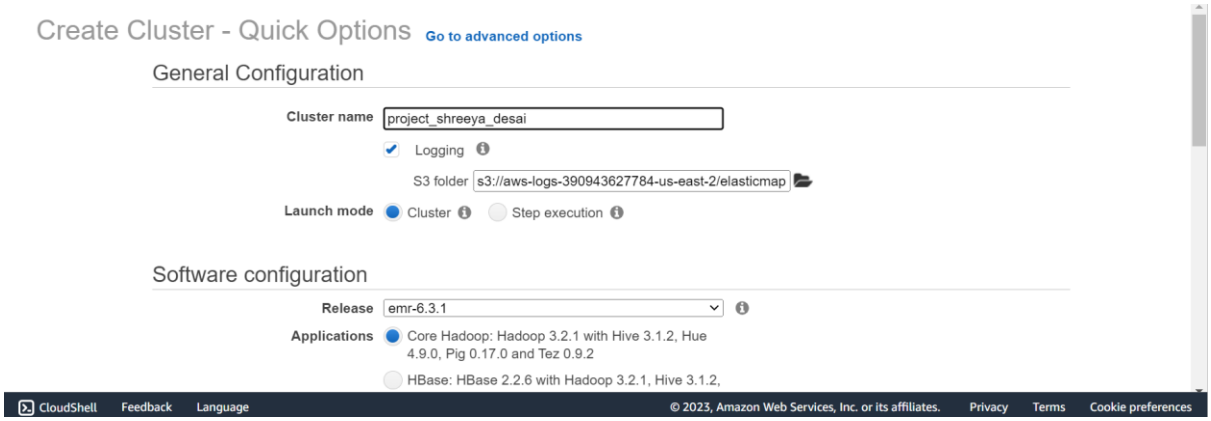
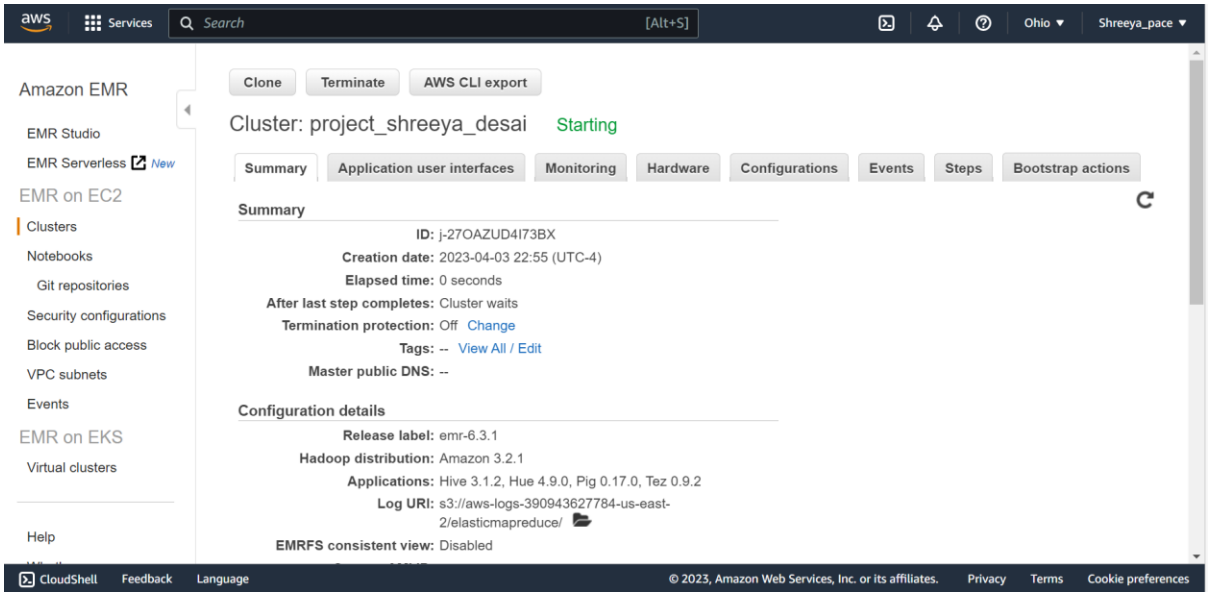
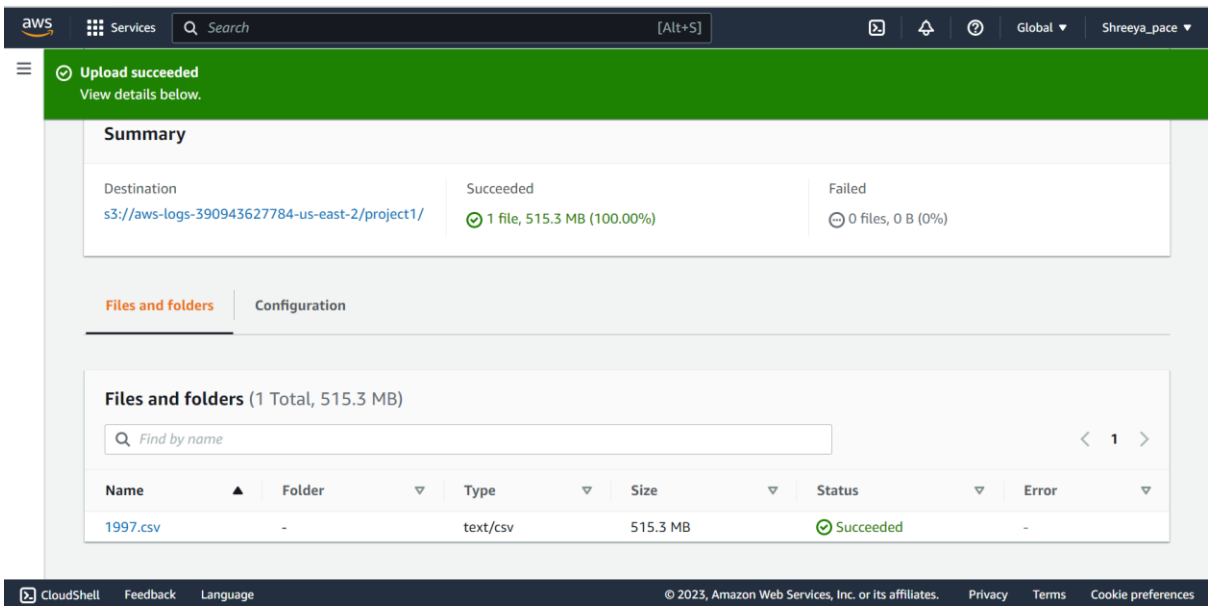
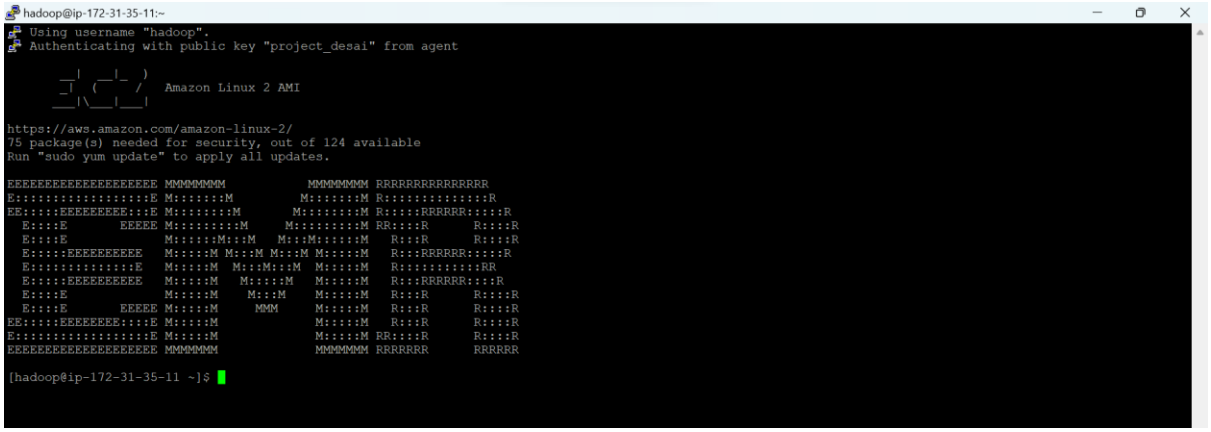
Snapshots of steps :










```
[hadoop@ip-172-31-35-11 ~]$ wget https://aws-logs-390943627784-us-east-2.s3.us-east-2.amazonaws.com/project1/1997.csv
--2023-04-04 03:10:49--  https://aws-logs-390943627784-us-east-2.s3.us-east-2.amazonaws.com/project1/1997.csv
Resolving aws-logs-390943627784-us-east-2.s3.us-east-2.amazonaws.com (aws-logs-390943627784-us-east-2.s3.us-east-2.amazonaws.com)... 3.5.133.138
Connecting to aws-logs-390943627784-us-east-2.s3.us-east-2.amazonaws.com (aws-logs-390943627784-us-east-2.s3.us-east-2.amazonaws.com)|3.5.133.138|:443... con
nected.
