2016

Big data analytics with vehicle data

Ashok Singamaneni

ABSTRACT

BIG DATA ANALYTICS WITH VEHICLE DATA

Ashok Singamaneni, M.S.
Department of Mechanical Engineering
Northern Illinois University, 2016
Dr. Abhijit Gupta, Director

Many companies have invested a lot over the past decade just to collect the data and store them in a cloud. However collection of such large amount of data will be justified only when there are some useful insights drawn from them. There is a lot of data collected from vehicles. The volume, velocity, variability and complexity of the data from various sensors are massive. Access to this type of data is only going to increase with time, so industries need appropriate methods to transform this raw data into insights and knowledge. Extraction of insights which were previously unknown or potentially useful patterns or knowledge from this kind of these massive amounts of data can only be achieved by using Big Data analytics. Conventional software cannot handle the robustness of these, so modern tools such as Hadoop and Knime were used in this thesis to analyze the data. Raw high resolution data was used and a model was developed to understand vehicle/customer behaviors and then compared and contrasted. This thesis involves found a proper method for identifying and calculating the principal attributes that accurately and efficiently characterize a vehicle's operation. Predicting the power of new vehicles and finding the similarities between new vehicles and old vehicles was the main goal of this thesis.

NORTHERN ILLINOIS UNIVERSITY
DE KALB, ILLINOIS

AUGUST 2016

BIG DATA ANALYTICS WITH VEHICLE DATA

BY

ASHOK SINGAMANENI

A THESIS SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

MASTER OF SCIENCE

DEPARTMENT OF MECHANICAL ENGINEERING

Thesis Director:
Dr. Abhijit Gupta

ACKNOWLEDGEMENTS

It gives me great joy to thank my advisor Dr.Abhijit Gupta who guided me in every single stride to finish this proposition work. He bolstered me with regards the research and grants. His incredible state of mind towards work empowered me to work without any obstacles.

I would also like to thank my committee members Dr. Majumdar and Dr. Ji-Chul Ryu who invested their time and offered suggestions to add value to my thesis.

A special grace to Mr. Brett Collins, who encouraged me to complete my work and boosted me to overcome the faults in developing my thesis and guiding me by being my mentor in completing my thesis.

It would have not been possible to complete my masters program without the support of my friends and family. I thank Sushma Gudipudi, K.BadriNath, Anuradha Surya Narayan Iyer and my roomates who always helped me balance my work and personal life.

DEDICATION

I dedicate my work to my parents and family who have dreamed and worked so hard to make me an engineer.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

We are in a century where technology is emerging faster that mankind has ever experienced. Engineers are not expected to be conventional and orthodox sticking to the old rules, but they need to think outside of the box and use today's emerging technologies to discover/invent new things that help make human life so easy. In this thesis, I have tried to extend my horizons to use big data technology to solve a mechanical engineering problem for a firm that deals with different types of vehicles. I got a lot of data from 90 vehicles and have analyzed it using different statistical techniques so as to let the business make decisions based on the results.

1.1 Big Data

The data which is collected in large amount of volume, velocity, variety and veracity is called big data.

"It is not the size which interests as hard ware to collect the data has become so cheap and every thing is being stored in the cloud, but what's really significant is that the different tools developed to access that enormous data with ease so as to analyze it and draw interesting insights out of it. Without the development of tools like Hadoop in contrast with traditional programming languages and databases, it's really difficult to analyze and visualize the data to get some insight" [8] [9].

1.2 Hadoop

Hadoop is a open source software for reliable, scalable, distributed computing.

"The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures"[1].

Though Apache has developed Hadoop, there are several firms which offer their packages with Hadoop distribution. The popular Hadoop distributions include from Cloudera, Horton Works, and MapR.

1.2.1 System requirements

I had got 17 GB of gun zipped files which when they were extracted has came to 215 GB. Ideally minimum of 500 GB of hard disc and 16 GB of RAM to process the data. But having four times of the size of data is always recommended.

1.2.2 Installation

First I installed Ubuntu 14.0 LTS as the operating system on my personal computer, and then I installed a single node cluster of Cloudera CDH5 using the following websites [2].

1.3 KNIME

"Knime Analytics Platform is the leading open solution for data-driven innovation, designed for discovering the potential hidden in data, mining for fresh insights, or predicting new futures. Organizations can take their collaboration, productivity and performance to the next level with a robust range of commercial extensions to our open source platform. For over a decade, a thriving community of data scientists in over 60 countries has been working with our platform on every kind of data: from numbers to images, molecules to humans, signals to complex networks, and simple statistics to big data analytics"[3].

