, just the sheer simplicity of Cloudera's Hadoop stack makes these Big Data technologies a forced to be reckoned with.

In our next analysis, I bring forward the basic visualization functionality in Cloudera. In figure 3.6, we can see a simple histogram that charts the frequency of the type of recalls that have incurred since 1975.

Figure 3.6 – Highest Type of Recalls.

Answer: 28,612 – Safety Recalls

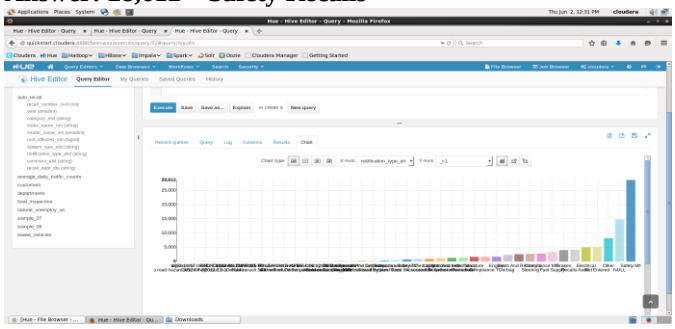
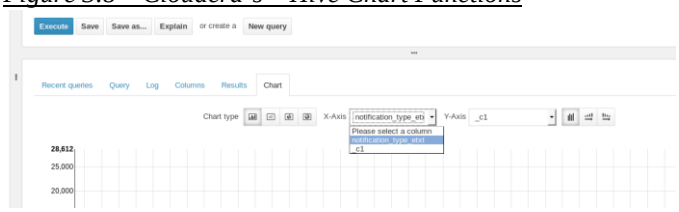


Figure 3.8 – Cloudera's – Hive Chart Functions



Using Cloudera's – Hive Query Editor's chart functionality we can plot further data points with relatively ease by selecting columns from the drop down list. We will plot additional data points to make further inferences below:

Figure 3.9 – Most Type of Vehicles Recalled

Answer: Trucks

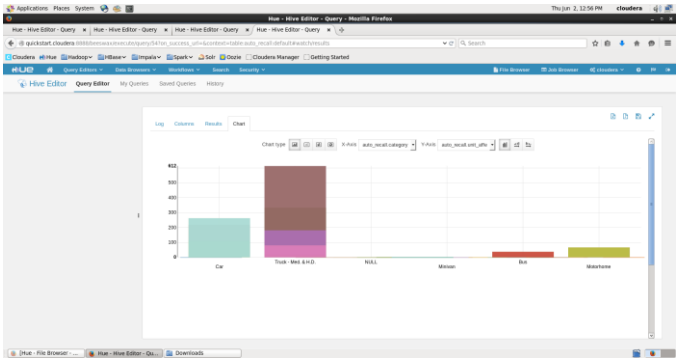
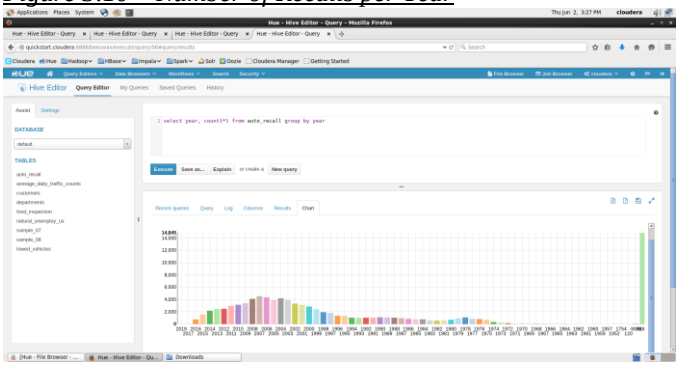


Figure 3.10 – Number of Recalls per Year



Interestingly, by looking at figure 3.10 we see the distribution of the number of recalls each year. We note that majority of the recalls have occurred from the years 2000 to 2016.

VII. Results & Observations

By using Cloudera's Hive Query editor I was able to systematically breakdown and understand theour vehicle recall database in an extremely efficient manger. Initially I brought the whole vehicle recall database into Hadoops HDFS. From this points I created a Hive table in a matter of seconds. Cloudera's Hive Editor ability to automatically define the datasets schema and build a table accordingly was truly a pleasure to work with. Lastly, I leveraged HiveQL and basic chart functionality in Hive to understand and plot the data

accordingly. These steps have brought forward several important inferences of our vehicle recall database that we will briefly touch upon.

