Big Data Assignment

1. Github Reference Link: https://github.com/sachithr7/248367H_SachithR_BigData_Assignment/tree/main/UoM_MapReduce-vs-Spark

2. Solutions:

D. The airline market has faced losses due to the flight delay and there are many reasons for delaying a flight. In this Analysis, you need to analyse the various delays that happen in airlines per year and run the queries as follows.

Approach 1: using EMR pyspark integration to analyse data
I have attached all pyspark outputfiles to the gitlab repository

Approach 2: Direct pyspark execution on the EMR server

1. Year wise carrier delay from 2003-2010

```
>>> df = spark.read.option("header", "true").csv("s3://bigdataassignment2024/data/flightdata.csv")
>>> spark.sql("SELECT Year as YEAR, avg((CarrierDelay/ArrDelay)*100) as CarrerDalyPresentage FROM delay_flights
WHERE Year BETWEEN 2003 AND 2010 GROUP BY Year").show()
+----+--------------------+
|YEAR|CarrerDalyPresentage|
+----+--------------------+
|2010|   21.89310246015957|
|2003|  24.557549755575373|
|2005|   28.01977637202288|
|2009|   28.33058554239575|
|2006|  30.453296261292596|
|2004|   43.64459443230066|
|2008|   28.88346981456985|
|2007|  19.850007017971283|
+----+--------------------+
>>>
```

2. Year wise NAS delay from 2003-2010

```
>>>
>>> spark.sql("SELECT Year as YEAR, avg((NASDelay/ArrDelay)*100) as CarrerDalyPresentage FROM delay_flights
WHERE Year BETWEEN 2003 AND 2010 GROUP BY Year").show()
+----+--------------------+
|YEAR|CarrerDalyPresentage|
+----+--------------------+
|2010|   33.87351363404217|
|2003|  29.686276314267346|
|2005|   16.63868805373129|
|2009|   37.63093330628511|
|2006|  18.119312329937703|
|2004|   18.24570061769958|
|2008|   30.16552562594132|
|2007|  30.625925917941924|
+----+--------------------+
>>>
```

3. Year wise Weather delay from 2003-2010

```
>>>
>>> spark.sql("SELECT Year as YEAR, avg((WeatherDelay/ArrDelay)*100) as CarrerDalyPresentage FROM delay_flights
WHERE Year BETWEEN 2003 AND 2010 GROUP BY Year").show()
+----+--------------------+
|YEAR|CarrerDalyPresentage|
+----+--------------------+
|2010|  2.9023312955584664|
|2003|  7.8319479664511205|
|2005|    5.85069715149616|
|2009| 0.45316615137982363|
|2006|   4.588604183967953|
|2004|  6.4475279976916555|
|2008|  3.7254490054008955|
|2007|   4.042975783210287|
+----+--------------------+
>>>
```

4. Year wise late aircraft delay from 2003-2010

```
>>>
>>> spark.sql("SELECT Year as YEAR, avg((LateAircraftDelay/ArrDelay)*100) as CarrerDalyPresentage FROM
delay_flights WHERE Year BETWEEN 2003 AND 2010 GROUP BY Year").show()
+----+--------------------+
|YEAR|CarrerDalyPresentage|
+----+--------------------+
|2010|  41.331052610239794|
|2003|  37.924225963706164|
|2005|  49.490838422749654|
|2009|  33.585314999939314|
|2006|  46.838787224801735|
|2004|  31.662176952308105|
|2008|   37.22555555408794|
|2007|  45.252432744291134|
+----+--------------------+
>>>
```

5. Year wise security delay from 2003-2010

```
>>>
>>> spark.sql("SELECT Year as YEAR, avg((SecurityDelay/ArrDelay)*100) as CarrerDalyPresentage FROM delay_flights
WHERE Year BETWEEN 2003 AND 2010 GROUP BY Year").show()
+----+--------------------+
|YEAR|CarrerDalyPresentage|
+----+--------------------+
|2010|                 0.0|
|2003|                 0.0|
|2005|                 0.0|
|2009|                 0.0|
|2006|                 0.0|
|2004|                 0.0|
|2008|                 0.0|
|2007| 0.22865853658536586|
+----+--------------------+
>>>
```

E. Run the query using Hadoop and Spark for 5 times and plot the graph in comparing both methods (running time vs iteration).

