IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 5, May 2020

Performance Comparison Between Hive-QL and Spark-SQL on Analysis of Airlines Dataset

Sanchita Chourawar[1]

Computer Science Engineering, GEC College, CSVTU University, Bilaspur, India[1]

Abstract: The problems faced by airlines market is not unique. But their large amount of unstructured data and incomplete information creates a problem for analytics to analyze these data. So, analyzing the complex unstructured data by using traditional tools and techniques is an expensive task. Airlines needed a proper analysis result to increase their market and reduce expenses. In this paper, the analysis of the airline data set is performed using Spark-SQL and hive which runs Hadoop in the background. HDFS is used for storing huge number of airlines data, Hive and spark have been used for querying the data in which hive uses HiveQL statements which runs on MapReduce framework and spark uses Spark-SQL which runs on spark framework. Data visualization has been done by extracting the output of the HIVE and SPARK query in excel and plotting the data using line and bar plot charts. The visualization of the data shows some patterns that exist different airlines delays caused by weather, security, NAS delay, late aircraft delay etc.

Keywords: Hadoop, airlines datasets, big data, MapReduce, HDFS, spark, data analysis, data visualization.

I. INTRODUCTION

Big data is very large amount of data provided by social site Airlines market continuously trying to increase their market and reduce expenses but the problems faced by airlines market is not unique [6]. But their large amount of unstructured data and incomplete information creates a problem for analytics to analyze these data. Traditional systems were used for airlines to store and process the unstructured data, but handling such data by tradition systems is too time-consuming and expensive too. The data generated by all operational enterprise systems are automatically archived and indexed. Airlines should also be capable of searching the entire corporate database to retrieve the relevant data. Airlines need the right information at the right time, with the right degree of accuracy. Typical attributes of airline data can be identified by three main attributes:

Volume – Airline data is huge and massive.
Velocity –Airline data is changes rapidly and arrives quickly so processing data in less time is very difficult.
Variety – Airline data have the different structure they are semi-structured or unstructured data.

II. LITERATURE REVIEW

[1] present the log data analysis by using spark through sql type queries, the web server logs and unstructured in nature and these data are analyse by using spark and hadoop framework, and they also compare both the framework on the basis of various parameters.

[2] Present the how the profit of airlines changes with time which is mainly driven by input price change which exhibits a similar pattern to output price change. In presence of productivity growth, the output price increase is lower than the input price increase suggesting that part of productivity gains are transferred from airlines to consumers connectivity, convenience and comfort), while the cost index incorporates three sub-indices (unit cost, aircraft and labor). They developed model enables the identification of the hybrid business models that are successfully pursuing an integrated cost leadership and differentiation strategy.

III. PROBLEM FORMULATION

Airlines market continuously trying to increase their market and reduce expenses but the problems faced by airlines market is not unique. But their large amount of unstructured data and incomplete information creates a problem for analytics to analyse these data. Traditional tools and techniques create the problem for storing and analysing the huge amount of airlines data called big data [2] because of its nature so it's the biggest challenges in big data to store and process such huge amount data [10].

DOI 10.17148/IJARCCE.2020.9537

IV. PROPOSED WORK

We are using Hadoop [5] for storing and processing a huge amount of data, for storing its uses HDFS (Hadoop Distributed File System) and for processing its uses MapReduce [8]. For analyzing we are using hive which uses hiveQL statement which runs over MapReduce framework to analyze the data. And we can also analyze the data by using spark [4] which uses Spark-SQL and runs over spark framework.

