**TRƯỜNG ĐẠI HỌC XÂY DỰNG**

**KHOA CÔNG NGHỆ THÔNG TIN  
------------o0o------------**

****

BÀI TẬP MÔN DỮ LIỆU LỚN

*ĐỀ TÀI:*

**CÀI ĐẶT HADOOP, MONGODB, REDIS**

***GV: Nguyễn Đình Quý***

**Nhóm sinh viên thực hiện (Nhóm 05):**

Hoàng Đình Long Nhật – 1658264 – 64CS2

Lương Thanh Tài – 173264 – 64CS2

Nguyễn Đình Huy – 1524964 – 64CS2

***Hà Nội, tháng 4 năm 2022.***

1. **Hadoop**
   1. **Triển khai Hadoop cluster**

Khởi tạo môi trường hadoop gồm 1 namenode và 2 datanode

Môi trường này đã được expose các port theo thứ tự tăng dần để tránh trùng lặp

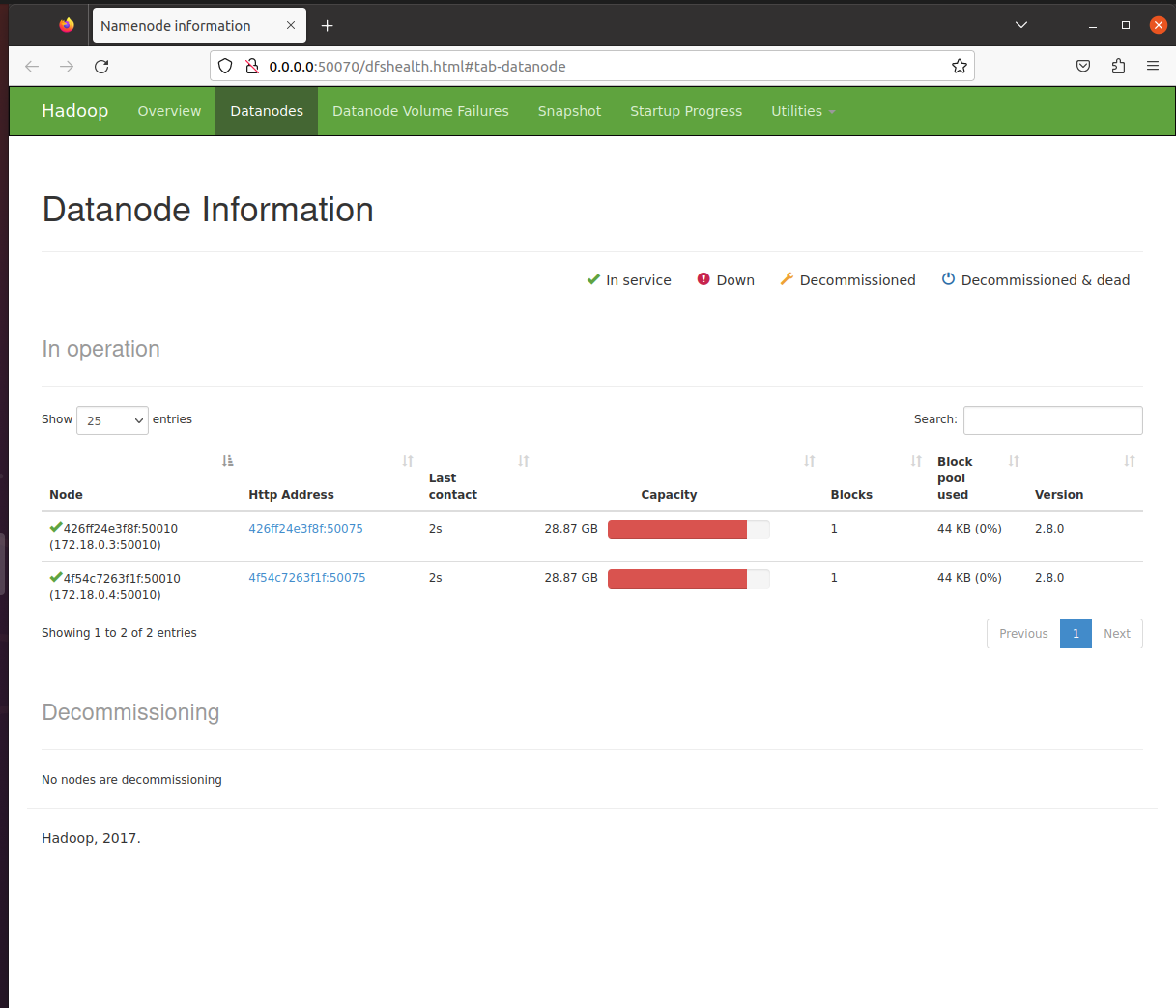
|  |
| --- |
| version: '3.8'  services:  namenode:  image: bde2020/hadoop-namenode:1.1.0-hadoop2.8-java8  container\_name: namenode  volumes:  - ./data/namenode:/hadoop/dfs/name  - ./testdata:/home/testdata  environment:  - CLUSTER\_NAME=test  env\_file:  - ./hadoop.env  ports:  - 50070:50070    datanode1:  image: bde2020/hadoop-datanode:1.1.0-hadoop2.8-java8  container\_name: datanode1  depends\_on:  - namenode  volumes:  - ./data/datanode1:/hadoop/dfs/data  env\_file:  - ./hadoop.env  expose:  - 50075  ports:  - 50075:50075  datanode2:  image: bde2020/hadoop-datanode:1.1.0-hadoop2.8-java8  container\_name: datanode2  depends\_on:  - namenode  volumes:  - ./data/datanode2:/hadoop/dfs/data  env\_file:  - ./hadoop.env  expose:  - 50076  ports:  - 50076:50075  # datanode3:  # image: bde2020/hadoop-datanode:1.1.0-hadoop2.8-java8  # container\_name: datanode3  # depends\_on:  # - namenode  # volumes:  # - ./data/datanode3:/hadoop/dfs/data  # env\_file:  # - ./hadoop.env  # ports:  # - 50077:50075 |

