Methodology

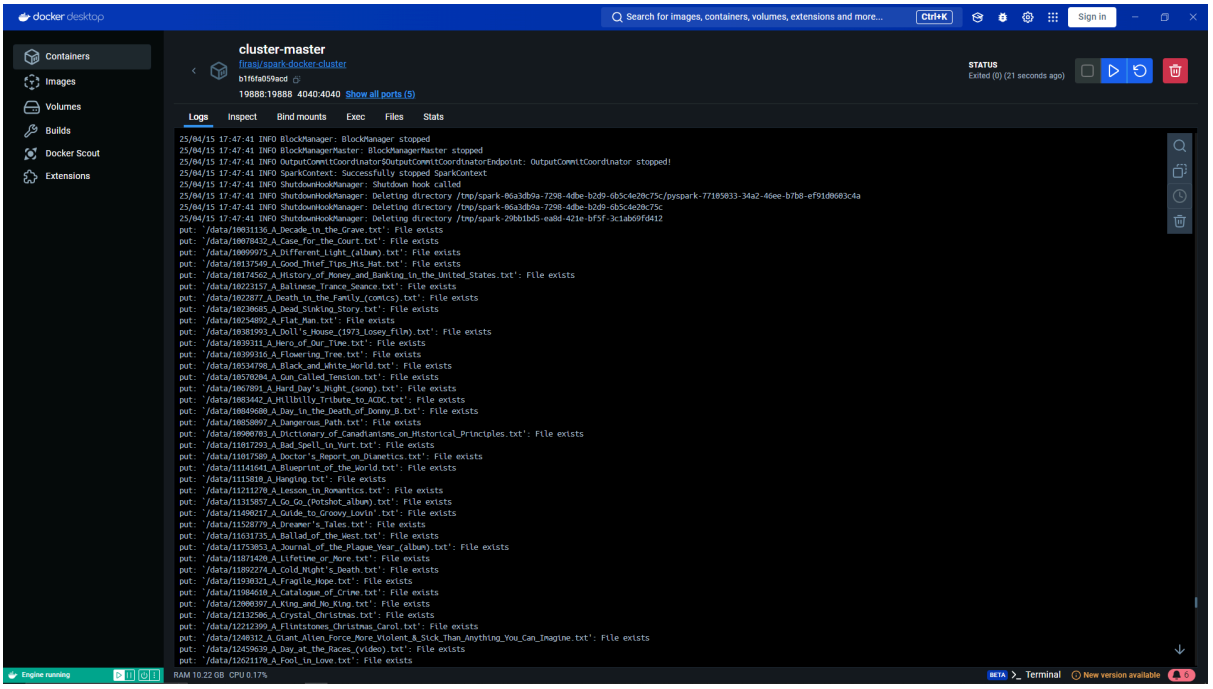
I used your initial scripts prepare_data.py and prepare_data.sh and just upload a.parquet file to the folder. Everything was fine with that code except I added executor-memory 4G and driver-memory 2G parameters because without them scripts were unable to complete due to the lack of memory
in the start_servisec.sh I changed hdfs dfsadmin -safemode leave to hdfs dfsadmin -safemode forceExit, because without it I was unable to put data inside a hdfs
In app.sh I initialize cassandra keyspace and tables inside that to store the data and call other scripts (prepare_data, index.sh, search.sh)
index.sh runs mapreduce scripts and that's it. Unfortunately, my code did not succeed to pass the mapper1.py file and the whole system doesn't work, I just obtain empty output for search queries.
In the query.py I just read data from cassandra, calculate bm25 and produce top documents (again, just empty output)