I used knime in my thesis to perform statistical analysis on my data and perform algorithms like linear correlation, predictive analytics, similarity search, and decision tree.

1.4 Vehicle Data

We had data from of 90 vehicles and the information about those vehicles.

The first data set consisted of 187 variables/columns and 226,380,164 rows. Among all 187 variables available for this thesis, I concentrated on vehicle information and the seven variables 'Vin, time stamp, speed, estimated power, brake switch, cruise state, and engine speed (rpm)'.

The second data set contained vehicle information with 19 variables/columns and 90 rows.

1.5 Literature Review

Sanjay Ghemawat *et al* [8] discussed the Google file system and how Google managed to work with such vast amounts of data available with them. This paper revolutionized the information technology world and led to the development of various technologies/tools like Hadoop from various distributions, and Tableau.

Barna Saha *et al* [10] discussed data quality from data itself and compared accuracy vs efficiency. The authors also discusses four V's of big data - volume, velocity, variety and veracity. The significance and role of veracity was mainly discussed in the paper.

1.6 Approach

The approach of this thesis can be defined using the following steps and they denote different stages in the project.

1. I extracted the data and uploaded it into Hadoop system.
2. I processed the data and eliminated unwanted variables so that the size of the data can be reduced.
3. I wrote the queries using impala, spark and hive to extract the data set appropriate for statistical analysis and save it in an Excel file.
4. I imported the data into KNIME and design the work flows to perform statistical analysis.
5. I showed the results and discussed about them.

CHAPTER 2

DATA EXPLORATION

This chapter discusses about uploading the data into Hadoop, processing the data into a data set which was used for statistical analysis and Hadoop components used for it.

2.1 Uploading Data into Hadoop

This was one of the biggest challenges as I had a computer with insufficient disc space. As discussed in 1.2.1, with 215 GB to upload into Hadoop I used Hue which is an user interface to manage sql editors, file system and tables in cloudera Hadoop. I created a table vehicle_data" and uploaded the data into it. As they were large amounts of data, it took lot of time and system performance.

2.2 Cleaning the data

As the data was so large, querying the data was not efficient as the size of the computer's hard disc was smaller than the data size. In order to overcome this issue, I started cleaning the data so that the size of it was reduced and querying efficiently reduce the burden on the computer. The following steps were performed to clean the data.

1. I removed all the variables except the seven variables mentioned in 1.4.

2. I did not consider the rows which had data that was not logically true. For example, I did not consider the speed values which were recorded as less than zero. There might have been a chance that sensor might have malfunctioned and recorded such values.

Note that I could delete the data from the editors available in Hadoop as I traditionally did in relational database systems. Only option available for me is to write a query in one of the data editors in Hadoop such that the cleaned data was stored in another new table.

The query used to do this task is given below. It was executed in impala

```
create table vehicle_data_filtered
row format delimited fields terminated by ','
as
select vin, timestamp, speed, estimated_power, brake_switch,
cruise_state, engine_speed_rpm from vehicle_data
where
vehicle_speed_mph_1 >=0 and
estimated_power_bhp>=0 and
engine_speed_rpm between 0 and 8031.87
```

This resulted in seven variables/columns and 196,202,196 rows. Before this step I also needed to check if there were any missing values in other columns which were necessary.

I needed to check if there were any values other than 0 to 6 in brake_switch and cruise_state. Below are the queries I ran in impala to check this

1. `select distinct cruise_state from vehicle_data_filtered`

2. `select distinct brake_switch from vehicle_data_filtered`

If the above queries gave any numbers other than 0,1,2,3,4,5,6 then I should have omitted those rows too, but luckily sensors recorded the data perfectly.

Hence, the data was processed and cleaned so that I could get the required data sets and perform analysis. This new table data was downloaded from hue into my local system with the name "vehicle_data_filtered" and the file size was 25.2 GB which is a lot less than the original size of the data mentioned in 1.2.1.

2.3 Preparing the dataset

As there are multiple rows for each vin number, my goal was to prepare the required dataset is to have a single row for each vin. So in total I expected 90 rows of data.

