**WQD7007 BIG DATA MANAGEMENT**

**GROUP PROJEVT REPORT**


**<u>GROUP MEMBERS:</u>**

**WQD170002 – GUNASEGARRAN**

**WQD170035 - MUHAMMAD NOORAIZAD**


# BIG DATA IN TELECOMMUNICATION INDUSTRY


**FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY**

**UNIVERSITY MALAYA**


**2018**

## 1. Introduction

A handful off industries needs to gain more from big data compared to telecommunications. For decades, communications service providers have delivered and captured huge volumes of information about calling patterns, wireless data usage, location data, network bandwidth statistics, and even the individual apps and webpages accessed by customer on their mobile devices, until recently, there is no effective and efficient way to dig value from it and to fund in storing is in a higher side, therefore, plenty of that data was discarded.

All that is changing, through the combination of streaming analytics and analytics at scale, Big data technologies are enabling telecommunications companies to uncover significant new insights about their infrastructure and their customers.

## 2. When the big data is needed?

Telecommunication industry have been moving numerous terabytes of information around their systems for some decades. These are the key point required when big data is needed :

### 2.1 Volume.

The volume of operational data generated with every call or session is increasing tenfold because of LTE/4G in mobile networks,. The frequent usage of GPS, location-based services, and social media is adding to the torrent of data. Finally, the advent of IPv6 will create as many IP addresses as possible, to ensure and allowing the number of Internet-connected devices to grow exponentially. This volume of data needs new real-time operational capabilities for function such as real-time charging and event-based marketing—and new tools for mediating, managing, and archiving data within available time frames needs increased data storage for compliance and potential future uses.

### 2.2 . Variety

Social media, mobile devices, and sensors that monitor and over view everything from utility use to medical compliance are telecommunication infrastructures with data in myriad formats. Before they can analyse it for significant subscriber insight and new business opportunities, it is important for telco's to enrich their CDR data with location-based services, financial information, and other unstructured data, then standardize it for business intelligence platforms.

### 2.3 Volume

Telecommunication provider must integrate their legacy in operational and business systems that still prolong in years of useful life with new environments. They must support batch to right- time data to access applications such as real-time CRM while delivering their own cloud based services and support from other vendors concurrently. They also need complex event processing systems that can handle data volumes that are substantial and complex for human response. And all this must be done while ensuring data quality and accessibility across multiple solutions for regulatory compliance.

## 3. Key obstacles faced by telecommunication industry.

### 3.1 Handle Large Volumes of Data

Transforming, analysing, and integrating the vast amounts of data generated by 4G networks, CDRs, clickstreams, IPv6 devices, location sensors, and machine-to-machine monitors in a single format information platform, therefore, Telco's require technology. The technology must integrate data in near real time, scale cost-effectively and integrate with legacy systems and technologies, and shrink batch windows for high performance.

### 3.2 Utilize the Variety of Data

Web, social, and machine monitor device data—and provide easy, consistent access to all types of interaction data and to ensure this, Telco's must have the capability to transform and analyse data from multiple sources and formats—including unstructured mobile.

### 3.3 Manage the Volume of Data

The data processing across platforms, integrating big data with legacy systems at the data level both on premise and in the cloud needs to be optimize by Telco's. They must prove that they are identifying, masking, and managing sensitive data for regulatory compliance at the same time.

## 4. Using Big Data Technology in Telecommunication Industry

With the advancement of technology, there are many tools developed to deal with big data problem. All these tools and software aim to solve the problems faced by telecommunication company. One of the widely used tool is Hadoop. Hadoop is an open source distributed processing framework that manages data processing and storage for big data applications running in clustered and parallel system. Hadoop is the focal point of big data storage technology. It is used to support advanced analytics initiatives, including predictive analytics, data mining and machine learning applications. Hadoop can handle variety of data types including structured and unstructured data type. By deploying Hadoop in their big data storage system, users gains more flexibility for collecting, processing and analysing data compared to relational databases and data warehouses.

One of the component in Hadoop is Hive. Hive serves as a data warehousing structure for Hadoop. Hive provides data summarization, query and analysis. It is used to store and analysis large datasets stored in Hadoop's Distributed File System (HDFS). Hive provides support for SQL-like query access to structured data. This is especially important in data mining process.

As telecommunication industry deals with large and vast amount of data, Hive can help in storing the data. Not only storing, Hive enable users to access and analyse the data fast. Hive enable analysis perform the calculations and processing at the database end at considerably fast time. The results can be transmit in an efficient, fast and secure manner. It also prevents unnecessary clogging of bandwidth which can be efficiently used for other more important network oriented operations.

Likewise, problem with handling large variety of data can be tackled with the implementation of Hive as a sole database framework. With this, business users can benefit from the fast and accurate results. Because sometimes they need the information to make decision. For example, to innovate and to design new product, it is important to analyse the usage history and forecast

the next generation of products which the customers are likely to expect and be ready with them as soon as the demand arises. The same goes to the problem of the complexity of the data, Hive enable fast integration between platforms.

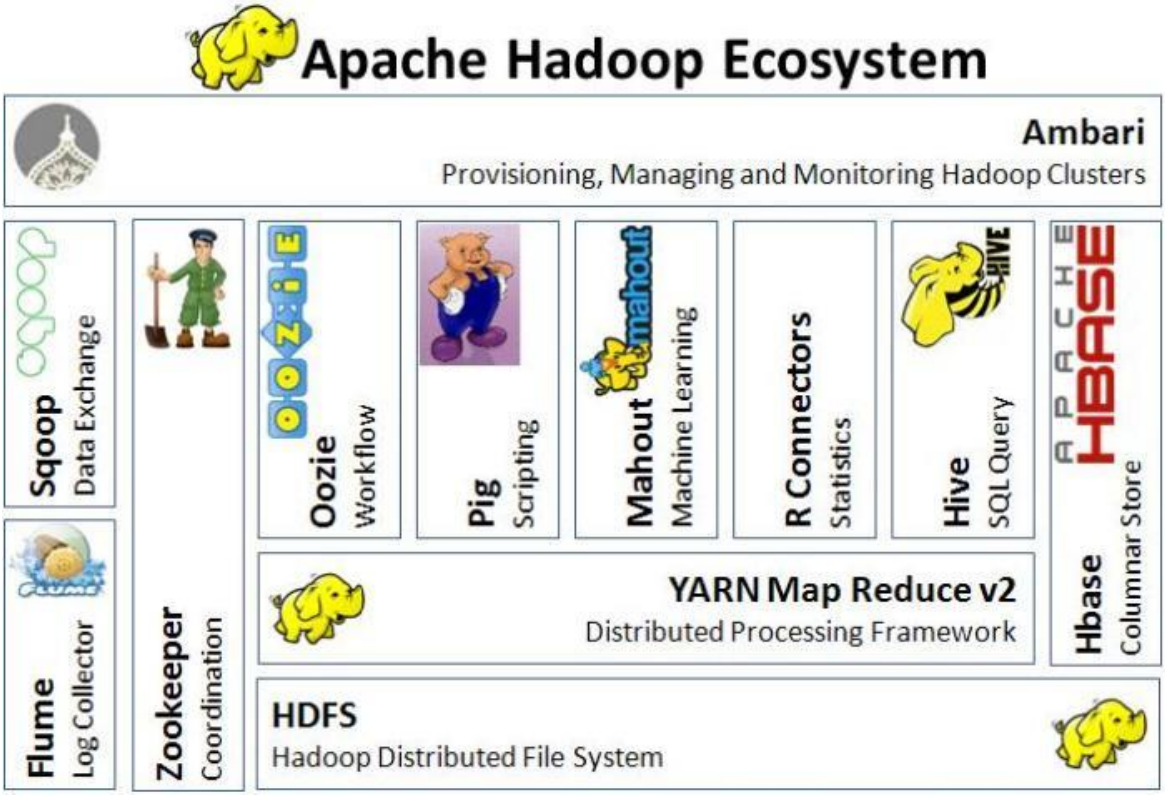



Figure 1: Apache Hadoop Ecosystem.

