FLIGHT DATA ANALYSIS USING HADOOP

A Mini Project Report Submitted by

KAUSHIK C RAJSHEKAR (4NM17CS083)

LESTON JOHN ALVA (4NM17CS092)

M. RAMYA PRABHU (4NM17CS094)

MALLIKA (4NM17CS098)

UNDER THE GUIDANCE OF

Mrs. Savitha Shetty

Assistant Professor, Grade II
Department of Computer Science and Engineering

in partial fulfilment of the requirements for the award of the Degree of

Bachelor of Engineering in
Computer Science & Engineering
from

Visvesvaraya Technological University, Belagavi

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

N.M.A.M. INSTITUTE OF TECHNOLOGY

(An Autonomous Institution under VTU, Belgaum) (AICTE approved, NBA Accredited, ISO 9001:2008 Certified) NITTE -574 110, Udupi District, KARNATAKA.

May 2020

NITTE EDUCATION TRUST | N.M.A.M. INSTITUTE OF TECHNOLOGY
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)
Nitte – 574 110, Karnataka, India
(ISO 9001:2015 Certified), Accredited with 'A' Grade by NAAC
☎: 08258 - 281039 – 281263, Fax: 08258 – 281265
Department of Computer Science and Engineering
B.E. CSE Program Accredited by NBA, New Delhi from 1-7-2018 to 30-6-2021

CERTIFICATE

"Flight Data Analysis Using Hadoop"

is a bonafide work carried out by

KAUSHIK C RAJSHEKAR	LESTON JOHN ALVA
(4NM17CS083)	(4NM11CS092)
M. RAMYA PRABHU	MALLIKA
(4NM17CS094)	(4NM17CS098)

in partial fulfilment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering prescribed by Visvesvaraya Technological University, Belagavi during the year 2019-2020.

It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report.

The Mini project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the Bachelor of Engineering Degree.

Signature of Guide Signature of HOD

ACKNOWLEDGEMENT

We believe that our project will be complete only after we thank the people who have contributed to make this project successful.

First and foremost, our sincere thanks to our beloved principal, **Dr. Niranjan N. Chiplunkar** for giving us an opportunity to carry out our project work at our college and providing us with all the needed facilities.

We sincerely thank **Dr. K.R. Udaya Kumar Reddy**, Head of Department of Computer Science and Engineering, Nitte Mahalinga Adyanthaya Memorial Institute of Technology, Nitte.

We express our deep sense of gratitude and indebtedness to our guide **Mrs. Savitha Shetty**, Assistant Professor, Department of Computer Science and Engineering, for her inspiring guidance, constant encouragement, support and suggestions for improvement during the course of our project.

We also thank all those who have supported us throughout the entire duration of our project.

Finally, we thank the staff members of the Department of Computer Science and Engineering and all our friends for their honest opinions and suggestions throughout the course of our project.

Kaushik C Rajshekar (4NM17CS083)
Leston John Alva (4NM17CS092)
M. Ramya Prabhu (4NM17CS094)
Mallika (4NM17CS099)

ABSTRACT

Big Data is a term which is used for the description of huge or large volume of data which cannot be stored or processed using traditional approach within the given time frame. The data could be structured and unstructured. Big data helps users to collect variety of data and analyse large and varied data sets. Data processing helps us to collect and organize raw data and to get a meaningful information. The Big data analytics tools offer a variety of analytics packages which gives different options to the users to implement.

TABLE OF CONTENTS

INTRODUCTION

Hadoop is a software library which allows the users to process large amount of distributed data across various heterogeneous computers using simple programming techniques. It is also an open source framework.

Hive is a data warehouse and it is based on hadoop for providing us the data with data analysis and querying techniques. Hive is designed in a way that it can give us a SQL-like interface to query data which will stored in heterogeneous databases and file systems that will integrate with hadoop. Traditional SQL queries must be put into efforts in the MapReduce Java API to execute SQL application and queries over the distributed data. Hive provides us with necessary SQL abstraction to put together SQL like queries into the Java without any need to implement queries in the low level Java API.

Since most data warehousing applications work with SQL based application to hadoop, hive supports analysis of large datasets which is stored in hadoop's HDFS (Hadoop Distributed File System) and compatible file system. Here are the certain features of Hive.

1) Indexing to provide acceleration, index type including compaction and bitmap index.