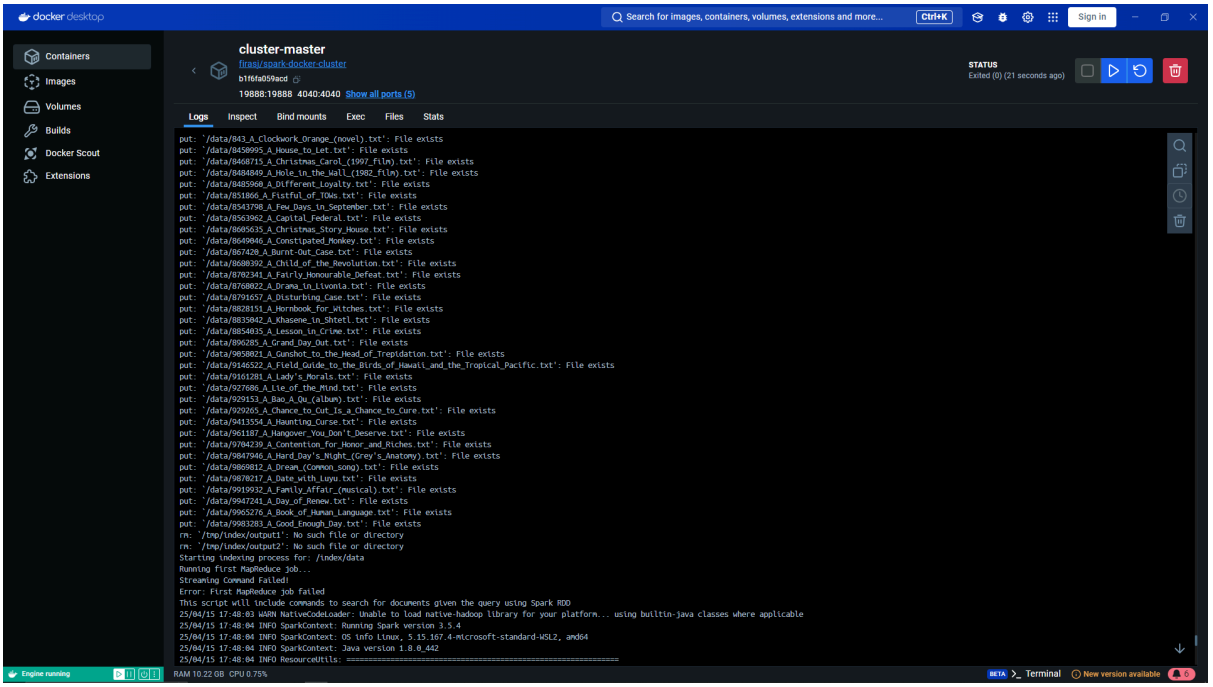
mapper and reducer were done using reference from lab5 and the whole system fails on mapper1.py for some reason. When I run it locally, not inside the docker, it works and processes input data, but in the container nothing works. I obtain mapreduce exception (I'll put screenshots)

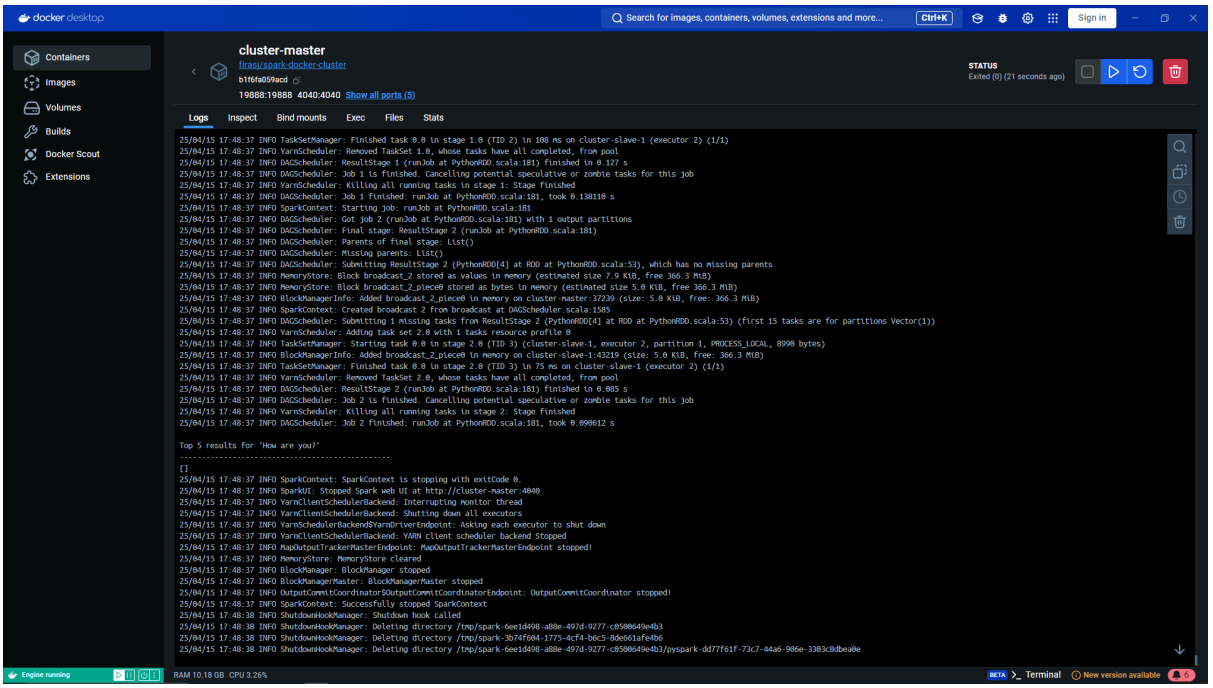
Also, I modify docker-compose file and provide port 9042 to connect to it