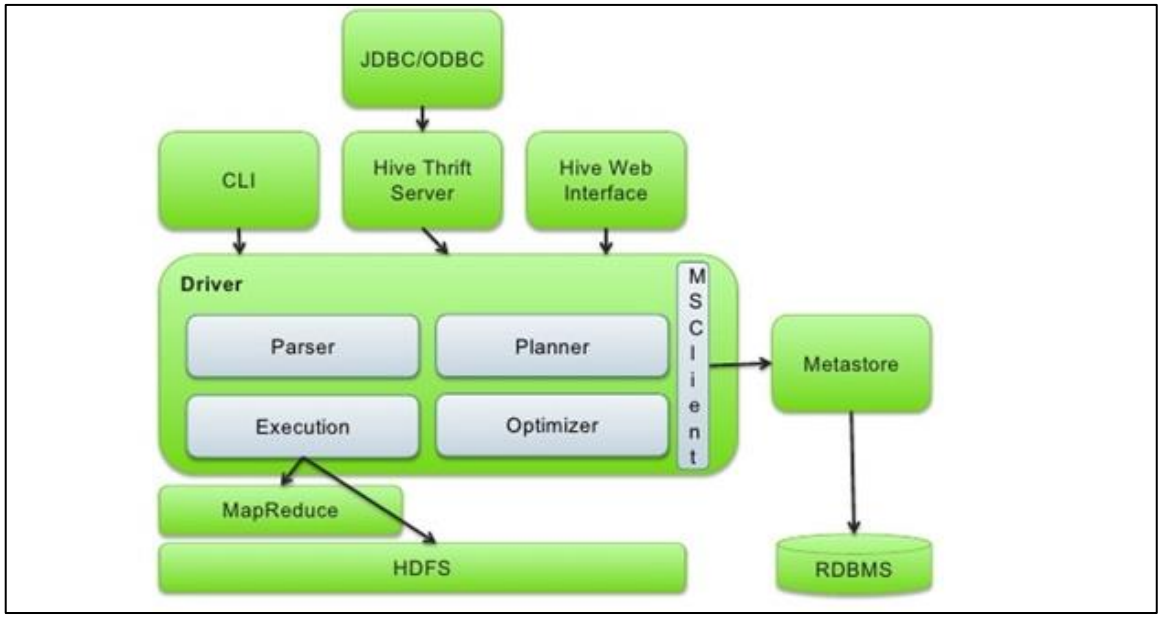



Figure 2 : Apache HIVE Architecture.

## 5. Sample Dataset

The sample dataset use analyst the problem statement, we have used Direct Carrier Billier dataset, owned by webe digital sdn bhd. The data set concise of 12 files of dataset in csv format, and all these file are been used for education purpose with permission of webe digital for only displaying the result of performance.

*(Gunasegarran)*

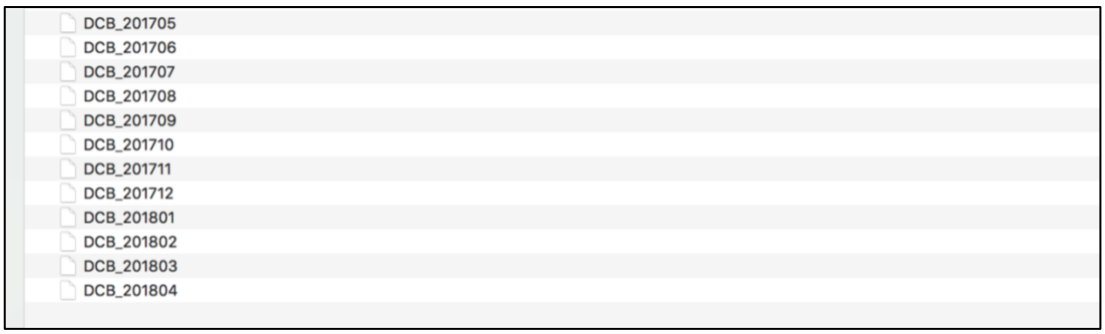
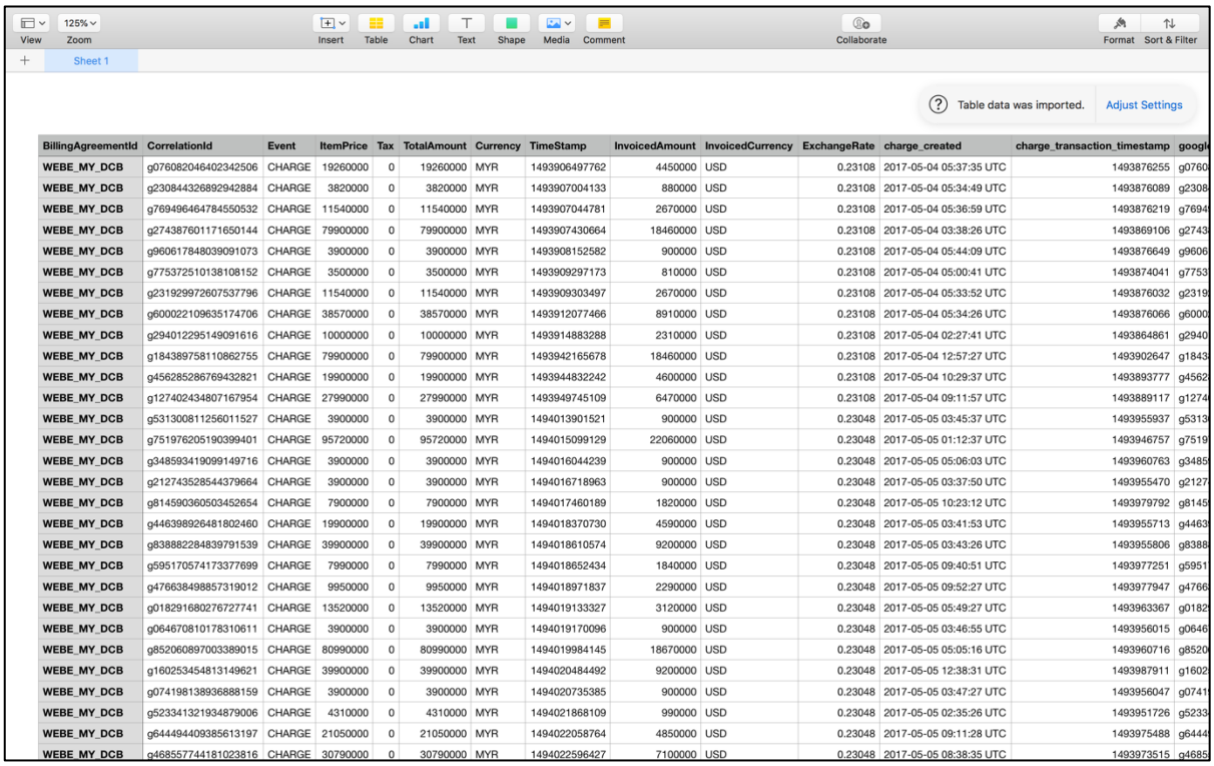