Biến môi trường sẽ được load vào thông qua config của docker compose

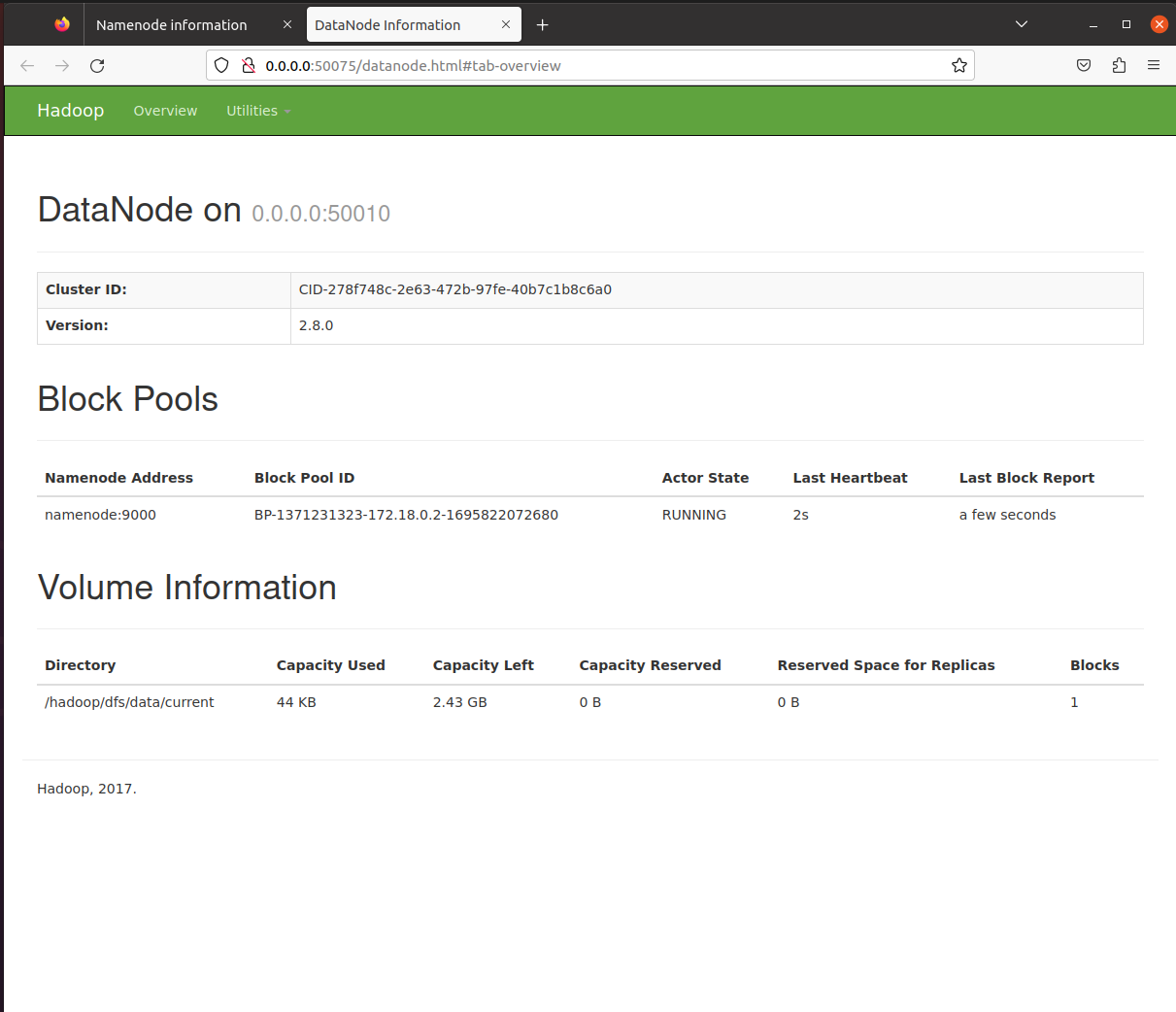
|  |
| --- |
| CORE\_CONF\_fs\_defaultFS=hdfs://namenode:9000  CORE\_CONF\_hadoop\_http\_staticuser\_user=root  CORE\_CONF\_hadoop\_proxyuser\_hue\_hosts=\*  CORE\_CONF\_hadoop\_proxyuser\_hue\_groups=\*  CORE\_CONF\_io\_compression\_codecs=org.apache.hadoop.io.compress.SnappyCodec  HDFS\_CONF\_dfs\_webhdfs\_enabled=true  HDFS\_CONF\_dfs\_permissions\_enabled=false  HDFS\_CONF\_dfs\_namenode\_datanode\_registration\_ip\_\_\_hostname\_\_\_check=false  YARN\_CONF\_yarn\_log\_\_\_aggregation\_\_\_enable=true  YARN\_CONF\_yarn\_log\_server\_url=http://historyserver:8188/applicationhistory/logs/  YARN\_CONF\_yarn\_resourcemanager\_recovery\_enabled=true  YARN\_CONF\_yarn\_resourcemanager\_store\_class=org.apache.hadoop.yarn.server.resourcemanager.recovery.FileSystemRMStateStore  YARN\_CONF\_yarn\_resourcemanager\_scheduler\_class=org.apache.hadoop.yarn.server.resourcemanager.scheduler.capacity.CapacityScheduler  YARN\_CONF\_yarn\_scheduler\_capacity\_root\_default\_maximum\_\_\_allocation\_\_\_mb=8192  YARN\_CONF\_yarn\_scheduler\_capacity\_root\_default\_maximum\_\_\_allocation\_\_\_vcores=4  YARN\_CONF\_yarn\_resourcemanager\_fs\_state\_\_\_store\_uri=/rmstate  YARN\_CONF\_yarn\_resourcemanager\_system\_\_\_metrics\_\_\_publisher\_enabled=true  YARN\_CONF\_yarn\_resourcemanager\_hostname=resourcemanager  YARN\_CONF\_yarn\_resourcemanager\_address=resourcemanager:8032  YARN\_CONF\_yarn\_resourcemanager\_scheduler\_address=resourcemanager:8030  YARN\_CONF\_yarn\_resourcemanager\_resource\_\_tracker\_address=resourcemanager:8031  YARN\_CONF\_yarn\_timeline\_\_\_service\_enabled=true  YARN\_CONF\_yarn\_timeline\_\_\_service\_generic\_\_\_application\_\_\_history\_enabled=true  YARN\_CONF\_yarn\_timeline\_\_\_service\_hostname=historyserver  YARN\_CONF\_mapreduce\_map\_output\_compress=true  YARN\_CONF\_mapred\_map\_output\_compress\_codec=org.apache.hadoop.io.compress.SnappyCodec  YARN\_CONF\_yarn\_nodemanager\_resource\_memory\_\_\_mb=16384  YARN\_CONF\_yarn\_nodemanager\_resource\_cpu\_\_\_vcores=8  YARN\_CONF\_yarn\_nodemanager\_disk\_\_\_health\_\_\_checker\_max\_\_\_disk\_\_\_utilization\_\_\_per\_\_\_disk\_\_\_percentage=98.5  YARN\_CONF\_yarn\_nodemanager\_remote\_\_\_app\_\_\_log\_\_\_dir=/app-logs  YARN\_CONF\_yarn\_nodemanager\_aux\_\_\_services=mapreduce\_shuffle  MAPRED\_CONF\_mapreduce\_framework\_name=yarn  MAPRED\_CONF\_mapred\_child\_java\_opts=-Xmx4096m  MAPRED\_CONF\_mapreduce\_map\_memory\_mb=4096  MAPRED\_CONF\_mapreduce\_reduce\_memory\_mb=8192  MAPRED\_CONF\_mapreduce\_map\_java\_opts=-Xmx3072m  MAPRED\_CONF\_mapreduce\_reduce\_java\_opts=-Xmx6144m  MAPRED\_CONF\_yarn\_app\_mapreduce\_am\_env=HADOOP\_MAPRED\_HOME=/opt/hadoop-3.2.1/  MAPRED\_CONF\_mapreduce\_map\_env=HADOOP\_MAPRED\_HOME=/opt/hadoop-3.2.1/  MAPRED\_CONF\_mapreduce\_reduce\_env=HADOOP\_MAPRED\_HOME=/opt/hadoop-3.2.1/  PYSPARK\_DRIVER\_PYTHON='jupyter'  PYSPARK\_DRIVER\_PYTHON\_OPTS='notebook --no-browser --port=8889' |

