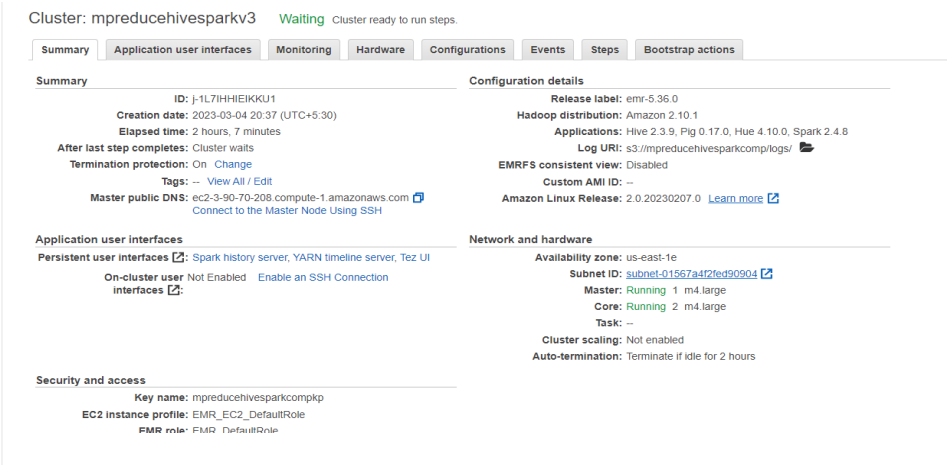
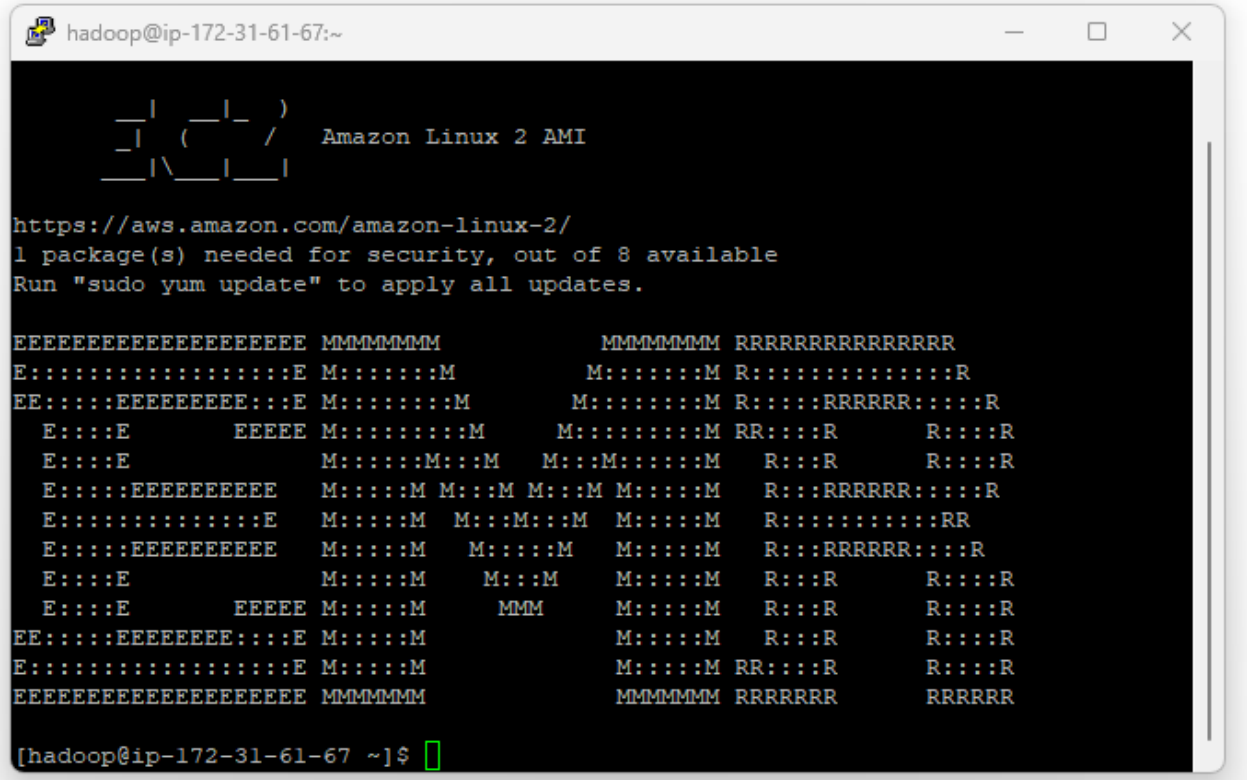
MapReduce & Apace Spark Comparison