Demonstration

to run the system just run docker compose up –build




docker desktop
Search for images, containers, volumes, extensions and more...
Ctrl+K
Sign in
Containers
Images
Volumes
Builds
Docker Scout
Extensions
cluster-master
firasj/spark-docker-cluster
b1f6fa059acd
19888:19888 4040:4040 Show all ports (5)
STATUS
Exited (0) (21 seconds ago)
Logs
Inspect
Bind mounts
Exec
Files
Stats
put: `/data/843_A_Clockwork_Orange_(novel).txt': File exists
put: `/data/8458995_A_House_to_Let.txt': File exists
put: `/data/8468715_A_Christmas_Carol_(1997_film).txt': File exists
put: `/data/8484849_A_Hole_in_the_Wall_(1982_film).txt': File exists
put: `/data/8485969_A_Different_Loyalty.txt': File exists
put: `/data/851866_A_Fistful_of_TDMs.txt': File exists
put: `/data/8543798_A_Few_Days_in_September.txt': File exists
put: `/data/8563962_A_Capital_Federal.txt': File exists
put: `/data/8605635_A_Christmas_Story_House.txt': File exists
put: `/data/8649046_A_Constipated_Monkey.txt': File exists
put: `/data/867420_A_Burnt-Out_Case.txt': File exists
put: `/data/8688392_A_Child_of_the_Revolution.txt': File exists
put: `/data/8702341_A_Fairly_Honourable_Defeat.txt': File exists
put: `/data/8768022_A_Drama_in_Livonia.txt': File exists
put: `/data/8791657_A_Disturbing_Case.txt': File exists
put: `/data/8828151_A_Hornbook_for_Witches.txt': File exists
put: `/data/8835042_A_Khasene_in_Shtetl.txt': File exists
put: `/data/8854035_A_Lesson_in_Crime.txt': File exists
put: `/data/896285_A_Grand_Day_Out.txt': File exists
put: `/data/9050821_A_Gunshot_to_the_Head_of_Trepidation.txt': File exists
put: `/data/9146522_A_Field_Guide_to_the_Birds_of_Hawaii_and_the_Tropical_Pacific.txt': File exists
put: `/data/9161281_A_Lady's_Morals.txt': File exists
put: `/data/927686_A_Lie_of_the_Mind.txt': File exists
put: `/data/929153_A_Bao_A_Qu_(album).txt': File exists
put: `/data/929265_A_Chance_to_Cut_Is_a_Chance_to_Cure.txt': File exists
put: `/data/9413594_A_Haunting_Curse.txt': File exists
put: `/data/961187_A_Hangover_You_Don't_Deserve.txt': File exists
put: `/data/9704239_A_Contention_for_Honor_and_Riches.txt': File exists
put: `/data/9847946_A_Hard_Day's_Night_(Grey's_Anatomy).txt': File exists
put: `/data/9869812_A_Dream_(Common_song).txt': File exists
put: `/data/9878217_A_Date_with_Luyu.txt': File exists
put: `/data/9919932_A_Family_Affair_(musical).txt': File exists
put: `/data/9947241_A_Day_of_Renew.txt': File exists
put: `/data/9965276_A_Book_of_Human_Language.txt': File exists
put: `/data/9983283_A_Good_Enough_Day.txt': File exists
rm: `/tmp/index/output1': No such file or directory
rm: `/tmp/index/output2': No such file or directory
Starting indexing process for: /index/data
Running first MapReduce job...
Streaming Command Failed!
Error: First MapReduce job failed
This script will include commands to search for documents given the query using Spark RDD
25/04/15 17:48:03 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
25/04/15 17:48:04 INFO SparkContext: Running Spark version 3.5.4
25/04/15 17:48:04 INFO SparkContext: OS info Linux, 5.15.167.4-microsoft-standard-WSL2, amd64
25/04/15 17:48:04 INFO SparkContext: Java version 1.8.0_442
25/04/15 17:48:04 INFO ResourceUtils: ==
Engine running
RAM 10.22 GB CPU 0.75%
BETA Terminal
New version available
6