1. Pre Task - Local Terminal - Copy shell script to EMR cluster

```
sacithrangana@Saciths-MacBook-Pro Fully_completed_final_solution_with_shell_scripting_spark_and_mapReduce % ls
-ltr
total 224
-r-x------@ 1 sacithrangana  staff  100974 Mar  3 00:15 flightdata.csv
-rw-r--r--  1 sacithrangana  staff    2358 Mar  4 06:03 248367H_spark_hadoop_iterative_comparison.sh
-rw-r--r--  1 sacithrangana  staff    7770 Mar  4 06:09 248367H_spark_hadoop_execution_for_multiple_queries.sh
sacithrangana@Saciths-MacBook-Pro Fully_completed_final_solution_with_shell_scripting_spark_and_mapReduce %
scp -i ~/Downloads/BIG_DATA_ASSIGNMENT_001.pem ./flightdata.csv
hadoop@ec2-44-220-247-28.compute-1.amazonaws.com:./
The authenticity of host 'ec2-44-220-247-28.compute-1.amazonaws.com (44.220.247.28)' can't be established.
ED25519 key fingerprint is SHA256:otW9+Cp3/Zaa58bjOpPl7tD/cidEsLxvN8CxaM+xdGo.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-44-220-247-28.compute-1.amazonaws.com' (ED25519) to the list of known hosts.
flightdata.csv
100%   99KB  63.1KB/s   00:01
sacithrangana@Saciths-MacBook-Pro Fully_completed_final_solution_with_shell_scripting_spark_and_mapReduce %
scp -i ~/Downloads/BIG_DATA_ASSIGNMENT_001.pem ./248367H_spark_hadoop_iterative_comparison.sh
hadoop@ec2-44-220-247-28.compute-1.amazonaws.com:./
248367H_spark_hadoop_iterative_comparison.sh
100% 2358     5.9KB/s   00:00
sacithrangana@Saciths-MacBook-Pro Fully_completed_final_solution_with_shell_scripting_spark_and_mapReduce %
scp -i ~/Downloads/BIG_DATA_ASSIGNMENT_001.pem ./248367H_spark_hadoop_execution_for_multiple_queries.sh
hadoop@ec2-44-220-247-28.compute-1.amazonaws.com:./
248367H_spark_hadoop_execution_for_multiple_queries.sh
100% 7770    18.5KB/s   00:00
sacithrangana@Saciths-MacBook-Pro Fully_completed_final_solution_with_shell_scripting_spark_and_mapReduce %
```

2. EMR Terminal - Execute iterative shell script on the server

```
[hadoop@ip-172-31-68-240 ~]$ ls -ltr
total 112
-r-x------. 1 hadoop hadoop 100974 Mar  4 02:02 flightdata.csv
-rw-r--r--. 1 hadoop hadoop   2358 Mar  4 02:02 248367H_spark_hadoop_iterative_comparison.sh
-rw-r--r--. 1 hadoop hadoop   7770 Mar  4 02:03 248367H_spark_hadoop_execution_for_multiple_queries.sh
[hadoop@ip-172-31-68-240 ~]$ chmod +x 248367H_spark_hadoop_iterative_comparison.sh
[hadoop@ip-172-31-68-240 ~]$ chmod +x 248367H_spark_hadoop_execution_for_multiple_queries.sh
[hadoop@ip-172-31-68-240 ~]$ ls -ltr
total 112
-r-x------. 1 hadoop hadoop 100974 Mar  4 02:02 flightdata.csv
-rwxr-xr-x. 1 hadoop hadoop   2358 Mar  4 02:02 248367H_spark_hadoop_iterative_comparison.sh
-rwxr-xr-x. 1 hadoop hadoop   7770 Mar  4 02:03 248367H_spark_hadoop_execution_for_multiple_queries.sh
[hadoop@ip-172-31-68-240 ~]$ ./248367H_spark_hadoop_iterative_comparison.sh
Running Spark query iteration 1...
Mar 04, 2024 2:06:10 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:06:22 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 02:06:40 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take
effect.
24/03/04 02:06:47 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This
behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
```