Cluster creation



EMR Start



Hive SQL

Creating table

```
hive> CREATE DATABASE flightdelay_db_hive;
OK
Time taken: 0.1 seconds
hive> CREATE EXTERNAL TABLE flightdelay_db_hive.delayedflight ( Year INT,Month I
NT,DayofMonth INT,DayOfWeek INT,DepTime INT,CRSDepTime INT,ArrTime INT,CRSArrTim
e INT,UniqueCarrier STRING,
    >    FlightNum INT,TailNum STRING,ActualElapsedTime INT,CRSElapsedTime INT,
    >    AirTime INT,
    >    ArrDelay INT,
    >    DepDelay INT,
    >    Origin STRING,
    >    Dest STRING,
    >    Distance INT,
    >    TaxiIn INT,
    >    TaxiOut INT,
    >    Cancelled INT,
    >    CancellationCode STRING,
    >    Diverted INT,
    >    CarrierDelay INT,
    >    WeatherDelay INT,
    >    NASDelay INT,
    >    SecurityDelay INT,
    >    LateAircraftDelay INT
    > );
OK
Time taken: 1.827 seconds
```

Data loading

```
hive>
    > load data inpath 's3://mpreducehivesparkcompp/DelayedFlights-updated.csv'
into table flightdelay_db_hive.delayedflight;
Loading data to table flightdelay_db_hive.delayedflight
OK
Time taken: 2.19 seconds
```

Script running (Hive SQL)

```
hive> SELECT Year, SUM(CarrierDelay) AS CarrierDelayYearly
    > FROM flightdelay_db_hive.delayedflightrevised2
    > WHERE Year BETWEEN 2003 AND 2010
    > GROUP BY Year;
Query ID = hadoop_20230304153245_aba1926d-5176-497f-8cd9-cae674c65320
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1677942823688_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 18.53 s
--------------------------------------------------------------------------------
OK
2003    1577
2007    1835
2009    1191
2010    1992
2004    2064
2005    1436
2006    1540
2008    1298
Time taken: 28.624 seconds, Fetched: 8 row(s)
```

Script running (Spark SQL)