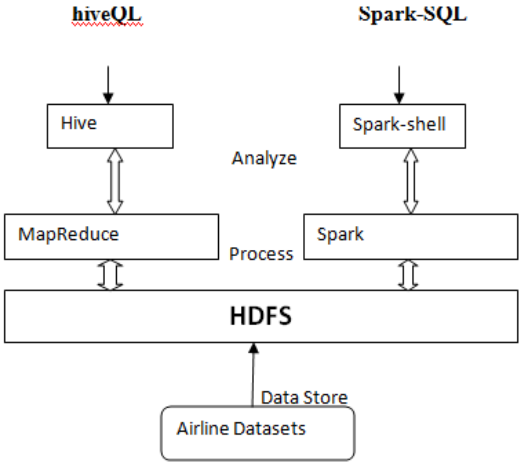
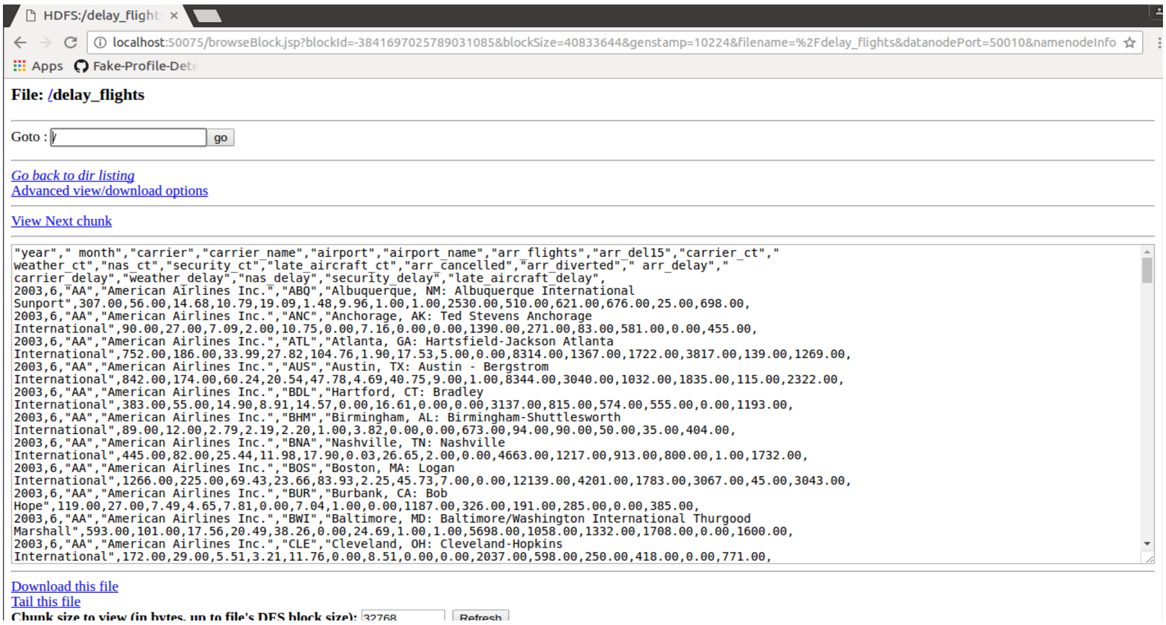


Fig. 1 Workflow Diagram

Loading and Storing: We can collect the airline's datasets which is very huge in size so we need a reliable storage so we can load the data into HDFS by using Hadoop put command through which we can load the data from the local file system into Hadoop file system. The dataset first split into multiple 64 MB blocks and each block is identified by unique block id and then these blocks are stored into HDFS. The airline's datasets are stored into HDFS are shown below.


HDFS:/delay_flight
localhost:50075/browseBlock.jsp?blockId=-3841697025789031085&blockSize=40833644&genstamp=10224&filename=%2Fdelay_flights&datanodePort=50010&namenodeInfo
Apps
Fake-Profile-Det