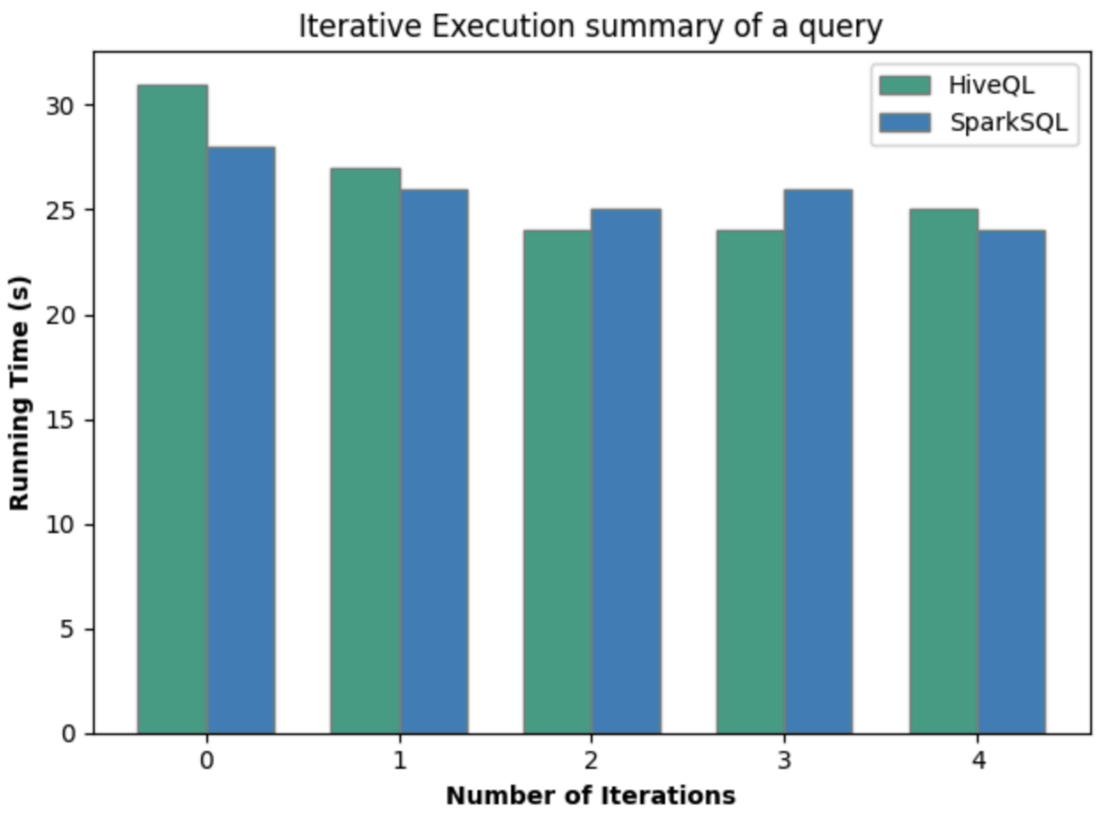
```
Mar 04, 2024 2:06:48 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but /usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:07:01 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
24/03/04 02:07:14 WARN HiveConf: HiveConf of name hive.server2.thrift.url does not exist
Mar 04, 2024 2:07:16 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but /usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:07:26 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
Running Hive query iteration 1...
Hive Session ID = ffb92f3c-8cd2-4efd-acd0-2a9de140c3eb
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
OK
Time taken: 0.838 seconds
Hive Session ID = ef48bcfb-0bbb-476a-914f-d07bcb089b3a
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Query ID = hadoop_20240304020806_0aea65da-4511-4b6a-84ec-3bb0ba1455da
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709517553576_0006)
Map 1: -/-	Reducer 2: 0/2
Map 1: -/-	Reducer 2: 0/2
Map 1: -/-	Reducer 2: 0(+2)/2
Map 1: -/-	Reducer 2: 2/2
OK
Time taken: 18.351 seconds
Running Spark query iteration 2...
Mar 04, 2024 2:08:25 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but /usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:08:36 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
24/03/04 02:08:49 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
24/03/04 02:08:56 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
Mar 04, 2024 2:08:57 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but /usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:09:07 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
24/03/04 02:09:22 WARN HiveConf: HiveConf of name hive.server2.thrift.url does not exist
Mar 04, 2024 2:09:24 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but /usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:09:34 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
Running Hive query iteration 2...
Hive Session ID = 20646936-ff11-46fd-8e0e-5ecc9a217e23
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
```

```
OK
Time taken: 0.492 seconds
Hive Session ID = ec6c18a2-ce48-4360-9ab1-0944d6684650
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Query ID = hadoop_20240304021009_ed47edec-5005-4635-b52a-3020e634d2bd
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709517553576_0012)
Map 1: -/-        Reducer 2: 0/2
Map 1: -/-        Reducer 2: 0/2
Map 1: -/-        Reducer 2: 0(+2)/2
Map 1: -/-        Reducer 2: 2/2
OK
Time taken: 16.361 seconds
Running Spark query iteration 3...
Mar 04, 2024 2:10:27 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:10:37 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 02:10:49 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take
effect.
24/03/04 02:10:58 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This
behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
Mar 04, 2024 2:10:58 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:11:09 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 02:11:22 WARN HiveConf: HiveConf of name hive.server2.thrift.url does not exist
Mar 04, 2024 2:11:24 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:11:35 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
Running Hive query iteration 3...
Hive Session ID = 8c12ddf8-e9ec-4524-9ea0-5087792e0d01
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
OK
Time taken: 0.676 seconds
Hive Session ID = 6d3f4b53-3014-4d76-9ae9-ef4f0859a15a
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Query ID = hadoop_20240304021209_995347bc-b6c9-42e1-aea2-4b95392fa817