Truy cập thông tin của các node:

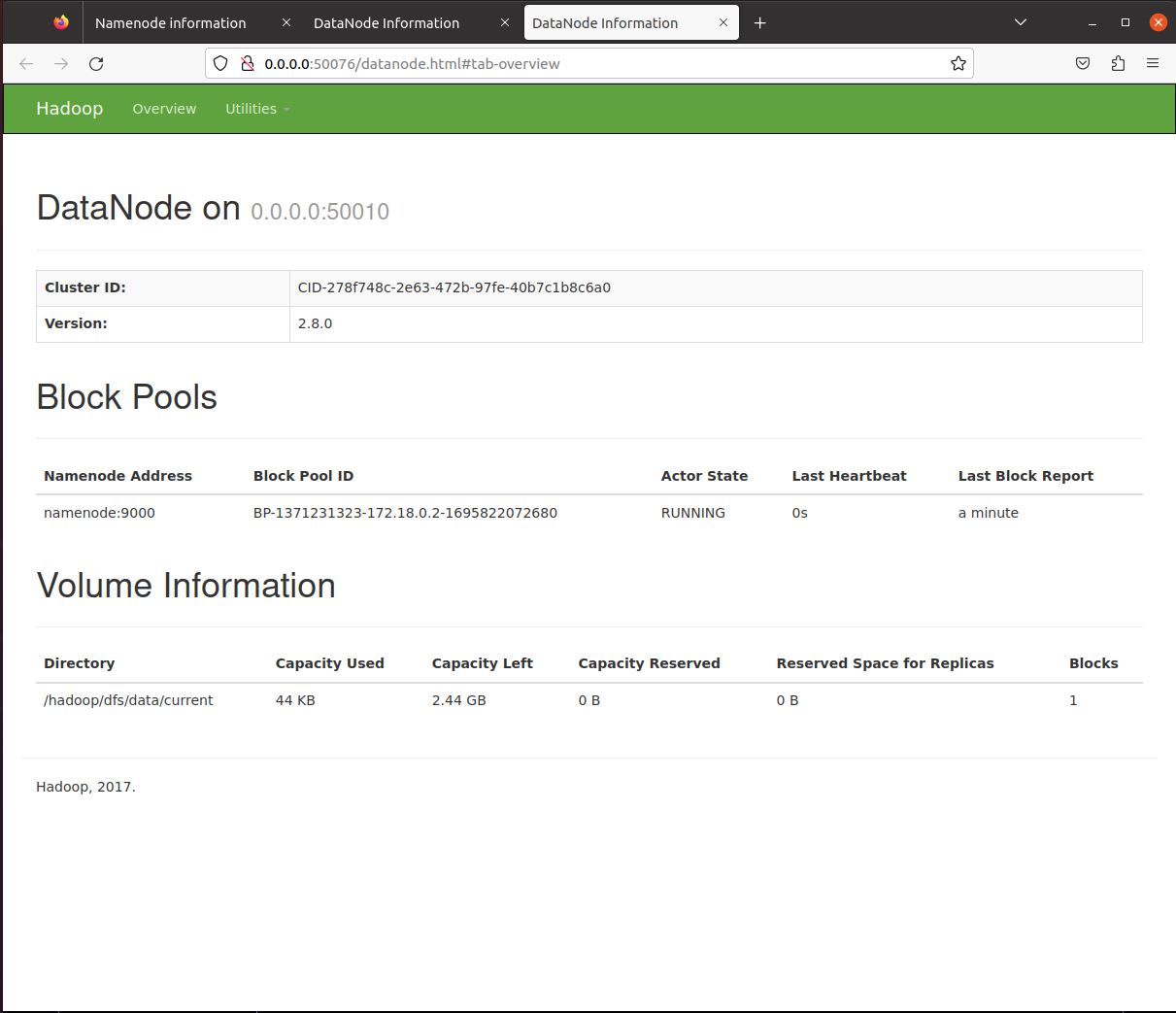
Namenode: <http://0.0.0.0:50070/dfshealth.html#tab-datanode>



Datanode1: <http://0.0.0.0:50075/datanode.html#tab-overview>



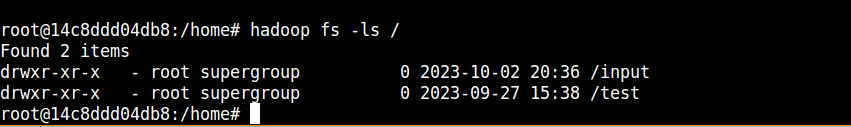
Datanode2: <http://0.0.0.0:50076/datanode.html#tab-overview>



Datanode3: <http://0.0.0.0:50077/datanode.html#tab-overview> (Trường hợp mà thêm node)

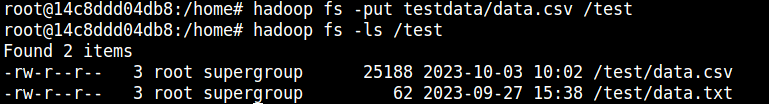
* 1. **Các tính năng của hadoop**

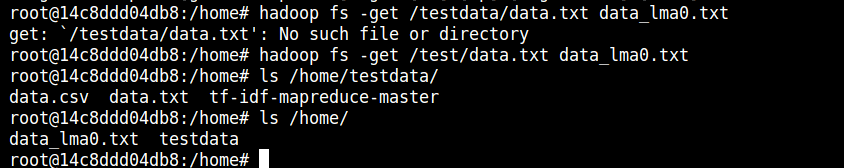
Xem các file trong hadoop



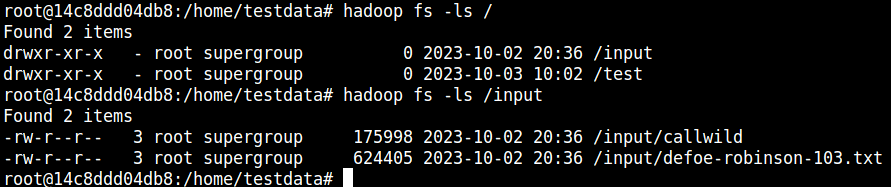
Xem đọc file trong hadoop



Tải file lên hadoop 

Tải file từ hadoop 

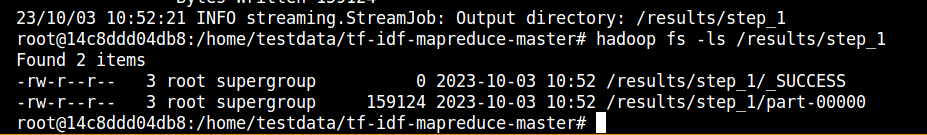
* 1. **TF-IDF trong hadoop**
* File input