File: /delay_flights

Goto : / go

Go back to dir listing
Advanced view/download options

View Next chunk

```
"year"," month","carrier","carrier_name","airport","airport_name","arr_flights","arr_del15","carrier_ct","
weather_ct","nas_ct","security_ct","late_aircraft_ct","arr_cancelled","arr_diverted"," arr_delay","
carrier_delay","weather_delay","nas_delay","security_delay","late_aircraft_delay",
2003,6,"AA","American Airlines Inc.","ABQ","Albuquerque, NM: Albuquerque International
Sunport",307.00,56.00,14.68,10.79,19.09,1.48,9.96,1.00,1.00,2530.00,510.00,621.00,676.00,25.00,698.00,
2003,6,"AA","American Airlines Inc.","ANC","Anchorage, AK: Ted Stevens Anchorage
International",90.00,27.00,7.09,2.00,10.75,0.00,7.16,0.00,0.00,1390.00,271.00,83.00,581.00,0.00,455.00,
2003,6,"AA","American Airlines Inc.","ATL","Atlanta, GA: Hartsfield-Jackson Atlanta
International",752.00,186.00,33.99,27.82,104.76,1.90,17.53,5.00,0.00,8314.00,1367.00,1722.00,3817.00,139.00,1269.00,
2003,6,"AA","American Airlines Inc.","AUS","Austin, TX: Austin - Bergstrom
International",842.00,174.00,60.24,20.54,47.78,4.69,40.75,9.00,1.00,8344.00,3040.00,1032.00,1835.00,115.00,2322.00,
2003,6,"AA","American Airlines Inc.","BDL","Hartford, CT: Bradley
International",383.00,55.00,14.90,8.91,14.57,0.00,16.61,0.00,0.00,3137.00,815.00,574.00,555.00,0.00,1193.00,
2003,6,"AA","American Airlines Inc.","BHM","Birmingham, AL: Birmingham-Shuttlesworth
International",89.00,12.00,2.79,2.19,2.20,1.00,3.82,0.00,0.00,673.00,94.00,90.00,50.00,35.00,404.00,
2003,6,"AA","American Airlines Inc.","BNA","Nashville, TN: Nashville
International",445.00,82.00,25.44,11.98,17.90,0.03,26.65,2.00,0.00,4663.00,1217.00,913.00,800.00,1.00,1732.00,
2003,6,"AA","American Airlines Inc.","BOS","Boston, MA: Logan
International",1266.00,225.00,69.43,23.66,83.93,2.25,45.73,7.00,0.00,12139.00,4201.00,1783.00,3067.00,45.00,3043.00,
2003,6,"AA","American Airlines Inc.","BUR","Burbank, CA: Bob
Hope",119.00,27.00,7.49,4.65,7.81,0.00,7.04,1.00,0.00,1187.00,326.00,191.00,285.00,0.00,385.00,
2003,6,"AA","American Airlines Inc.","BWI","Baltimore, MD: Baltimore/Washington International Thurgood
Marshall",593.00,101.00,17.56,20.49,38.26,0.00,24.69,1.00,1.00,5698.00,1058.00,1332.00,1708.00,0.00,1600.00,
2003,6,"AA","American Airlines Inc.","CLE","Cleveland, OH: Cleveland-Hopkins
International",172.00,29.00,5.51,3.21,11.76,0.00,8.51,0.00,0.00,2037.00,598.00,250.00,418.00,0.00,771.00,
```

Download this file
Tail this file
Chunk size to view (in bytes, up to file's DFS block size): 32768 Refresh

Fig. 2 Data Stored into HDFS

Analysis: Data is analyzed by using Hive [11] and Spark-SQL [7] both use SQL like queries to analyze the data which is stored in HDFS. The hive was developed by Facebook to analyze the huge amount of unstructured data and its work on top of the Hadoop while Apache Spark is also a lightning-fast cluster computing technology, designed for fast computation.

Research Questions: Some of the problems identify in airlines data and the analysis is done in this paper, the airlines market has been faced losses due to the flight delay and there are many reasons for delaying a flight, In this paper, we analyze the various delay happens in airlines per year. We can analyze the delay in this paper are:

DOI 10.17148/IJARCCE.2020.9537

1. Year wise carrier delay from 2003-2017
2. Year wise NAS delay from 2003-2017
3. Year wise Weather delay from 2003-2017
4. Year wise late aircraft delay from 2003-2017
5. Year wise security delay from 2003-2017

Hadoop: we can use Pseudo-distributed Hadoop cluster mode in which all the Hadoop daemons are running on a single machine [9].