The columns which I needed to get from my original dataset were vin, average_speed, average_power, average_rpm, average_cruise_sspeed, average_cruise_power, average_cruise_rpm, stop_count, and brake_count

The following query in Impala gives all the variables except stop_count and brake_count

```
select t1.vin, t1.speed, t1.power, t1.rpm, t2.cruise_speed,
t2.cruise_power, t2.cruise_rpm from
(select a.vin, avg(a.speed) as speed, avg(a.power) as ‘power‘,
 avg(a.rpm) as rpm from(
select vin, to_date(‘timestamp‘), avg(speed) as speed,
avg(power) as ‘power‘, avg(rpm) as rpm from
vehicle_data_filtered
where speed !=0 or power !=0 or rpm !=0
group by vin, to_date(‘timestamp‘)) a group by a.vin) t1
LEFT OUTER JOIN
(select a.vin, avg(a.cruise_speed) as cruise_speed,
avg(a.cruise_power) as cruise_power,avg(a.cruise_rpm) as cruise_rpm
```

```
from(
select vin, to_date(‘timestamp‘), avg(speed) as cruise_speed,
avg(‘power‘) as cruise_power, avg(rpm) as cruise_rpm from
vehicle_data_filtered
where cruise !=0
group by vin, to_date(‘timestamp‘)) a group by a.vin) t2
on t1.vin = t2.vin
```

I saved this table into an Excel file named vehicle_dataset_1.

In order to get stop_count and brake_count, I needed to create a work-flow in KNIME as per the steps shown in Figure 2.1

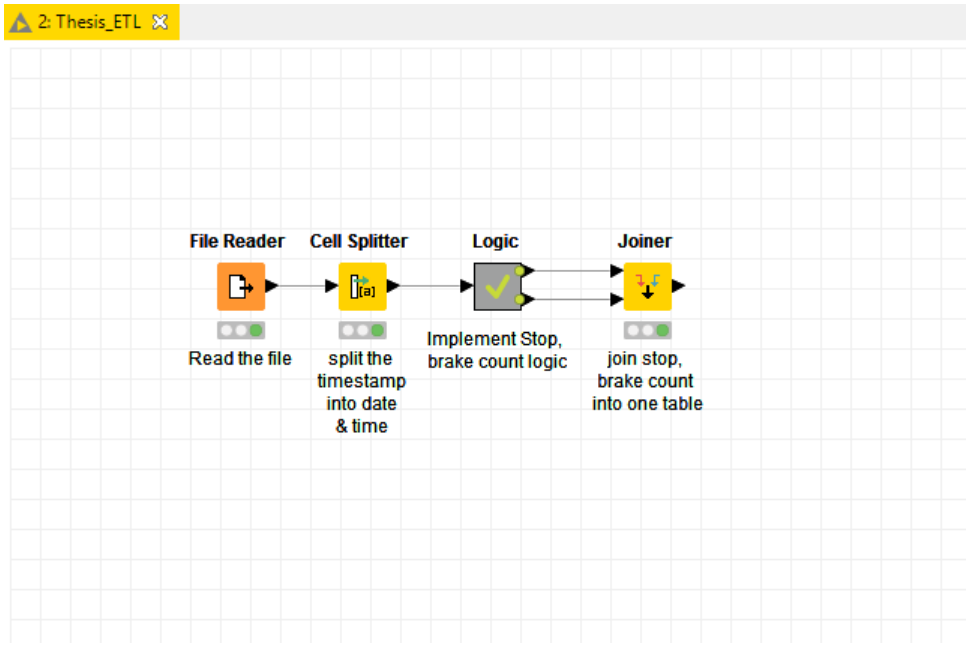

Figure 2.1: Work flow - To get brake and stop count in KNIME

1. File Reader: I uploaded the file which had ”vehicle_data_filtered” table data into Knime.

2. Cell Splitter: I split the timestamp into date and time as I needed to take the average of brake and stop count for each date and then over all average with respect to each vin.

3. Logic: The main logic to get the brake and stop count was implemented in this metanode. The same is shown in Figure 2.2. The Java snippet is used to write the code to get

