

گام اول:

موارد زیر را در فایل گزارش نمایش دهید:

- نمایش کانتینرهای ایجاد شده
- توضیح وظیفه هرکدام از کانتینرها در Hadoop
- با استفاده از دستور `jps` در هر کانتینر، صحت نقش آن کانتینر در Hadoop را بررسی کنید و اسکرین شات آن را بیاورید.

نمایش کانتینر های ایجاد شده:

`docker compose ps`

```

+ hadoop git:(main) # docker compose up -d
[+] Building 0.0s (0/0)
[+] Running 0/6
  Network hadoop_default    Created                                0.1s
  Container resourcemanager Started                                0.6s
  Container nodemanager     Started                                0.6s
  Container namenode        Started                                0.6s
  Container datanode         Started                                0.9s
  Container historyserver   Started                                0.9s
+ hadoop git:(main) # docker compose ps

```

NAME	IMAGE	COMMAND	SERVICE	CREATED	STATUS	PORTS
datanode	arminzolfagharid/hadoop-datanode:v2-hadoop3.2.1	"/entrypoint.sh /run_"	datanode	10 seconds ago	Up 8 seconds (health: starting)	9864/tcp
historyserver	arminzolfagharid/hadoop-historyserver:v2-hadoop3.2.1	"/entrypoint.sh /run_"	historyserver	10 seconds ago	Up 8 seconds (health: starting)	8188/tcp
namenode	arminzolfagharid/hadoop-namenode:v2-hadoop3.2.1	"/entrypoint.sh /run_"	namenode	10 seconds ago	Up 8 seconds (health: starting)	0.0.0.0:9000->9000/tcp, :::9000->9000/tcp
resourcemanager	arminzolfagharid/hadoop-resource-manager:v2-hadoop3.2.1	"/entrypoint.sh /run_"	resourcemanager	10 seconds ago	Up 8 seconds (health: starting)	8042/tcp
nodemanager	arminzolfagharid/hadoop-nodemanager:v2-hadoop3.2.1	"/entrypoint.sh /run_"	nodemanager	10 seconds ago	Up 8 seconds (health: starting)	8088/tcp

```

+ hadoop git:(main) # |

```

همان گونه که مشخص است کانتینر های ما همه Up می باشند.

توضیح وظایف هر کدام از کانتینر ها در hadoop:

ResourceManager:

پروسه اصلی است که resource های مختلف را مدیریت میکند و آنها را بین application های مختلف یا همان job ها توزیع میکند.

NodeManager:

به ازای هر نود یک پروسه وجود دارد که resource های آن نود را محاسبه میکند.

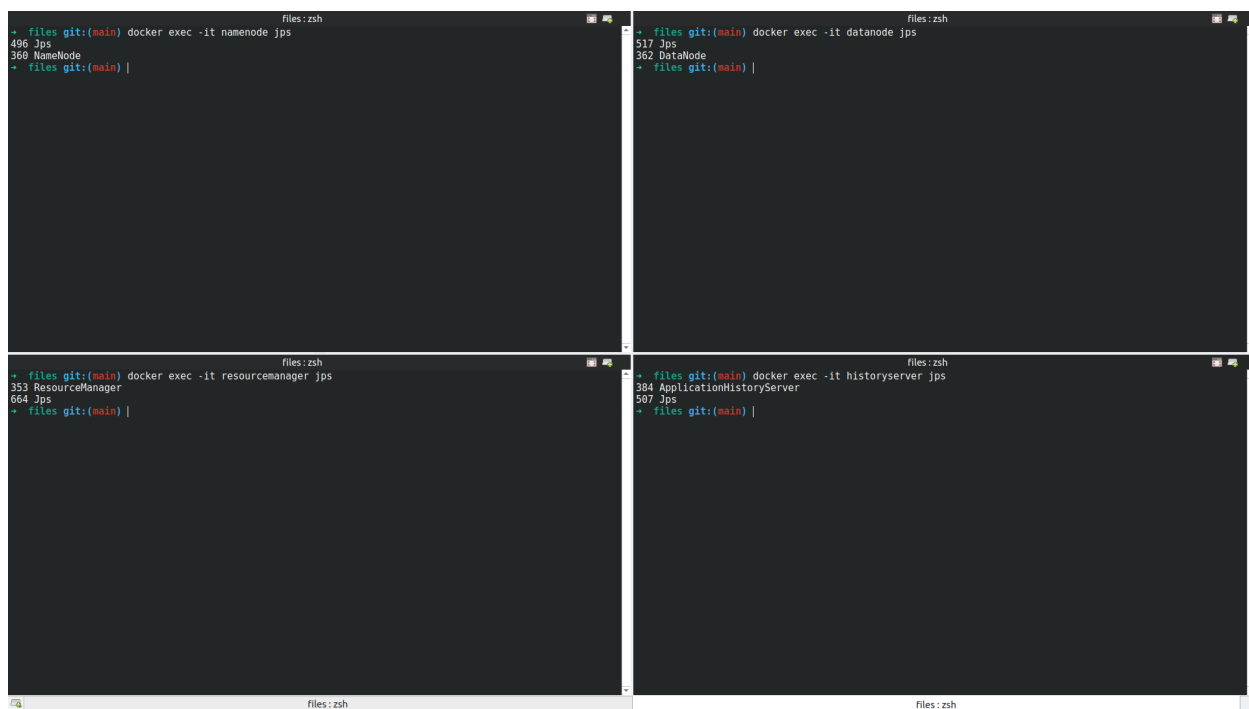
NameNode:

قلب HDFS filesystem است، این کانتینر متادیتا های سیستم را مدیریت و manage میکند، یعنی کدام بلاک ها یک فایل را تشکیل می دهند، و این بلاک ها روی کدام datanode ها ذخیره شده اند.

DataNode:

جایی است که HDFS دیتای واقعی را ذخیره میکند، اغلب چند dataNode وجود دارد.

با استفاده از دستور `jps` در هر کانتینر، صحت نقش آن کانتینر در Hadoop را بررسی کنید و اسکرین شات آن را بیاورید.



نمایش WebUI:

The screenshot shows the Hadoop WebUI Overview page for namenode:9000 (active). The page has a green header with navigation tabs: Hadoop, Overview (selected), Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. The main content area is titled "Overview 'namenode:9000' (active)".

Started:	Sun Jun 11 01:14:16 +0330 2023
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled:	Tue Sep 10 20:26:00 +0430 2019 by rohitsharmaks from branch-3.2.1
Cluster ID:	CID-d37b91a5-3c93-458a-9020-7d7a0fb4f872
Block Pool ID:	BP-235333020-172.18.0.4-1686431918609

Summary

Security is off.

Safe mode is ON. Resources are low on NN. Please add or free up more resources then turn off safe mode manually. NOTE: If you turn off safe mode before adding resources, the NN will immediately return to safe mode. Use "hdfs dfsadmin -safemode leave" to turn safe mode off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 96.92 MB of 289.5 MB Heap Memory. Max Heap Memory is 1.58 GB.

Non Heap Memory used 46.86 MB of 48.31 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	8.59 GB
Configured Remote Capacity:	0 B
DFS Used:	28 KB (0%)
Non DFS Used:	8.51 GB

در قسمت Utilities > Browse the file system داریم:

The screenshot shows the Hadoop WebUI Browse Directory page. The header is the same as the Overview page. The main content area is titled "Browse Directory".

At the top, there is a text input field containing "/", a "Go!" button, and three icons: a folder, an upload arrow, and a trash can.

Below the input field, there is a "Show" dropdown menu set to "25" and a "Search:" text input field.

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Jun 11 20:00	0	0 B	rmstate	

Showing 1 to 1 of 1 entries

Previous 1 Next

Hadoop, 2019.

گام دوم:

برای استفاده از dataset، فایل csv را داخل namenode قرار دهید:

```
+ files git:(main) docker compose up -d
[+] Building 0.0s (0/0)
[+] Running 5/5
✓ Container resourcemanager Started 0.7s
✓ Container nodemanager Started 0.3s
✓ Container datanode Started 0.7s
✓ Container namenode Started 0.7s
✓ Container historyserver Started 0.7s
+ files git:(main) docker volume ls
DRIVER VOLUME NAME
local 290c217dd92055d802666ccca273b6ab59ac15520d0beaf48ac0599925f4fcf1
local c859677ef10229378ff64870c5bf0297ef00c7674907a4901a77651e2a464832
local files_hadoop_datanode
local files_hadoop_historyserver
local files_hadoop_namenode
local hadoop_hadoop_datanode
local hadoop_hadoop_historyserver
local hadoop_hadoop_namenode
+ files git:(main) docker volume inspect files_hadoop_namenode
[
  {
    "CreatedAt": "2023-06-11T19:57:39+03:30",
    "Driver": "local",
    "Labels": {
      "com.docker.compose.project": "files",
      "com.docker.compose.version": "2.18.1",
      "com.docker.compose.volume": "hadoop_namenode"
    },
    "Mountpoint": "/home/ali/docker/volumes/files_hadoop_namenode/_data",
    "Name": "files_hadoop_namenode",
    "Options": null,
    "Scope": "local"
  }
]
+ files git:(main) sudo cp dataset.csv /home/ali/docker/volumes/files_hadoop_namenode/_data
[sudo] password for ali:
Sorry, try again.
[sudo] password for ali:
+ files git:(main) docker exec -it namenode sh
# ls
KEYS bin boot dev entrypoint.sh etc hadoop hadoop-data home lib lib64 media mnt opt proc root run run.sh/sbin srv sys tmp usr var
# cd hadoop
# ls
dfs
# cd dfs
# ls
name
# cd name
# ls
current dataset.csv in_use.lock
#
```

گام سوم:

1. با استفاده از **HDFS CLI**، پوشه‌ی **user/root/input** را در **HDFS** ایجاد کنید.
 2. فایل **dataset.csv** را با استفاده از **HDFS CLI** در **HDFS** در پوشه‌ی **input** قرار دهید.
- حال قرار است سه برنامه **MapReduce** بنویسید که از دیتاستی که در مسیر **HDFS** قرار داده شده است، استفاده کند.

1 و 2:

```
hdfs dfs -mkdir -p /user/root/input
```

```
hdfs dfs -ls /user/root/input
```

```
hdfs dfs -put dataset.csv /user/root/input
```

برنامه اولی که اجرا کردم این بود:

```
sudo cp * /home/ali/docker/volumes/files_hadoop_namenode/_data
```

فایل های **mapper.py**, **reducer.py** رو کپی کردم توی **namenode** کانتینر، بعدش با دستور زیر اجراش کردم:

```
hadoop jar /opt/hadoop-3.2.1/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar \
```

```
-file /hadoop/dfs/name/mapper.py -mapper /hadoop/dfs/name/mapper.py \
```

```
-file /hadoop/dfs/name/reducer.py -reducer /hadoop/dfs/name/reducer.py \
```

```
-input /user/root/input/dataset.csv -output /hadoop/dfs/name/output.txt
```