Figure 3: List of dataset used form webe digital.




| BillingAgreementId | CorrelationId | Event | ItemPrice | Tax | TotalAmount | Currency | TimeStamp | InvoicedAmount | InvoicedCurrency | ExchangeRate | charge_created | charge_transaction_timestamp | googl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| WEBE_MY_DCB | g076082046402342506 | CHARGE | 19260000 | 0 | 19260000 | MYR | 1493906497762 | 4450000 | USD | 0.23108 | 2017-05-04 05:37:35 UTC | 1493876255 | g0760 |
| WEBE_MY_DCB | g230844326892942884 | CHARGE | 3820000 | 0 | 3820000 | MYR | 1493907004133 | 880000 | USD | 0.23108 | 2017-05-04 05:34:49 UTC | 1493876089 | g2308 |
| WEBE_MY_DCB | g769496464784550532 | CHARGE | 11540000 | 0 | 11540000 | MYR | 1493907044781 | 2670000 | USD | 0.23108 | 2017-05-04 05:36:59 UTC | 1493876219 | g7694 |
| WEBE_MY_DCB | g274387601171650144 | CHARGE | 79900000 | 0 | 79900000 | MYR | 1493907430664 | 18460000 | USD | 0.23108 | 2017-05-04 03:38:26 UTC | 1493869106 | g2743 |
| WEBE_MY_DCB | g960617848039091073 | CHARGE | 3900000 | 0 | 3900000 | MYR | 1493908152582 | 900000 | USD | 0.23108 | 2017-05-04 05:44:09 UTC | 1493876649 | g9606 |
| WEBE_MY_DCB | g775372510138108152 | CHARGE | 3500000 | 0 | 3500000 | MYR | 1493909297173 | 810000 | USD | 0.23108 | 2017-05-04 05:00:41 UTC | 1493874041 | g7753 |
| WEBE_MY_DCB | g231929972607537796 | CHARGE | 11540000 | 0 | 11540000 | MYR | 1493909303497 | 2670000 | USD | 0.23108 | 2017-05-04 05:33:52 UTC | 1493876032 | g2319 |
| WEBE_MY_DCB | g600022109635174706 | CHARGE | 38570000 | 0 | 38570000 | MYR | 1493912077466 | 8910000 | USD | 0.23108 | 2017-05-04 05:34:26 UTC | 1493876066 | g6000 |
| WEBE_MY_DCB | g294012295149091616 | CHARGE | 10000000 | 0 | 10000000 | MYR | 1493914883288 | 2310000 | USD | 0.23108 | 2017-05-04 02:27:41 UTC | 1493864861 | g2940 |
| WEBE_MY_DCB | g184389758110862755 | CHARGE | 79900000 | 0 | 79900000 | MYR | 1493942165678 | 18460000 | USD | 0.23108 | 2017-05-04 12:57:27 UTC | 1493902647 | g1843 |
| WEBE_MY_DCB | g456285286769432821 | CHARGE | 19900000 | 0 | 19900000 | MYR | 1493944832242 | 4600000 | USD | 0.23108 | 2017-05-04 10:29:37 UTC | 1493893777 | g4562 |
| WEBE_MY_DCB | g127402434807167954 | CHARGE | 27990000 | 0 | 27990000 | MYR | 1493949745109 | 6470000 | USD | 0.23108 | 2017-05-04 09:11:57 UTC | 1493889117 | g1274 |
| WEBE_MY_DCB | g531300811256011527 | CHARGE | 3900000 | 0 | 3900000 | MYR | 1494013901521 | 900000 | USD | 0.23048 | 2017-05-05 03:45:37 UTC | 1493955937 | g5313 |
| WEBE_MY_DCB | g751976205190399401 | CHARGE | 95720000 | 0 | 95720000 | MYR | 1494015099129 | 22060000 | USD | 0.23048 | 2017-05-05 01:12:37 UTC | 1493946757 | g7519 |
| WEBE_MY_DCB | g348593419099149716 | CHARGE | 3900000 | 0 | 3900000 | MYR | 1494016044239 | 900000 | USD | 0.23048 | 2017-05-05 05:06:03 UTC | 1493960763 | g3485 |
| WEBE_MY_DCB | g212743528544379664 | CHARGE | 3900000 | 0 | 3900000 | MYR | 1494016718963 | 900000 | USD | 0.23048 | 2017-05-05 03:37:50 UTC | 1493955470 | g2127 |
| WEBE_MY_DCB | g814590360503452654 | CHARGE | 7900000 | 0 | 7900000 | MYR | 1494017460189 | 1820000 | USD | 0.23048 | 2017-05-05 10:23:12 UTC | 1493979792 | g8145 |
| WEBE_MY_DCB | g446398926481802460 | CHARGE | 19900000 | 0 | 19900000 | MYR | 1494018370730 | 4590000 | USD | 0.23048 | 2017-05-05 03:41:53 UTC | 1493955713 | g4463 |
| WEBE_MY_DCB | g838882284839791539 | CHARGE | 39900000 | 0 | 39900000 | MYR | 1494018610574 | 9200000 | USD | 0.23048 | 2017-05-05 03:43:26 UTC | 1493955806 | g8388 |
| WEBE_MY_DCB | g595170574173377699 | CHARGE | 7990000 | 0 | 7990000 | MYR | 1494018652434 | 1840000 | USD | 0.23048 | 2017-05-05 09:40:51 UTC | 1493977251 | g5951 |
| WEBE_MY_DCB | g476638498857319012 | CHARGE | 9950000 | 0 | 9950000 | MYR | 1494018971837 | 2290000 | USD | 0.23048 | 2017-05-05 09:52:27 UTC | 1493977947 | g4766 |
| WEBE_MY_DCB | g018291680276727741 | CHARGE | 13520000 | 0 | 13520000 | MYR | 1494019133327 | 3120000 | USD | 0.23048 | 2017-05-05 05:49:27 UTC | 1493963367 | g0182 |
| WEBE_MY_DCB | g064670810178310611 | CHARGE | 3900000 | 0 | 3900000 | MYR | 1494019170096 | 900000 | USD | 0.23048 | 2017-05-05 03:46:55 UTC | 1493956015 | g0646 |
| WEBE_MY_DCB | g852060897003389015 | CHARGE | 80990000 | 0 | 80990000 | MYR | 1494019984145 | 18670000 | USD | 0.23048 | 2017-05-05 05:05:16 UTC | 1493960716 | g8520 |
| WEBE_MY_DCB | g160253454813149621 | CHARGE | 39900000 | 0 | 39900000 | MYR | 1494020484492 | 9200000 | USD | 0.23048 | 2017-05-05 12:38:31 UTC | 1493987911 | g1602 |
| WEBE_MY_DCB | g074198138936888159 | CHARGE | 3900000 | 0 | 3900000 | MYR | 1494020735385 | 900000 | USD | 0.23048 | 2017-05-05 03:47:27 UTC | 1493956047 | g0741 |
| WEBE_MY_DCB | g523341321934879006 | CHARGE | 4310000 | 0 | 4310000 | MYR | 1494021868109 | 990000 | USD | 0.23048 | 2017-05-05 02:35:26 UTC | 1493951726 | g5233 |
| WEBE_MY_DCB | g644494409385613197 | CHARGE | 21050000 | 0 | 21050000 | MYR | 1494022058764 | 4850000 | USD | 0.23048 | 2017-05-05 09:11:28 UTC | 1493975488 | g6444 |
| WEBE_MY_DCB | g468557744181023816 | CHARGE | 30790000 | 0 | 30790000 | MYR | 1494022596427 | 7100000 | USD | 0.23048 | 2017-05-05 08:38:35 UTC | 1493973515 | g4685 |

