



National University
of computer and emerging sciences

BDA LAB PROJECT

Name: Jiyad Khan

Roll No: 19I-1771

Section: BS-DS (N)

Course's Name: Big Data Analytics

Course's Instructor: Sir Saad Naeem

Due Date: 28th, May 2021

Master Node

Requirement 1:

There need to be minimum 3 nodes configured for multi node setup.

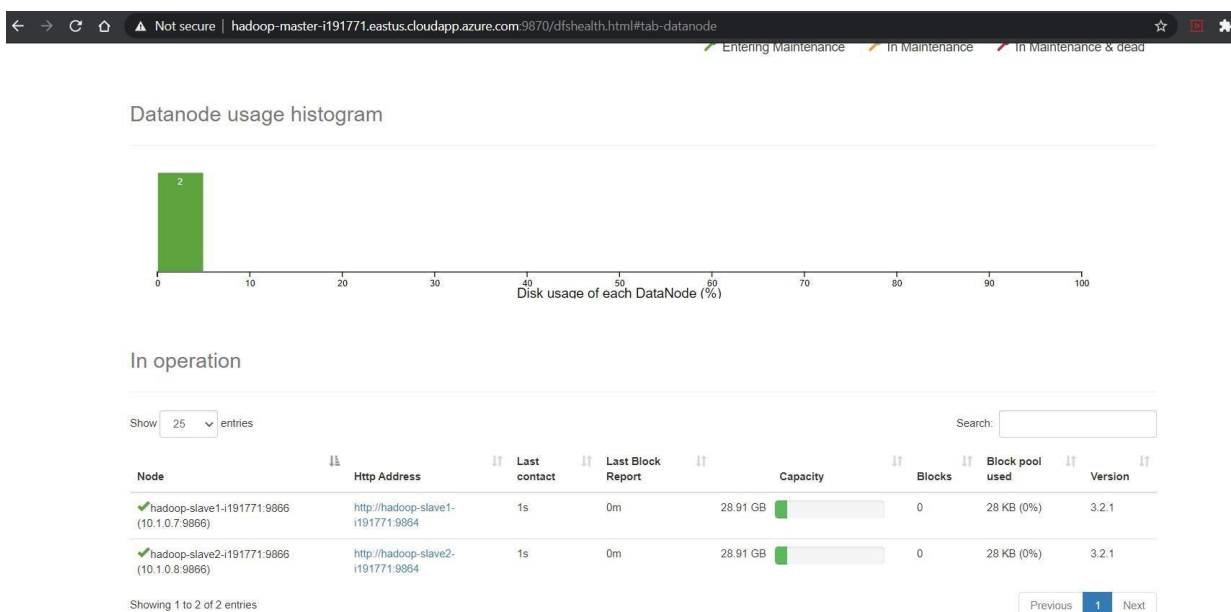
Numbers of the Machine	Name of the Machine	IP address of the Machine
Machine 1	hadoop-masters	10.1.0.6
Machine 2	hadoop-slave1-i191771	10.1.0.7
Machine 3	hadoop-slave2-i191771	10.1.0.8

Requirement 2:

The cluster is to be evaluated on the text file provided (“books_large_p1.txt” 2.52/2.3 GB).

Requirement 3: Data Node

A Screenshot displaying all the data nodes that are up and sending their heartbeat to the Master Node.



Slaves Nodes

Requirement 1: Node States (here showing 32 GB Pooled Memory 8GB each)

A Screen shot showing the cluster metrics including total Pooled Memory and the number of Active Nodes.

The screenshot displays the Hadoop cluster management interface. The top navigation bar includes the Hadoop logo and the title "Nodes of the cluster". The left sidebar contains a menu with options: Cluster, About Nodes, Node Labels, Applications, NEW, NEW SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED, Scheduler, and Tools. The main content area shows the "Nodes of the cluster" page. It includes a "Cluster Metrics" table with columns for Apps Submitted, Apps Pending, Apps Running, Apps Completed, Containers Running, Memory Used, Memory Total, Memory Reserved, VCoers Used, VCoers Total, and VCoers Reserved. Below this is a "Cluster Nodes Metrics" table with columns for Active Nodes, Decommissioning Nodes, Decommissioned Nodes, Lost Nodes, Unhealthy Nodes, Rebooted Nodes, and Shutdown Nodes. The "Scheduler Metrics" section shows the Scheduler Type (Capacity Scheduler), Scheduling Resource Type (memory-mb (unit=Mi), vcores), Minimum Allocation (<memory:1024, vCores:1>), Maximum Allocation (<memory:8192, vCores:4>), and Maximum Cluster Application Priority (0). The "Nodes" table lists individual nodes with columns for Node Labels, Rack, Node State, Node Address, Node HTTP Address, Last health-update, Health-report, Containers, Allocation Tags, Mem Used, Mem Avail, VCoers Used, VCoers Avail, and Version. The table shows two nodes in a "RUNNING" state, both with 8 GB of memory and 4 vcores. The bottom of the page indicates "Showing 1 to 2 of 2 entries" and includes navigation links for First, Previous, 1, Next, and Last.