Data set: The airline's data is using which is in .csv format means each record in the datasets are separated by a new line and the fields in each record are separated by a comma.

Tools & Technology used:

1. Hadoop
2. Spark
3. Hive

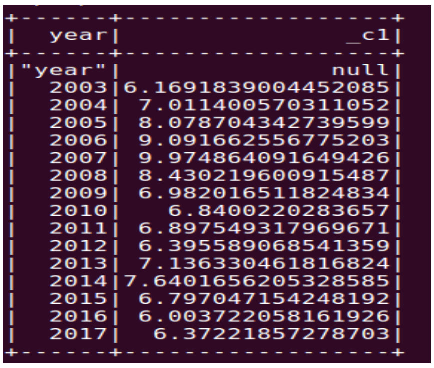
V. EXPERIMENTAL FINDING

Analyzing using spark

Input: Airlines Dataset Output: Year wise carrier delay

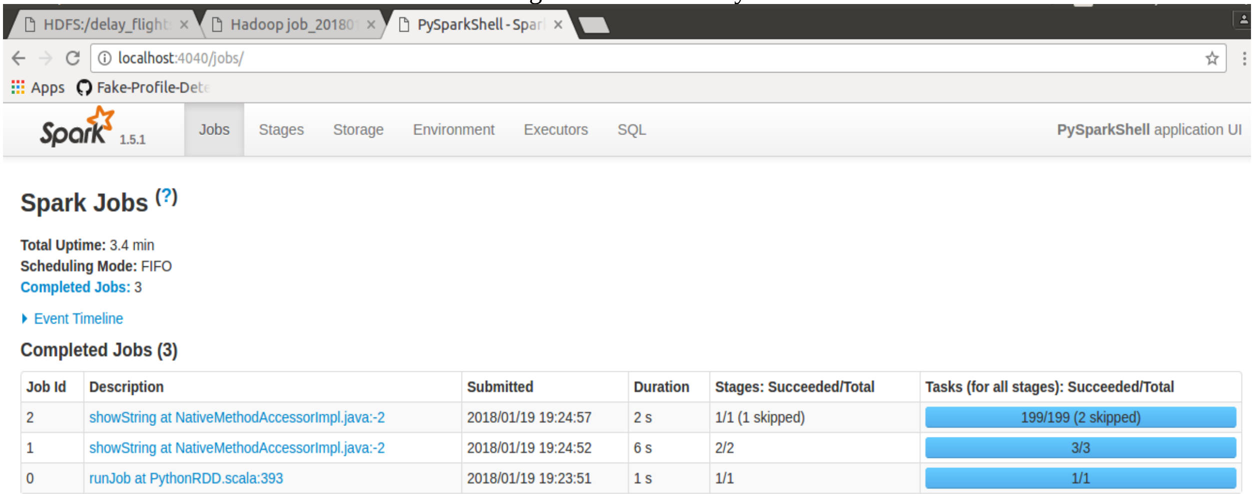
Algorithm used in spark

```
1.          import SQLContext and  row
2.          load the data set and split the records
parts = lines.map(lambda l: l.split(","))
3.          construct the Rows by passing a list of key/value pairs
4.          Create the DataFrame and register it has Table
            schema=sqlContext.createDataFrame(delay  _flights)
            schema.registerTempTable("delay_flights")
5.          run the query for getting the required result
result=sqlContext.sql("select year, avg((carrier_ct /arr_del)*100) from delay_flights group by year").show()
```



```
+------+-------------------+
|  year|                _c1|
+------+-------------------+
|"year"|               null|
|  2003|6.1691839004452085|
|  2004| 7.011400570311052|
|  2005| 8.078704342739599|
|  2006| 9.091662556775203|
|  2007| 9.974864091649426|
|  2008| 8.430219600915487|
|  2009| 6.982016511824834|
|  2010|    6.8400220283657|
|  2011| 6.897549317969671|
|  2012| 6.395589068541359|
|  2013| 7.136330461816824|
|  2014|7.6401656205328585|
|  2015| 6.797047154248192|
|  2016| 6.003722058161926|
|  2017|  6.37221857278703|
+------+-------------------+
```

Fig. 3 Year wise delay



Spark Jobs (?)