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709517553576_0018)
Map 1: -/-        Reducer 2: 0/2
Map 1: -/-        Reducer 2: 0/2
Map 1: -/-        Reducer 2: 0(+1)/2
Map 1: -/-        Reducer 2: 0(+2)/2
Map 1: -/-        Reducer 2: 2/2
OK
Time taken: 13.9 seconds
Running Spark query iteration 4...
Mar 04, 2024 2:12:24 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
```

```
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:12:36 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 02:12:49 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take
effect.
24/03/04 02:12:56 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This
behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
Mar 04, 2024 2:12:57 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:13:07 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 02:13:21 WARN HiveConf: HiveConf of name hive.server2.thrift.url does not exist
Mar 04, 2024 2:13:23 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:13:33 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
Running Hive query iteration 4...
Hive Session ID = 2cbde2a7-2e7b-4ad5-aa1b-196d4988dd83
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
OK
Time taken: 0.629 seconds
Hive Session ID = 3cfc8347-8b21-464a-9261-d40f3c89e41a
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Query ID = hadoop_20240304021407_1a12ca76-b3d1-4a90-a03a-d81ef19c02d8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709517553576_0024)
Map 1: -/-        Reducer 2: 0/2
Map 1: -/-        Reducer 2: 0/2
Map 1: -/-        Reducer 2: 0(+2)/2
Map 1: -/-        Reducer 2: 2/2
OK
Time taken: 14.357 seconds
Running Spark query iteration 5...
Mar 04, 2024 2:14:22 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:14:32 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 02:14:45 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take
effect.
24/03/04 02:14:53 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This
behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
Mar 04, 2024 2:14:54 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:15:04 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 02:15:17 WARN HiveConf: HiveConf of name hive.server2.thrift.url does not exist
```

```
Mar 04, 2024 2:15:19 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:15:30 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
Running Hive query iteration 5...
Hive Session ID = b0c9da58-b937-4932-8303-4e1e54e640ba
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
OK
Time taken: 0.636 seconds
Hive Session ID = 29760cf1-1a8f-4686-9b6e-d94cb8439687
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Query ID = hadoop_20240304021603_418c0e1b-73b6-4733-9c84-b9df425217ed
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709517553576_0030)
Map 1: -/-          Reducer 2: 0/2
Map 1: -/-          Reducer 2: 0/2
Map 1: -/-          Reducer 2: 0(+2)/2
Map 1: -/-          Reducer 2: 1(+1)/2
Map 1: -/-          Reducer 2: 2/2
OK
Time taken: 15.068 seconds
[hadoop@ip-172-31-68-240 ~]$ ls -ltr
total 116
-r-x------. 1 hadoop hadoop 100974 Mar  4 02:02 flightdata.csv
-rwxr-xr-x. 1 hadoop hadoop   2358 Mar  4 02:02 248367H_spark_hadoop_iterative_comparison.sh
-rwxr-xr-x. 1 hadoop hadoop   7770 Mar  4 02:03 248367H_spark_hadoop_execution_for_multiple_queries.sh
-rw-r--r--. 1 hadoop hadoop     91 Mar  4 02:16 execution_times_common_query_iterator.csv
[hadoop@ip-172-31-68-240 ~]$ cat execution_times_common_query_iterator.csv
Iteration,Spark Execution Time,Hive Execution Time
1,28,31
2,26,27
3,25,24
4,26,24
5,24,25
[hadoop@ip-172-31-68-240 ~]$
```