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCoers Used	VCoers Total	VCoers Reserved
0	0	0	0	0	0 B	16 GB	0 B	0	16	0

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
2	0	0	0	0	0	0

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Node Labels	Rack	Node State	Node Address	Node HTTP Address	Last health-update	Health-report	Containers	Allocation Tags	Mem Used	Mem Avail	VCoers Used	VCoers Avail	Version
/default-rack		RUNNING	hadoop-slave2-191771:36691	hadoop-slave2-191771:8042	Thu May 27 17:31:45 +0000 2021		0		0 B	8 GB	0	8	3.2.1
/default-rack		RUNNING	hadoop-slave1-191771:42993	hadoop-slave1-191771:8042	Thu May 27 17:31:45 +0000 2021		0		0 B	8 GB	0	8	3.2.1

Requirement 2: Success on 2.5 GB Text file

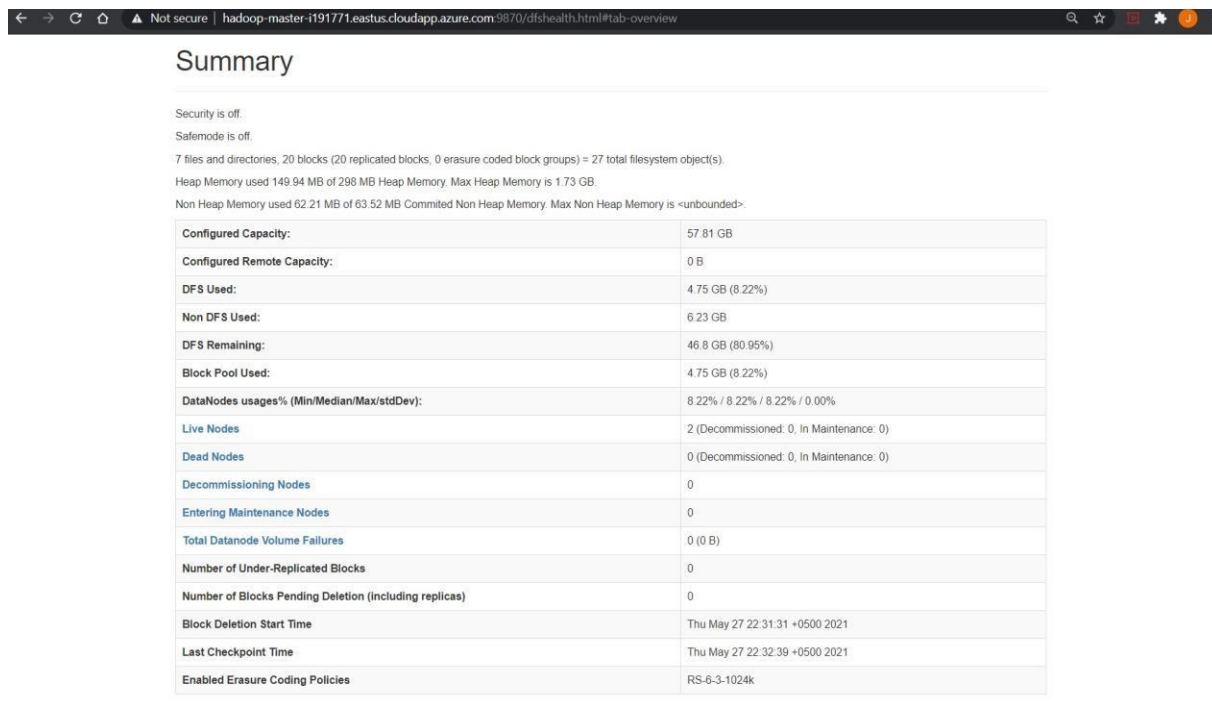
A Screenshot displaying the directory of the cluster.