Figure 4: Raw sample of dataset for Direct Carrier Billing (DCB)

## 6. Data Preparation and Data Cleaning using Apache Hadoop and Apache HIVE

For the this assessment we use Apache Hadoop in Hortanwork Data Flow sandbox, under virtual studio. While Apache HIVE in the Hadoop environment to perform our data preparation and data cleanings. Below are the step been used to perform our works:

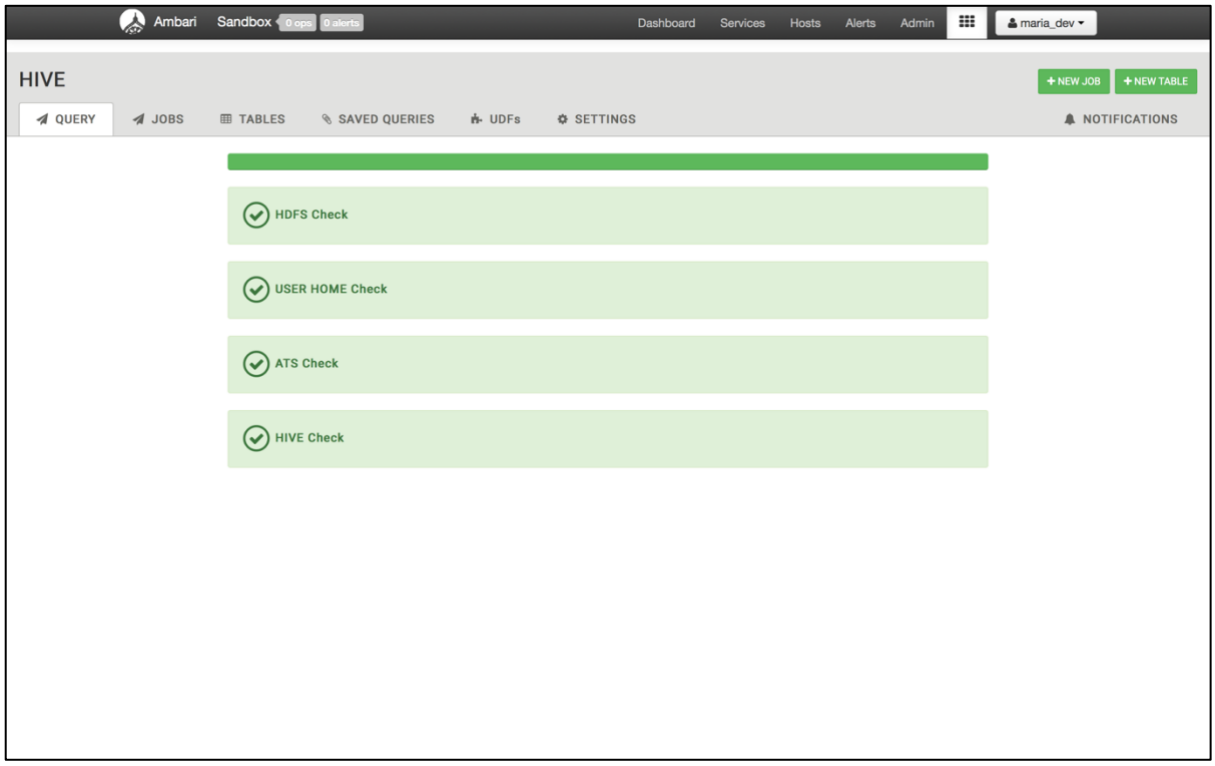



Figure 5: Launched Apache Hive 2.0 in Hadoop Enviroment, once the servers startup.

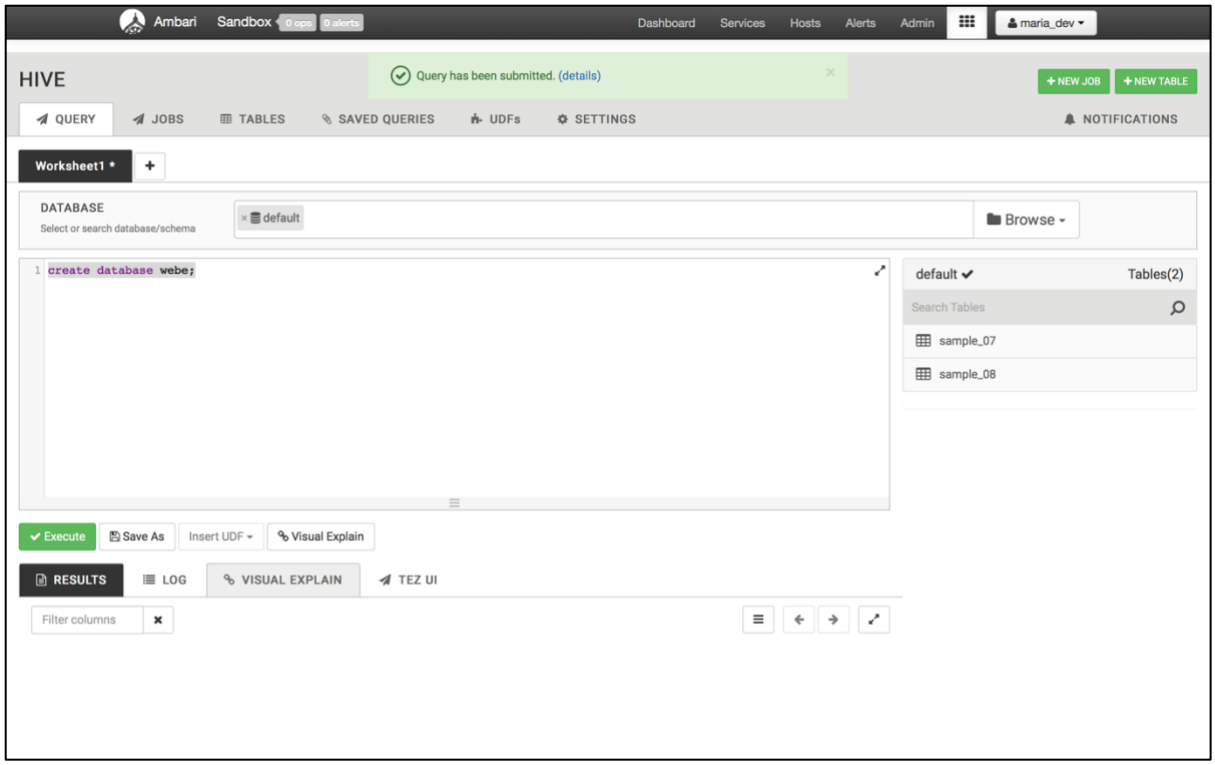



Figure 6: Created a new database ‘webe’ for initialization.