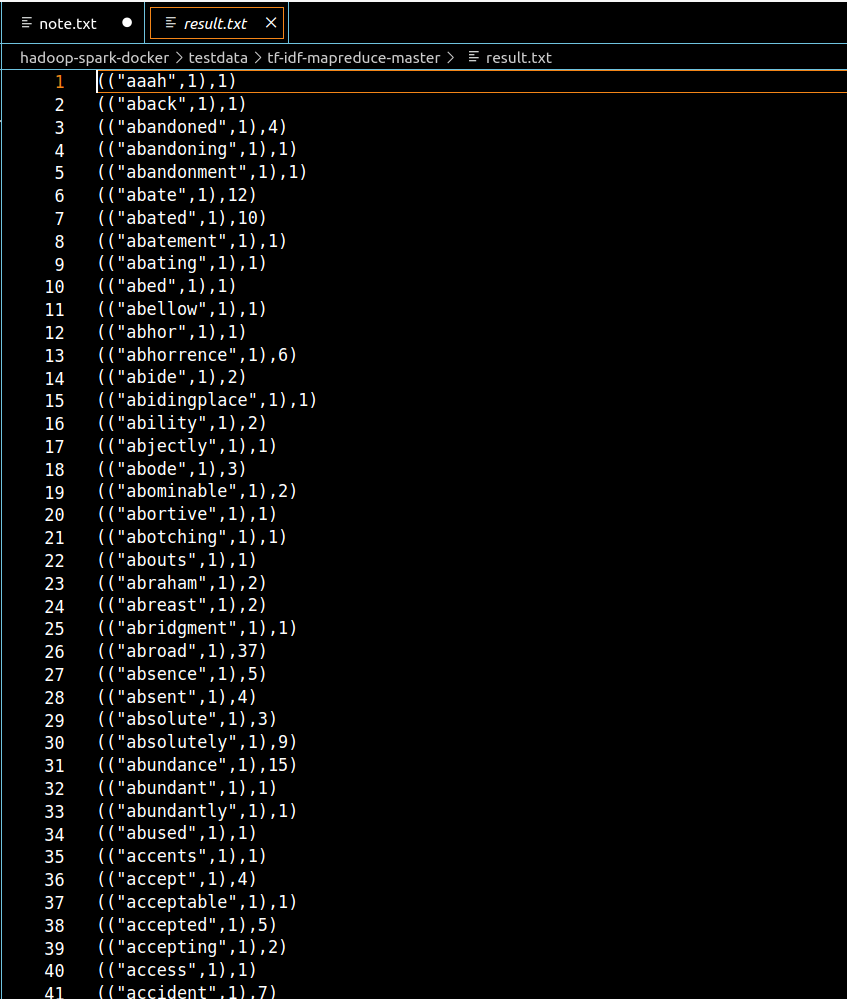
* Thực hiện worldcount

|  |
| --- |
| root@14c8ddd04db8:/home/testdata/tf-idf-mapreduce-master# hadoop jar hadoop-streaming-2.8.0.jar -input /input/\* -output /results/step\_1 -mapper 1\_wordcount\_mapper.py -reducer 1\_wordc  ount\_reducer.py  23/10/03 10:52:19 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id  23/10/03 10:52:19 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=  23/10/03 10:52:19 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized  23/10/03 10:52:19 INFO mapred.FileInputFormat: Total input files to process : 2  23/10/03 10:52:19 INFO mapreduce.JobSubmitter: number of splits:2  23/10/03 10:52:19 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_local1389664555\_0001  23/10/03 10:52:19 INFO mapreduce.Job: The url to track the job: http://localhost:8080/  23/10/03 10:52:19 INFO mapred.LocalJobRunner: OutputCommitter set in config null  23/10/03 10:52:19 INFO mapreduce.Job: Running job: job\_local1389664555\_0001  23/10/03 10:52:19 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter  23/10/03 10:52:19 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1  23/10/03 10:52:19 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup \_temporary folders under output directory:false, ignore cleanup failures: false  23/10/03 10:52:19 INFO mapred.LocalJobRunner: Waiting for map tasks  23/10/03 10:52:19 INFO mapred.LocalJobRunner: Starting task: attempt\_local1389664555\_0001\_m\_000000\_0  23/10/03 10:52:19 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1  23/10/03 10:52:19 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup \_temporary folders under output directory:false, ignore cleanup failures: false  23/10/03 10:52:19 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]  23/10/03 10:52:19 INFO mapred.MapTask: Processing split: hdfs://namenode:9000/input/defoe-robinson-103.txt:0+624405  23/10/03 10:52:19 INFO mapred.MapTask: numReduceTasks: 1  23/10/03 10:52:19 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)  23/10/03 10:52:19 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100  23/10/03 10:52:19 INFO mapred.MapTask: soft limit at 83886080  23/10/03 10:52:19 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600  23/10/03 10:52:19 INFO mapred.MapTask: kvstart = 26214396; length = 6553600  23/10/03 10:52:19 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer  23/10/03 10:52:19 INFO streaming.PipeMapRed: PipeMapRed exec [/home/testdata/tf-idf-mapreduce-master/./1\_wordcount\_mapper.py]  23/10/03 10:52:19 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir  23/10/03 10:52:19 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start  23/10/03 10:52:19 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap  23/10/03 10:52:19 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id  23/10/03 10:52:19 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id  23/10/03 10:52:19 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir  23/10/03 10:52:19 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file  23/10/03 10:52:19 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords  23/10/03 10:52:19 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length  23/10/03 10:52:19 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id  23/10/03 10:52:19 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name  23/10/03 10:52:19 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition  23/10/03 10:52:19 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]  23/10/03 10:52:19 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]  23/10/03 10:52:19 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]  23/10/03 10:52:20 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]  23/10/03 10:52:20 INFO streaming.PipeMapRed: Records R/W=2182/1  23/10/03 10:52:20 INFO streaming.PipeMapRed: R/W/S=10000/25021/0 in:NA [rec/s] out:NA [rec/s]  23/10/03 10:52:20 INFO streaming.PipeMapRed: MRErrorThread done  23/10/03 10:52:20 INFO streaming.PipeMapRed: mapRedFinished  23/10/03 10:52:20 INFO mapred.LocalJobRunner:  23/10/03 10:52:20 INFO mapred.MapTask: Starting flush of map output  23/10/03 10:52:20 INFO mapred.MapTask: Spilling map output  23/10/03 10:52:20 INFO mapred.MapTask: bufstart = 0; bufend = 688949; bufvoid = 104857600  23/10/03 10:52:20 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26077696(104310784); length = 136701/6553600  23/10/03 10:52:20 INFO zlib.ZlibFactory: Successfully loaded & initialized native-zlib library  23/10/03 10:52:20 INFO compress.CodecPool: Got brand-new compressor [.deflate]  23/10/03 10:52:20 INFO mapred.MapTask: Finished spill 0  23/10/03 10:52:20 INFO mapred.Task: Task:attempt\_local1389664555\_0001\_m\_000000\_0 is done. And is in the process of committing  23/10/03 10:52:20 INFO mapred.LocalJobRunner: Records R/W=2182/1  23/10/03 10:52:20 INFO mapred.Task: Task 'attempt\_local1389664555\_0001\_m\_000000\_0' done.  23/10/03 10:52:20 INFO mapred.LocalJobRunner: Finishing task: attempt\_local1389664555\_0001\_m\_000000\_0  23/10/03 10:52:20 INFO mapred.LocalJobRunner: Starting task: attempt\_local1389664555\_0001\_m\_000001\_0  23/10/03 10:52:20 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1  23/10/03 10:52:20 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup \_temporary folders under output directory:false, ignore cleanup failures: false  23/10/03 10:52:20 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]  23/10/03 10:52:20 INFO mapred.MapTask: Processing split: hdfs://namenode:9000/input/callwild:0+175998  23/10/03 10:52:20 INFO mapred.MapTask: numReduceTasks: 1  23/10/03 10:52:20 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)  23/10/03 10:52:20 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100  23/10/03 10:52:20 INFO mapred.MapTask: soft limit at 83886080  23/10/03 10:52:20 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600  23/10/03 10:52:20 INFO mapred.MapTask: kvstart = 26214396; length = 6553600  23/10/03 10:52:20 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer  23/10/03 10:52:20 INFO streaming.PipeMapRed: PipeMapRed exec [/home/testdata/tf-idf-mapreduce-master/./1\_wordcount\_mapper.py]  23/10/03 10:52:20 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]  23/10/03 10:52:20 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]  23/10/03 10:52:20 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]  23/10/03 10:52:20 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]  23/10/03 10:52:20 INFO mapreduce.Job: Job job\_local1389664555\_0001 running in uber mode : false  23/10/03 10:52:20 INFO mapreduce.Job: map 100% reduce 0%  23/10/03 10:52:20 INFO streaming.PipeMapRed: Records R/W=2070/1  23/10/03 10:52:20 INFO streaming.PipeMapRed: MRErrorThread done  23/10/03 10:52:20 INFO streaming.PipeMapRed: mapRedFinished  23/10/03 10:52:20 INFO mapred.LocalJobRunner:  23/10/03 10:52:20 INFO mapred.MapTask: Starting flush of map output  23/10/03 10:52:20 INFO mapred.MapTask: Spilling map output  23/10/03 10:52:20 INFO mapred.MapTask: bufstart = 0; bufend = 252674; bufvoid = 104857600  23/10/03 10:52:20 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26164248(104656992); length = 50149/6553600  23/10/03 10:52:20 INFO mapred.MapTask: Finished spill 0  23/10/03 10:52:20 INFO mapred.Task: Task:attempt\_local1389664555\_0001\_m\_000001\_0 is done. And is in the process of committing  23/10/03 10:52:20 INFO mapred.LocalJobRunner: Records R/W=2070/1  23/10/03 10:52:20 INFO mapred.Task: Task 'attempt\_local1389664555\_0001\_m\_000001\_0' done.  23/10/03 10:52:20 INFO mapred.LocalJobRunner: Finishing task: attempt\_local1389664555\_0001\_m\_000001\_0  23/10/03 10:52:20 INFO mapred.LocalJobRunner: map task executor complete.  