docker desktop
Search for images, containers, volumes, extensions and more...
Ctrl+K
Sign in
Containers
Images
Volumes
Builds
Docker Scout
Extensions
cluster-master
firasj/spark-docker-cluster
b1f6fa059acd
19888:19888 4040:4040 Show all ports (5)
STATUS
Exited (0) (21 seconds ago)
Logs
Inspect
Bind mounts
Exec
Files
Stats
25/04/15 17:48:37 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2) in 108 ms on cluster-slave-1 (executor 2) (1/1)
25/04/15 17:48:37 INFO YarnScheduler: Removed TaskSet 1.0, whose tasks have all completed, from pool
25/04/15 17:48:37 INFO DAGScheduler: ResultStage 1 (runJob at PythonRDD.scala:181) finished in 0.127 s
25/04/15 17:48:37 INFO DAGScheduler: Job 1 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/15 17:48:37 INFO YarnScheduler: Killing all running tasks in stage 1: Stage finished
25/04/15 17:48:37 INFO DAGScheduler: Job 1 finished: runJob at PythonRDD.scala:181, took 0.138110 s
25/04/15 17:48:37 INFO SparkContext: Starting job: runJob at PythonRDD.scala:181
25/04/15 17:48:37 INFO DAGScheduler: Got job 2 (runJob at PythonRDD.scala:181) with 1 output partitions
25/04/15 17:48:37 INFO DAGScheduler: Final stage: ResultStage 2 (runJob at PythonRDD.scala:181)
25/04/15 17:48:37 INFO DAGScheduler: Parents of final stage: List()
25/04/15 17:48:37 INFO DAGScheduler: Missing parents: List()
25/04/15 17:48:37 INFO DAGScheduler: Submitting ResultStage 2 (PythonRDD[4] at RDD at PythonRDD.scala:53), which has no missing parents
25/04/15 17:48:37 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 7.9 KiB, free 366.3 MiB)
25/04/15 17:48:37 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 5.0 KiB, free 366.3 MiB)
25/04/15 17:48:37 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on cluster-master:37239 (size: 5.0 KiB, free: 366.3 MiB)
25/04/15 17:48:37 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1585
25/04/15 17:48:37 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 2 (PythonRDD[4] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions Vector(1))
25/04/15 17:48:37 INFO YarnScheduler: Adding task set 2.0 with 1 tasks resource profile 0
25/04/15 17:48:37 INFO TaskSetManager: Starting task 0.0 in stage 2.0 (TID 3) (cluster-slave-1, executor 2, partition 1, PROCESS_LOCAL, 8990 bytes)
25/04/15 17:48:37 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on cluster-slave-1:43219 (size: 5.0 KiB, free: 366.3 MiB)
25/04/15 17:48:37 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 3) in 75 ms on cluster-slave-1 (executor 2) (1/1)
25/04/15 17:48:37 INFO YarnScheduler: Removed TaskSet 2.0, whose tasks have all completed, from pool
25/04/15 17:48:37 INFO DAGScheduler: ResultStage 2 (runJob at PythonRDD.scala:181) finished in 0.085 s
25/04/15 17:48:37 INFO DAGScheduler: Job 2 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/15 17:48:37 INFO YarnScheduler: Killing all running tasks in stage 2: Stage finished
25/04/15 17:48:37 INFO DAGScheduler: Job 2 finished: runJob at PythonRDD.scala:181, took 0.090612 s
Top 5 results for 'How are you?'
[]
25/04/15 17:48:37 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/15 17:48:37 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/15 17:48:37 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/15 17:48:37 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/15 17:48:37 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
25/04/15 17:48:37 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
25/04/15 17:48:37 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/15 17:48:37 INFO MemoryStore: MemoryStore cleared
25/04/15 17:48:37 INFO BlockManager: BlockManager stopped
25/04/15 17:48:37 INFO BlockManagerMaster: BlockManagerMaster stopped
25/04/15 17:48:37 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/15 17:48:37 INFO SparkContext: Successfully stopped SparkContext
25/04/15 17:48:38 INFO ShutdownHookManager: Shutdown hook called
25/04/15 17:48:38 INFO ShutdownHookManager: Deleting directory /tmp/spark-6ee1d498-a88e-497d-9277-c0500649e4b3
25/04/15 17:48:38 INFO ShutdownHookManager: Deleting directory /tmp/spark-3b74f604-1775-4cf4-b6c5-8de661afe4b6
25/04/15 17:48:38 INFO ShutdownHookManager: Deleting directory /tmp/spark-6ee1d498-a88e-497d-9277-c0500649e4b3/pyspark-dd77f61f-73c7-44a6-906e-3383c8dbea0e
Engine running
RAM 10.18 GB CPU 3.26%
BETA Terminal
New version available
6