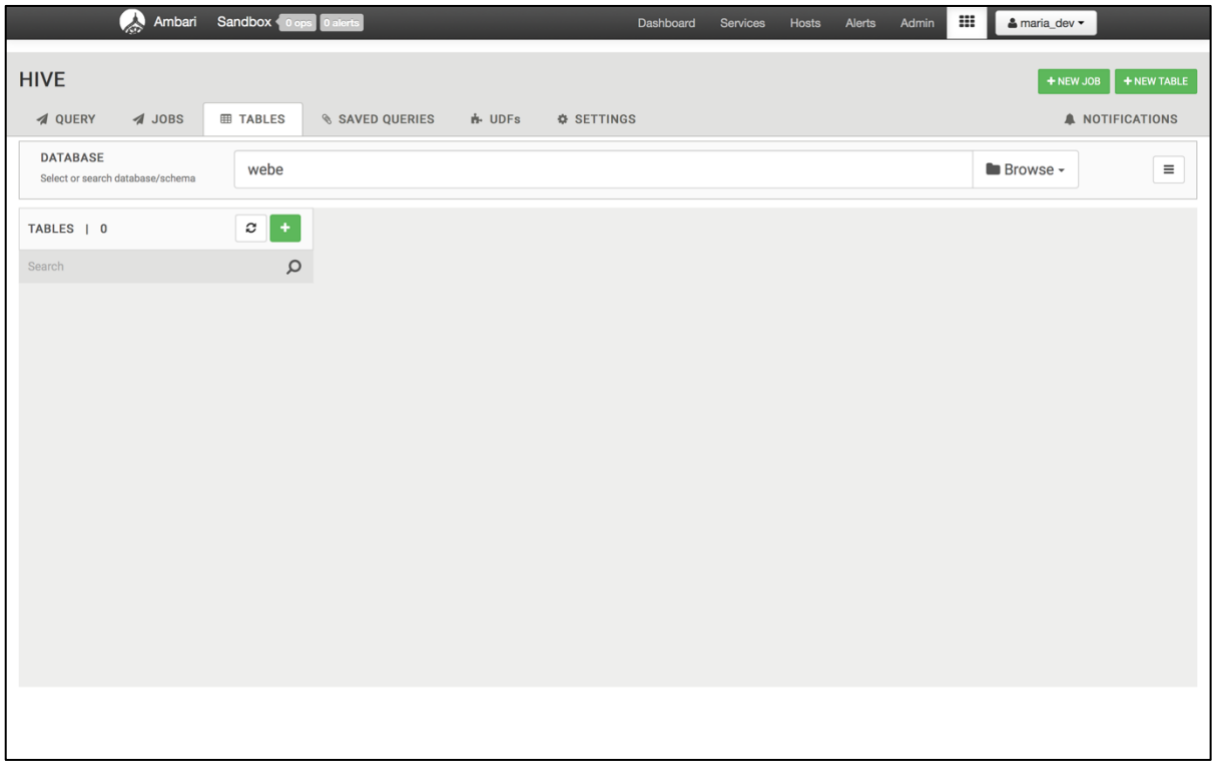



Figure 7: The database ‘webe’ created with empty tables.

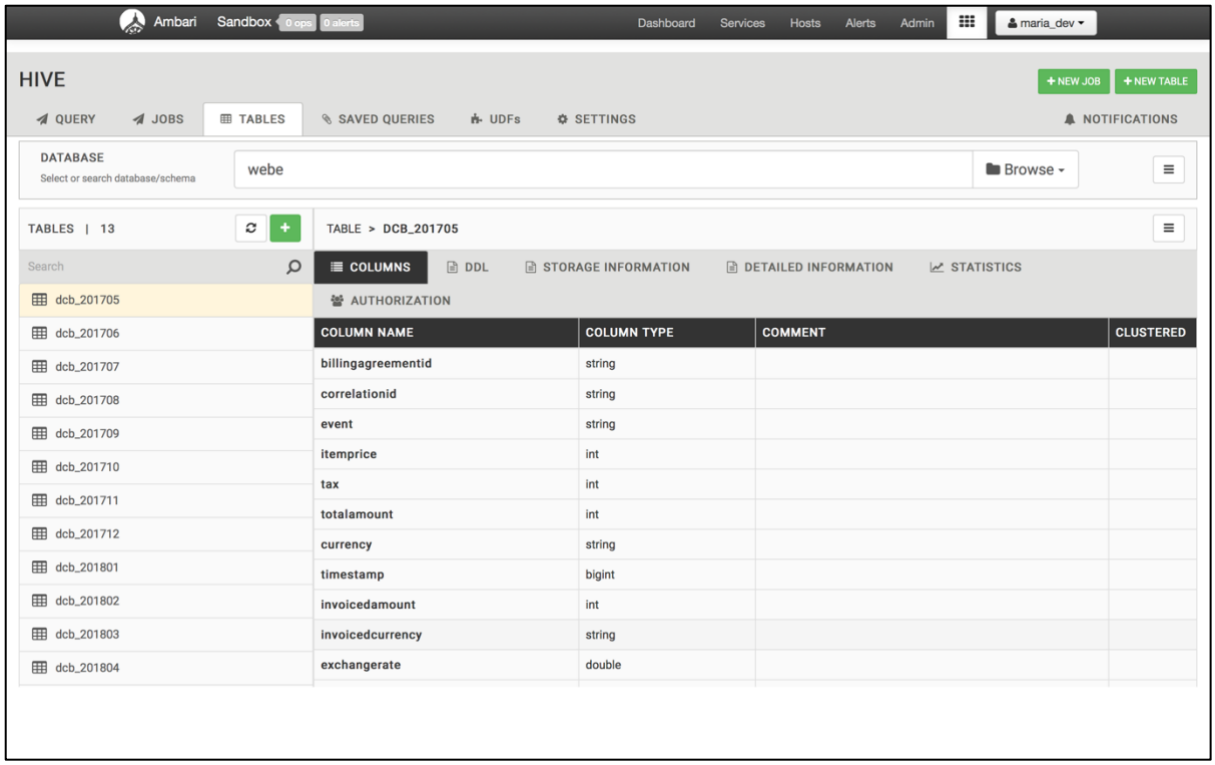



Figure 8: Uploaded all the 12 record csv files, in range 201705 to 201804 as table into ‘webe’ database