Iteration	Spark-SQL	HiveQL
1	28	31
2	26	27
3	25	24
4	26	24
5	24	25



F. Similarly process all queries and plot the time-comparison graphs as shown above.

```
[hadoop@ip-172-31-68-240 ~]$ ls -ltr
total 116
-r-x------. 1 hadoop hadoop 100974 Mar  4 02:02 flightdata.csv
-rwxr-xr-x. 1 hadoop hadoop   2358 Mar  4 02:02 248367H_spark_hadoop_iterative_comparison.sh
-rw-r--r--. 1 hadoop hadoop     91 Mar  4 02:16 execution_times_common_query_iterator.csv
-rwxr-xr-x. 1 hadoop hadoop   7739 Mar  4 02:58 248367H_spark_hadoop_execution_for_multiple_queries.sh
[hadoop@ip-172-31-68-240 ~]$
[hadoop@ip-172-31-68-240 ~]$ ./248367H_spark_hadoop_execution_for_multiple_queries.sh
Running Spark and hadoop for multiple query no 1 ...
Mar 04, 2024 2:59:37 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 02:59:47 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
24/03/04 02:59:59 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
24/03/04 03:00:07 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
Running spark query no 1 ...
Mar 04, 2024 3:00:08 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 03:00:17 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
```

```
24/03/04 03:00:30 WARN HiveConf: HiveConf of name hive.server2.thrift.url does not exist
Mar 04, 2024 3:00:32 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 03:00:43 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
Running Hive query 1 ...
Hive Session ID = 8f4417b5-be42-4df3-a21c-b4714593f280
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
OK
Time taken: 0.527 seconds
Hive Session ID = 0f3aad2a-599a-4369-98aa-289a2bd1b65a
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Query ID = hadoop_20240304030115_7d16f1e1-857b-49c1-863d-b3df067f1275
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709517553576_0093)
Map 1: -/-        Reducer 2: 0/2
Map 1: -/-        Reducer 2: 0/2
Map 1: -/-        Reducer 2: 0(+2)/2
Map 1: -/-        Reducer 2: 2/2
OK
Time taken: 13.479 seconds
Running Spark and hadoop for multiple query no 2 ...
Mar 04, 2024 3:01:29 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 03:01:39 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 03:01:51 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take
effect.
24/03/04 03:01:58 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This
behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
Running spark query no 2 ...
Mar 04, 2024 3:01:59 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 03:02:08 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 03:02:21 WARN HiveConf: HiveConf of name hive.server2.thrift.url does not exist
Mar 04, 2024 3:02:23 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 03:02:33 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
Running Hive query 2 ...
Hive Session ID = 08fea578-7069-4271-948b-a19efb4468e9
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
OK
Time taken: 0.613 seconds
Hive Session ID = 830f5f98-09dd-4e71-8b20-e6dc37bb3ae3
```