2) Different storage types.

3) Metadata storage in a relational database management system.

4) Operating on compressed data stored into the hadoop ecosystem using algorithms.

5) Built in user defined functions.

6) SQL like queries which are converted into MapReduce.

1.1 Objective

Our mini project is about analyzing flight data and querying it accordingly. Flight data consists of huge amount of data every day. Unlike other transportation platforms, the people using aircrafts, fly on daily basis and their data is quite private and extractable. Using Big Data analytics it is easy to fetch the data and perform necessary operations on it.

In Aircraft, there are approximately millions of monthly active travelers. Aircraft travelling allows the people to constantly make business travels, family tours, pilgrimage visit etc.

In our project we will calculate the delays of the flights and also the customer feedback based on that delays and the passenger willing to travel again or not based on their feedback on these delays.

In our project we identify the flight data which is nothing but the characterized data of the travelers, which is actually private but can be easily extracted. The feedback of the passengers could be positive or negative or even neutral.

We fetch the positive feedbacks and query them accordingly. Similarly we fetch the negative feedbacks and query them according to the user's requirement.

1.2 Methodology

We have extracted the Flight data and stored it as one .csv file. We have then copied the file from Windows to Hadoop using WinSCP. We created one database and altered the Flight data file in the form of a table, where we have columns regarding the flight delays and the feedback of the customer.

We have used Hive to query our Flight data. The below figure shows the working of how the tweets would be stored and analyzed using Hive.