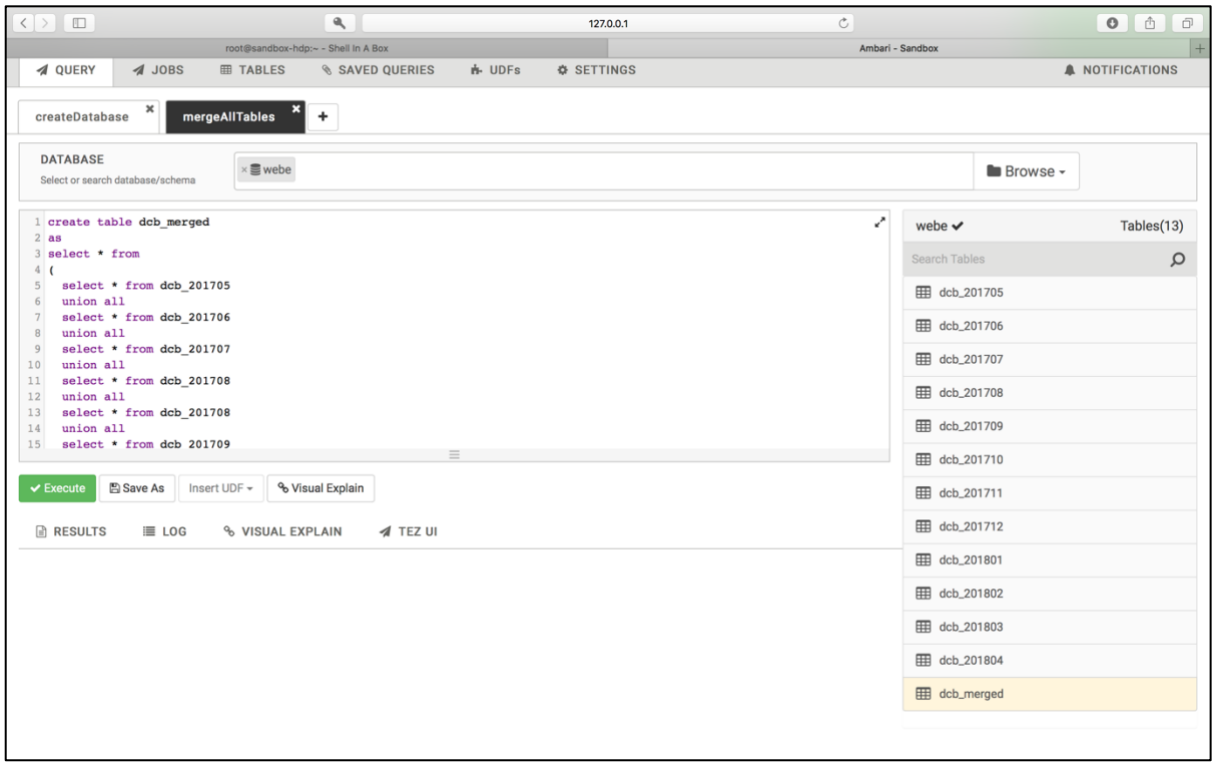



Figure 9: Merging all the 12 tables in range 201705 to 201804 in to a new table called ‘dcb_merged’.

merged_dcb

| msisdn | chargeItemCode | purchaseType | purchaseDescription | amount | status | startTime | endTime |
|---|---|---|---|---|---|---|---|
| 60122199015 | SPP_GPLAY | (PES 2017 PRO EVOLUTION SOCCER) | 株式会社コナミデジタルエンタテインメント - 1,050 myClub Coin | 39.9 | SUCCESS | 24/08/2017 16:20 | 24/08/2017 16:20 |
| 60105534536 | SPP_GPLAY | (Hotel Dash) | Glu Games Inc. - Bistro Bay Cruise | 3.9 | SUCCESS | 24/08/2017 16:30 | 24/08/2017 16:30 |
| 60105534536 | SPP_GPLAY | (Hotel Dash) | Glu Games Inc. - Tiki Palace | 3.9 | SUCCESS | 24/08/2017 16:30 | 24/08/2017 16:30 |
| 601110282827 | SPP_GPLAY | (Mobile Legends: Bang bang) | YoungJoy Technology Limited - 500 Diamonds | 39.9 | SUCCESS | 24/08/2017 16:34 | 24/08/2017 16:34 |
| 601110084241 | SPP_GPLAY | (Mobile Legends: Bang bang) | YoungJoy Technology Limited - 250 Diamonds | 19.9 | SUCCESS | 24/08/2017 16:34 | 24/08/2017 16:34 |
| 60198517958 | SPP_GPLAY | (Mobile Legends: Bang bang) | YoungJoy Technology Limited - 50 Diamonds | 3.9 | SUCCESS | 24/08/2017 16:42 | 24/08/2017 16:42 |
| 601110024532 | SPP_GPLAY | (Smule Sing!) | Smule, Inc. - All Access Pass - 1 Month | 3.49 | FAIL | 24/08/2017 16:43 | 24/08/2017 16:43 |
| 60192627266 | SPP_GPLAY | (Smule Sing!) | Smule, Inc. - All Access Pass - 1 Month | 3.49 | SUCCESS | 24/08/2017 16:49 | 24/08/2017 16:49 |
| 60102249968 | SPP_GPLAY | (Mobile Legends: Bang bang) | YoungJoy Technology Limited - 250 Diamonds | 19.9 | SUCCESS | 24/08/2017 16:51 | 24/08/2017 16:51 |
| 601110240506 | SPP_GPLAY | (Mobile Legends: Bang bang) | YoungJoy Technology Limited - 250 Diamonds | 19.9 | SUCCESS | 24/08/2017 16:54 | 24/08/2017 16:54 |
| 601110238769 | SPP_GPLAY | (Mobile Legends: Bang bang) | YoungJoy Technology Limited - 50 Diamonds | 3.9 | SUCCESS | 24/08/2017 17:13 | 24/08/2017 17:13 |
| 601110084562 | SPP_GPLAY | (Pokémon GO) | Niantic, Inc. - 100 PokéCoins | 1.9 | SUCCESS | 25/08/2017 3:43 | 25/08/2017 3:43 |
| 601110201203 | SPP_GPLAY | (Tencent Poker-Texas Hold'em) | Tencent Mobile International Ltd. - diamond1_wechat_android | 4 | SUCCESS | 25/08/2017 4:01 | 25/08/2017 4:01 |
| 60106524040 | SPP_GPLAY | (War Machines: Free Multiplayer Tank Shooting Games) | Fun Games For Free Limited - Small Hand of Diamonds | 8.49 | SUCCESS | 25/08/2017 4:30 | 25/08/2017 4:30 |
| 60193629569 | SPP_DCBNT | Gamebasics: 700 Boss Coins | Gamebasics: 700 Boss Coins | 50 | FAIL | 06/09/2017 17:20 | 06/09/2017 17:20 |
| 60194780427 | SPP_GPLAY | (黒道風雲) | game168 - 禮包 | 20.99 | SUCCESS | 15/09/2017 23:33 | 15/09/2017 23:33 |
| 60135255557 | SPP_DCBNT | GOOGLE - Despicable Me 3 | GOOGLE - Despicable Me 3 | 14 | SUCCESS | 04/10/2017 10:41 | 04/10/2017 10:41 |
| 601110307403 | SPP_GPLAY | (Choices: Stories You Play) | Pixelberry Studios - Diamonds - Bag | 19.99 | SUCCESS | 24/08/2017 16:09 | 24/08/2017 16:09 |
| 601110140484 | SPP_GPLAY | (Mobile Legends: Bang bang) | YoungJoy Technology Limited - 250 Diamonds | 19.9 | SUCCESS | 24/08/2017 16:20 | 24/08/2017 16:20 |
| 601110056303 | SPP_GPLAY | (Zynga Poker – Texas Holdem) | Zynga Inc - Medium Gold Stack | 6.31 | SUCCESS | 24/08/2017 16:26 | 24/08/2017 16:26 |
| 60105534536 | SPP_GPLAY | (Hotel Dash) | Glu Games Inc. - Lovely Suites Hotel | 3.9 | SUCCESS | 24/08/2017 16:31 | 24/08/2017 16:31 |
| 60104565981 | SPP_GPLAY | (BIGO LIVE - Live Stream) | BIGO TECHNOLOGY PTE. LTD. - 42 Diamonds | 3.99 | SUCCESS | 24/08/2017 16:39 | 24/08/2017 16:40 |
| 60134844032 | SPP_GPLAY | (Clash of Clans) | Supercell - Builder Pack | 19.9 | SUCCESS | 24/08/2017 16:48 | 24/08/2017 16:48 |
| 60102649206 | SPP_GPLAY | (Smule Sing!) | Smule, Inc. - All Access Pass - 1 Month | 3.49 | FAIL | 24/08/2017 16:55 | 24/08/2017 16:55 |
| 601110141973 | SPP_GPLAY | (Top Eleven 2017 - Be a Soccer Manager) | Nordeus - 37 Tokens | 19.71 | SUCCESS | 24/08/2017 16:58 | 24/08/2017 16:58 |
| 601110090813 | SPP_GPLAY | (Mobile Legends: Bang bang) | YoungJoy Technology Limited - 50 Diamonds | 3.9 | SUCCESS | 25/08/2017 3:32 | 25/08/2017 3:32 |
| 60187811265 | SPP_GPLAY | (JOOX Music - Free Streaming) | Tencent Mobility Limited - JOOX VIP service (1 month) | 14.9 | SUCCESS | 25/08/2017 3:56 | 25/08/2017 3:56 |
| 601110201203 | SPP_GPLAY | (Tencent Poker-Texas Hold'em) | Tencent Mobile International Ltd. - diamond2_wechat_android | 18 | SUCCESS | 25/08/2017 4:32 | 25/08/2017 4:32 |
| 601110215305 | SPP_GPLAY | (Mobile Legends: Bang bang) | YoungJoy Technology Limited - 500 Diamonds | 39.9 | SUCCESS | 25/08/2017 4:56 | 25/08/2017 4:56 |