23/10/03 10:52:20 INFO mapred.LocalJobRunner: Waiting for reduce tasks  23/10/03 10:52:20 INFO mapred.LocalJobRunner: Starting task: attempt\_local1389664555\_0001\_r\_000000\_0  23/10/03 10:52:20 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1  23/10/03 10:52:20 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup \_temporary folders under output directory:false, ignore cleanup failures: false  23/10/03 10:52:20 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]  23/10/03 10:52:20 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@7b6b6f31  23/10/03 10:52:20 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10  23/10/03 10:52:20 INFO reduce.EventFetcher: attempt\_local1389664555\_0001\_r\_000000\_0 Thread started: EventFetcher for fetching Map Completion Events  23/10/03 10:52:20 INFO compress.CodecPool: Got brand-new decompressor [.deflate]  23/10/03 10:52:20 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt\_local1389664555\_0001\_m\_000000\_0 decomp: 757303 len: 38038 to MEMORY  23/10/03 10:52:20 INFO reduce.InMemoryMapOutput: Read 757303 bytes from map-output for attempt\_local1389664555\_0001\_m\_000000\_0  23/10/03 10:52:20 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 757303, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->757303  23/10/03 10:52:20 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt\_local1389664555\_0001\_m\_000001\_0 decomp: 277752 len: 26128 to MEMORY  23/10/03 10:52:20 INFO reduce.InMemoryMapOutput: Read 277752 bytes from map-output for attempt\_local1389664555\_0001\_m\_000001\_0  23/10/03 10:52:20 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 277752, inMemoryMapOutputs.size() -> 2, commitMemory -> 757303, usedMemory ->1035055  23/10/03 10:52:20 WARN io.ReadaheadPool: Failed readahead on ifile  EBADF: Bad file descriptor  at org.apache.hadoop.io.nativeio.NativeIO$POSIX.posix\_fadvise(Native Method)  at org.apache.hadoop.io.nativeio.NativeIO$POSIX.posixFadviseIfPossible(NativeIO.java:267)  at org.apache.hadoop.io.nativeio.NativeIO$POSIX$CacheManipulator.posixFadviseIfPossible(NativeIO.java:146)  at org.apache.hadoop.io.ReadaheadPool$ReadaheadRequestImpl.run(ReadaheadPool.java:208)  at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)  at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)  at java.lang.Thread.run(Thread.java:745)  23/10/03 10:52:20 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning  23/10/03 10:52:20 INFO mapred.LocalJobRunner: 2 / 2 copied.  23/10/03 10:52:20 INFO reduce.MergeManagerImpl: finalMerge called with 2 in-memory map-outputs and 0 on-disk map-outputs  23/10/03 10:52:20 INFO mapred.Merger: Merging 2 sorted segments  23/10/03 10:52:20 INFO mapred.Merger: Down to the last merge-pass, with 2 segments left of total size: 1035012 bytes  23/10/03 10:52:21 INFO reduce.MergeManagerImpl: Merged 2 segments, 1035055 bytes to disk to satisfy reduce memory limit  23/10/03 10:52:21 INFO reduce.MergeManagerImpl: Merging 1 files, 54158 bytes from disk  23/10/03 10:52:21 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce  23/10/03 10:52:21 INFO mapred.Merger: Merging 1 sorted segments  23/10/03 10:52:21 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1035034 bytes  23/10/03 10:52:21 INFO mapred.LocalJobRunner: 2 / 2 copied.  23/10/03 10:52:21 INFO streaming.PipeMapRed: PipeMapRed exec [/home/testdata/tf-idf-mapreduce-master/./1\_wordcount\_reducer.py]  23/10/03 10:52:21 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  23/10/03 10:52:21 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps  23/10/03 10:52:21 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]  23/10/03 10:52:21 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]  23/10/03 10:52:21 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]  23/10/03 10:52:21 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]  23/10/03 10:52:21 INFO streaming.PipeMapRed: Records R/W=6508/1  23/10/03 10:52:21 INFO streaming.PipeMapRed: R/W/S=10000/660/0 in:NA [rec/s] out:NA [rec/s]  23/10/03 10:52:21 INFO streaming.PipeMapRed: MRErrorThread done  23/10/03 10:52:21 INFO streaming.PipeMapRed: mapRedFinished  23/10/03 10:52:21 INFO mapred.Task: Task:attempt\_local1389664555\_0001\_r\_000000\_0 is done. And is in the process of committing  23/10/03 10:52:21 INFO mapred.LocalJobRunner: 2 / 2 copied.  23/10/03 10:52:21 INFO mapred.Task: Task attempt\_local1389664555\_0001\_r\_000000\_0 is allowed to commit now  23/10/03 10:52:21 INFO output.FileOutputCommitter: Saved output of task 'attempt\_local1389664555\_0001\_r\_000000\_0' to hdfs://namenode:9000/results/step\_1/\_temporary/0/task\_local1389664555\_0001\_r\_000000  23/10/03 10:52:21 INFO mapred.LocalJobRunner: Records R/W=6508/1 > reduce  23/10/03 10:52:21 INFO mapred.Task: Task 'attempt\_local1389664555\_0001\_r\_000000\_0' done.  23/10/03 10:52:21 INFO mapred.LocalJobRunner: Finishing task: attempt\_local1389664555\_0001\_r\_000000\_0  23/10/03 10:52:21 INFO mapred.LocalJobRunner: reduce task executor complete.  23/10/03 10:52:21 INFO mapreduce.Job: map 100% reduce 100%  23/10/03 10:52:21 INFO mapreduce.Job: Job job\_local1389664555\_0001 completed successfully  23/10/03 10:52:21 INFO mapreduce.Job: Counters: 35  File System Counters  FILE: Number of bytes read=521339  FILE: Number of bytes written=1615733  FILE: Number of read operations=0  FILE: Number of large read operations=0  FILE: Number of write operations=0  HDFS: Number of bytes read=2225211  HDFS: Number of bytes written=159124  HDFS: Number of read operations=25  HDFS: Number of large read operations=0  HDFS: Number of write operations=5  Map-Reduce Framework  Map input records=13490  Map output records=46714  Map output bytes=941623  Map output materialized bytes=64166  Input split bytes=188  Combine input records=0  Combine output records=0  Reduce input groups=8172  Reduce shuffle bytes=64166  Reduce input records=46714  Reduce output records=8172  Spilled Records=93428  Shuffled Maps =2  Failed Shuffles=0  Merged Map outputs=2  GC time elapsed (ms)=174  Total committed heap usage (bytes)=1071644672  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=800403  File Output Format Counters  Bytes Written=159124  23/10/03 10:52:21 INFO streaming.StreamJob: Output directory: /results/step\_1 |