Total Uptime: 3.4 min
Scheduling Mode: FIFO
Completed Jobs: 3

Event Timeline

Completed Jobs (3)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
2	showString at NativeMethodAccessorImpl.java:-2	2018/01/19 19:24:57	2 s	1/1 (1 skipped)	199/199 (2 skipped)
1	showString at NativeMethodAccessorImpl.java:-2	2018/01/19 19:24:52	6 s	2/2	3/3
0	runJob at PythonRDD.scala:393	2018/01/19 19:23:51	1 s	1/1	1/1

Figure-4. Time taken by spark framework

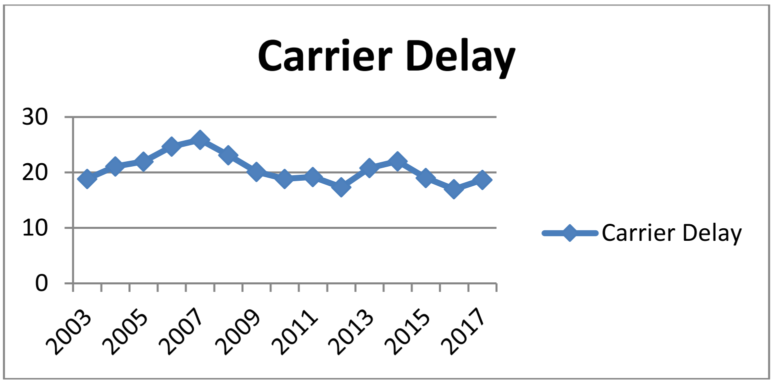


Figure 5. Year wise carrier delay

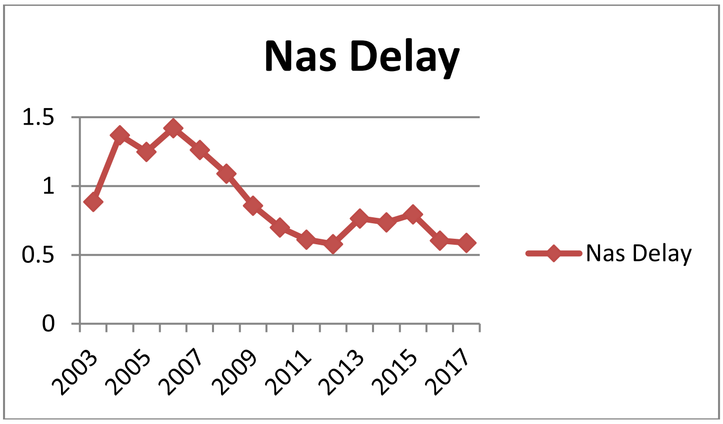


Figure 6. Year wise NAS delay

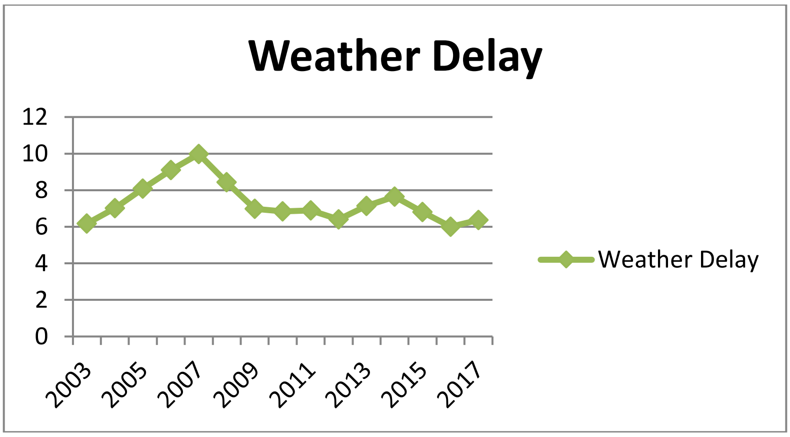


Figure 7. Year-wise weather delay

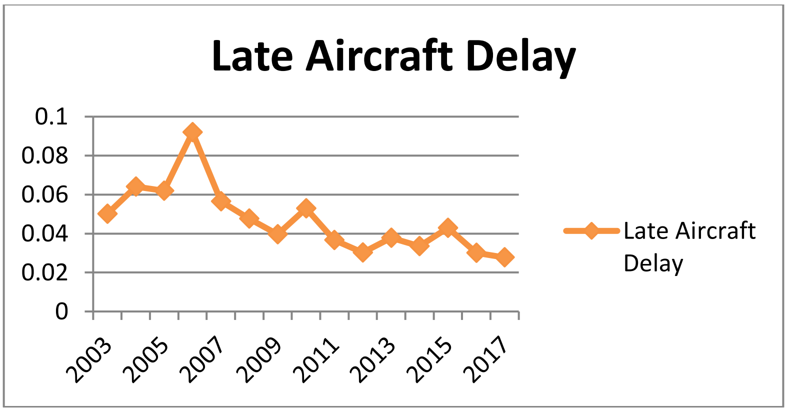


Figure 8. Year-wise Late Aircraft delay

DOI 10.17148/IJARCCE.2020.9537

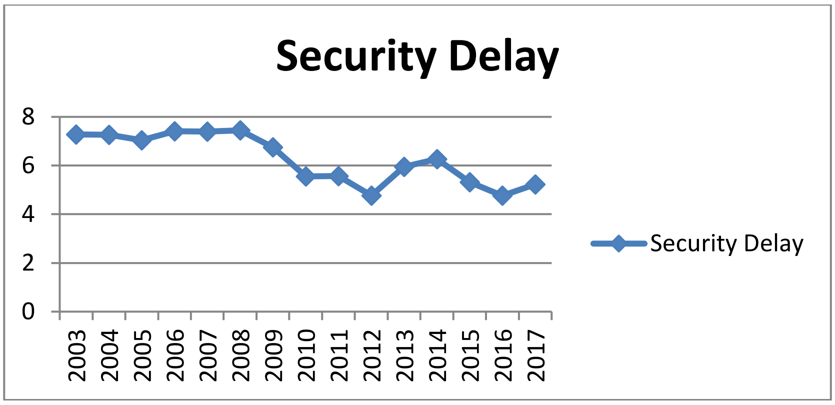
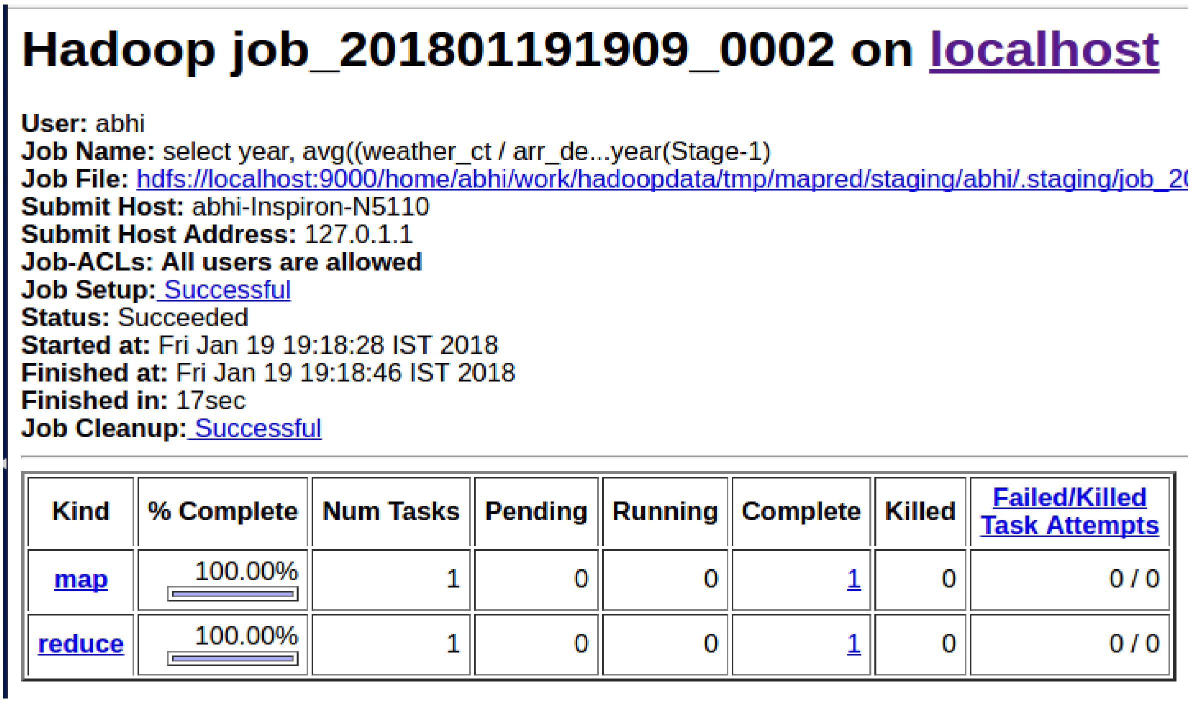


