DS 644101 - COURSE PROJECT
FLIGHT ANALYSIS REPORT

Instructor:
YAJUAN LI

Report by:
Ankush Ranapure (ar2653@njit.edu)
Dheeraj Ananth Kumar (da559njit.edu)
Jaya Sathwika Gadi (jg899njit.edu)
Nivedita Reddy Addulla (na566njit.edu)

Table of context

Table of Figures

1. STRUCTURE OF OOZIE WORKFLOW:

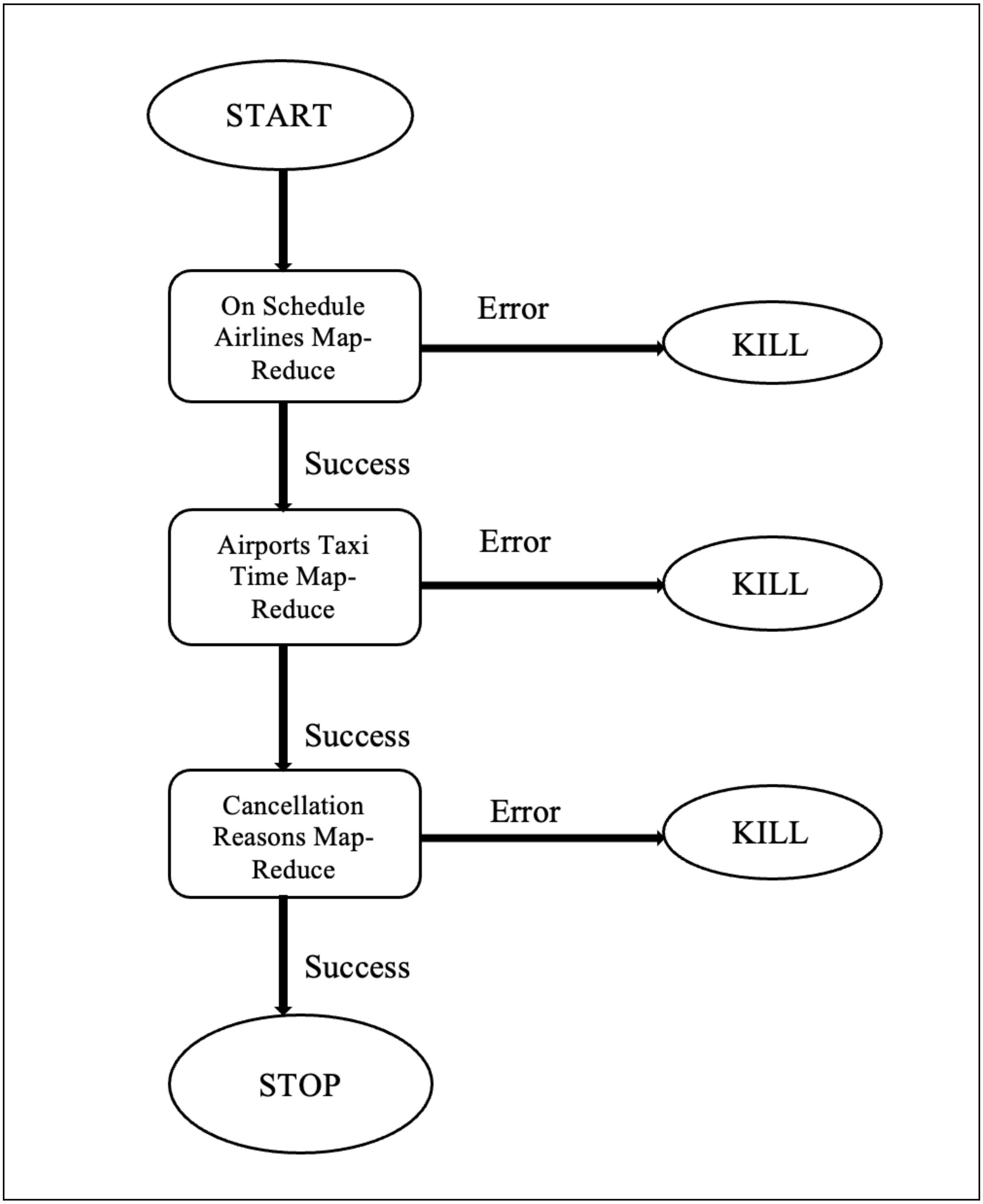

START
On Schedule Airlines Map-Reduce
Error
KILL
Success
Airports Taxi Time Map-Reduce
Error
KILL
Success
Cancellation Reasons Map-Reduce
Error
KILL
Success
STOP