The screenshot displays the Hadoop cluster management interface, specifically the "Browse Directory" page. The top navigation bar includes the Hadoop logo and the title "Browse Directory". The left sidebar contains a menu with options: Hadoop, Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. The main content area shows the "Browse Directory" page. It includes a search bar with the text "/output_191771" and a "Go!" button. Below the search bar is a "Show" dropdown menu set to "25" entries. The "Nodes" table lists individual nodes with columns for Permission, Owner, Group, Size, Last Modified, Replication, Block Size, and Name. The table shows two entries: a file named "_SUCCESS" with a size of 0 B, and a file named "part-r-00000" with a size of 13.38 MB. The bottom of the page indicates "Showing 1 to 2 of 2 entries" and includes navigation links for Previous, 1, Next, and Last.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoopuser	supergroup	0 B	May 27 23:02	2	128 MB	_SUCCESS
-rw-r--r--	hadoopuser	supergroup	13.38 MB	May 27 23:02	2	128 MB	part-r-00000

Requirement 3: Cluster Overview

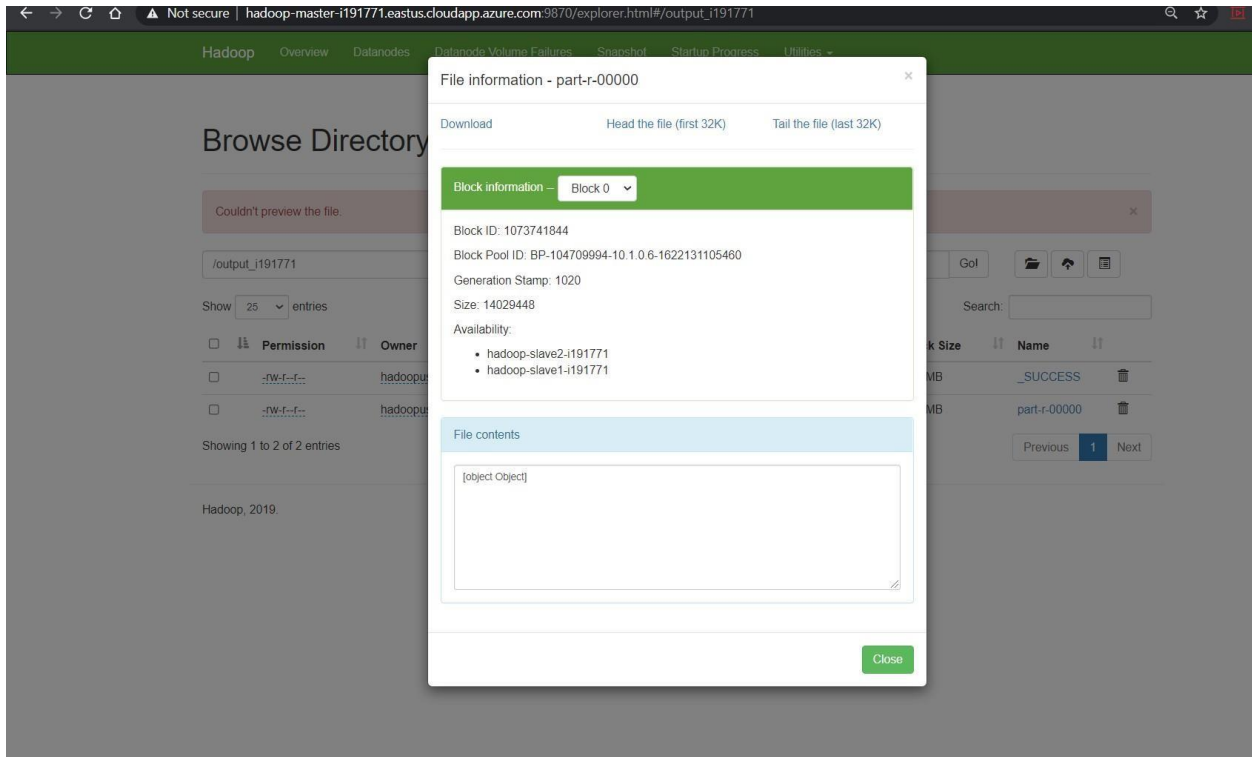
A Screenshot showing live nodes, total configured capacity etc.

A screenshot of a web browser displaying the 'Summary' page of a Hadoop DFS Health Overview. The browser's address bar shows the URL 'hadoop-master-i191771.eastus.cloudapp.azure.com:9870/dfshealth.html#tab-overview'. The page title is 'Summary'. Below the title, there are several status messages: 'Security is off.', 'Safemode is off.', '7 files and directories, 20 blocks (20 replicated blocks, 0 erasure coded block groups) = 27 total filesystem object(s).', 'Heap Memory used 149.94 MB of 298 MB Heap Memory. Max Heap Memory is 1.73 GB.', and 'Non Heap Memory used 62.21 MB of 63.52 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>'. Below these messages is a table with 2 columns: a label and a value. The table contains the following data: Configured Capacity: 57.81 GB; Configured Remote Capacity: 0 B; DFS Used: 4.75 GB (8.22%); Non DFS Used: 6.23 GB; DFS Remaining: 46.8 GB (80.95%); Block Pool Used: 4.75 GB (8.22%); DataNodes usages% (Min/Median/Max/stdDev): 8.22% / 8.22% / 8.22% / 0.00%; Live Nodes: 2 (Decommissioned: 0, In Maintenance: 0); Dead Nodes: 0 (Decommissioned: 0, In Maintenance: 0); Decommissioning Nodes: 0; Entering Maintenance Nodes: 0; Total Datanode Volume Failures: 0 (0 B); Number of Under-Replicated Blocks: 0; Number of Blocks Pending Deletion (including replicas): 0; Block Deletion Start Time: Thu May 27 22:31:31 +0500 2021; Last Checkpoint Time: Thu May 27 22:32:39 +0500 2021; Enabled Erasure Coding Policies: RS-6-3-1024k.