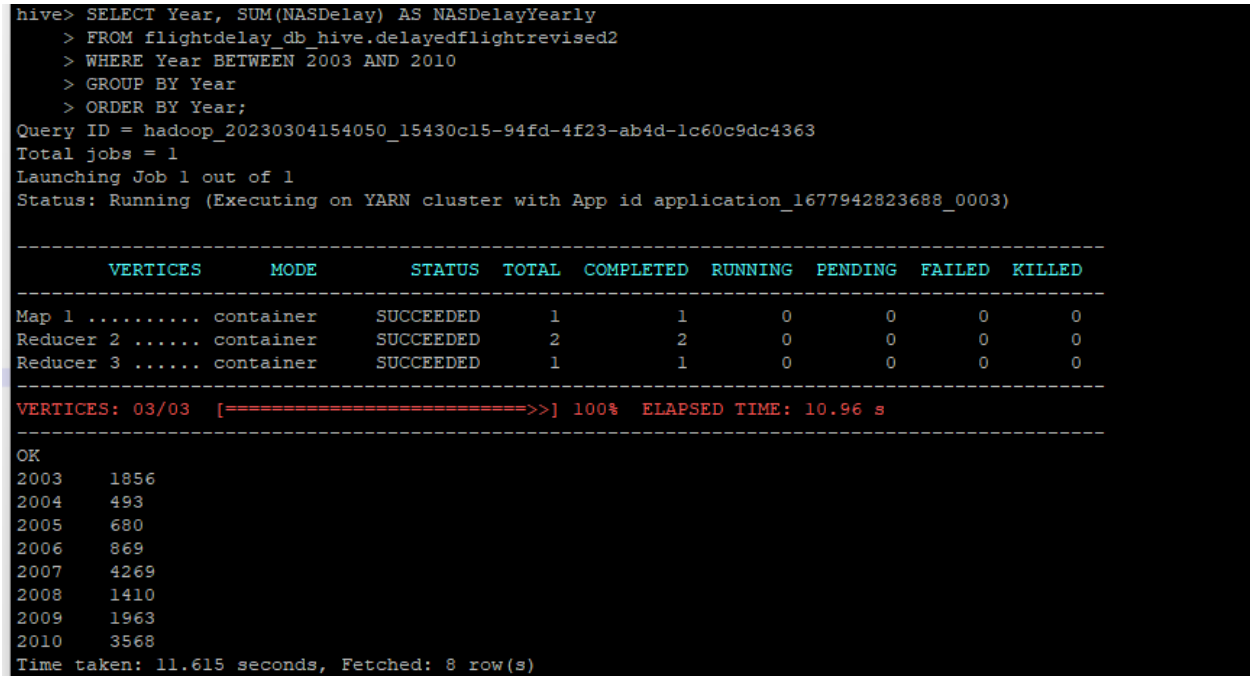
```
# run the queries
CarrierDelayYearly = spark.sql("SELECT Year, SUM(CarrierDelay) as CarrierDelayYearly FROM flightdelay_spark_tempview WHERE Year BETWEEN 2003
CarrierDelayYearly.show()
```

Last executed at 2023-03-04 23:14:07 in 7.54s

▸ Spark Job Progress

```
+----+-----------------+
|Year|CarrierDelayYearly|
+----+-----------------+
|2003|           1577.0|
|2004|           2064.0|
|2005|           1436.0|
|2006|           1540.0|
|2007|           1835.0|
|2008|           1298.0|
|2009|           1191.0|
|2010|           1992.0|
+----+-----------------+
```

Script running (Hive SQL)



```
hive> SELECT Year, SUM(NASDelay) AS NASDelayYearly
    > FROM flightdelay_db_hive.delayedflightrevised2
    > WHERE Year BETWEEN 2003 AND 2010
    > GROUP BY Year
    > ORDER BY Year;
Query ID = hadoop_20230304154050_15430c15-94fd-4f23-ab4d-1c60c9dc4363
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1677942823688_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 10.96 s
--------------------------------------------------------------------------------
OK
2003    1856
2004    493
2005    680
2006    869
2007    4269
2008    1410
2009    1963
2010    3568
Time taken: 11.615 seconds, Fetched: 8 row(s)
```

Script running (Spark SQL)

```
NASDelay = spark.sql("SELECT Year, SUM(NASDelay) as NASDelay FROM flightdelay_spark_tempview WHERE Year BETWEEN 2003 AND 2010 GROUP BY Year
NASDelay.show()
```

Last executed at 2023-03-04 23:18:43 in 17.61s

▸ Spark Job Progress

```
+----+--------+
|Year|NASDelay|
+----+--------+
|2003|  1856.0|
|2004|   493.0|
|2005|   680.0|
|2006|   869.0|
|2007|  4269.0|
|2008|  1410.0|
|2009|  1963.0|
|2010|  3568.0|
+----+--------+
```