Figure 1.1 - Oozie Workflow

2. ALGORITHM:

First Map-Reduce: On Schedule Airlines

1. Mapper Function: <Key,Value> : <UniqueCarrier, 1 or 0>
   For each input line (ignoring the first line and NA data):
   - Read the UniqueCarrier and ArrDelay columns.
   - If ArrDelay is less than or equal to 10 minutes, output: `<UniqueCarrier, 1>`.
   - Otherwise, output: `<UniqueCarrier, 0>`.

2. Reducer Function: <Key,Value> : <UniqueCarrier, Probability>
   Initialize variables: `sum_1 = 0`, `sum_0_1 = 0`, `sum_0 = 0`.
   For each input from the mapper:
   - Sum the values based on the UniqueCarrier key.
   - If the value is 1, increment `sum_1`.
   - Increment `sum_0_1` for each value (both 1 and 0).
   - Calculate the probability for each UniqueCarrier: `Probability = sum_1 / sum_0_1`.

   Sort the probabilities using a comparator function.

3. Output Handling:
   Output the top 3 and bottom 3 UniqueCarriers with their probabilities.
   If the data is NULL, output: "There is no value used, so no output."

Second Map-Reduce: On Airport Taxi Time

1. Mapper Algorithm Read Data: <key,value>:<IATA airport code, TaxiTime>:<Origin,TaxiOut>or<Dest,Taxi>
   - Read the dataset line by line, skipping the first line (header).
   - Check Taxi Time: For each line, check if the TaxiIn or TaxiOut value is not 'N/A'.
   - Output Key-Value Pair: If the taxi time is valid, output a key-value pair: <IATA airport code, TaxiTime>.

2. Reducer Algorithm: <key,value>:<IATA airport code, Average TaxiTime>
   - Sum Taxi Times: For each unique key (IATA airport code), sum the taxi times from all its associated values.
   - Count Occurrences: Count the total number of times each key appears.

    - Calculate Average: For each key, calculate the average taxi time using the formula: average = total taxi time / number of occurrences.
    - Sort Results: Sort the airports by their average taxi times. Output Top and Bottom

3. Output Handling:
   Output the three airports with the longest and three with the shortest avg t.
   If the data is NULL, then output: There is no value used, so no output.

Third Map-Reduce: On Cancellation Reasons

1. Mapper Function: <Key,Value> : <CancellationCode, 1>
    - Input: Key-Value pairs (CancellationCode, 1)
    - Read the data line by line.
    - Ignore the first line.
    - If the value of Canceled is 1 and CancellationCode is not "N/A", output a new key-value pair: (CancellationCode, 1)

2. Reducer Function: <key,value>:<CancellationCode, sum of 1s>
    - Input: Key-Value pairs (CancellationCode, list of 1s)
    - Sum the values (count of 1s) for each CancellationCode.
    - Use a comparator function to sort the CancellationCode based on the sum of 1s.
    - Output: The most common reason for flight cancellations.

