علی مرادزاده ۹۸۳۱۰۵۸

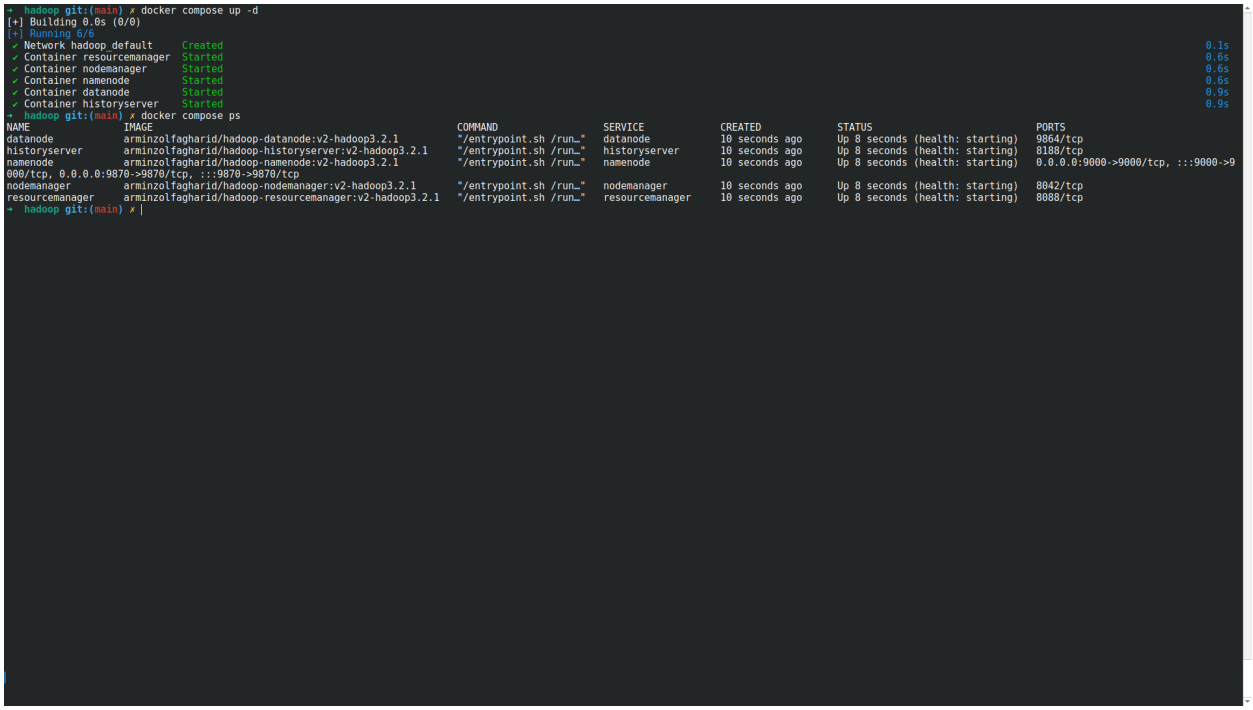
گام اول:

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 ایجاده شده:

docker compose ps



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.

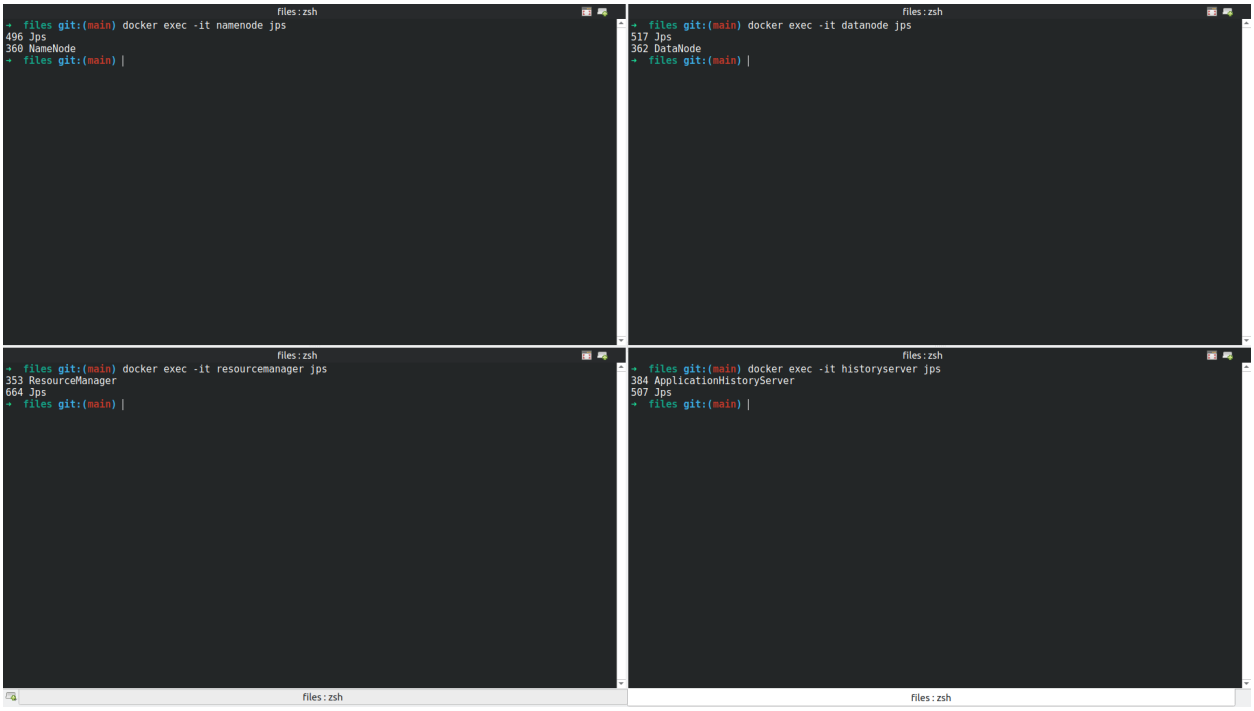
NameNode:

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.

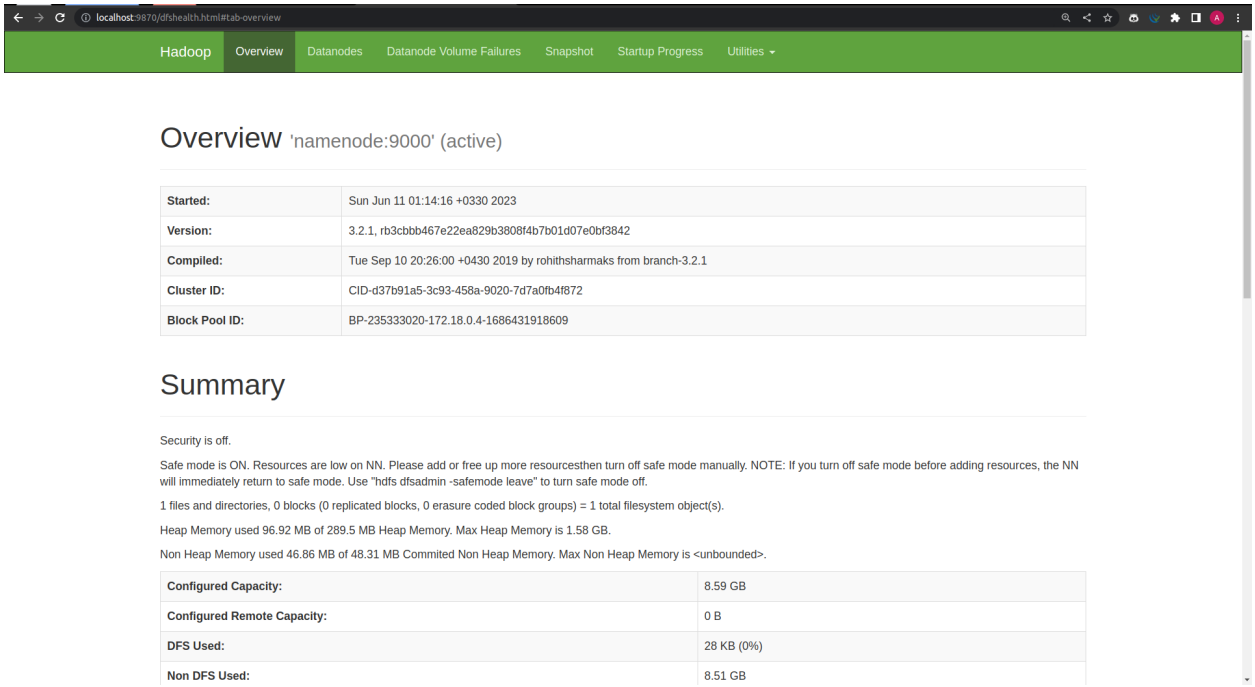
DataNode:

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

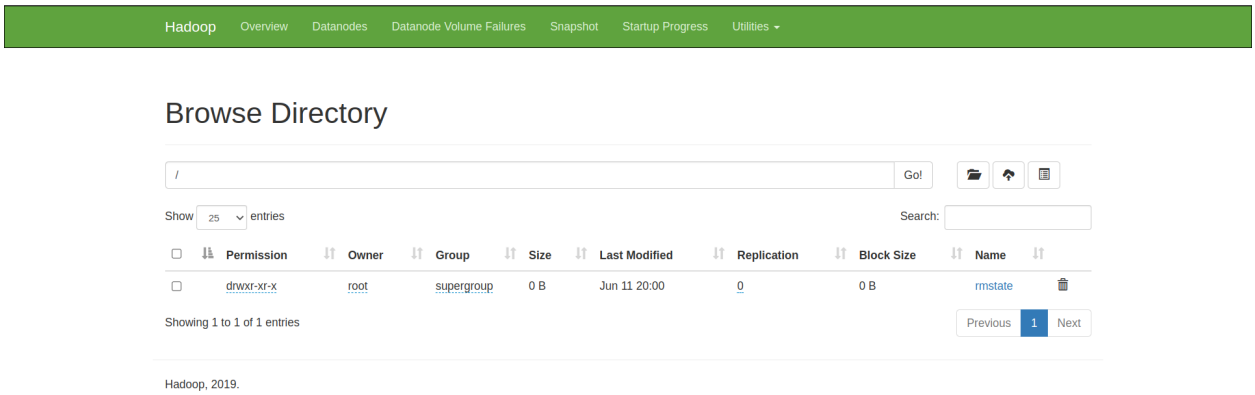
با استفاده از دستور jps در هر کانتینر، صحت نقش آن کانتینر در Hadoop را بررسی کنید و اسکرین شات آن را بیاورید.



نمایش WebUI:



در قسمت Utilities > Browse the file system داریم:



گام دوم:

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

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:

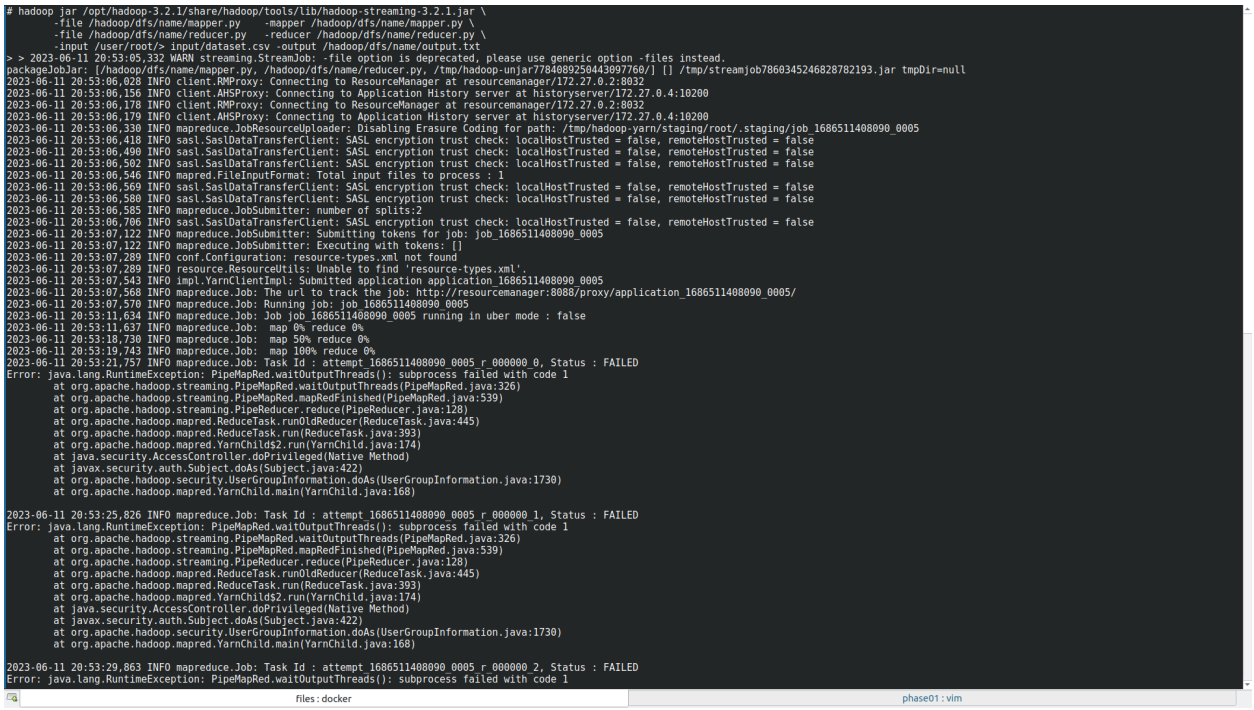
hdfs dfs -mkdir -p /user/root/input
hdfs dfs -ls /user/root/input
hdfs dfs -put dataset.csv /user/root/input