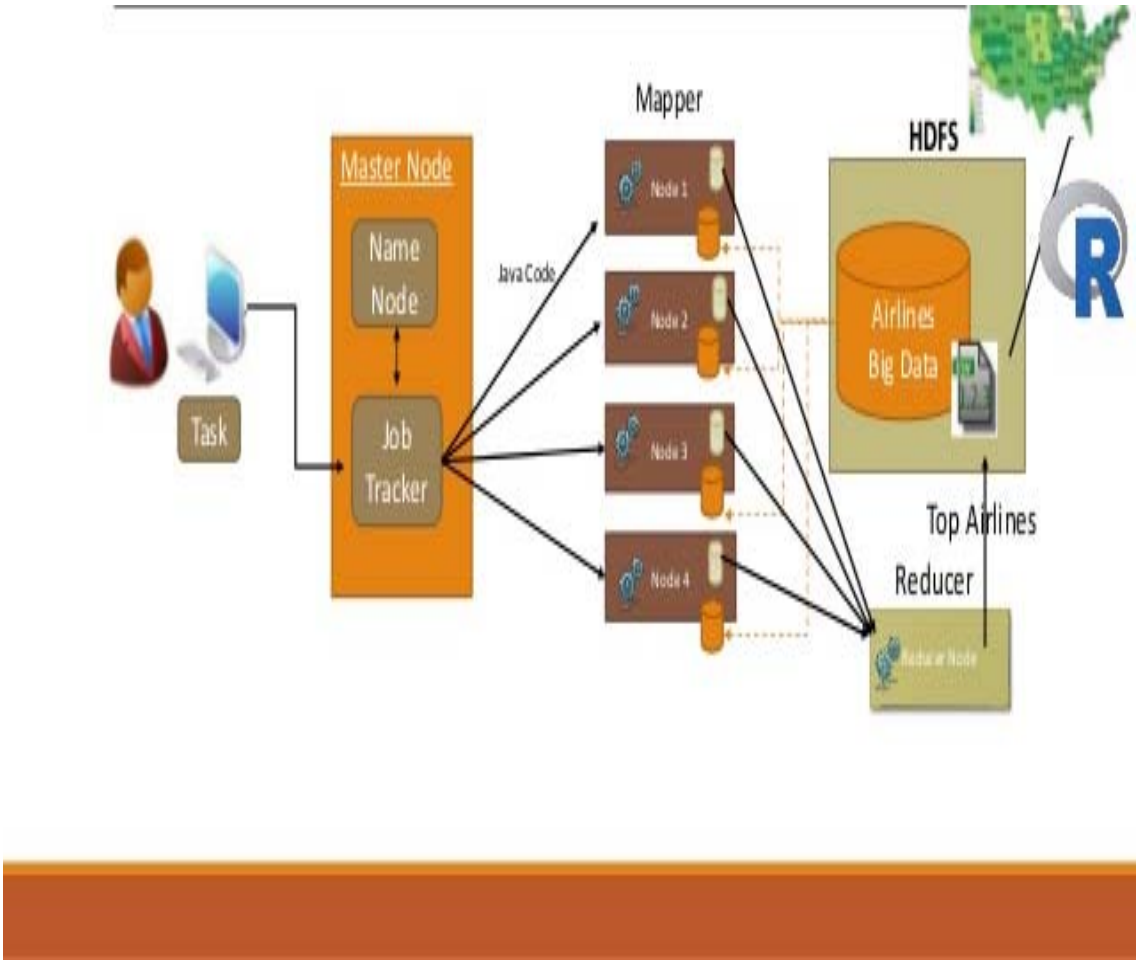


Fig. (1)

CHAPTER 2

SYSTEM ANALYSIS AND REQUIREMENTS

2.1 Functional Requirements

A functional requirement defines a function of a system or its component. A function is described as a set of inputs, the behavior and outputs. Functional requirements may be calculations, technical details, data manipulation and processing and other specific functionality that define what a system is supposed to accomplish. Functional requirements are made up of business features and technical requirements for support of particular features.

2.1.1 Hardware Requirements

- 4 GB or 8 GB RAM
- Intel i3 or above processor
- 2 GB or above storage

2.1.2 Software Requirements

- Horton works Sandbox
- Hadoop
- WinSCP
- VMWare Workstation
- Windows Host Operating System