Script running (Hive SQL)

```
hive> SELECT Year, SUM(WeatherDelay) AS WeatherDelayYearly
    > FROM flightdelay_db_hive.delayedflightrevised2
    > WHERE Year BETWEEN 2003 AND 2010
    > GROUP BY Year
    > ORDER BY Year;
Query ID = hadoop_20230304154521_51cc1519-5aa0-43a5-bcc7-a46af005aefe
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1677942823688_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 11.39 s
--------------------------------------------------------------------------------
OK
2003    660
2004    287
2005    267
2006    384
2007    452
2008    305
2009    24
2010    240
Time taken: 12.033 seconds, Fetched: 8 row(s)
```

Script running (Spark SQL)

```
WeatherDelay = spark.sql("SELECT Year, SUM(WeatherDelay) as WeatherDelay FROM flightdelay_spark_tempview WHERE Year BETWEEN 2003 AND 2010 GRO
WeatherDelay.show()
```

Last executed at 2023-03-04 23:22:34 in 867ms

▸ Spark Job Progress

```
+----+------------+
|Year|WeatherDelay|
+----+------------+
|2003|       660.0|
|2004|       287.0|
|2005|       267.0|
|2006|       384.0|
|2007|       452.0|
|2008|       305.0|
|2009|        24.0|
|2010|       240.0|
+----+------------+
```

Script running (Hive SQL)

```
hive> SELECT Year, SUM(LateAircraftDelay) AS LateAircraftDelayYearly
    > FROM flightdelay_db_hive.delayedflightrevised2
    > WHERE Year BETWEEN 2003 AND 2010
    > GROUP BY Year
    > ORDER BY Year;
Query ID = hadoop_20230304154925_d9b3b9f9-9b05-4fa4-8ed2-aca10fdf80d2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1677942823688_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 11.63 s
--------------------------------------------------------------------------------
OK
2003    1986
2004    1451
2005    2332
2006    2295
2007    5726
2008    2069
2009    1819
2010    4233
Time taken: 12.24 seconds, Fetched: 8 row(s)
```

Script running (Spark SQL)

```
LateAircraftDelay = spark.sql("SELECT Year, SUM(LateAircraftDelay) as LateAircraftDelay FROM flightdelay_spark_tempview WHERE Year BETWEEN 2
LateAircraftDelay.show()
```

Last executed at 2023-03-04 23:25:21 in 866ms

▸ Spark Job Progress

```
+----+----------------+
|Year|LateAircraftDelay|
+----+----------------+
|2003|          1986.0|
|2004|          1451.0|
|2005|          2332.0|
|2006|          2295.0|
|2007|          5726.0|
|2008|          2069.0|
|2009|          1819.0|
|2010|          4233.0|
+----+----------------+
```

Script running (Hive SQL)

```
hive> SELECT Year, SUM(SecurityDelay) AS SecurityDelayYearly
    > FROM flightdelay_db_hive.delayedflightrevised2
    > WHERE Year BETWEEN 2003 AND 2010
    > GROUP BY Year
    > ORDER BY Year;
Query ID = hadoop_20230304155239_827ae101-0983-4b2d-9d17-0c70edba5833
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1677942823688_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 10.80 s
--------------------------------------------------------------------------------
OK
2003    0
2004    0
2005    0
2006    0
2007    12
2008    0
2009    0
2010    0
Time taken: 11.394 seconds, Fetched: 8 row(s)
```

Script running (Spark SQL)

```
SecurityDelay = spark.sql("SELECT Year, SUM(SecurityDelay) as SecurityDelay FROM flightdelay_spark_tempview WHERE Year BETWEEN 2003 AND
SecurityDelay.show()
```

Last executed at 2023-03-04 23:27:29 in 885ms

▸ Spark Job Progress

```
+----+-------------+
|Year|SecurityDelay|
+----+-------------+
|2003|          0.0|
|2004|          0.0|
|2005|          0.0|
|2006|          0.0|
|2007|         12.0|
|2008|          0.0|
|2009|          0.0|
|2010|          0.0|
+----+-------------+
```