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Query ID = hadoop_20240304030305_199e60c7-8c88-41fe-bf8a-83927ab67977
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709517553576_0099)
Map 1: -/-          Reducer 2: 0/2
Map 1: -/-          Reducer 2: 0/2
Map 1: -/-          Reducer 2: 0(+2)/2
Map 1: -/-          Reducer 2: 2/2
OK
Time taken: 13.976 seconds
Running Spark and hadoop for multiple query no 3 ...
Mar 04, 2024 3:03:20 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 03:03:31 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 03:03:43 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take
effect.
24/03/04 03:03:49 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This
behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
Running spark query 3 ...
Mar 04, 2024 3:03:50 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 03:04:00 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 03:04:13 WARN HiveConf: HiveConf of name hive.server2.thrift.url does not exist
Mar 04, 2024 3:04:15 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 03:04:24 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 03:04:33 WARN YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to request executors before the
AM has registered!
Running Hive query 3 ...
Hive Session ID = e1e1d530-5917-4a22-a3cd-3a3d14ed9fd9
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
OK
Time taken: 0.405 seconds
Hive Session ID = 9017c0df-6c94-4bb5-b390-121692c39a85
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Query ID = hadoop_20240304030455_9d301a38-ae3e-4aeb-9209-c94d2e06357b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709517553576_0105)
Map 1: -/-          Reducer 2: 0/2
Map 1: -/-          Reducer 2: 0/2
Map 1: -/-          Reducer 2: 0(+2)/2
Map 1: -/-          Reducer 2: 2/2
OK
Time taken: 14.189 seconds
Running Spark and hadoop for multiple query no 4 ...
```

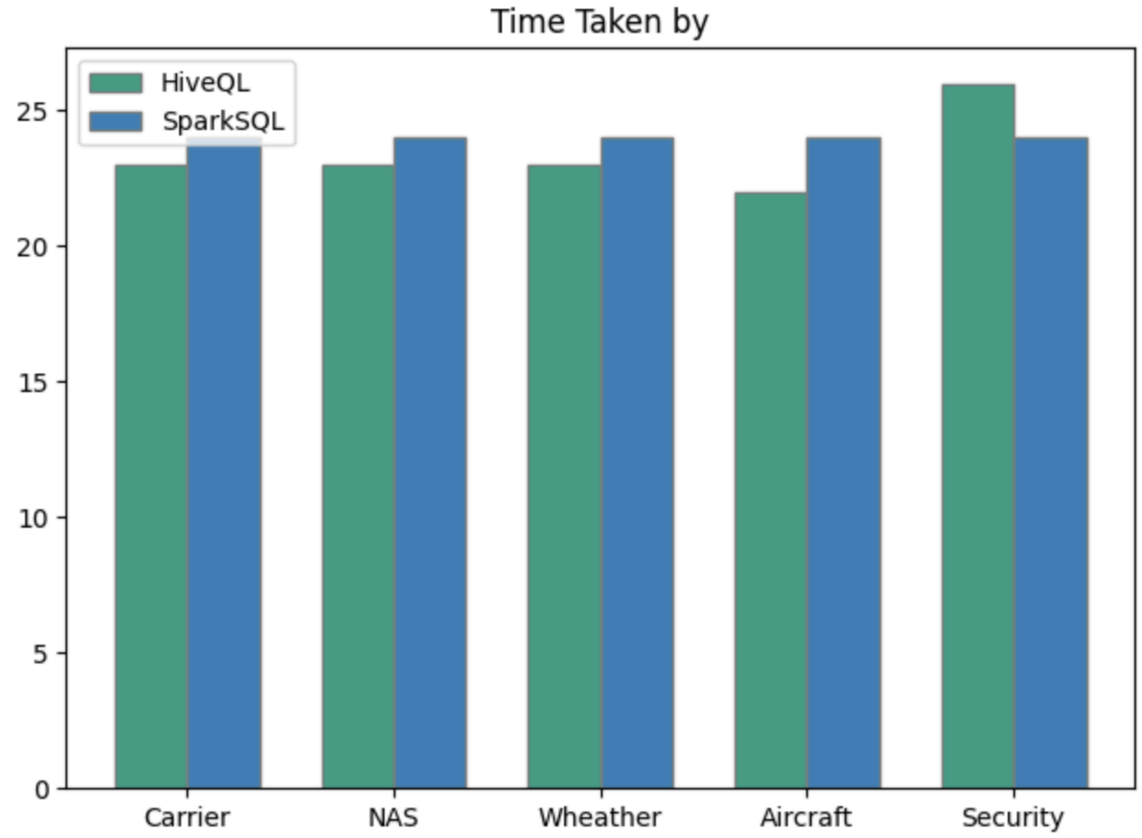
```
Mar 04, 2024 3:05:10 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 03:05:19 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 03:05:32 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take
effect.
24/03/04 03:05:38 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This
behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
Running spark query 4 ...
Mar 04, 2024 3:05:39 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 03:05:49 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 03:06:01 WARN HiveConf: HiveConf of name hive.server2.thrift.url does not exist
Mar 04, 2024 3:06:03 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 03:06:12 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
Running Hive query 4 ...
Hive Session ID = 33589c8d-5dce-4de1-b4e3-387e6920a611
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
OK
Time taken: 0.52 seconds
Hive Session ID = 01d1b39d-f346-4ba7-bdf5-3c0abb95c9e8
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Query ID = hadoop_20240304030643_4e34c000-c46a-4597-8117-42f618af5a06
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709517553576_0111)
Map 1: -/-          Reducer 2: 0/2
Map 1: -/-          Reducer 2: 0/2
Map 1: -/-          Reducer 2: 0(+2)/2
Map 1: -/-          Reducer 2: 1(+1)/2
Map 1: -/-          Reducer 2: 2/2
OK
Time taken: 13.059 seconds
Running Spark and hadoop for multiple query no 5 ...
Mar 04, 2024 3:06:56 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 03:07:07 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 03:07:19 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take
effect.
24/03/04 03:07:27 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This
behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
Running spark query 5 ...
Mar 04, 2024 3:07:28 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
```