3. Output Handling:
    - If the data is NULL, output: "There is no common reason for flight cancellation."

3. A performance measurement plot that compares the workflow execution time in response to an increasing number of VMs used for processing the entire data set (22 years) and an in-depth discussion on the observed performance comparison results

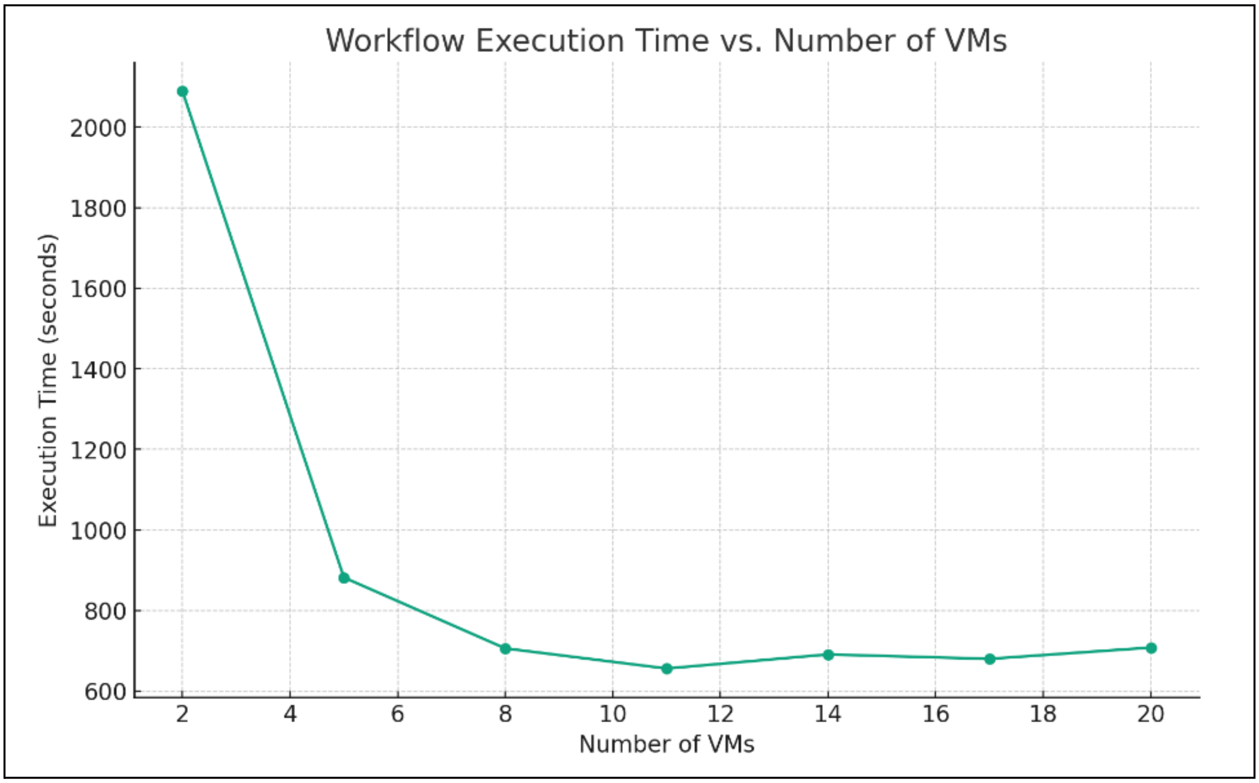


Figure - 1.2 - Workflow Execution Time vs Number of VMs

In Figure 1.2, an increase in the number of virtual machines (VMs) leads to a reduction in workflow execution time. This enhancement is attributed to the heightened processing capacity of the Hadoop cluster, facilitated by the parallel handling of data across multiple datanodes.

Consequently, the execution time for each map-reduce job diminishes, resulting in a shorter overall duration for the Oozie workflow. It's important to note, however, that the execution time for dealing with the same data size may not consistently decrease with a rise in the number of VMs.

Once the execution time reaches a certain threshold, further attempts to augment the number of VMs may not yield additional time savings. This occurs due to the increased information interaction time among datanodes within the Hadoop cluster as the number of VMs escalates.