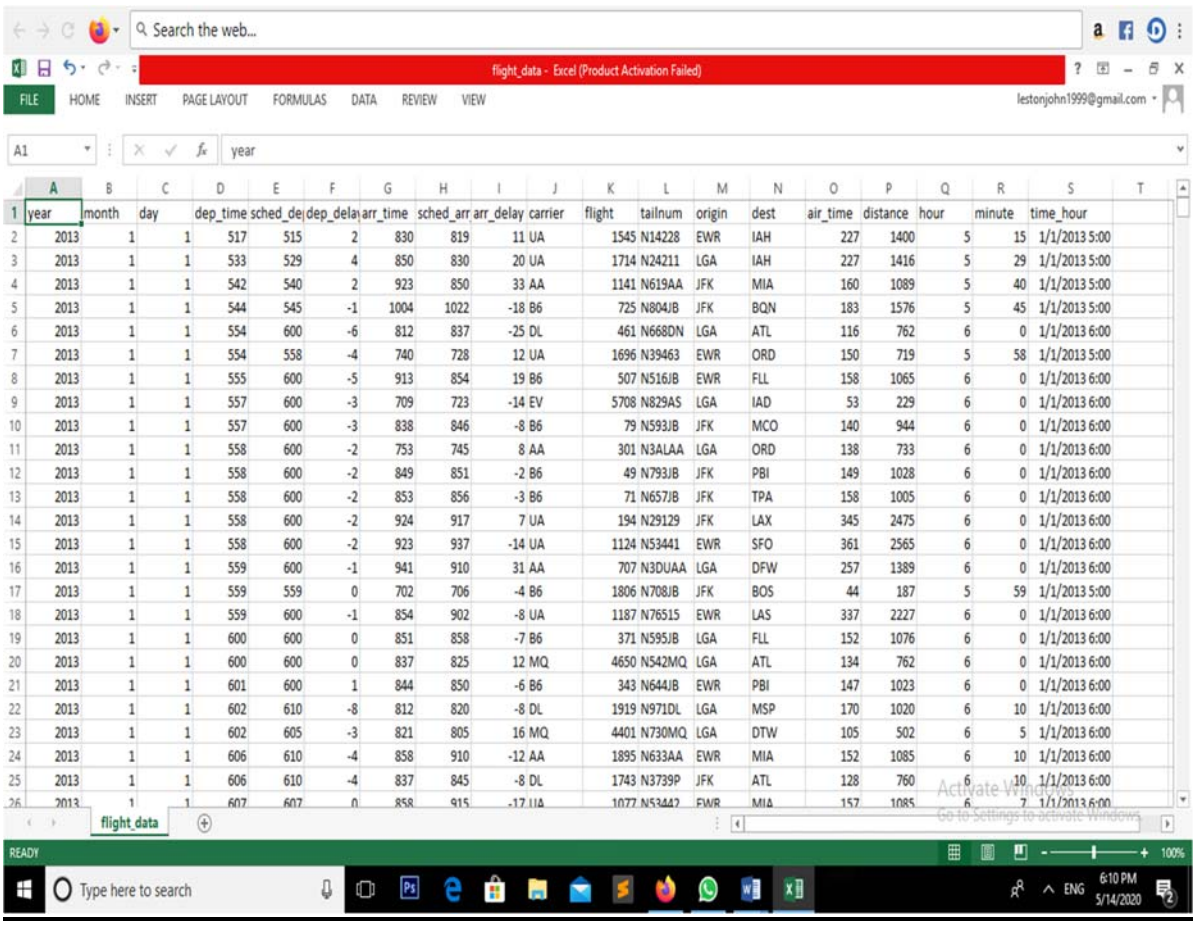
2.2 Non-functional Requirements

The non-functional requirements have the capabilities that are offered by the framework. They are also known as quality requirements. It specifies the criteria that can be used to judge the operation of a system, rather than specific behavior. Non-functional requirements address features of a system that are not isolated to the ability of the user application administrator to carry out a particular operation within the system.

CHAPTER 3

IMPLEMENTATION

We have included the snapshot of the Flight dataset that we have used to do the required query using Hive. In this file we have mainly many fields. The field consists of the day of travel, month of travel, departure time, arrival based on which we will be calculating the delays and based on these delays we will be analyzing the data to get the feedback of the customer.



year	month	day	dep_time	sched_de	dep_dela	arr_time	sched_arr	arr_delay	carrier	flight	tailnum	origin	dest	air_time	distance	hour	minute	time_hour
2013	1	1	517	515	2	830	819	11	UA	1545	N14228	EWR	IAH	227	1400	5	15	1/1/2013 5:00
2013	1	1	533	529	4	850	830	20	UA	1714	N24211	LGA	IAH	227	1416	5	29	1/1/2013 5:00
2013	1	1	542	540	2	923	850	33	AA	1141	N619AA	JFK	MIA	160	1089	5	40	1/1/2013 5:00
2013	1	1	544	545	-1	1004	1022	-18	B6	725	N804JB	JFK	BQN	183	1576	5	45	1/1/2013 5:00
2013	1	1	554	600	-6	812	837	-25	DL	461	N668DN	LGA	ATL	116	762	6	0	1/1/2013 6:00
2013	1	1	554	558	-4	740	728	12	UA	1696	N39463	EWR	ORD	150	719	5	58	1/1/2013 5:00
2013	1	1	555	600	-5	913	854	19	B6	507	N516JB	EWR	FLL	158	1065	6	0	1/1/2013 6:00
2013	1	1	557	600	-3	709	723	-14	EV	5708	N829AS	LGA	IAD	53	229	6	0	1/1/2013 6:00
2013	1	1	557	600	-3	838	846	-8	B6	79	N593JB	JFK	MCO	140	944	6	0	1/1/2013 6:00
2013	1	1	558	600	-2	753	745	8	AA	301	N3ALAA	LGA	ORD	138	733	6	0	1/1/2013 6:00
2013	1	1	558	600	-2	849	851	-2	B6	49	N793JB	JFK	PBI	149	1028	6	0	1/1/2013 6:00
2013	1	1	558	600	-2	853	856	-3	B6	71	N657JB	JFK	TPA	158	1005	6	0	1/1/2013 6:00
2013	1	1	558	600	-2	924	917	7	UA	194	N29129	JFK	LAX	345	2475	6	0	1/1/2013 6:00
2013	1	1	558	600	-2	923	937	-14	UA	1124	N53441	EWR	SFO	361	2565	6	0	1/1/2013 6:00
2013	1	1	559	600	-1	941	910	31	AA	707	N3DUAA	LGA	DFW	257	1389	6	0	1/1/2013 6:00
2013	1	1	559	559	0	702	706	-4	B6	1806	N708JB	JFK	BOS	44	187	5	59	1/1/2013 5:00
2013	1	1	559	600	-1	854	902	-8	UA	1187	N76515	EWR	LAS	337	2227	6	0	1/1/2013 6:00
2013	1	1	600	600	0	851	858	-7	B6	371	N595JB	LGA	FLL	152	1076	6	0	1/1/2013 6:00
2013	1	1	600	600	0	837	825	12	MQ	4650	N542MQ	LGA	ATL	134	762	6	0	1/1/2013 6:00
2013	1	1	601	600	1	844	850	-6	B6	343	N644JB	EWR	PBI	147	1023	6	0	1/1/2013 6:00
2013	1	1	602	610	-8	812	820	-8	DL	1919	N971DL	LGA	MSP	170	1020	6	10	1/1/2013 6:00
2013	1	1	602	605	-3	821	805	16	MQ	4401	N730MQ	LGA	DTW	105	502	6	5	1/1/2013 6:00
2013	1	1	606	610	-4	858	910	-12	AA	1895	N633AA	EWR	MIA	152	1085	6	10	1/1/2013 6:00
2013	1	1	606	610	-4	837	845	-8	DL	1743	N3739P	JFK	ATL	128	760	6	10	1/1/2013 6:00
2013	1	1	607	607	0	858	915	-17	UA	1077	N53442	EWR	MIA	157	1085	6	7	1/1/2013 6:00