خروجی برنامه این بود، فکر میکنم مشکلات ایجادي به خاطر python 3.5.3 باشه:

```
# hadoop jar /opt/hadoop-3.2.1/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar \
  -file /hadoop/dfs/name/mapper.py -mapper /hadoop/dfs/name/mapper.py \
  -file /hadoop/dfs/name/reducer.py -reducer /hadoop/dfs/name/reducer.py \
  -input /user/root/2 input/dataset.csv -output /hadoop/dfs/name/output.txt
> > 2023-06-11 20:53:06,332 WARN streaming.StreamJob: file option is deprecated, please use generic option -files instead.
packageJobJar: (/hadoop/dfs/name/mapper.py, /hadoop/dfs/name/reducer.py, /tmp/hadoop-unjar7784089250443097760/) [] /tmp/streamjob7860345246828782193.jar tmpDir=null
2023-06-11 20:53:06,628 INFO client.RMProxy: Connecting to ResourceMgr at resourcemanager/172.27.0.2:8032
2023-06-11 20:53:06,156 INFO client.AMProxy: Connecting to Application History server at historyserver/172.27.0.4:10200
2023-06-11 20:53:06,178 INFO client.RMProxy: Connecting to ResourceMgr at resourcemanager/172.27.0.2:8032
2023-06-11 20:53:06,179 INFO client.AMProxy: Connecting to Application History server at historyserver/172.27.0.4:10200
2023-06-11 20:53:06,338 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1686511408090_0005
2023-06-11 20:53:06,410 INFO sasL.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-06-11 20:53:06,490 INFO sasL.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-06-11 20:53:06,582 INFO sasL.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-06-11 20:53:06,546 INFO mapred.FileInputFormat: Total input files to process : 1
2023-06-11 20:53:06,560 INFO sasL.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-06-11 20:53:06,580 INFO sasL.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-06-11 20:53:06,585 INFO mapreduce.JobSubmitter: number of splits:2
2023-06-11 20:53:06,706 INFO sasL.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-06-11 20:53:07,122 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1686511408090_0005
2023-06-11 20:53:07,122 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-06-11 20:53:07,289 INFO conf.Configuration: resource-types.xml not found
2023-06-11 20:53:07,289 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-06-11 20:53:07,543 INFO impl.YarnClientImpl: Submitted application 1686511408090_0005
2023-06-11 20:53:07,568 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1686511408090_0005/
2023-06-11 20:53:07,570 INFO mapreduce.Job: Running job: job_1686511408090_0005
2023-06-11 20:53:11,634 INFO mapreduce.Job: Job job_1686511408090_0005 running in uber mode : false
2023-06-11 20:53:11,637 INFO mapreduce.Job: map 0% reduce 0%
2023-06-11 20:53:18,730 INFO mapreduce.Job: map 50% reduce 0%
2023-06-11 20:53:19,743 INFO mapreduce.Job: map 100% reduce 0%
2023-06-11 20:53:21,757 INFO mapreduce.Job: Task Id : attempt 1686511408090_0005_r_0000000_0, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 1
    at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:326)
    at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:539)
    at org.apache.hadoop.streaming.PipeReducer.reduce(PipeReducer.java:128)
    at org.apache.hadoop.mapred.ReduceTask.runOldReducer(ReduceTask.java:445)
    at org.apache.hadoop.mapred.ReduceTask.run(ReduceTask.java:393)
    at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:174)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1730)
    at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:168)

2023-06-11 20:53:25,826 INFO mapreduce.Job: Task Id : attempt 1686511408090_0005_r_0000000_1, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 1
    at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:326)
    at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:539)
    at org.apache.hadoop.streaming.PipeReducer.reduce(PipeReducer.java:128)
    at org.apache.hadoop.mapred.ReduceTask.runOldReducer(ReduceTask.java:445)
    at org.apache.hadoop.mapred.ReduceTask.run(ReduceTask.java:393)
    at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:174)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1730)
    at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:168)

2023-06-11 20:53:29,863 INFO mapreduce.Job: Task Id : attempt 1686511408090_0005_r_0000000_2, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 1
#
```

```
    at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:174)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1730)
    at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:168)

2023-06-11 20:53:34,901 INFO mapreduce.Job: map 100% reduce 100%
2023-06-11 20:53:34,922 INFO mapreduce.Job: Job job_1686511408090_0005 failed with state FAILED due to: Task failed task_1686511408090_0005_r_000000
Job failed as tasks failed. failedMaps:0 failedReduces:1 killedMaps:0 killedReduces: 0

2023-06-11 20:53:35,015 INFO mapreduce.Job: Counters: 40
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=834840
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=103282660
  HDFS: Number of bytes written=0
  HDFS: Number of read operations=6
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=0
  HDFS: Number of bytes read erasure-coded=0
Job Counters
  Failed reduce tasks=4
  Launched map tasks=2
  Launched reduce tasks=4
  Rack-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=29476
  Total time spent by all reduces in occupied slots (ms)=55928
  Total time spent by all map tasks (ms)=7369
  Total time spent by all reduce tasks (ms)=6991
  Total vcore-milliseconds taken by all map tasks=7369
  Total vcore-milliseconds taken by all reduce tasks=6991
  Total megabyte-milliseconds taken by all map tasks=30183424
  Total megabyte-milliseconds taken by all reduce tasks=57270272
Map-Reduce Framework
  Map input records=208000
  Map output records=19973
  Map output bytes=7614025
  Map output materialized bytes=369158
  Input split bytes=200
  Combine input records=0
  Spilled Records=19973
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=01
  CPU time spent (ms)=2800
  Physical memory (bytes) snapshot=570494076
  Virtual memory (bytes) snapshot=1021682176
  Total committed heap usage (bytes)=522715136
  Peak Map Physical memory (bytes)=285622272
  Peak Map Virtual memory (bytes)=5108641792
File Input Format Counters
  Bytes Read=103282460
2023-06-11 20:53:35,015 ERROR streaming.StreamJob: Job not successful!
Streaming Command Failed!
#
```