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 ایجادی به خاطر python 3.5.3 باشه:

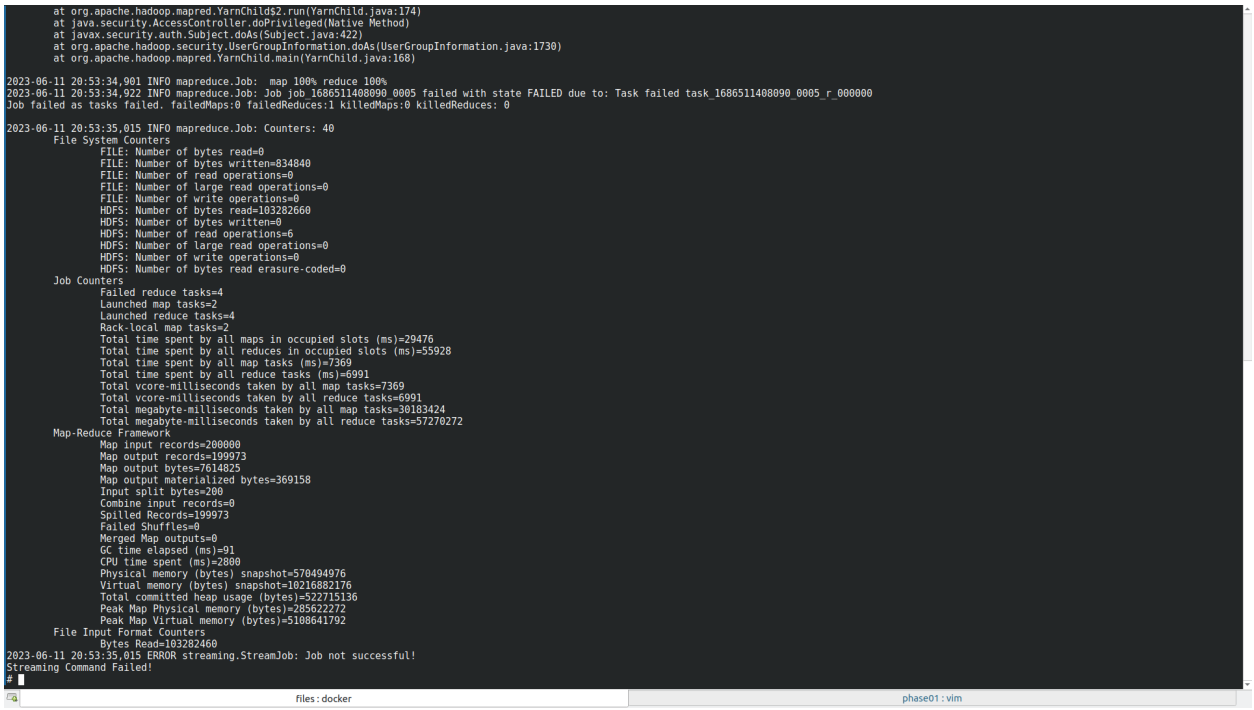


```
# hadoop jar /opt/hadoop-3.2.1/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar \
        -file /hadoop/dfs/name/mapper.py    -mapper /hadoop/dfs/name/mapper.py \
        -file /hadoop/dfs/name/reducer.py   -reducer /hadoop/dfs/name/reducer.py \
        -input /user/root/> input/dataset.csv -output /hadoop/dfs/name/output.txt
> > 2023-06-11 20:53:05,332 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 20:53:06,028 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.27.0.2:8032
2023-06-11 20:53:06,156 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.27.0.4:10200
2023-06-11 20:53:06,178 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.27.0.2:8032
2023-06-11 20:53:06,179 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.27.0.4:10200
2023-06-11 20:53:06,330 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1686511408090_0005
2023-06-11 20:53:06,418 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 20:53:06,502 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 20:53:06,569 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 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_000000_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_000000_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_000000_2, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 1
```

files : docker | phase01 : vim



```
        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=200000
                Map output records=199973
                Map output bytes=7614825
                Map output materialized bytes=369158
                Input split bytes=200
                Combine input records=0
                Spilled Records=199973
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=91
                CPU time spent (ms)=2800
                Physical memory (bytes) snapshot=570494976
                Virtual memory (bytes) snapshot=10216882176
                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!
#
```

files : docker | phase01 : vim