Fig. (2)

CHAPTER 4

RESULTS

We have executed several Hive queries to obtain the Flight data and to analyze the data based on the required sentiments. Following are the snapshots of the result that we have obtained.

Airport and Flight Information:

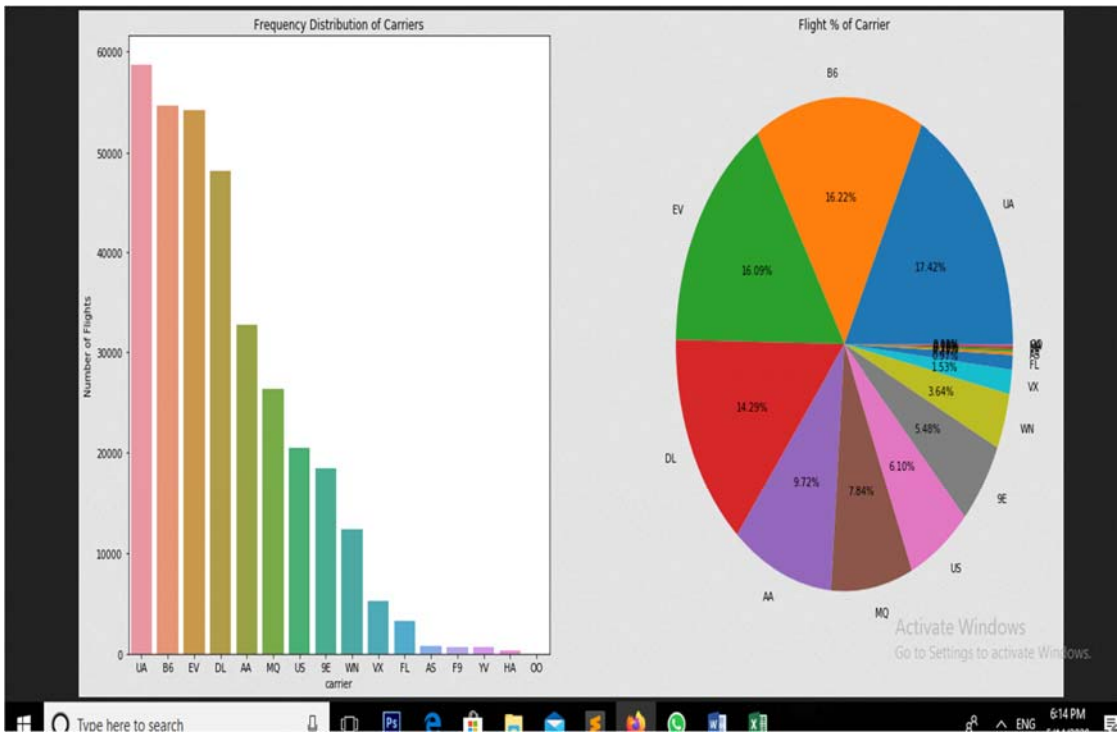


Fig. (3)

Delay Information:

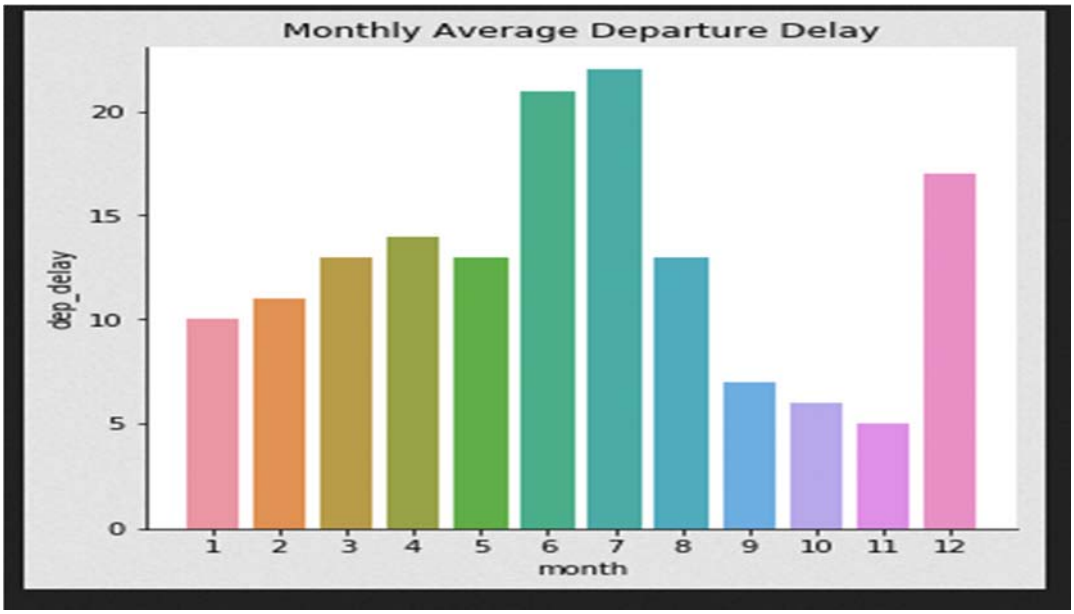


Fig. (4)

On-Time Departure and Arrival Analysis:

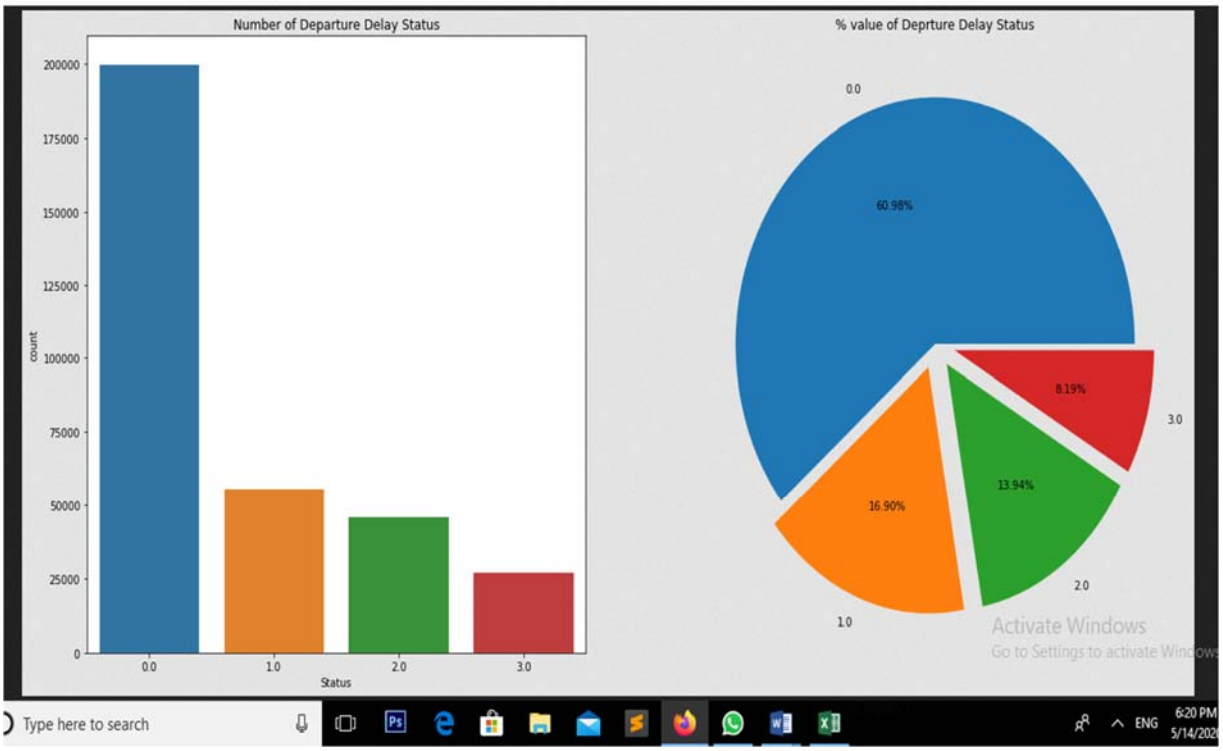


Fig. (5)

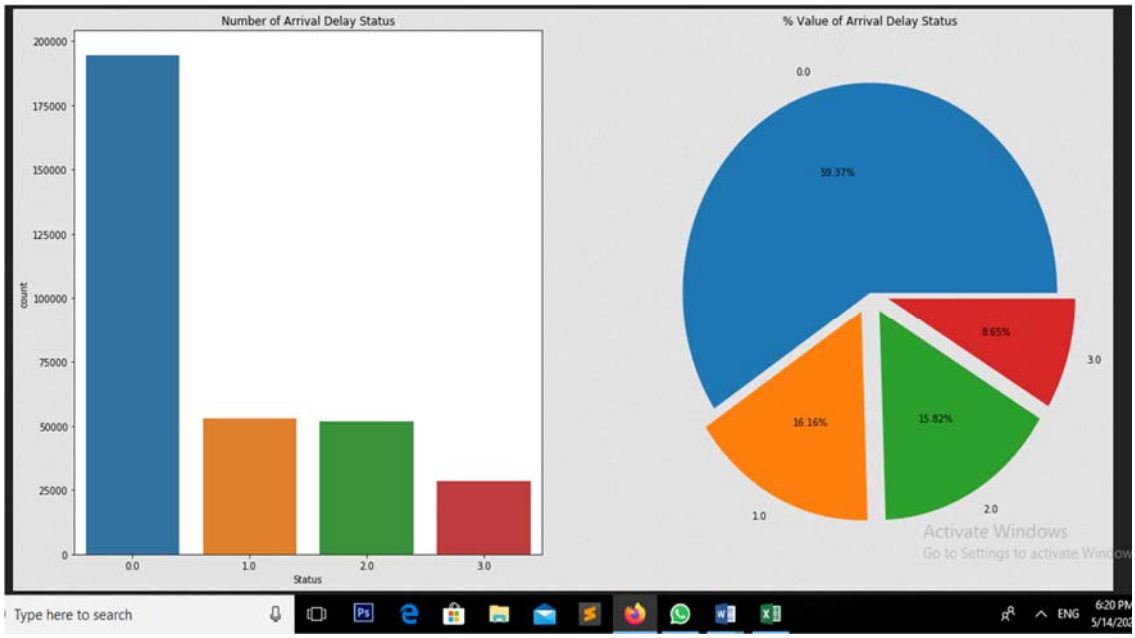


Fig. (6)

Performance Analysis:

```
Sched_dep:  336776
Operated:  328521
Cancelled:  8255


Ratio operated flights over scheduled flights: 97.54881583010666
Ratio of cancelled flights: 2.451184169893338
```

Fig. (7)

CHAPTER 5

CONCLUSION

Though the dataset doesn't offer reasons for delays and missing important data such as taxi in and out, flight diversion, chocks on and off timing, and fuel consumption. So, it is clear that the dataset doesn't provide clear understanding of delay issues, which may be supportive to look into delays that can be controlled or reduced. Using Hive, we have analyzed the Flight Data Analysis. If traditional DBMS would have been used, querying on this unstructured data would not have been possible and feasible. The efficient use of Hadoop, to query on big or large volumes of data, helps improve understandability to the user.

REFERENCES

[1] https://www.kaggle.com/lampubhutia/nyc-flight-data-analysis

[2] https://github.in/flight_data_analysis

[3] Big Data for Dummies – Text Book.

[4] https://acadgild.com/blog/data-analysis-on-tweets-with-apache-hive-using-afinn-dictionary

[5] https://xebia.com/blog/flight-analysis-using-apache-hive/