HTTP request sent, awaiting response... 200 OK
Length: 540347861 (515M) [text/csv]
Saving to: '1997.csv'

100%[=============================================================================================================>] 540,347,861 62.0MB/s   in 8.0s

2023-04-04 03:10:57 (64.7 MB/s) - '1997.csv' saved [540347861/540347861]

[hadoop@ip-172-31-35-11 ~]$
```

```
2023-04-04 03:10:57 (64.7 MB/s) - '1997.csv' saved [540347861/540347861]

[hadoop@ip-172-31-35-11 ~]$ hadoop fs -mkdir /shreeya
[hadoop@ip-172-31-35-11 ~]$ hadoop fs -mkdir /shreeya/csv
[hadoop@ip-172-31-35-11 ~]$ hadoop fs -put 1997.csv /shreeya/csv
[hadoop@ip-172-31-35-11 ~]$
```

```
hive> load data inpath '/shreeya/csv/1997.csv' into table AIRLINESPROJECT_SHREEYA;
Loading data to table default.airlinesproject_shreeya
OK
Time taken: 0.741 seconds
hive>
```

```
hive> SELECT * FROM AIRLINESPROJECT_SHREEYA LIMIT 15;
OK
1997    1       28      2       1615    1615    1728    1725    WN      1159    N683    73      70      60      3       0       ABQ     PHX     328     4       9
0       NA      0       NULL    NULL    NULL    NULL    NULL
1997    1       29      3       1624    1615    1735    1725    WN      1159    N626    71      70      56      10      9       ABQ     PHX     328     4       1
1       0       NA      0       NULL    NULL    NULL    NULL    NULL
1997    1       30      4       1626    1615    1730    1725    WN      1159    N321    64      70      52      5       11      ABQ     PHX     328     2       1
0       0       NA      0       NULL    NULL    NULL    NULL    NULL
1997    1       31      5       1628    1615    1729    1725    WN      1159    N383    61      70      50      4       13      ABQ     PHX     328     4       7
0       NA      0       NULL    NULL    NULL    NULL    NULL
1997    1       14      2       1503    1425    1619    1535    WN      1221    N513    76      70      64      44      38      ABQ     PHX     328     4       8
0       NA      0       NULL    NULL    NULL    NULL    NULL
1997    1       1       3       1029    1020    1127    1130    WN      1293    N369    58      70      52      -3      9       ABQ     PHX     328     2       4
0       NA      0       NULL    NULL    NULL    NULL    NULL
1997    1       2       4       1020    1020    1130    1130    WN      1293    N395    70      70      57      0       0       ABQ     PHX     328     7       6
0       NA      0       NULL    NULL    NULL    NULL    NULL
1997    1       3       5       1029    1020    1145    1130    WN      1293    N694    76      70      64      15      9       ABQ     PHX     328     3       9
0       NA      0       NULL    NULL    NULL    NULL    NULL
1997    1       5       7       1020    1020    1141    1130    WN      1293    N601    81      70      61      11      0       ABQ     PHX     328     5       1
5       0       NA      0       NULL    NULL    NULL    NULL    NULL
1997    1       6       1       1030    1020    1139    1130    WN      1293    N321    69      70      61      9       10      ABQ     PHX     328     2       6
0       NA      0       NULL    NULL    NULL    NULL    NULL
1997    1       7       2       1020    1020    1123    1130    WN      1293    N351    63      70      56      -7      0       ABQ     PHX     328     2       5
0       NA      0       NULL    NULL    NULL    NULL    NULL
1997    1       8       3       1107    1020    1205    1130    WN      1293    N677    58      70      46      35      47      ABQ     PHX     328     3       9
0       NA      0       NULL    NULL    NULL    NULL    NULL
1997    1       9       4       1020    1020    1130    1130    WN      1293    N623    70      70      59      0       0       ABQ     PHX     328     3       8
0       NA      0       NULL    NULL    NULL    NULL    NULL
1997    1       10      5       1020    1020    1123    1130    WN      1293    N632    63      70      56      -7      0       ABQ     PHX     328     2       5
0       NA      0       NULL    NULL    NULL    NULL    NULL
1997    1       12      7       1020    1020    1134    1130    WN      1293    N375    74      70      65      4       0       ABQ     PHX     328     2       7