The overall vehicle recall database that was leveraged from the Government of Canada recall website, brought forward one key hurdles that Cloudera's big data technology was able to resolve. One of the first inferences that I analyzed was the amount of missing values in our dataset. This is a natural characteristics with many databases that contain data which have spanned for several years. With a simple query we were able to view the dirty data aspect of the vehicle recall database.

I then briefly put together a HiveQL query that helped bring forward the year with the highest number of recalls which was in 2007 and the manufacturer that had the highest recalls which was Chevrolet. Thereafter I looked at the most type of recalls that had occurred since 1975 which was a safety manufacturer recall and the least type of recall was a fuel leak.

Lastly, I used Hive's Query editor to plot some of the queries that we had put together. The simplistic functionality of the charting function helped us create plot in a matter of seconds. I was able to understand that our database contained all types of vehicles and not only cars. It was interesting to see that the most type of vehicles that were affected by recalls were Trucks. This is interesting as you don't really hear Trucks that were recalled in the news. For our final inference I plotted the frequency over recalls over time. This was probably one of the most important plot as it brought into perspective that majority of the recalls since 1975 have happened from the year 2001. This brings back to our initial point of how technology over the past decade has really put an emphasis on safety. Moreover, big data tool like Cloudera's Hadoop stack make it that much easier to analyze a vast array of data.

It's important to note the vehicle recalls in general have increased over the past decade and this is due to tools and technologies that have significantly evolved since 1975. Increased safety regulation and additional sensor technology that have been implemented in many vehicles on the road today increased the safety magnifying glass. Moreover, tools like Hadoop make the handling and analysis of large vehicle data much easier and seamless.

VIII. Conclusion

The process and results of this brief analysis has brought forward many interesting facts from our vehicle recall database. My conclusion in general, is that vehicle recalls have significantly increased over the past decade. This increase has led to believe that a part of it is solely due to increased safety regulation and technology innovation in both the automotive industry but also the management of large sets of data with Hadoop.

This project has demonstrated that Cloudera's Hadoop stack has a high degree of ease of use and sheer processing power with the use of Hadoops distributed file system. The ability store and analyze large quantities of data has become seamless. Using specific Hadoop components such as Hive we have answered very basic yet important questions of our vehicle recall database. One fascinating fact again was that majority of large vehicle recalls have just happened recently. This leads us to believe that technology innovation such as Hadoop has been an enabler to these recalls as it makes it that much easier to analyze car data.

The goal and the overall objective of this study was to showcase key big data technologies and apply it to a large data set in a timely manner. In terms of next step, this process can be applied to any large database for quick inferences. More importantly, we can now learn at a much faster rate from our past in order to identify and apply appropriate safety measures to limit vehicle recalls in the future.

References

[1] Andersen etl al. (2015). Deloitte – Big Data and Anlaytics in the automotive industry. Retrieved from http://www2.deloitte.com/content/dam/Deloitte/us/Documents/manufacturing/us-auto-automotive-news-supplement.pdf

[2] Damodaran B. (2015). Addressing automotive safety issues with analytics Retrieved from http://www.ibmbigdatahub.com/blog/addressing-automotive-safety-issues-analytics

[3] Squire S. (2014). GM recalls: The numbers tell a surprising story Retrieved from http://projects.marketwatch.com/2014/gm-recalls-the-numbers-tell-a-surprising-story/

[4] Boagey R. (2014). Automotive World: Big Data – big solution of big problem? Retrieved from http://www.automotiveworld.com/analysis/big-data-big-solution-big-problem/

[5] McNulty E. (2014). Dataconomy: Understanding Big Data: The ecosystem Retrieved from http://dataconomy.com/understanding-big-data-ecosystem/

[6] Dave P. (2013). Big Data – What is big data – 3Vs of Big Data – Volume, Velocity and Variety Retrieved from http://blog.sqlauthority.com/2013/10/02/big-data-what-is-big-data-3-vs-of-big-data-volume-velocity-and-variety-day-2-of-21/