Big Data Processing using Parallelism Techniques

Shazia Zaman

MSDS 7333 Quantifying the World, 4/20/2017

ABSTRACT

In order to process and analyze Big Data, different techniques have been introduced to perform parallel processing and loading segments of data into memory to optimize the resource management. One of the techniques to optimize processing time and resource management is Task Parallelism and Data Parallelism [1].

INTRODUCTION

In this case study, big data parallel processing technique has been used to analyze data for on-time flight performance for domestic flights being operated by major US airlines [2]. Split-Apply-Combine technique has been use in the case study in order to perform Map Reduce methodology [1]. For resource management, data has been transformed into binary data in order to access it fast and to avoid all the data into memory at once.

BACKGROUND

In Split-Apply-Combine technique, data is processed in three stages as discussed in the asynchronous material [1] as following:

- Split: Data will be split into groups [1]
- Apply: Choice of aggregation technique will be applied on each group separately [1]
- Combine: Groups will be combined again with reduced output [1]

If data is huge as big data, then loading data in virtual memory is cumbersome even with high performance machines. In this case, working with data close to physical memory on the machine can speed up the process [3]. There are number of packages are available in R that are actually written in C/C++ to deal with big data. One of such packages is bigmemory that is being used in this case study.

METHODS

I have setup the environment in AWS cloud to analyze Big Data for Airport and Flight on-time performance data. Cloud is becoming a popular medium to work with Big Data as enough memory can set for analysis time and then release later for cost efficiency.

Data acquisition and readiness

For this case study, data has been downloaded from http://stat-computing.org/dataexpo/2009/%d.csv.bz2 and then uncompressed.

In order to analyze the data in R, all the alpha or alpha-numeric values has been replaced with numeric values. The numeric value assignment for alpha or alpha-numeric fields were handled by using unique-values for following columns:

- Origin, - 3-letter airport code that flight is flying from.
- Destination – 3-letter airport code that flight in flying to.
- TailNumber – Unique alpha-numeric id provided to each plane (equipment).
  Cancellation – Indicator if flight was cancelled. Four different indicator have been used to identify if flight was cancelled due to Airline operations (A), Weather(B), Security(D) or NAS(C) [4]
- Unique Carrier – Carrier code identifying each carrier. The data is available for 29 different carrier for this case study.

Binary Data and memory allocation

Once data has been replaced from Alpha and Alpha-numeric types to numeric types, then it has been stored as binary data on the disk. For further analysis, binary data is accessed through the program for faster load and processing time. However, there is one flaw that I have noticed with mapping file unique_mapping.p. As instructed in the asynchronous material [1], unique values for origin, destination, cancellation code, tail number and unique carrier code were saved into mapping file for later access. I have noticed that either library for pickle, cPickle or _pickle based on python working is not providing correct mapping. As a result, it was mapping the airport correctly. I have used the unique values directly from .csv files and keep them in cache for later used in the analysis, and then it is displaying the correct airport ORD being the busiest airport to Origin.

RESULTS

After preparing the data, I have collected some initial statistics about the airline on-time performance data as following:

- Number of unique Origins = 347
- Number of unique Destinations = 352
- Number of unique planes (equipment wit unique Tail Number) = 13537
- Data acquisition for the years – 1987 - 2008

Overall top ten busiest airport for domestic flight departures:

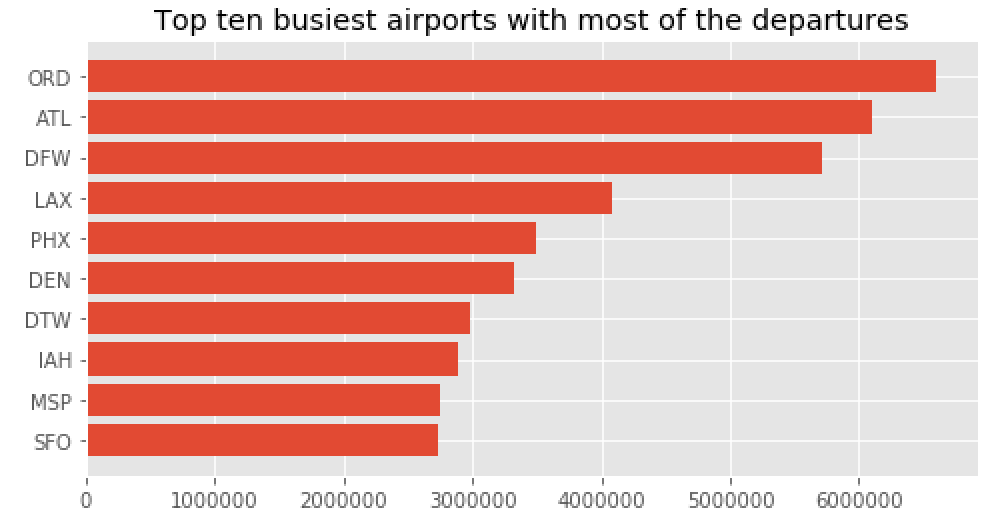
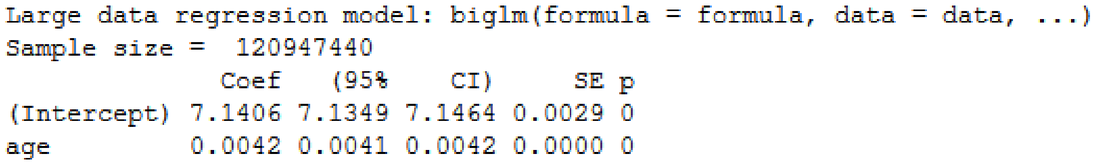


Figure 1

From the chart in Figure1, it is clear that ORD, Chicago O'Hare airport, is the busiest airport for domestic flight departures based on the report collected over year 1987 to 2008, followed by ATL (Atlanta) and DFW (Dallas/Fort Worth).

Additional statistics are collected as instructed in the asynchronous material [1] a following:

- Youngest plane that has started flying in 2009 and has a tail number of N824SK.
- Association between Age and arrival delay status of the plane:



```
Large data regression model: biglm(formula = formula, data = data, ...)
Sample size =  120947440
                Coef    (95%     CI)      SE p
(Intercept) 7.1406 7.1349 7.1464 0.0029 0
age         0.0042 0.0041 0.0042 0.0000 0
```

Figure 2

As it has not factored out other delay condition as weather related, delayed departure from origin airport due to security or NAS, this regression model is showing weak relationship between age and delayed arrival status.

This leads to the next question being asked for this case study. Further analysis should be performed on Origin(s) where flights are departing from and getting delayed to arrive at the destinations.

Which airports are most likely to be delayed flying out of or into?

In order to answer this questions, I have analyzed the data in couple of ways.

First I have checked the airports with most for the flight departure delays and come up with following results:

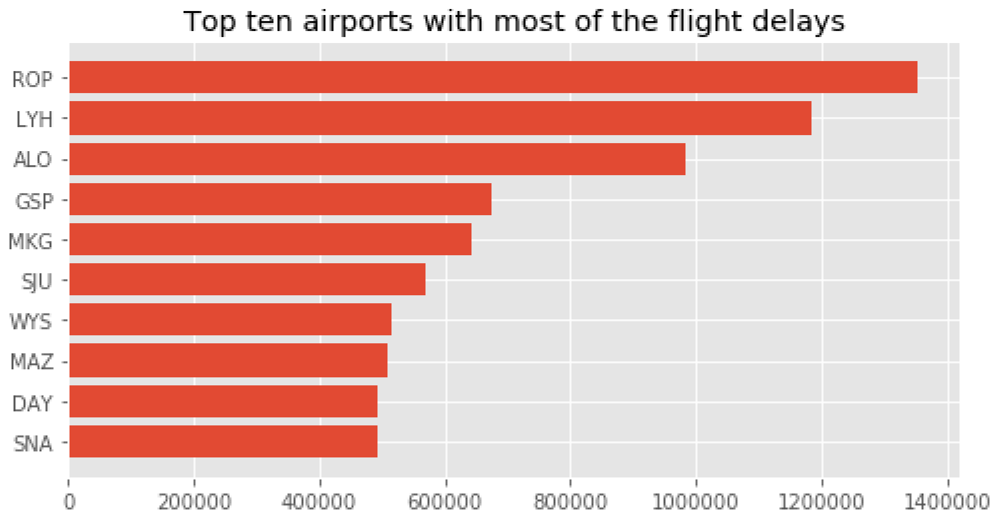


Figure 3

And then follow the results with airports with least number of flight departure delays as following:

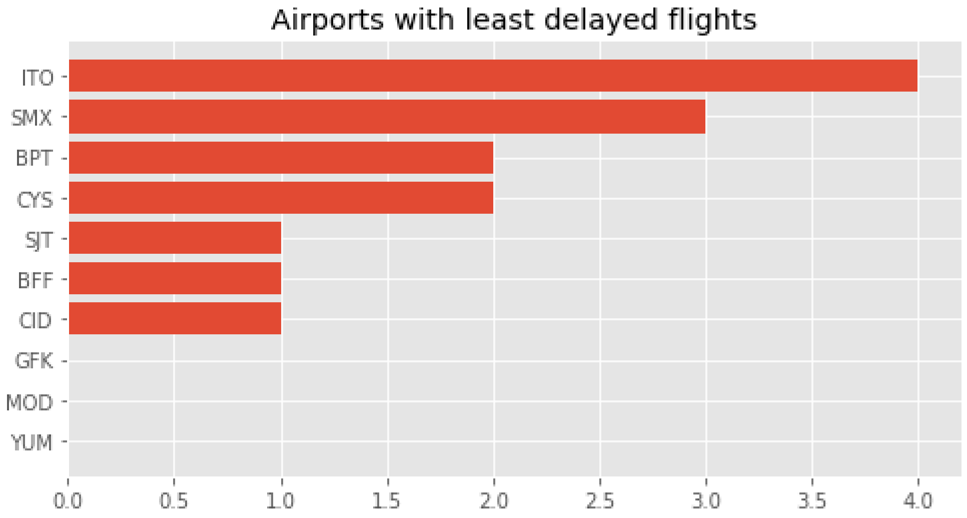


Figure 4

However, in figure 3 and 4, the airports being reported are not very popular airport and the overall operation on these airports might be significantly lower than other busiest or more popular airport in USA. The only relevant result from above chart for airports with high number of flight departure delays is displayed for airports SNA (Orange County, CA) and DAY (Dayton, Ohio).

In order to perform more analysis for possible delays from airport, I will analyze the result for airports with high number of flight operations and then compare the results among them. Going back to Figure1 that displayed the airports with highest number of flight departure, I will analyze the flight departure delays among those airport.