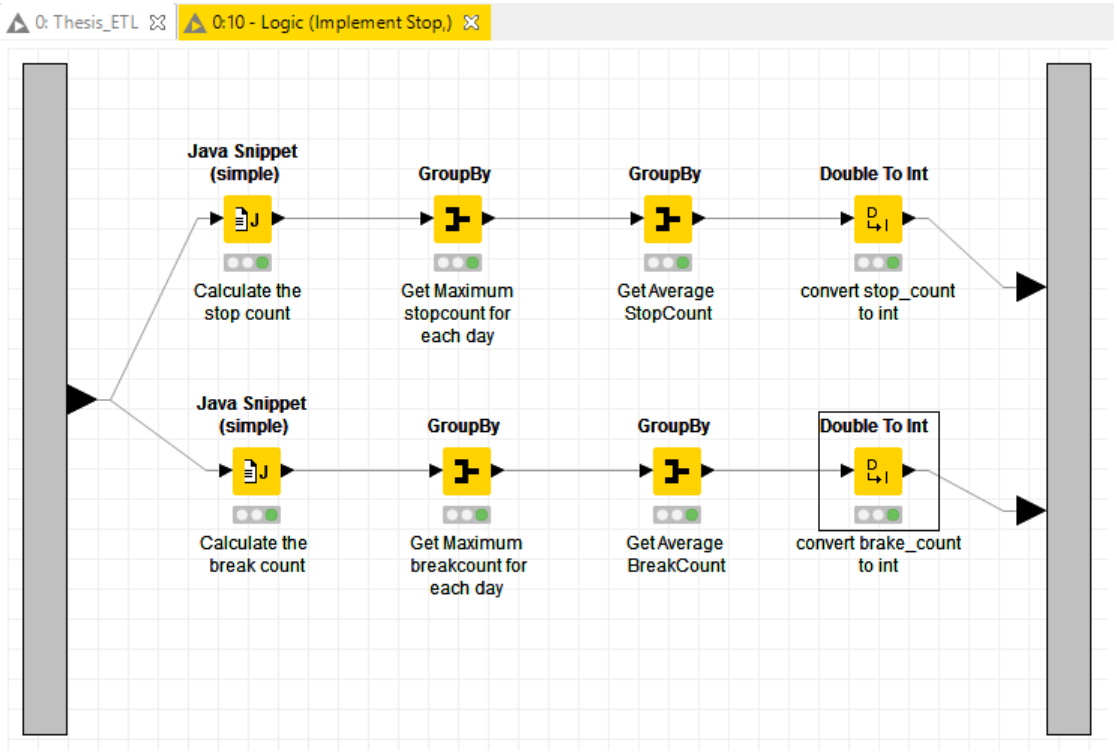

Figure 2.2: Work flow - break and stop count logic

brake and stop count. The Figures 2.3 and 2.4 shows the logic and code implemented.

4. After I got the brake and stop count, I used the joiner node in knime to join the brake and stop count which I had obtained and the excel file which I stored as vehicle_dataset_1 in my local system.

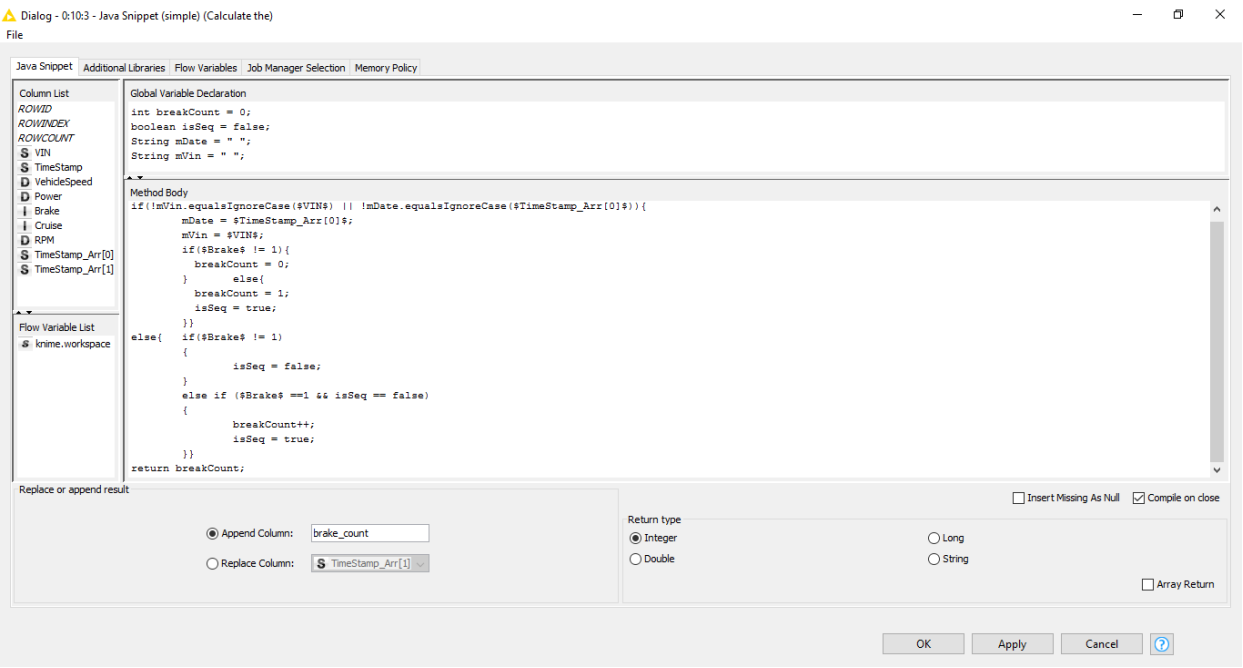

Figure 2.3: Break count Java Snippet

I joined this intermediate data set with second data set mentioned in 1.4 using the joiner node in Knime with "vin" as key. The outcome of this joiner was my required data set, and I saved that into an Excel file with name "vehicle_dataset" in my local system.