Summary	
Security is off.	
Safemode is off.	
7 files and directories, 20 blocks (20 replicated blocks, 0 erasure coded block groups) = 27 total filesystem object(s).	
Heap Memory used 149.94 MB of 298 MB Heap Memory. Max Heap Memory is 1.73 GB.	
Non Heap Memory used 62.21 MB of 63.52 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.	
Configured Capacity:	57.81 GB
Configured Remote Capacity:	0 B
DFS Used:	4.75 GB (8.22%)
Non DFS Used:	6.23 GB
DFS Remaining:	46.8 GB (80.95%)
Block Pool Used:	4.75 GB (8.22%)
DataNodes usages% (Min/Median/Max/stdDev):	8.22% / 8.22% / 8.22% / 0.00%
Live Nodes	2 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Thu May 27 22:31:31 +0500 2021
Last Checkpoint Time	Thu May 27 22:32:39 +0500 2021
Enabled Erasure Coding Policies	RS-6-3-1024k

Requirement 4: GUI Results with Active Nodes

A Screenshot that shows the head of the results with “Availability” parameter showing how many slave nodes were involved in processing these results.



Requirement 5: CAT out the results in terminal

A Screenshot of the output using cat command running on "part-00000".

```
hadoopuser@hadoop-master: /usr/local/hadoop
~$ cat /usr/local/hadoop/output_191771/part-r-00000
~s~ 26
~the 34
~titlemagic.blogspot.com~ 1
~transcendental 1
~trystan~ 74
~tyrstan~ 2
~weathercock 2
~what 1
~whoops~ 2
~wild 1
~will 4
~william 2
~will~ 7
~winter 2
~wood 1
~word 1
~wunder~ 1
~you 2
~your 3
~ziah~ 2
~ 302
~*****~ 2
~*~ 104
~book 1
~i 1
~instead 1
~my 1
~other 2
~tearing 1
~ 1722
~*****~ 2
~000~ 33
~ooo~ 18
~ 320
~*****~ 8
~*****~ 34
~ 315
~ 45
~ 12
~ 1
~ 15
~ 157
~ 1
~ 1
~ 3
~ 3
~couple 1
~ 2
~ 1
~ 1
~ 1
```

Prove Task

Requirement 7:

- Prove that the 2.5 GB file was chunked, distributed, and processed on multiplenodes.
- Change the number of reducers to be equal to the number of Data Nodes.
- Prove how many mappers and reducers are configured on the cluster. By default,there is only one reducer which becomes the bottleneck of the whole process.

```
hadoopuser@hadoop-master: /usr/local/hadoop
Pending deletion blocks: 0
Erasure Coded Block groups:
Low redundancy block groups: 0
Block groups with corrupt internal blocks: 0
Missing block groups: 0
Low redundancy blocks with highest priority to recover: 0
Pending deletion blocks: 0

-----
Live datanodes (2):

Name: 10.1.0.7:9866 (hadoop-slave1-i191771)
Hostname: hadoop-slave1-i191771
Decommission Status : Normal
Configured Capacity: 31036686336 (28.91 GB)
DFS Used: 2550718464 (2.38 GB)
Non DFS Used: 3321954304 (3.09 GB)
DFS Remaining: 25147236352 (23.42 GB)
DFS Used%: 8.22%
DFS Remaining%: 81.02%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Thu May 27 18:22:14 UTC 2021
Last Block Report: Thu May 27 17:31:35 UTC 2021
Num of Blocks: 20

Name: 10.1.0.8:9866 (hadoop-slave2-i191771)
Hostname: hadoop-slave2-i191771
Decommission Status : Normal
Configured Capacity: 31036686336 (28.91 GB)
DFS Used: 2550714368 (2.38 GB)
Non DFS Used: 3368800256 (3.14 GB)
DFS Remaining: 25100394496 (23.38 GB)
DFS Used%: 8.22%
DFS Remaining%: 80.87%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Thu May 27 18:22:14 UTC 2021
Last Block Report: Thu May 27 17:33:22 UTC 2021
Num of Blocks: 20

hadoopuser@hadoop-master: /usr/local/hadoop$
```