

## **2CS702 Big Data Analytics**

### **Lab-3 Task**

**Submitted by: Labdhi Sheth 18BCE101**

**Aim: Setup single-node Hadoop cluster and apply HDFS commands on single-node Hadoop Cluster.**

#### **Methodology followed:**

The installation process as shared in the drive was followed.

#### **Output:**

CMD run as administrator

1. Checking the java and Hadoop version

[illegible]

## 2. Starting dfs and yarn

```

Administrator Command Prompt
2021-09-14 12:22:24,272 INFO namenode.FSDirectory: XATTR serial map: bits=24 maxEntries=16777215
2021-09-14 12:22:24,288 INFO util.GSet: Computing capacity for map INodeMap
2021-09-14 12:22:24,288 INFO util.GSet: VM type = 64-bit
2021-09-14 12:22:24,289 INFO util.GSet: 1.0% max memory 889 MB = 8.9 MB
2021-09-14 12:22:24,290 INFO util.GSet: capacity = 2^20 = 1048576 entries
2021-09-14 12:22:24,290 INFO namenode.FSDirectory: ACLs enabled? false
2021-09-14 12:22:24,291 INFO namenode.FSDirectory: POSIX ACL inheritance enabled? true
2021-09-14 12:22:24,291 INFO namenode.FSDirectory: Xattrs enabled? true
2021-09-14 12:22:24,292 INFO namenode.NameNode: Caching file names occurring more than 10 times
2021-09-14 12:22:24,300 INFO snapshot.SnapshotManager: loaded config captureOpenFiles: false, skipCaptureAccessTimeOnlyChange: false, snapshotDiffAllowSnapRootDescendant: true, maxSnapshotLimit: 65536
2021-09-14 12:22:24,303 INFO snapshot.SnapshotManager: Skiplist is disabled
2021-09-14 12:22:24,309 INFO util.GSet: Computing capacity for map cachedBlocks
2021-09-14 12:22:24,309 INFO util.GSet: VM type = 64-bit
2021-09-14 12:22:24,310 INFO util.GSet: 0.25% max memory 889 MB = 2.2 MB
2021-09-14 12:22:24,310 INFO util.GSet: capacity = 2^18 = 262144 entries
2021-09-14 12:22:24,322 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2021-09-14 12:22:24,322 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2021-09-14 12:22:24,323 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.minutes = 1.5,25
2021-09-14 12:22:24,327 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2021-09-14 12:22:24,328 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
2021-09-14 12:22:24,338 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2021-09-14 12:22:24,338 INFO util.GSet: VM type = 64-bit
2021-09-14 12:22:24,339 INFO util.GSet: 0.029999999932944774% max memory 889 MB = 273.1 KB
2021-09-14 12:22:24,340 INFO util.GSet: capacity = 2^15 = 32768 entries
2021-09-14 12:22:24,517 INFO namenode.FSImage: Allocated new BlockPoolId: BP-2023101307-192.168.2.5-1631602344469
2021-09-14 12:22:24,655 INFO common.Storage: Storage directory \tmp\hadoop-labdh\dfs\name has been successfully formatted.
2021-09-14 12:22:24,758 INFO namenode.FSImageFormatProtobuf: Saving image file \tmp\hadoop-labdh\dfs\name\current\fsimage.cpt 00000000000000000000 using no compression
2021-09-14 12:22:24,997 INFO namenode.FSImageFormatProtobuf: Image file \tmp\hadoop-labdh\dfs\name\current\fsimage.cpt 00000000000000000000 of size 400 bytes saved in 0 seconds .
2021-09-14 12:22:25,044 INFO namenode.MSStorageRetentionManager: Going to retain 1 images with txid >= 0
2021-09-14 12:22:25,054 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2021-09-14 12:22:25,055 INFO namenode.NameNode: SHUTDOWN_MSG:
=====
SHUTDOWN_MSG: Shutting down NameNode at LAPTOP-Q037V1L1/192.168.2.5
=====
D:\nirma\7th sem\Big Data Analytics\labwork\start-dfs.cmd
D:\nirma\7th sem\Big Data Analytics\labwork\start-yarn.cmd
starting yarn daemons
D:\nirma\7th sem\Big Data Analytics\labwork\jps
20068 NameNode
20644 Jps
26268 ResourceManager
2872 NodeManager
21532 DataNode
D:\nirma\7th sem\Big Data Analytics\labwork_

```

### 3. Checking the clusters and working on localhost

Non Heap Memory used 47.93 MB of 49.02 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.	
Configured Capacity:	465.76 GB
Configured Remote Capacity:	0 B
DFS Used:	148 B (0%)
Non DFS Used:	96.76 GB
DFS Remaining:	369 GB (79.23%)
Block Pool Used:	148 B (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (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	Tue Sep 14 12:22:37 +0530 2021
Last Checkpoint Time	Tue Sep 14 12:22:24 +0530 2021
Enabled Erasure Coding Policies	RS-6-3-1024k

Hadoop Overview Utilities						
DataNode on 192.168.2.5:9866						
Cluster ID:	CID-82e3b47f-b10e-4055-b5c0-e06f5248bb97					
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842					
Block Pools						
Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)	
localhost:9000	BP-2023101307-192.168.2.5-1631602344469	RUNNING	1s	a few seconds	0 B (64 MB)	
Volume Information						
Directory	StorageType	Capacity Used	Capacity Left	Capacity Reserved	Reserved Space for Replicas	Blocks
D:\tmp\hadoop-labdhdfs\data	DISK	148 B	369 GB	0 B	0 B	0

**All Applications**

**Cluster Metrics**

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total
0	0	0	0	0	0 B	8 GB	0 B	0	8

**Cluster Nodes Metrics**

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
1	0	0	0	0	0

**Scheduler Metrics**

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

**Applications**

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress
No data available in table																		