4. A performance measurement plot that compares the workflow execution time in response to an increasing data size (from 1 year to 22 years) and an in-depth discussion on the observed performance comparison results

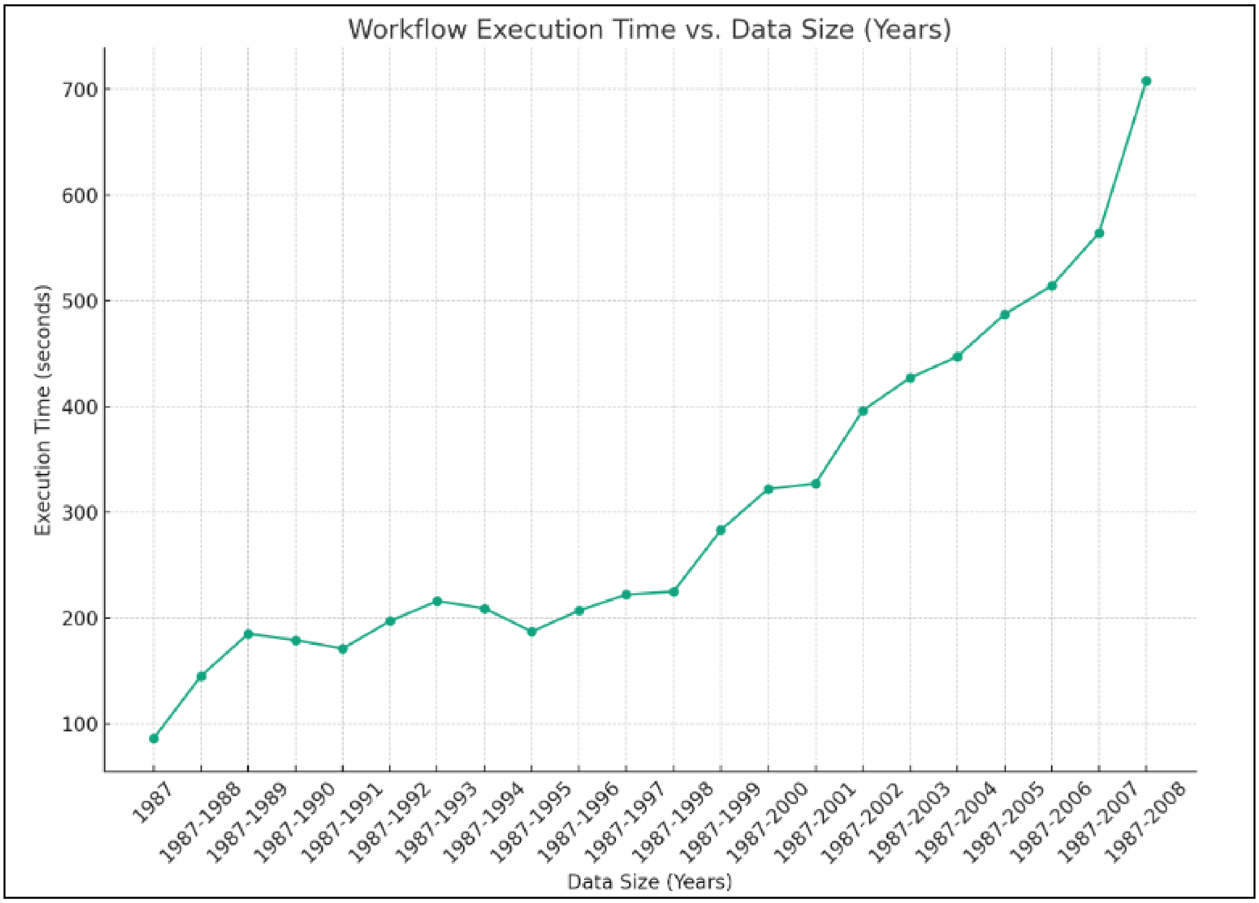
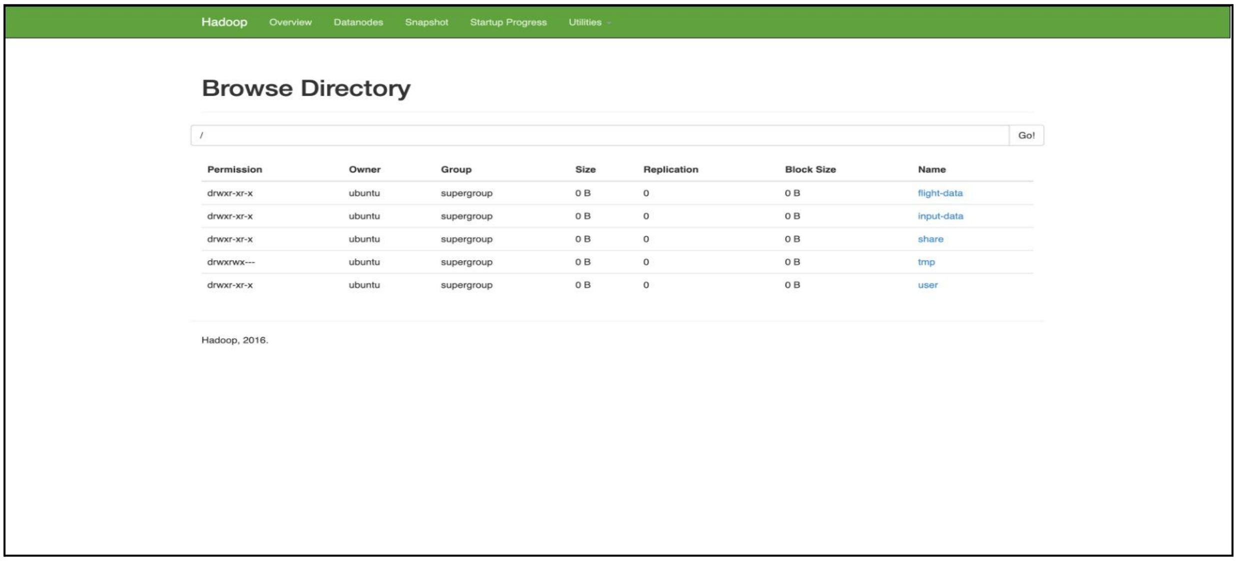