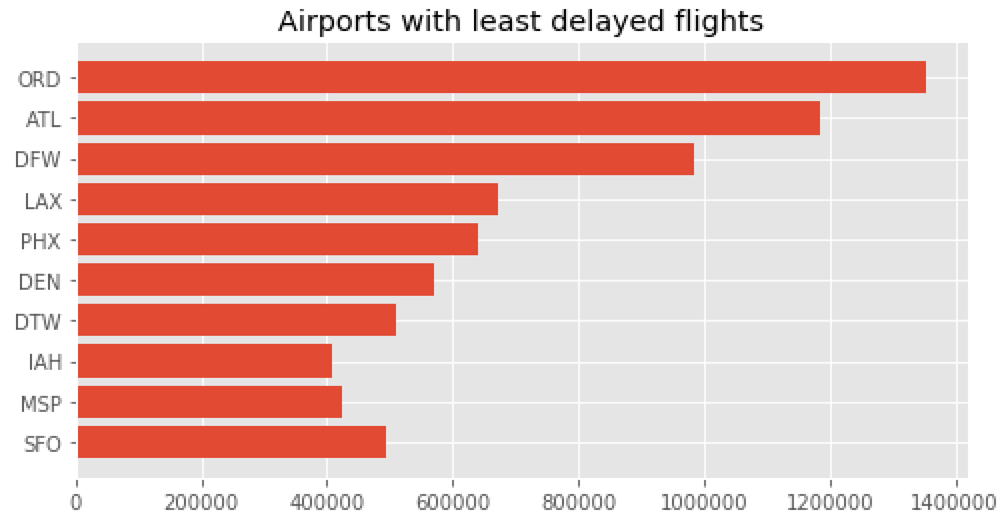


Figure 5

Figure 5 shows the flight departure delay statistics from top ten busiest airport. However, it is better to compare the number of flight departure delays with total number of flight departures from these airports as showing in figure 6 below:

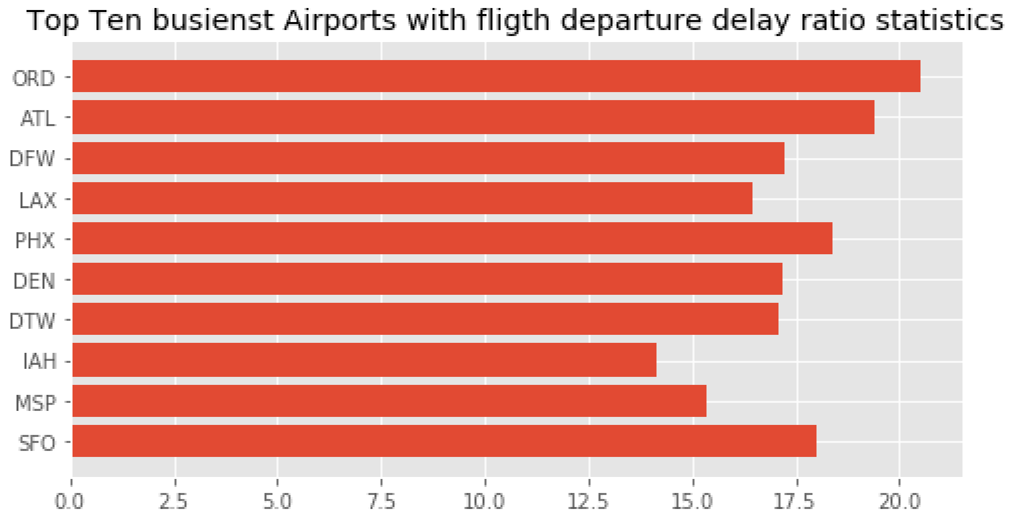


Figure 6

From Figure 6, it is evident from historical data that >20% of flights departing from ORD (Chicago) were delayed followed by next highest departure delay ratio from ATL (Atlanta). It is also evident from this chart, that even though DFW (Dallas/Fort Worth) is the 3rd busiest airport from Figure 1, it is not the 3rd airport with highest ratio of departure delays. PHX (Phoenix) is showing up as 3rd airport in regards to departure delay ratio.

So, it is more likely if a passenger is flying from ORD, his/her flight will be delayed as compare to one flying from IAH (Houston).

Which flights with same origin and destination are most likely to be delayed?

In order to solve this problem, I will consider the flights that are flying into the airport and then taking off from same airport. Most of the time these are connecting flights. As if flight is schedule from JFK (New York) to DFW (Dallas/Fort worth) and then heading out to IAH (Houston). In other case, there might be flight shuttle service between two airport, for example between DFW (Dallas/Fort Worth) and ORD (Chicago). Almost all the flights that are arriving at the airport will depart from the airport unless if the airplane (equipment) will be taken off the service for maintenance.