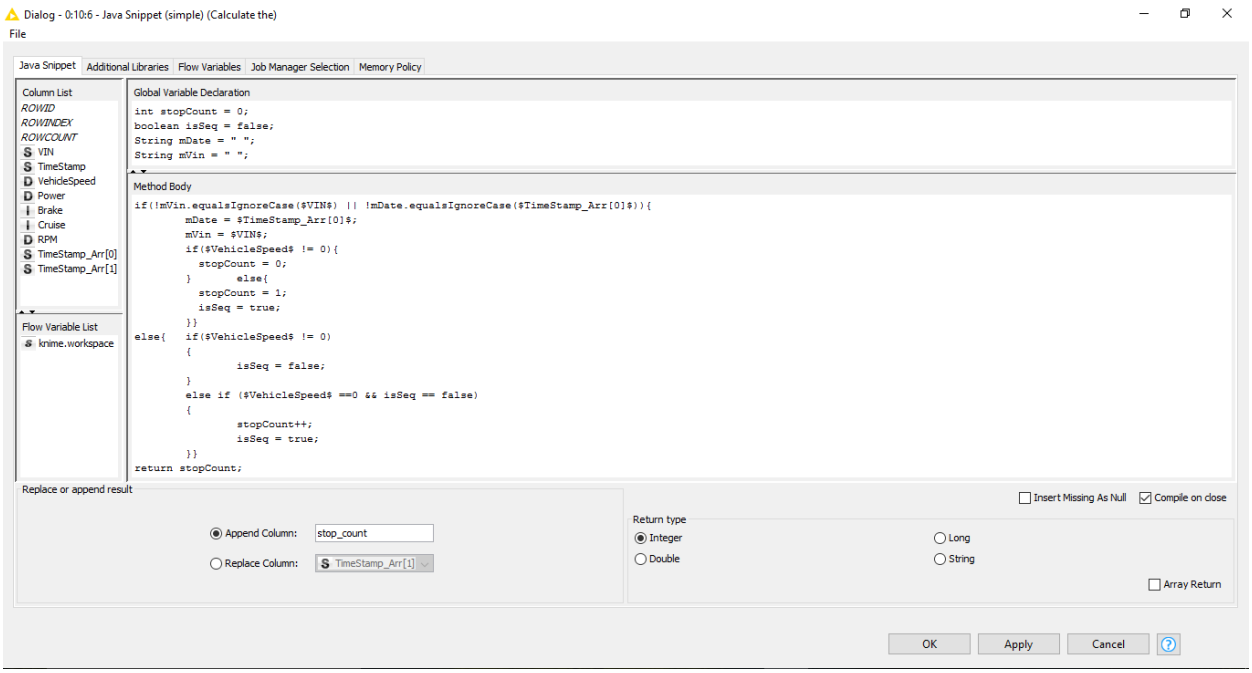

Figure 2.4: Stop count Java Snippet

CHAPTER 3

STATISTICAL ANALYSIS

After the preparation of the dataset, I needed to perform the statistical analysis to get some business insights from it. My main gain in this thesis was to find the similarity of a new vehicle if I had all the vehicle information when compared to the dataset which I had produced. I used knime for this statistical analysis.

3.1 Linear Correlation

This is a statistical method where one can find the relation or association between two variables or multiple variables. The measurement of correlation is between -1 to +1, -1 states the maximum negative correlation, +1 states the maximum positive correlation, and 0 states that there is no relation between the variables.[4]

This method is used to find the relation between all the variables and power and find the results in 3.1 I used Linear correlation node in the knime to find the association of power with other variables. Squared table view showing the pair-wise correlation values of all columns. The color range varies from dark red (strong negative correlation), over white (no correlation) to dark blue (strong positive correlation). If a correlation value for a pair of column is not available, the corresponding cell contains a missing value (shown as cross in the color view)[5].