Figure 1.3 - Workflow Execution Time vs Data Size (in Years)

In Figure 1.3, as the data size grows, the execution time of the oozie workflow consistently rises. Initially, the increase in time consumption is gradual due to the relatively moderate growth in data during the early years.

However, after 1998, the time consumption experienced a rapid surge, indicating a steeper slope compared to the earlier years. This increase is due to a faster growth in flight data from 1998 to 2008, reflecting a rising trend in the number of individuals wanting for air travel.

Miscellaneous Screenshots:



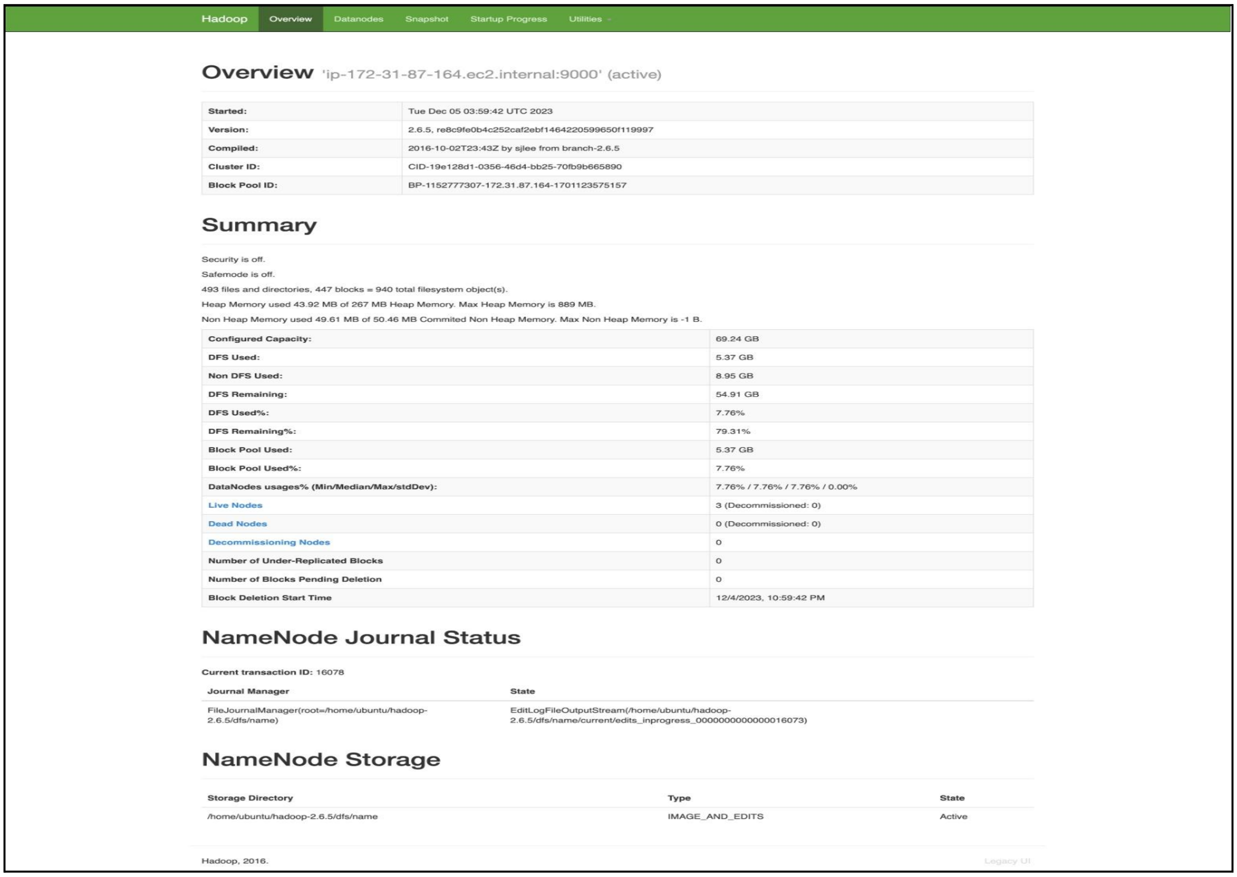
Browse Directory

/ Go!

Permission	Owner	Group	Size	Replication	Block Size	Name
drwxr-xr-x	ubuntu	supergroup	0 B	0	0 B	flight-data
drwxr-xr-x	ubuntu	supergroup	0 B	0	0 B	input-data
drwxr-xr-x	ubuntu	supergroup	0 B	0	0 B	share
drwxrwx---	ubuntu	supergroup	0 B	0	0 B	tmp
drwxr-xr-x	ubuntu	supergroup	0 B	0	0 B	user

Hadoop, 2016.

Figure 2.1 - Hadoop Directory screenshot



Overview 'ip-172-31-87-164.ec2.internal:9000' (active)

Started:	Tue Dec 05 03:59:42 UTC 2023
Version:	2.6.5, re8c9fe0b4c252caf2ebf1464220599650f119997
Compiled:	2016-10-02T23:43Z by sjlee from branch-2.6.5
Cluster ID:	CID-19e128d1-0356-46d4-bb25-70fb9b665890
Block Pool ID:	BP-1152777307-172.31.87.164-1701123575157

Summary

Security is off.

Safemode is off.

493 files and directories, 447 blocks = 940 total filesystem object(s).

Heap Memory used 43.92 MB of 267 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 49.61 MB of 50.46 MB Commited Non Heap Memory. Max Non Heap Memory is -1 B.