The concern is that if flight is arriving late at the airport, is it already a victim of delayed departure for next destination? Most of the time, the airline airport crew at the airport, both ground crew and gate agents, works efficiently to turn back the flight for next departure without causing any more delay. However, it is also based on other factors, as security, weather, and flight crew availability.

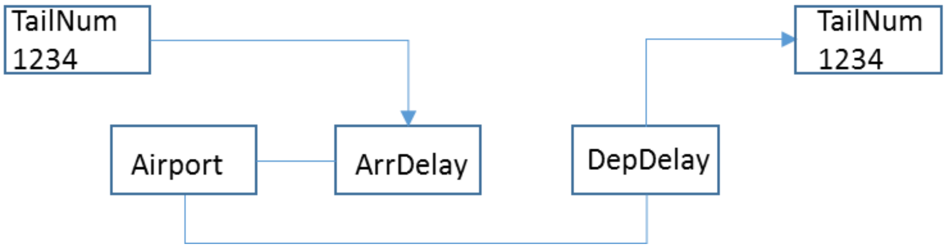
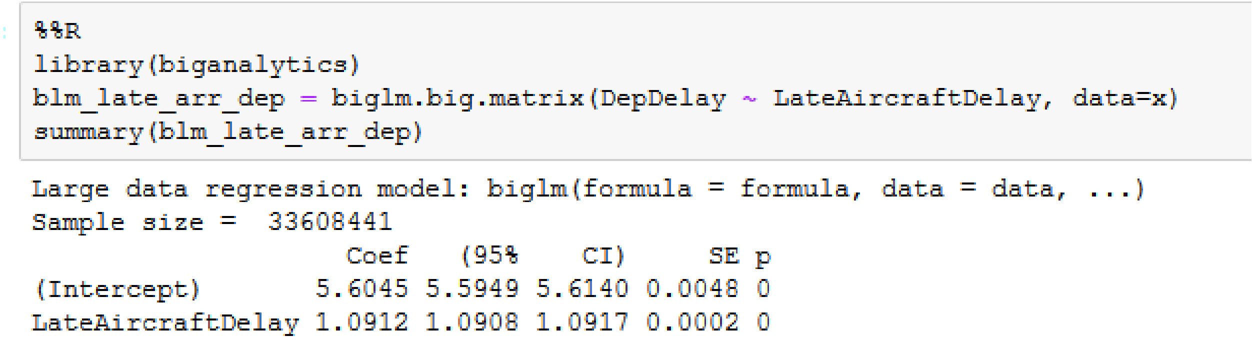


Figure 7

There is a column for LateAircraftDelay in the dataset that defines the delay time in minutes when the aircraft has arrived into the airport before next departure from the airport.



```
%%R
library(biganalytics)
blm_late_arr_dep = biglm.big.matrix(DepDelay ~ LateAircraftDelay, data=x)
summary(blm_late_arr_dep)
```

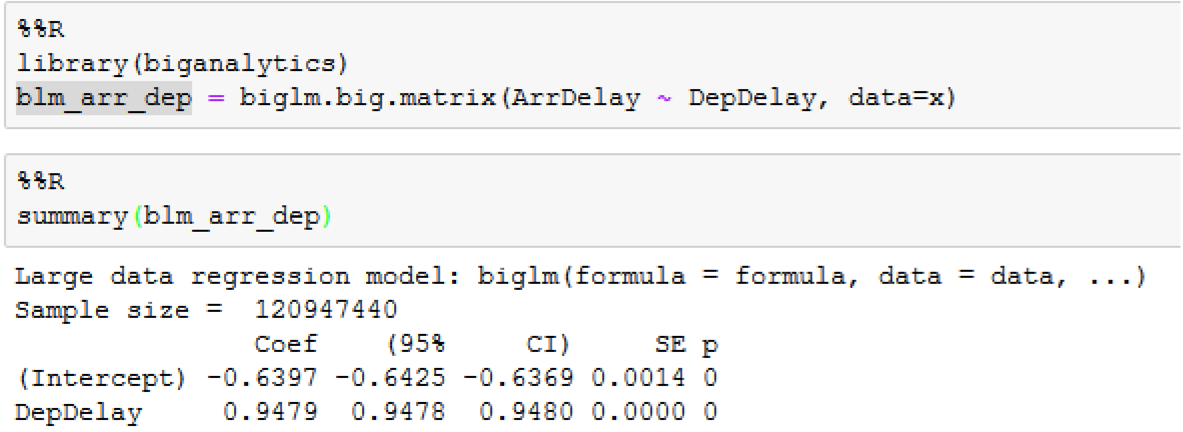
```
Large data regression model: biglm(formula = formula, data = data, ...)
Sample size =  33608441
                    Coef   (95%    CI)     SE p
(Intercept)       5.6045 5.5949 5.6140 0.0048 0
LateAircraftDelay 1.0912 1.0908 1.0917 0.0002 0
```