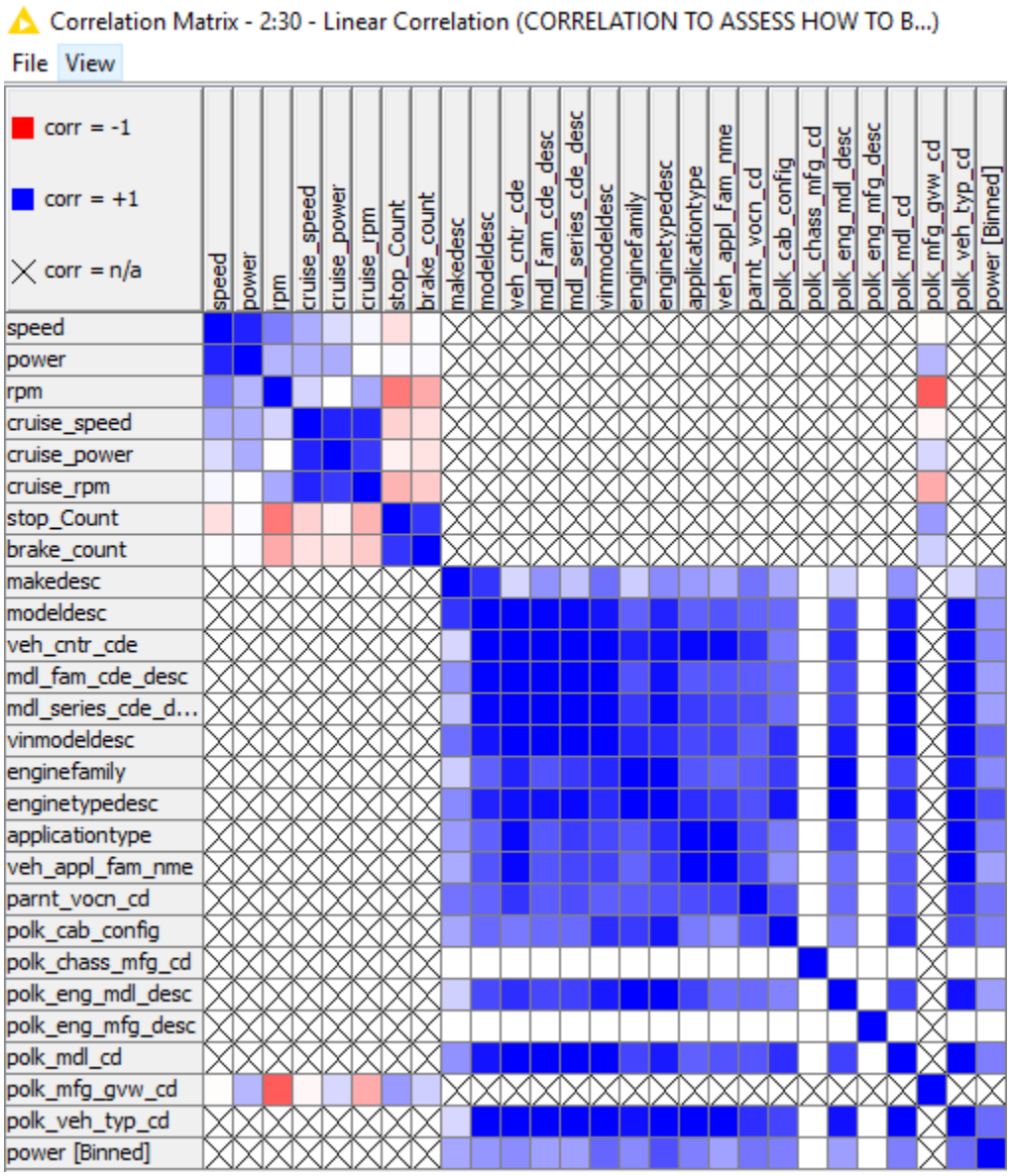

Figure 3.1: Linear Correlation Matrix between power and other variables

3.2 Polynomial Regression Predictor

This function fits a polynomial regression model to powers of a single predictor by the method of linear least squares. Interpolation and calculation of areas under the curve are also given. As a polynomial model is appropriate for our study I used the below function to fit a k order/degree polynomial to your data: $Y = b_0 + b_1X + b_2X^2 + ... + b_kX^k$ - where Y caret is the predicted outcome value for the polynomial model with regression coefficients b1 to k for each degree and Y intercept b0. The model is simply a general linear regression model[7] with k predictors raised to the power of i where i=1 to k. A second order (k=2) polynomial forms a quadratic expression (parabolic curve), a third order (k=3) polynomial forms a cubic expression and a fourth order (k=4) polynomial forms a quartic expression.[6]

I choose k=3 as it gives the best results least square error. The model is shown in 3.2