Figure 9. Year-wise security delay

Analysing using Apache Hive

After analyzing the datasets using Spark-SQL over spark framework we can also analyze the dataset by using Hive-QL which is also same as SQL like language, Hive-QL is running over MapReduce framework and after analyzing the datasets by Hive-QL we can find the same result as we get from Spark-SQL, so from which we can say that both are very accurate in terms of result, but the time taken by Hive-QL over MapReduce framework is shown in figure 13.



Hadoop job_201801191909_0002 on localhost

User: abhi
Job Name: select year, avg((weather_ct / arr_de...year(Stage-1)
Job File: hdfs://localhost:9000/home/abhi/work/hadoopdata/tmp/mapred/staging/abhi/.staging/job_2(
Submit Host: abhi-Inspiron-N5110
Submit Host Address: 127.0.1.1
Job-ACLs: All users are allowed
Job Setup: Successful
Status: Succeeded
Started at: Fri Jan 19 19:18:28 IST 2018
Finished at: Fri Jan 19 19:18:46 IST 2018
Finished in: 17sec
Job Cleanup: Successful

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	100.00%	1	0	0	1	0	0 / 0
reduce	100.00%	1	0	0	1	0	0 / 0

Figure 13. Time taken by Hive-QL

Performance Comparison of Spark-SQL and Hive-QL

Table.1 Execution time taken by hiveQL and spark-SQL

Time taken by query (in sec)	HiveQL	Spark-SQL
Carrier delay query	17	8
NAS delay query	19	11
Weather delay query	19	10
Late aircraft delay query	22	15
Security delay query	23	15