Figure 8

From the model above, the late arrival of aircraft at the airport is not highly significant to cause the departure delay for next flight as P-value is 0. So, the case that airline usually makeup for late arrived flight to get it ready for next on-time departure from the airport.

Can you regress how delayed a flight will be before it is delayed?

First and foremost response is that if flight is late to depart from origin airport then it will late to arrive at destination airport. In order to proof this theory, I did run the regression model for ArrDelay vs. DepDelay as following:



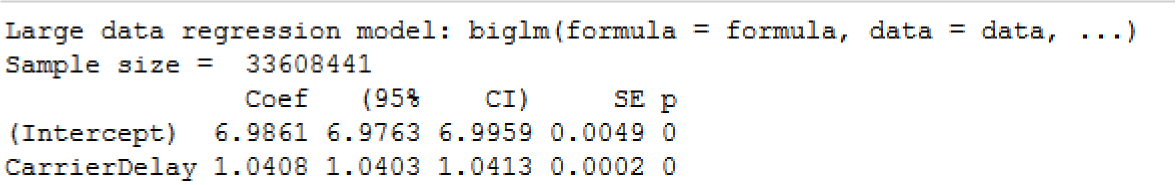
```
%%R
library(biganalytics)
blm_arr_dep = biglm.big.matrix(ArrDelay ~ DepDelay, data=x)
```

```
%%R
summary(blm_arr_dep)
```

```
Large data regression model: biglm(formula = formula, data = data, ...)
Sample size =  120947440
               Coef    (95%     CI)     SE p
(Intercept) -0.6397 -0.6425 -0.6369 0.0014 0
DepDelay     0.9479  0.9478  0.9480 0.0000 0
```

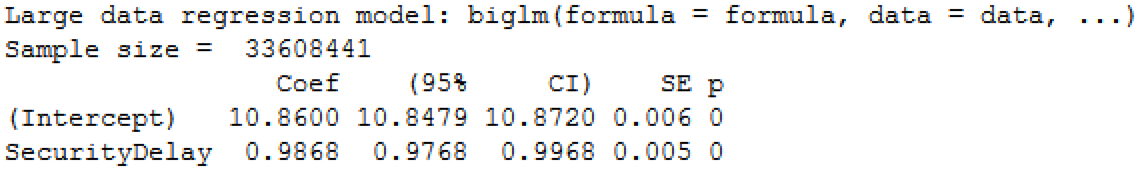
Figure 9

Next question is to measure if >= 15 minutes departure delay at the departure station (origin) will cause same number of minutes delay at the arrival station (destination). In order to test this, I have created a regression model for depDelay (departure delay) vs. LateAircraftDelay that is number of minutes that aircraft has arrived late at the airport for next flight. As mentioned in the last subsection, the departure delay is highly correlated the late aircraft arrival at the airport as mentioned in Figure 9. Next I have checked the correlation of flight departure with carrier delay, security delay, weather related delay and NAS delay and came up with following results:



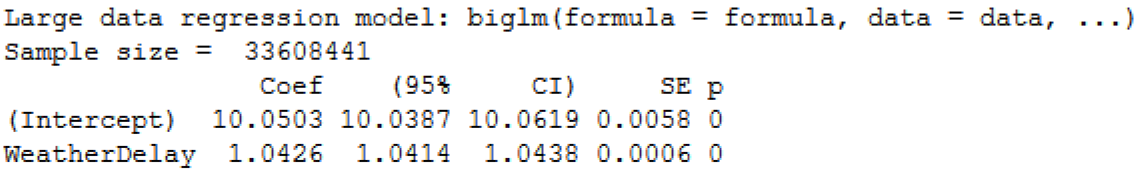
```
Large data regression model: biglm(formula = formula, data = data, ...)
Sample size =  33608441
               Coef   (95%    CI)     SE p
(Intercept)  6.9861 6.9763 6.9959 0.0049 0
CarrierDelay 1.0408 1.0403 1.0413 0.0002 0
```

Figure 10



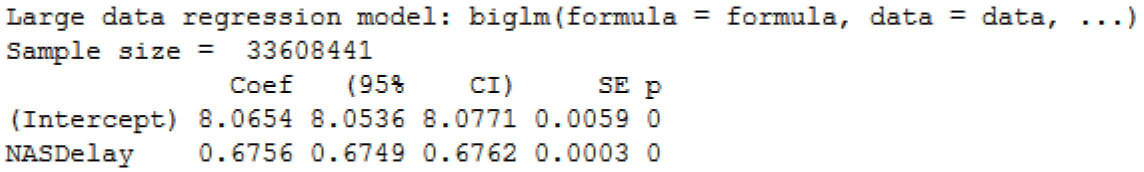
```
Large data regression model: biglm(formula = formula, data = data, ...)
Sample size =  33608441
                 Coef    (95%     CI)    SE p
(Intercept)   10.8600 10.8479 10.8720 0.006 0
SecurityDelay  0.9868  0.9768  0.9968 0.005 0
```

Figure 11



```
Large data regression model: biglm(formula = formula, data = data, ...)
Sample size =  33608441
                Coef    (95%     CI)     SE p
(Intercept)  10.0503 10.0387 10.0619 0.0058 0
WeatherDelay  1.0426  1.0414  1.0438 0.0006 0
```

Figure 12

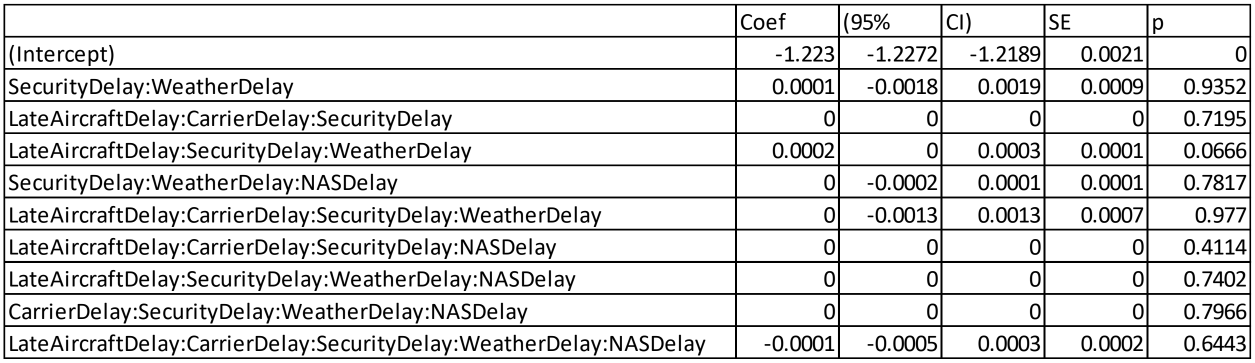


```
Large data regression model: biglm(formula = formula, data = data, ...)
Sample size =  33608441
              Coef   (95%    CI)     SE p
(Intercept) 8.0654 8.0536 8.0771 0.0059 0
NASDelay    0.6756 0.6749 0.6762 0.0003 0
```

Figure 13

So, from above test, flight departure delay is not significantly correlated to any single delay caused as discussed above. However, it might be combination of different caused that can result into flight departure delay.