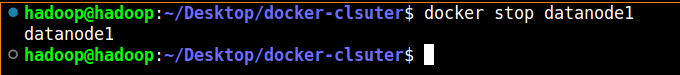
* Kết quả

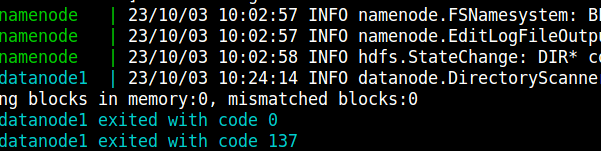




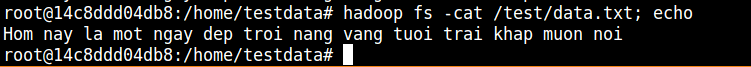


* 1. **Dừng 1 hoặc triển khai thêm 1 node trong cluster hadoop**
* Dừng 1 node

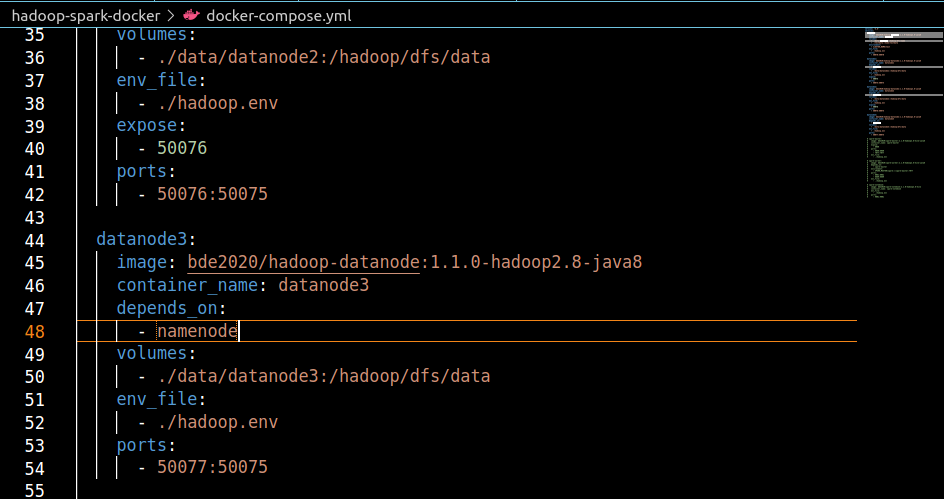


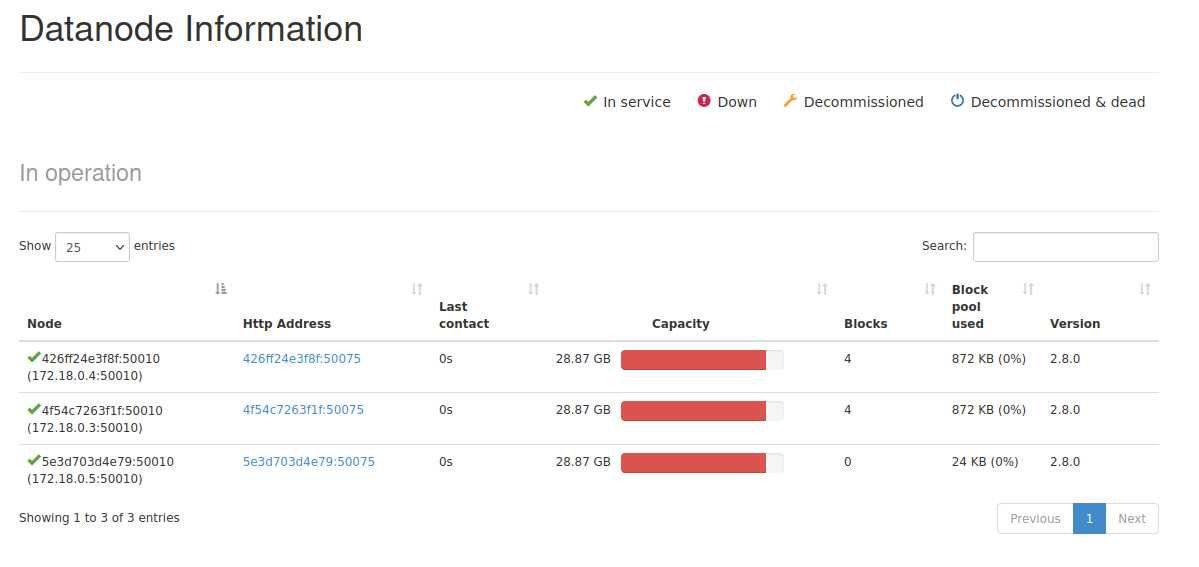


Vẫn đọc được file



* Triển khai thêm 1 node





1. **Mongodb Cluster**
   1. **Triển khai mongodb cluster**

Khởi tạo môi trường mongodb

|  |
| --- |
| version: '3.8'  services:  mongo1:  container\_name: mongo1  image: mongo:4.4  volumes:  - ./scripts/rs-init.sh:/scripts/rs-init.sh  - ./scripts/init.js:/scripts/init.js  - ./data/mongo1:/data/db  networks:  - mongo-network  expose:  - 27017  ports:  - 27017:27017  depends\_on:  - mongo2  - mongo3  links:  - mongo2  - mongo3  restart: always  entrypoint: [ "/usr/bin/mongod", "--bind\_ip\_all", "--replSet", "dbrs" ]  mongo2:  container\_name: mongo2  image: mongo:4.4  volumes:  - ./data/mongo2:/data/db  networks:  - mongo-network  ports:  - 27018:27017  restart: always  entrypoint: [ "/usr/bin/mongod", "--bind\_ip\_all", "--replSet", "dbrs" ]  mongo3:  container\_name: mongo3  image: mongo:4.4  volumes:  - ./data/mongo3:/data/db  networks:  - mongo-network  expose:  - 27019  ports:  - 27019:27017  restart: always  entrypoint: [ "/usr/bin/mongod", "--bind\_ip\_all", "--replSet", "dbrs" ]  networks:  mongo-network:  driver: bridge |

Truy cập vào container mongo1 và nạp config để khởi tạo sharding

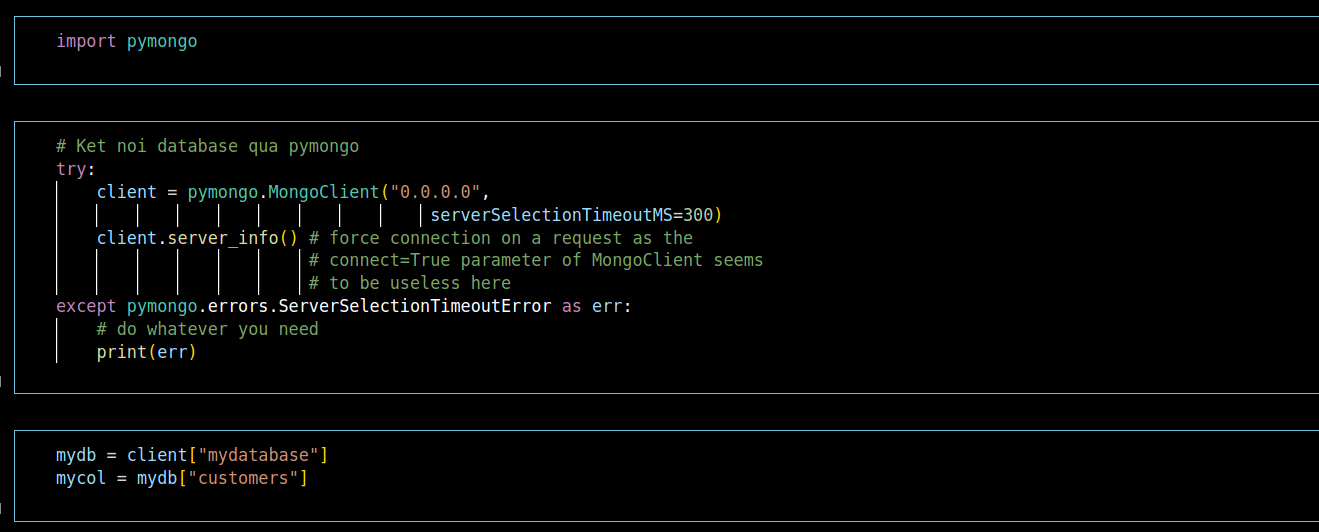
|  |
| --- |
| config = {  "\_id": "dbrs",  "version": 1,  "members": [  {  "\_id": 1,  "host": "mongo1:27017",  "priority": 2  },  {  "\_id": 2,  "host": "mongo2:27017",  "priority": 1  },  {  "\_id": 3,  "host": "mongo3:27017",  "priority": 1  }  ]  };  rs.initiate(config, { force: true }); |

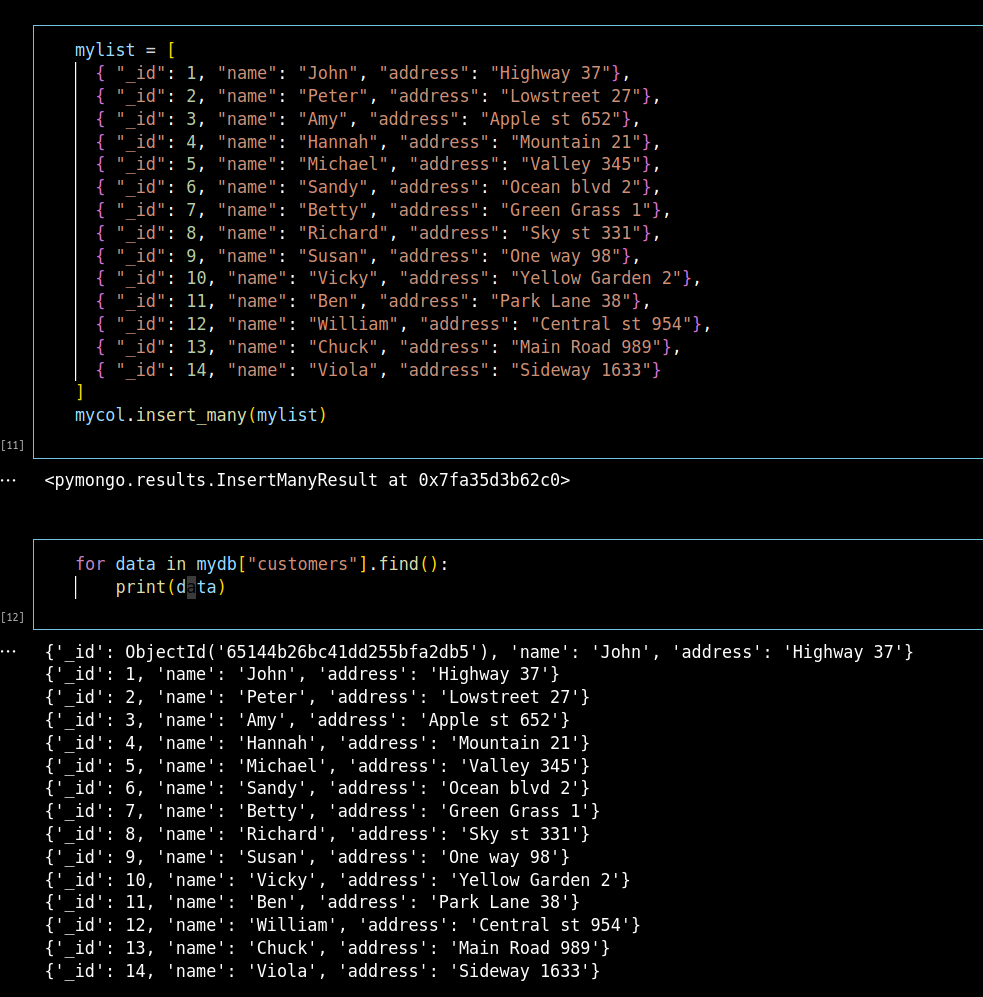
Kiểm tra kết quả status trả về của cluster mongo