0       NA      0       NULL    NULL    NULL    NULL    NULL
Time taken: 0.24 seconds, Fetched: 15 row(s)
hive>
```

```
hive> SELECT Origin AS Airport, SUM((ArrDelay + DepDelay) / 60.0) AS DelayHours FROM AIRLINESPROJECT_SHREEYA WHERE Origin IS NOT NULL AND ArrDelay IS NOT NUL
L AND DepDelay IS NOT NULL GROUP BY Origin ORDER BY DelayHours DESC LIMIT 3;
Query ID = hadoop_20230404034219_01bd0086-e276-462a-9f6e-910df977db84
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1680577291903_0006)

--------------------------------------------------------------------------------
        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: 12.61 s
--------------------------------------------------------------------------------
OK
ATL     88807.616761
ORD     84955.366468
DFW     67215.133298
Time taken: 15.293 seconds, Fetched: 3 row(s)
hive>
    >
```

```
hive>
    > SELECT UniqueCarrier AS Carrier, SUM(IF(ArrDelay > DepDelay, ArrDelay, DepDelay)) / 60.0 AS DelayHours FROM AIRLINESPROJECT_SHREEYA WHERE ArrDelay IS N
OT NULL AND DepDelay IS NOT NULL GROUP BY UniqueCarrier ORDER BY DelayHours DESC LIMIT 3;
Query ID = hadoop_20230404034314_9cc7c08b-e9c4-4d1c-bb4e-ec3f3b61379c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1680577291903_0006)

--------------------------------------------------------------------------------
        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: 7.86 s
--------------------------------------------------------------------------------
OK
DL      205310.200000
UA      162923.200000
WN      133664.450000
Time taken: 8.986 seconds, Fetched: 3 row(s)
hive>
```

```
hive> SELECT 'Arrival Delay' AS DelayType, SUM(ArrDelay) AS TotalDelayMinutes FROM AIRLINESPROJECT_SHREEYA WHERE ArrDelay IS NOT NULL UNION ALL SELECT 'Depar
ture Delay' AS DelayType, SUM(DepDelay) AS TotalDelayMinutes FROM AIRLINESPROJECT_SHREEYA WHERE DepDelay IS NOT NULL ORDER BY TotalDelayMinutes DESC LIMIT 1;

Query ID = hadoop_20230404042830_1ffbd209-bd54-4c6e-839c-5d9ecfe76fb1
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1680577291903_0013)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ......   container     SUCCEEDED      1          1        0        0       0       0
Map 5 ..........   container     SUCCEEDED      1          1        0        0       0       0
Reducer 6 ......   container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 7.51 s
--------------------------------------------------------------------------------
OK
Departure Delay 43764457
Time taken: 13.081 seconds, Fetched: 1 row(s)
hive>
```

Commands used:

1. wget (s3 path)
2. hadoop fs -mkdir /shreeya
3. hadoop fs -mkdir /shreeya/csv
4. hadoop fs -put 1997.csv /shreeya/csv
5. hive
6. CREATE EXTERNAL TABLE IF NOT EXISTS AIRLINESPROJECT_SHREEYA (Year INT, Month INT, DayofMonth INT, DayOfWeek INT, DepTime INT, CRSDepTime INT, ArrTime INT, CRSArrTime 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) ROW FORMAT DELIMITED FIELDS TERMINATED BY "," tblproperties("skip.header.line.count"="1");
7. load data inpath '/shreeya/csv/1997.csv' into table AIRLINESPROJECT_SHREEYA;
8. SELECT * FROM AIRLINESPROJECT_SHREEYA LIMIT 15;

9. **QUERY TO DETERMINE THE THREE AIRPORTS WITH THE HIGHEST DELAY TIME (IN HOURS):**
   SELECT Origin AS Airport, SUM((ArrDelay + DepDelay) / 60.0) AS DelayHours FROM AIRLINESPROJECT_SHREEYA WHERE Origin IS NOT NULL AND ArrDelay IS NOT NULL AND DepDelay IS NOT NULL GROUP BY Origin ORDER BY DelayHours DESC LIMIT 3;

10. **QUERY TO determine the three carriers with the highest delay time (in hours):**
    SELECT UniqueCarrier AS Carrier, SUM((ArrDelay + DepDelay) / 60.0) AS DelayHours FROM AIRLINESPROJECT_SHREEYA WHERE UniqueCarrier IS NOT NULL AND ArrDelay IS NOT NULL AND DepDelay IS NOT NULL GROUP BY UniqueCarrier ORDER BY DelayHours DESC LIMIT 3;

11. **QUERY TO determine overall which type of delay (arrivals or departures) is the largest for airports:**
    SELECT 'Arrival Delay' AS DelayType, SUM(ArrDelay) AS TotalDelayMinutes FROM AIRLINESPROJECT_SHREEYA WHERE ArrDelay IS NOT NULL UNION ALL SELECT 'Departure Delay' AS DelayType, SUM(DepDelay) AS TotalDelayMinutes FROM AIRLINESPROJECT_SHREEYA WHERE DepDelay IS NOT NULL ORDER BY TotalDelayMinutes DESC LIMIT 1;