```
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 03:07:37 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
24/03/04 03:07:50 WARN HiveConf: HiveConf of name hive.server2.thrift.url does not exist
Mar 04, 2024 3:07:52 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but
/usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/03/04 03:08:03 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries
under SPARK_HOME.
Running Hive query 5 ...
Hive Session ID = 511a642a-ddd4-4c66-b662-c9acae25adb4
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
OK
Time taken: 0.466 seconds
Hive Session ID = e42736be-2713-42a1-a524-01213c4dfcf8
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Query ID = hadoop_20240304030835_d0dd1f38-dbab-4bb2-8cc4-7e73c2dcc40d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709517553576_0117)
Map 1: -/-          Reducer 2: 0/2
Map 1: -/-          Reducer 2: 0/2
Map 1: -/-          Reducer 2: 0(+2)/2
Map 1: -/-          Reducer 2: 2/2
OK
Time taken: 15.119 seconds
[hadoop@ip-172-31-68-240 ~]$ ls -ltr
total 120
-r-x------. 1 hadoop hadoop 100974 Mar  4 02:02 flightdata.csv
-rwxr-xr-x. 1 hadoop hadoop   2358 Mar  4 02:02 248367H_spark_hadoop_iterative_comparison.sh
-rw-r--r--. 1 hadoop hadoop     91 Mar  4 02:16 execution_times_common_query_iterator.csv
-rwxr-xr-x. 1 hadoop hadoop   7739 Mar  4 02:58 248367H_spark_hadoop_execution_for_multiple_queries.sh
-rw-r--r--. 1 hadoop hadoop    179 Mar  4 03:08 execution_times_for_each_query.csv
[hadoop@ip-172-31-68-240 ~]$ cat execution_times_for_each_query.csv
Time taken by query in(sec),HiveQL,Spark-SQL
Carrier Delay Query,23,24
NAS Delay Query,23,24
Weather Delay Query,23,24
Late Aircraft Delay Query,22,24
Security Delay Query,26,24
[hadoop@ip-172-31-68-240 ~]$
```

Time taken by query in (sec)	HiveQL	Spark-SQL
Carrier Delay Query	23	24
NAS Delay Query	23	24
Weather Delay Query	23	24
Late Aircraft Delay Query	22	24

Security Delay Query	26	24

G. Calculate average time taken by MapReduce and Spark for each query and plot the graph.



H. Combine all your coding and screenshots in your GitHub account and share the link.

- Create directory named, 'UoM_MapReduce-vs-Spark' - done
- Create folders within the above directory named, 'MapReduce' and 'Spark' - done
- Put coding and screenshots for results inside the respective folders. - done