DOI 10.17148/IJARCCE.2020.9537

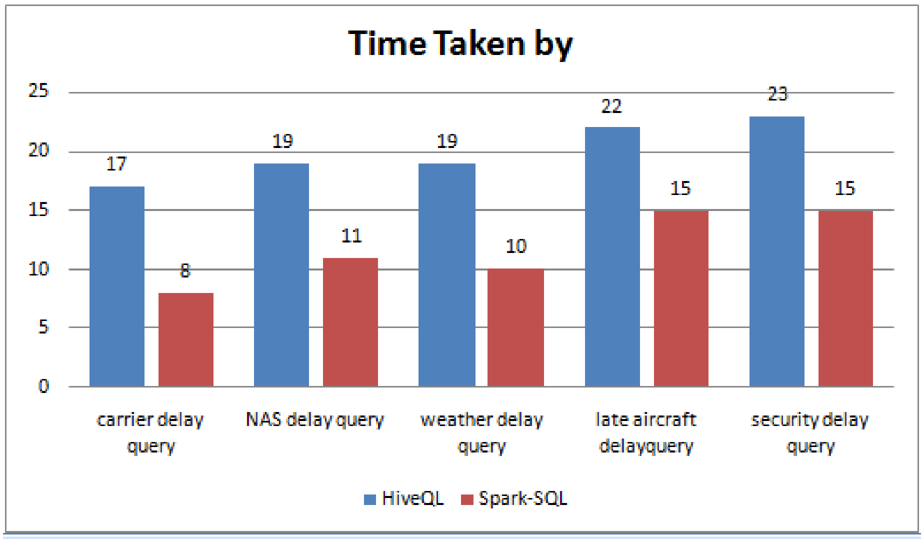


Figure-14 Execution time taken by Hive-QL and Spark-SQL

VI. CONCLUSION

On analyzing complete scenario regarding the analysis of big data we say that using the traditional analytical tool we cannot perform analysis on such huge and complex data, so we use a new powerful tool which is designed for deep analysis called Hadoop. HDFS is used for storing huge amount of airlines data, Hive and spark have been used for querying the data in which hive uses Hive-QL statements which runs on MapReduce framework and spark uses Spark-SQL which runs on spark framework. Data visualization has been done by extracting the output of the HIVE and SPARK query in excel and plotting the data using line and bar plot charts. The visualization of the data shows some patterns that exist different airlines delays caused by weather, security, NAS delay, late aircraft delay etc. In future, we can analyze the data of different sector.

REFERENCES

[1] Ilias Mavridisa, Eleni Karatzab ," Performance evaluation of cloud-based log file analysis with Apache Hadoop and Apache Spark" in *The Journal of Systems & Software* (2016), doi: 10.1016/j.jss.2016.11.037, Elsevier.
[2] Davide Scotti a, Nicola Volta , "Profitability change in the global airline industry"in 2017 Elsevier.
[3] Loren Moir, Gui Lohmann, " A quantitative means of comparing competitive advantage among airlines with heterogeneous business models: Analysis of U.S. airlines" in 2018 Elsevier Ltd
[4] Simon Mulwa Kiio, Elisha O. Abade, “Apache Spark based Big Data Analytics for Social Network Cybercrime Forensics” in International Journal of Computer Applications (0975 – 8887) Volume 179 – No.8, December 2017.
[5] Dave Jaffe “Three Approaches to Data Analysis with Hadoop”
[6] S. K. Pushpa , Manjunath T. N., Srividhya, “Analysis of Airport Data using Hadoop-Hive: A Case Study” in International Journal of Computer Applications (0975 – 8887) National Conference on “Recent Trends in Information Technology” (NCRTIT-2016)
[7] “Fast and Interactive Analytics over Hadoop Data with Spark” in august 2012.
[8] Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce",6-8 Dec. 2012.
[9] Michael G. Noll, Applied Research, Big Data, Distributed Systems, Open Source, "Running Hadoop on Ubuntu Linux (Single-Node Cluster)", [online], available at http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/
[10] Sagiroglu,S., & Sinanc,D, “Big data: A review”, IEEE International Conf on Collaboration Technologies and Systems (CTS), 2013, pp 42-47.
[11] https://hive.apache.org/”

BIOGRAPHY

Miss Sanchita Chourawar pursed B.E from University of Bhilai in 2013 and Master of technology from RGPV University in year 2018. She is currently working as Assistant Professor in Department of Computer Science Engineering, Bilaspur(C.G.) s. She is a research work focuses on Cloud Security and Privacy, Big Data Analytics, Data Mining, IoT and Computational Intelligence based education. She has 2.5 years of teaching experience.

DOI 10.17148/IJARCCE.2020.9537