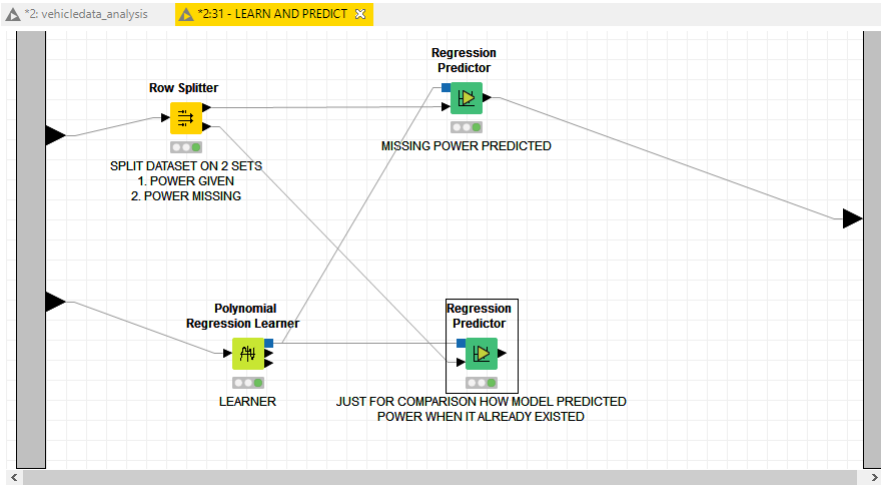

Figure 3.2: Model which shows the Power prediction

The old dataset is sent to Polynomial Regression learner, where speed, rpm, cruise_speed and cruise_power were chosen as variables for predicting the power. These variables were chosen as they are in strong positive correlation with power 3.1 and this will help the model to predict the power perfectly.

The results from regression learner and the new dataset with no power was sent to the polynomial regression predictor and the results were obtained and are shown in 3.3

Predicted data - 2:31:9 - Regression Predictor (MISSING POWER PREDICTED)

File

Table "default" - Rows: 7 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	speed	rpm	cruise_speed	cruise_power	power	Prediction (power)
A6x	13.09	1,118.935	63.356	208.851	?	60.728
A8x	35.733	1,428.519	59.109	143.353	?	99.469
A9x	19.459	1,169.331	63.512	170.436	?	61.899
A10x	13.041	1,144.657	0	0	?	48.918
A12x	30.031	1,305.751	60.601	153.422	?	88.813
A14x	39.781	1,119.723	64.382	238.776	?	141.797
A17x	39.727	1,131.973	66.787	220.107	?	137.398

Figure 3.3: Predicted Power Results

In order to check the prediction accuracy, I have sent the rows of new data set to polynomial prediction predictor in which power is already present. This helps us to directly see how power is predicted and the results obtained are shown in 3.4

Predicted data - 2:31:11 - Regression Predictor (JUST FOR COMPARISON HOW MODEL ...)

File

Table "default" - Rows: 12 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	speed	rpm	cruise_speed	cruise_power	power	Prediction (power)
A1x	40.18	1,423.196	63.105	156.864	103.701	110.924
A2x	15.092	1,155.039	55.96	165.319	60.852	60.34
A3x	39.086	1,133.145	64.402	175.787	109.351	120.225
A4x	34.32	1,460.431	67.808	208.802	107.376	111.239
A5x	39.016	1,562.131	61.06	150.019	96.275	97.136
A7x	18.935	1,199.992	55.248	135.854	53.762	61.968
A11x	21.525	1,297.468	62.494	195.776	97.239	78.226
A13x	36.272	1,310.067	60.353	181.96	120.577	118.886
A15x	39.886	1,152.916	59.273	209.813	146.01	142.206
A16x	43.53	1,629.375	60.775	195.946	154.004	119.53
A18x	23.704	1,260.135	74.752	230.74	65.884	102.057
A19x	12.091	1,090.654	57.117	108.82	40.213	36.433

Figure 3.4: Compare the predicted power with actual power

If we compare the results between predicted power and original power which is already present in the original dataset then we can see the difference. I have performed the square error of the power and performed the sum of the error which came out to be 315.615. The power of the polynomial (k value in 3.2) for which the sum of error is less is the best value to predict the values. When performed the same prediction with multiple values of k, the

least sum of error is for k=3. Hence for this power prediction we need to use polynomial regression predictor with power 3.

3.3 Similarity Search

The main goal is to find the similarity between the new vehicles and the old vehicles. This helps the business to take make many decisions, for example to quote, how much warranty to be given to the new vehicles based on the maintenance of the old vehicles which are similar to the new vehicles.