Configured Capacity:	69.24 GB
DFS Used:	5.37 GB
Non DFS Used:	8.95 GB
DFS Remaining:	54.91 GB
DFS Used%:	7.76%
DFS Remaining%:	79.31%
Block Pool Used:	5.37 GB
Block Pool Used%:	7.76%
DataNodes usages% (Min/Median/Max/stdDev):	7.76% / 7.76% / 7.76% / 0.00%
Live Nodes	3 (Decommissioned: 0)
Dead Nodes	0 (Decommissioned: 0)
Decommissioning Nodes	0
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0
Block Deletion Start Time	12/4/2023, 10:59:42 PM

NameNode Journal Status

Current transaction ID: 16078

Journal Manager	State
FileJournalManager(root=/home/ubuntu/hadoop-2.6.5/dfs/name)	EditLogFileOutputStream(/home/ubuntu/hadoop-2.6.5/dfs/name/current/edits_inprogress_0000000000000016073)

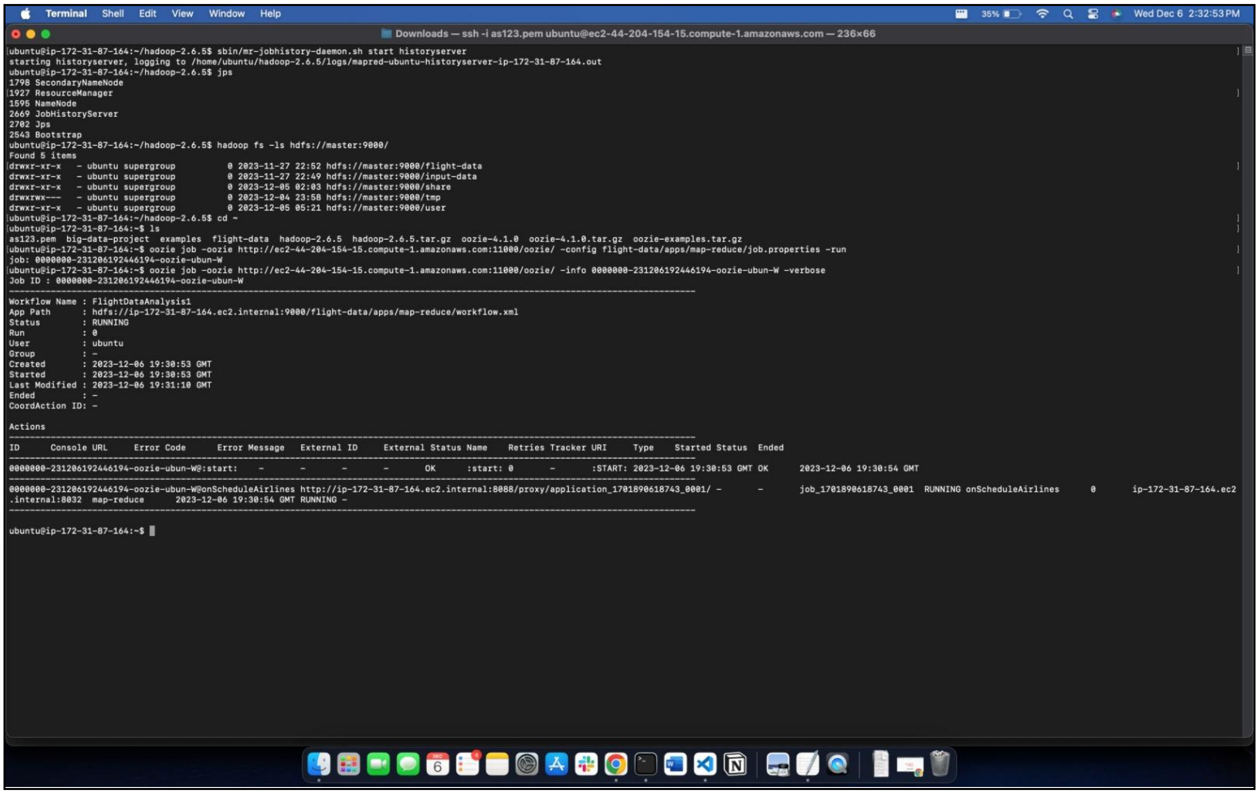
NameNode Storage

Storage Directory	Type	State
/home/ubuntu/hadoop-2.6.5/dfs/name	IMAGE_AND_EDITS	Active

Hadoop, 2016.