Showing 0 to 0 of 0 entries

## 4. Mkdir command

```

Administration: Command Prompt - hadoop jar d:\hadoop-3.2.1\share\hadoop\mapreduce\hadoop-mapreduce-examples-3.2.1.jar grep /input/output 'dfs[a-z]+'
D:\nirma\7th sem\Big Data Analytics\labwork>start-dfs.cmd
D:\nirma\7th sem\Big Data Analytics\labwork>start-yarn.cmd
starting yarn daemons
D:\nirma\7th sem\Big Data Analytics\labwork>jps
28068 NameNode
28644 Jps
22628 ResourceManager
2872 NodeManager
21532 DataNode
D:\nirma\7th sem\Big Data Analytics\labwork>hdfs dfs -ls /
D:\nirma\7th sem\Big Data Analytics\labwork>hdfs dfs -mkdir /input
D:\nirma\7th sem\Big Data Analytics\labwork>hdfs dfs -put d:\hadoop-3.2.1\etc\hadoop\*.xml /input
put: ".": No such file or directory: "hdfs://localhost:9000/user/labdh"
D:\nirma\7th sem\Big Data Analytics\labwork>hdfs dfs -put d:\hadoop-3.2.1\etc\hadoop\*.xml /input
2021-09-14 12:42:19,983 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-09-14 12:42:20,586 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-09-14 12:42:20,674 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-09-14 12:42:20,745 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-09-14 12:42:20,813 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-09-14 12:42:20,914 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-09-14 12:42:21,027 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-09-14 12:42:21,141 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-09-14 12:42:21,220 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
D:\nirma\7th sem\Big Data Analytics\labwork>hdfs dfs -ls /
Found 1 items
drwxr-xr-x  - labdh supergroup          0 2021-09-14 12:42 /input
D:\nirma\7th sem\Big Data Analytics\labwork>hadoop jar d:\hadoop-3.2.1\share\hadoop\mapreduce\hadoop-mapreduce-examples-3.2.1.jar grep /input/output 'dfs[a-z]+'
2021-09-14 12:44:31,079 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2021-09-14 12:44:32,166 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/labdh/.staging/job_1631602376065_0001
2021-09-14 12:44:32,348 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-09-14 12:44:32,541 INFO InputFileInputFormat: Total input files to process = 9
2021-09-14 12:44:32,783 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-09-14 12:44:32,860 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-09-14 12:44:32,882 INFO mapreduce.JobSubmitter: number of splits:9
2021-09-14 12:44:33,083 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-09-14 12:44:33,117 INFO mapreduce.JobSubmitter: Submitting tokens for job: Job_1631602376065_0001
2021-09-14 12:44:33,117 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-09-14 12:44:33,330 INFO conf.Configuration: resource-types.xml not found
2021-09-14 12:44:33,330 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-09-14 12:44:33,786 INFO impl.YarnClientImpl: Submitted application application_1631602376065_0001
2021-09-14 12:44:33,862 INFO mapreduce.Job: The url to track the job: http://LAP10P-0Q3JV1L1:8088/proxy/application_1631602376065_0001/
2021-09-14 12:44:33,863 INFO mapreduce.Job: Running job: job_1631602376065_0001
  
```

## 5. Grep command

```
Administrator: Command Prompt

Total time spent by all reduce tasks (ms)=2948
Total vcore-milliseconds taken by all map tasks=3459
Total vcore-milliseconds taken by all reduce tasks=2948
Total megabyte-milliseconds taken by all map tasks=3542016
Total megabyte-milliseconds taken by all reduce tasks=3018752


Map-Reduce Framework
  Map input records=0
  Map output records=0
  Map output bytes=0
  Map output materialized bytes=6
  Input split bytes=131
  Combine input records=0
  Combine output records=0
  Reduce input groups=0
  Reduce shuffle bytes=6
  Reduce input records=0
  Reduce output records=0
  Spilled Records=0
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=97
  CPU time spent (ms)=1060
  Physical memory (bytes) snapshot=491020288
  Virtual memory (bytes) snapshot=807473152
  Total committed heap usage (bytes)=432537600
  Peak Map Physical memory (bytes)=298971116
  Peak Map Virtual memory (bytes)=459620352
  Peak Reduce Physical memory (bytes)=192090112
  Peak Reduce Virtual memory (bytes)=347987968

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0

File Input Format Counters
  Bytes Read=86
  File Output Format Counters
  Bytes Written=0

D:\nirma\7th sem\Big Data Analytics\labwork>hdfs dfs -ls /output
Found 2 items
-rw-r--r-- 1 labdh supergroup          0 2021-09-14 12:45 /output/ SUCCESS
-rw-r--r-- 1 labdh supergroup          0 2021-09-14 12:45 /output/part-r-00000

D:\nirma\7th sem\Big Data Analytics\labwork>hdfs dfs -cat /output/*
D:\nirma\7th sem\Big Data Analytics\labwork>
```



Cluster

About

Nodes

Node Labels

Applications

NEW

NEW SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Tools

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved
2	0	0	2	0	0 B	8 GB	0 B

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes
1	0	0	0	0

Scheduler Metrics

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

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Reserved CPU VCore
application_1631602376065_0002	labdh	grep-sort	MAPREDUCE	default	0	Tue Sep 14 12:45:12 +0550 2021	Tue Sep 14 12:45:17 +0550 2021	Tue Sep 14 12:45:38 +0550 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A
application_1631602376065_0001	labdh	grep-search	MAPREDUCE	default	0	Tue Sep 14 12:44:33 +0550 2021	Tue Sep 14 12:44:34 +0550 2021	Tue Sep 14 12:45:10 +0550 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A

Showing 1 to 2 of 2 entries