After combing all the variables as discussed above, I have created a model and came up with following results that have p-value > 0.05:



	Coef	(95%	CI)	SE	p
(Intercept)	-1.223	-1.2272	-1.2189	0.0021	0
SecurityDelay:WeatherDelay	0.0001	-0.0018	0.0019	0.0009	0.9352
LateAircraftDelay:CarrierDelay:SecurityDelay	0	0	0	0	0.7195
LateAircraftDelay:SecurityDelay:WeatherDelay	0.0002	0	0.0003	0.0001	0.0666
SecurityDelay:WeatherDelay:NASDelay	0	-0.0002	0.0001	0.0001	0.7817
LateAircraftDelay:CarrierDelay:SecurityDelay:WeatherDelay	0	-0.0013	0.0013	0.0007	0.977
LateAircraftDelay:CarrierDelay:SecurityDelay:NASDelay	0	0	0	0	0.4114
LateAircraftDelay:SecurityDelay:WeatherDelay:NASDelay	0	0	0	0	0.7402
CarrierDelay:SecurityDelay:WeatherDelay:NASDelay	0	0	0	0	0.7966
LateAircraftDelay:CarrierDelay:SecurityDelay:WeatherDelay:NASDelay	-0.0001	-0.0005	0.0003	0.0002	0.6443

Figure 14

For most of the combinations of variables the coefficient is zero, so those combination will be out of the model. As a result the flight departure delay minutes can be predicted with following model:

DepDelay = -1.223 + 0.0001(SecurityDelay * WeatherDelay) + 0.0002(LateAircraftDelay * SecurityDelay * WeatherDelay) -0.0001(LateAircraftDelay * CarrierDelay * SecurityDelay * WeatherDelay * NASDelay)

I am not able to create train and test data for big data, so I am not able to test the model and then perform cross validation.

CONCLUSION

Analysis and processing of Big Data requires special resources and techniques. As a resource management, I will recommend to use the cloud preferably AWS cloud in order to allocate enough resources during the analysis and then release them once analysis is complete. Using special packages as bigmemory helps a lot to perform analysis in faster processing time. In the start of the analysis, data acquisition and readiness takes a while, however it is really beneficial to store the data as binary file and then use it to continue with analysis. In this case study for airline on-time performance, the combination for late aircraft arrival, security delay, and weather related delay and to some extent NAS delay play a role in order to cause departure delay for next flight. Interestingly, carrier delay is not included in this model.

REFERENCES

[1] E. Larson, "MSDS7333 Quatifying the World - Unit 13,14," SMU, 08 01 2017. [Online]. [Accessed 2017].

[2] "Airline On-Time Statistics and Delay Causes," Bureau of Transportation Statistics, [Online]. Available: https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp.

[3] M. J. Kane, P. Haverty, J. W. Emerson and C. J. Determan, "Manage Massive Matrices with Shared Memory and Memory-Mapped," 28 3 2016. [Online]. Available: https://cran.r-project.org/web/packages/bigmemory/bigmemory.pdf.

[4] "Data expo 09 - Get the data," Statistical Computing Statistical Graphics, [Online]. Available: http://stat-computing.org/dataexpo/2009/the-data.html.

[5] D. Temple, "Code for Case Study Chapters," duncan@r-project.org, 2015.

[6] T. Duncan and N. Deborah, "Strategies for Analyzing a 12-Gigabyte Data Set: Airline Flight Delays," in *Data Science in R*, Boca Raton, CRC Press, 2015, p. 539.