Legacy UI

Figure 2.2 - HDFS Overview


Terminal
Shell
Edit
View
Window
Help
35%
Wed Dec 6 2:32:53 PM
Downloads — ssh -i as123.pem ubuntu@ec2-44-204-154-15.compute-1.amazonaws.com — 236×66

```
ubuntu@ip-172-31-87-164:~/hadoop-2.6.5$ sbin/mr-jobhistory-daemon.sh start historyserver
starting historyserver, logging to /home/ubuntu/hadoop-2.6.5/logs/mapred-ubuntu-historyserver-ip-172-31-87-164.out
ubuntu@ip-172-31-87-164:~/hadoop-2.6.5$ jps
1798 SecondaryNameNode
1927 ResourceManager
1595 NameNode
2669 JobHistoryServer
2702 Jps
2543 Bootstrap
ubuntu@ip-172-31-87-164:~/hadoop-2.6.5$ hadoop fs -ls hdfs://master:9000/
Found 5 items
drwxr-xr-x   - ubuntu supergroup          0 2023-11-27 22:52 hdfs://master:9000/flight-data
drwxr-xr-x   - ubuntu supergroup          0 2023-11-27 22:49 hdfs://master:9000/input-data
drwxr-xr-x   - ubuntu supergroup          0 2023-12-05 02:03 hdfs://master:9000/share
drwxrwx---   - ubuntu supergroup          0 2023-12-04 23:58 hdfs://master:9000/tmp
drwxr-xr-x   - ubuntu supergroup          0 2023-12-05 05:21 hdfs://master:9000/user
ubuntu@ip-172-31-87-164:~/hadoop-2.6.5$ cd ~
ubuntu@ip-172-31-87-164:~$ ls
as123.pem  big-data-project  examples  flight-data  hadoop-2.6.5  hadoop-2.6.5.tar.gz  oozie-4.1.0  oozie-4.1.0.tar.gz  oozie-examples.tar.gz
ubuntu@ip-172-31-87-164:~$ oozie job -oozie http://ec2-44-204-154-15.compute-1.amazonaws.com:11000/oozie/ -config flight-data/apps/map-reduce/job.properties -run
job: 0000000-231206192446194-oozie-ubun-W
ubuntu@ip-172-31-87-164:~$ oozie job -oozie http://ec2-44-204-154-15.compute-1.amazonaws.com:11000/oozie/ -info 0000000-231206192446194-oozie-ubun-W -verbose
Job ID : 0000000-231206192446194-oozie-ubun-W
------------------------------------------------------------------------------------------------------------------------------------
Workflow Name : FlightDataAnalysis1
App Path      : hdfs://ip-172-31-87-164.ec2.internal:9000/flight-data/apps/map-reduce/workflow.xml
Status        : RUNNING
Run           : 0
User          : ubuntu
Group         : -
Created       : 2023-12-06 19:30:53 GMT
Started       : 2023-12-06 19:30:53 GMT
Last Modified : 2023-12-06 19:31:10 GMT
Ended         : -
CoordAction ID: -

Actions
------------------------------------------------------------------------------------------------------------------------------------
ID      Console URL     Error Code      Error Message   External ID     External Status Name    Retries Tracker URI     Type    Started Status  Ended
------------------------------------------------------------------------------------------------------------------------------------
0000000-231206192446194-oozie-ubun-W@:start:    -       -       -       -       OK      :start: 0       -       :START: 2023-12-06 19:30:53 GMT OK      2023-12-06 19:30:54 GMT
------------------------------------------------------------------------------------------------------------------------------------
0000000-231206192446194-oozie-ubun-W@onScheduleAirlines http://ip-172-31-87-164.ec2.internal:8088/proxy/application_1701890618743_0001/ -       -       job_1701890618743_0001  RUNNING onScheduleAirlines      0       ip-172-31-87-164.ec2
.internal:8032  map-reduce      2023-12-06 19:30:54 GMT RUNNING -
------------------------------------------------------------------------------------------------------------------------------------

ubuntu@ip-172-31-87-164:~$
```

Figure 2.3 - Oozie Job running screenshot