Figure 10: Overview of 'dcb_merged'

The total records under 'dcb_merged' were 3000 records. Meanwhile the study or the visualization from the cleaned and merged will be used Tableau to be presented under Chapter 7. Analysing Processed Data.

*(Gunasegarran)*

## 7. Analysing Processed Data

After we have done with processing our table, we proceed to analyse our datasets. It is important to gather meaningful insight from this data. Especially in big data which involves massive amount of data, business users need to adapt and scale with rapid and emerging technology. In recent years, we have seen a lot of tools and software aim to help business to understand and make use of the data that they have. One of the tools that is widely used especially in the field of data visualisation is Tableau. Tableau enable users to create interactive business dashboard that can cater business users requirement. They are the leader in data visualisation field.

In big data solution, Tableau aid business to see and understand their data. One way they help in big data is to answers questions that arise from datasets whether they are big or small datasets. Tableau is well known for easy integration and connection with various databases. With this, users are able to see and gather insights from their data. This is important in the later process of making decision. Besides, Tableau enable users to visualise their data. This help optimize the resources that a business has in order to make greater impact on the business. Sometimes, without a proper tool, users can missed a critical insights that is actually important to the business. With Tableau, users will have the capability to view their data in professional and meaningful way. Below is the image of Tableau interface:

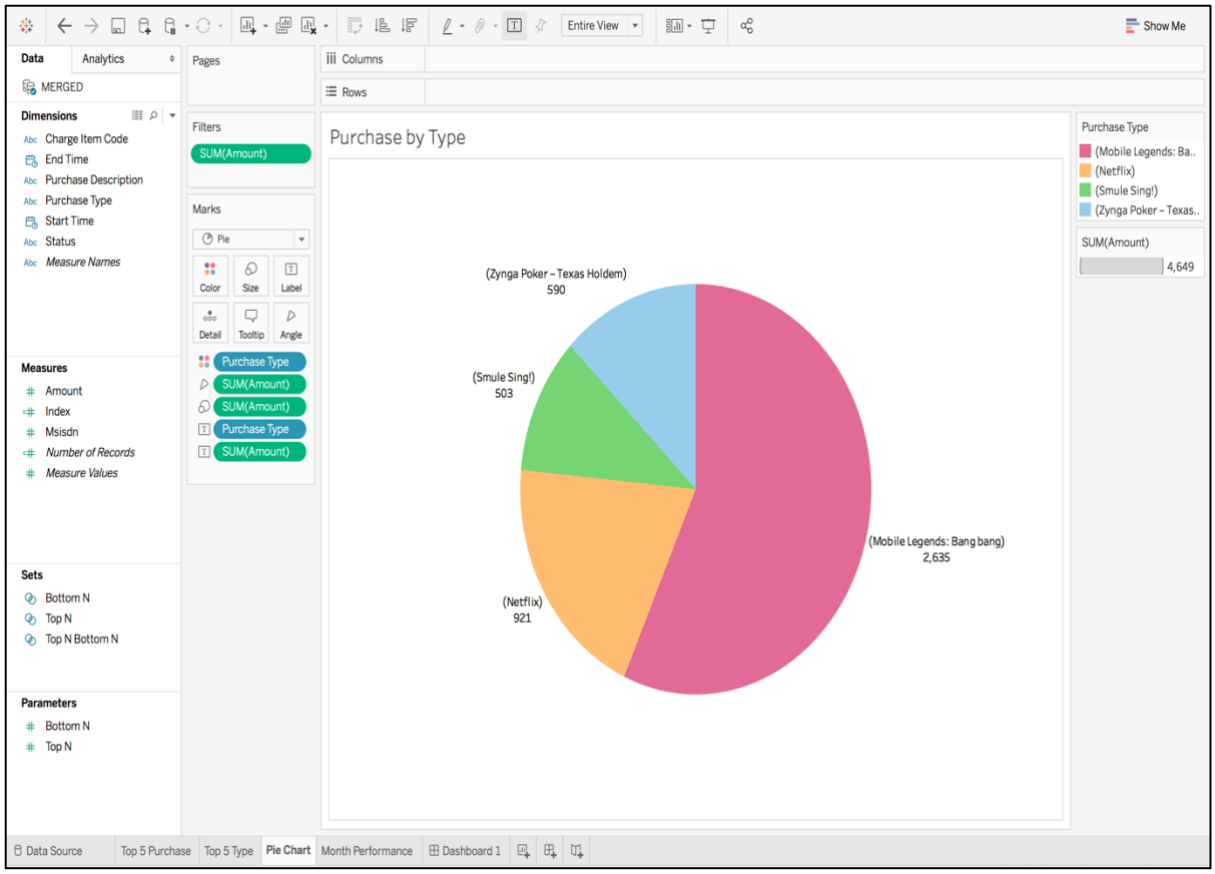



Figure 11: Tableau Visualization for Purchase by Type

By using Tableau, we are able to develop a business dashboard that helps us to understand the performance of the business based on the customer data. After the process of cleaning and transforming the datasets with Hive, we connect our data with Tableau. Later, we build several charts with different parameters based on function we believed important to be seen from the datasets. Below is the dashboard that we managed to developed:

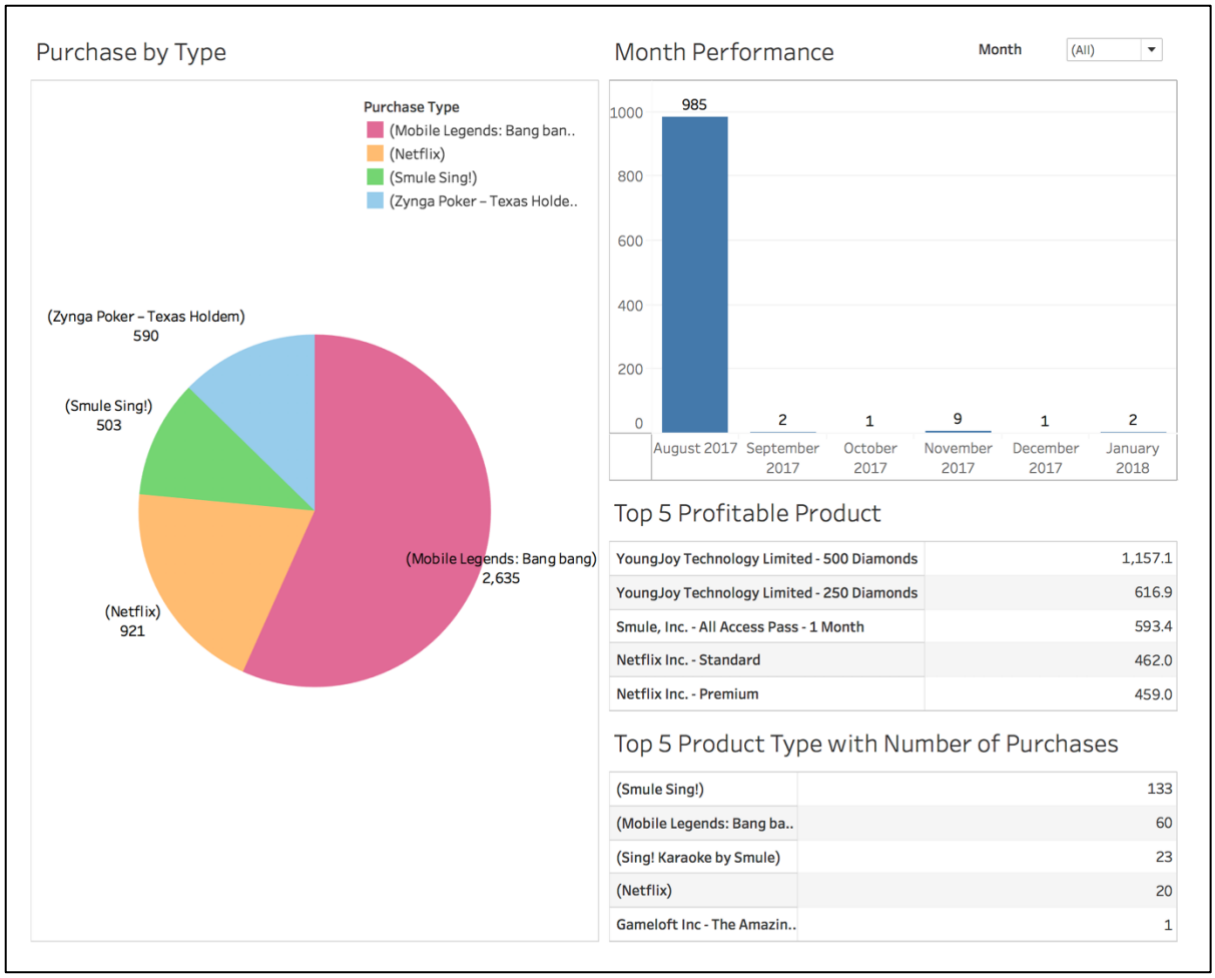



Figure 12: Tableau Visualization for Month Performance

The first chart is a pie chart which describes the number of purchase by each products. With this chart, we can infer the portions of a product type that we offered to our customer. The second chart is a bar chart which shows the number of product sold in each month. The third is a view of top 5 most profitable product sold by the company. Lastly, is the view of the product type with top 5 highest number of purchases. Each chart plays an important role in order to help business users to understand the performance of their business. Not only they can see the top 5 most profitable, the view can also be adjusted to view the bottom 5 of the product type. Then, business user can deduce which type of product need to be focused more compared to others.

In conclusion, Tableau helps us to understand our data better. One way Tableau surpass its competitors is the ability to work with many type of data and many of data sources. This features helps users in making swift analysis especially when urgent decision need to be made. Some tools may take longer time to configure compared to Tableau.

*(Muhammad Nooraizad)*

## 8. Reference

BigDataMadeSimple – Baiju NT, 11 Interesting Big Data Cases studies in Teleom. August 2017. http://bigdata-madesimple.com/11-interesting-big-data-case-studies-in-telecom/

Booz & Company, Benefiting from big data, A new approach for the telecom industry. 2013. https://www.strategyand.pwc.com/media/file/Strategyand_Benefiting-from-Big-Data_A-New-Approach-for-the-Telecom-Industry.pdf

Gunasegarran, Direct Carrier Billier. webe digital sdn bhd (a TM Company). 2018

Hewlett Packard Enterprise, MTS India relies on HPE Vertica in a highly competitive telecom market. How analytics keeps customers from moving to the competition. November, 2015. https://www.vertica.com/wp-content/uploads/2017/01/4AA5-2844ENW.pdf

IBM Global Service, Analytics: The real world use of big data. How innovative enterprises extract value from uncertain data. October, 2012. https://www.bdvc.nl/images/Rapporten/GBE03519USEN.pdf

Infocepts, Re-engineering a Telecom Market Share Analytical Application. October, 2013. http://www.infocepts.com/pdf/Industries/Telecommunications/Re-engineering_a_Telecom_Market_Share_Analytical_Application-Case_Study.pdf

Zafar Gilani and Salman Ul Haq, Analyzing large datasets with Hive. July, 2013. https://www.ibm.com/developerworks/library/bd-hiveanalyze/index.html