In order to find the similarity, we use similarity search node and several operations are done before that as shown in 3.5

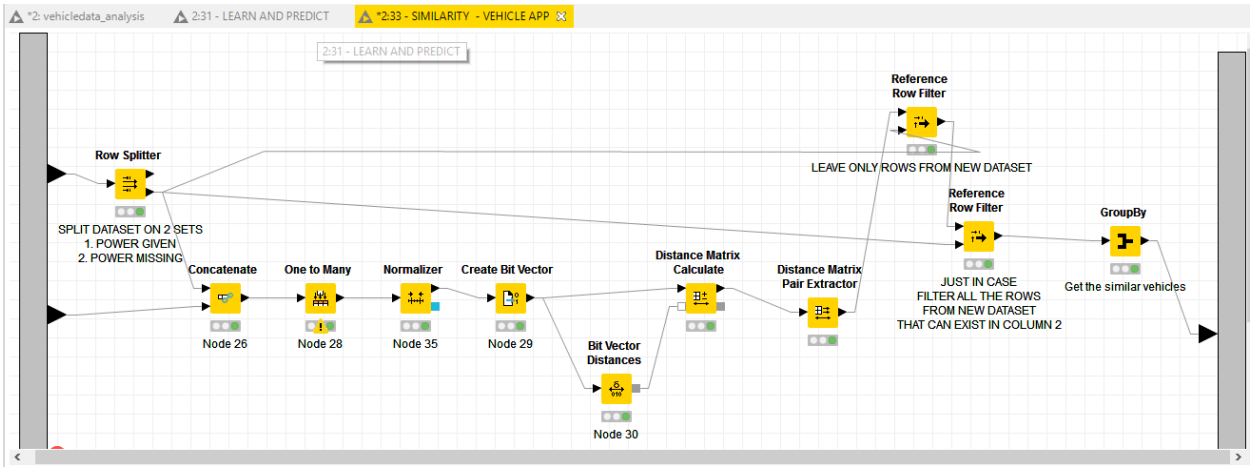

Figure 3.5: Work Flow for the similarity search

As there are many columns with string and in order to find the distance, first I needed to convert the each rows into a bit vector and then calculate the distance between the bit vectors. I have used "Tversky Distance (Tanim0to/Dice)" to calculate the distance between each new vehicle to all old vehicles which are similar to it.

The results are shown in 3.6.

Group table - 0:33:42 - GroupBy

File

Table "default" - Rows: 11 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	Object1	Unique concatenate(Object2)
Row0	A11x	A73
Row1	A13x	A44, A49
Row2	A15x	A32, A41, A68, A79, A81, A85
Row3	A16x	A39
Row4	A18x	A25, A53, A54, A61, A67
Row5	A19x	A24, A25, A33, A61, A67, A80
Row6	A1x	A23, A24, A25, A31, A33, A46, A48, A53, A54, A61, A67
Row7	A2x	A22, A45
Row8	A3x	A29, A32, A41, A52, A63, A77, A79, A81, A85
Row9	A4x	A23, A24, A25, A31, A33, A48, A53, A54, A61, A67, A80
Row10	A5x	A36, A66, A83, A89

Figure 3.6: The results of similarity search

The first column shows the new vehicles and the second column shows all the old vehicles which are similar to that respective new vehicle. The vehicles which are more similar are shown first and the order follows as the distance increases.

CHAPTER 4

CONCLUSION

This is just the beginning of analyzing the vehicle data. In this thesis I found the correlation between different variables, predicted the missing power of the new vehicles and found the similarities of the new vehicles with old vehicles which could be useful to predict the maintenance of new vehicles etc.

For the future work, we should be able to predict the power for reading of the vehicle speed and get the drive cycle out of it. We can do live monitoring of the vehicles data and see how the drive cycle is changing.

REFERENCES

[1] "Apache Hadoop. What is Apache Hadoop?" http://hadoop.apache.org/

[2] "Install Hadoop using Cloudera manager" https://www.youtube.com/watch?v=z0l87tArUDk and https://weidongzhou.wordpress.com/category/cloudera/cloudera-manager/

[3] "What is KNIME?" https://www.knime.org/about

[4] "Introduction to Correlation and Regression Analysis" http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable5.html

[5] "Linear Correlation Model Description in KNIME." https://www.knime.org/files/nodedetails/_statistics_Linear_Correlation.html

[6] "Polynomial Regression" http://www.statsdirect.com/help/default.htm#regression_and_correlation/polynomial.htm

[7] "General Linear Regression" http://www.statsdirect.com/help/content/regression_and_correlation/multiple_linear.htm

[8] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung., "The Google File System" http://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf

[9] 'The collection of papers on Big Data' http://bigdata-madesimple.com/research-papers-that-changed-the-world-of-big-data/

[10] Barna Saha, D. S., "Data Quality : The other Face of Big Data," AT&T Labs-Research.Pg1