|  |
| --- |
| dbrs:PRIMARY> rs.status()  {  "set" : "dbrs",  "date" : ISODate("2023-09-27T14:38:00.652Z"),  "myState" : 1,  "term" : NumberLong(1),  "syncSourceHost" : "",  "syncSourceId" : -1,  "heartbeatIntervalMillis" : NumberLong(2000),  "majorityVoteCount" : 2,  "writeMajorityCount" : 2,  "votingMembersCount" : 3,  "writableVotingMembersCount" : 3,  "optimes" : {  "lastCommittedOpTime" : {  "ts" : Timestamp(1695825474, 3),  "t" : NumberLong(1)  },  "lastCommittedWallTime" : ISODate("2023-09-27T14:37:54.922Z"),  "readConcernMajorityOpTime" : {  "ts" : Timestamp(1695825474, 3),  "t" : NumberLong(1)  },  "readConcernMajorityWallTime" : ISODate("2023-09-27T14:37:54.922Z"),  "appliedOpTime" : {  "ts" : Timestamp(1695825474, 3),  "t" : NumberLong(1)  },  "durableOpTime" : {  "ts" : Timestamp(1695825474, 3),  "t" : NumberLong(1)  },  "lastAppliedWallTime" : ISODate("2023-09-27T14:37:54.922Z"),  "lastDurableWallTime" : ISODate("2023-09-27T14:37:54.922Z")  },  "lastStableRecoveryTimestamp" : Timestamp(1695825430, 1),  "electionCandidateMetrics" : {  "lastElectionReason" : "electionTimeout",  "lastElectionDate" : ISODate("2023-09-27T14:34:10.706Z"),  "electionTerm" : NumberLong(1),  "lastCommittedOpTimeAtElection" : {  "ts" : Timestamp(0, 0),  "t" : NumberLong(-1)  },  "lastSeenOpTimeAtElection" : {  "ts" : Timestamp(1695825239, 1),  "t" : NumberLong(-1)  },  "numVotesNeeded" : 2,  "priorityAtElection" : 2,  "electionTimeoutMillis" : NumberLong(10000),  "numCatchUpOps" : NumberLong(0),  "newTermStartDate" : ISODate("2023-09-27T14:34:10.748Z"),  "wMajorityWriteAvailabilityDate" : ISODate("2023-09-27T14:34:11.436Z")  },  "members" : [  {  "\_id" : 1,  "name" : "mongo1:27017",  "health" : 1,  "state" : 1,  "stateStr" : "PRIMARY",  "uptime" : 606,  "optime" : {  "ts" : Timestamp(1695825474, 3),  "t" : NumberLong(1)  },  "optimeDate" : ISODate("2023-09-27T14:37:54Z"),  "lastAppliedWallTime" : ISODate("2023-09-27T14:37:54.922Z"),  "lastDurableWallTime" : ISODate("2023-09-27T14:37:54.922Z"),  "syncSourceHost" : "",  "syncSourceId" : -1,  "infoMessage" : "",  "electionTime" : Timestamp(1695825250, 1),  "electionDate" : ISODate("2023-09-27T14:34:10Z"),  "configVersion" : 1,  "configTerm" : 1,  "self" : true,  "lastHeartbeatMessage" : ""  },  {  "\_id" : 2,  "name" : "mongo2:27017",  "health" : 1,  "state" : 2,  "stateStr" : "SECONDARY",  "uptime" : 240,  "optime" : {  "ts" : Timestamp(1695825474, 3),  "t" : NumberLong(1)  },  "optimeDurable" : {  "ts" : Timestamp(1695825474, 3),  "t" : NumberLong(1)  },  "optimeDate" : ISODate("2023-09-27T14:37:54Z"),  "optimeDurableDate" : ISODate("2023-09-27T14:37:54Z"),  "lastAppliedWallTime" : ISODate("2023-09-27T14:37:54.922Z"),  "lastDurableWallTime" : ISODate("2023-09-27T14:37:54.922Z"),  "lastHeartbeat" : ISODate("2023-09-27T14:37:58.806Z"),  "lastHeartbeatRecv" : ISODate("2023-09-27T14:37:59.796Z"),  "pingMs" : NumberLong(0),  "lastHeartbeatMessage" : "",  "syncSourceHost" : "mongo1:27017",  "syncSourceId" : 1,  "infoMessage" : "",  "configVersion" : 1,  "configTerm" : 1  },  {  "\_id" : 3,  "name" : "mongo3:27017",  "health" : 1,  "state" : 2,  "stateStr" : "SECONDARY",  "uptime" : 240,  "optime" : {  "ts" : Timestamp(1695825474, 3),  "t" : NumberLong(1)  },  "optimeDurable" : {  "ts" : Timestamp(1695825474, 3),  "t" : NumberLong(1)  },  "optimeDate" : ISODate("2023-09-27T14:37:54Z"),  "optimeDurableDate" : ISODate("2023-09-27T14:37:54Z"),  "lastAppliedWallTime" : ISODate("2023-09-27T14:37:54.922Z"),  "lastDurableWallTime" : ISODate("2023-09-27T14:37:54.922Z"),  "lastHeartbeat" : ISODate("2023-09-27T14:37:58.806Z"),  "lastHeartbeatRecv" : ISODate("2023-09-27T14:37:59.796Z"),  "pingMs" : NumberLong(0),  "lastHeartbeatMessage" : "",  "syncSourceHost" : "mongo1:27017",  "syncSourceId" : 1,  "infoMessage" : "",  "configVersion" : 1,  "configTerm" : 1  }  ],  "ok" : 1,  "$clusterTime" : {  "clusterTime" : Timestamp(1695825474, 3),  "signature" : {  "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAA="),  "keyId" : NumberLong(0)  }  },  "operationTime" : Timestamp(1695825474, 3)  } |

Ta kiểm tra thấy kết quả 3 node đã online

Nạp dữ liệu nháp vào mongo cluster





* 1. **Tắt một node**
* Tắt một node có sẵn trên mongo





Vẫn đọc được kết quả từ trước đó

