Egor Lebedev AAI-01
Report A2

Methodology

important!!!!
the system has many points of failure during initialization (simply because there are a lot of tools), so if you see a problem, please restart the build by clearing the cache in docker, it always helps. The search itself is stable

app.sh:
Orchestrates workflow: Starts services → Prepares data → Runs indexing → Ranck for ex → Keeps container alive (tail -f /dev/null).

start-services.sh:
not changed

init-cassandra.sh:
Waits for Cassandra availability → Creates tables via cassandra/cassandra-init.cql.

prepare_data.sh:
Downloads a.parquet (if missing) → Transfers data to HDFS → Runs prepare_data.py.

prepare_data.py:
Processes Parquet → Generates .txt files → Writes metadata to HDFS /index/data.

index.sh:
Runs two MapReduce pipelines:
Pipeline 1: Builds inverted index and fill the tables (mapper1.py, reducer1.py).
Pipeline 2: Computes document stats (mapper2.py, reducer2.py).

search.sh:
Executes search via Spark → Passes query to query.py with the necessary dependencies → Outputs top 10 results.

query.py:
Implements BM25 with stability fixes → Ranks documents → Returns IDs & titles.

query.py

this script connects to cassandra and downloads the necessary data for the request, after that it calculates bm_25, ranks and outputs a list of files as requested in the task.

I used a different version of bm_25, it's just protected from division by 0 problems, we went through the formula in the IR course and I checked its performance in that course (and in general, it's just adding constants, which does not affect the ranking)

```
def calculate_bm25(row):
    term = row.term_text
    tf = row.tf
    doc_length = row.length
    df = df_bc.value.get(term, 0)
    idf = math.log((total_docs + 1) / (df + 0.5))
    score = idf * (tf * (k1 + 1)) / (tf + k1 * (1 - b + b * (doc_length / avg_length)))
    return (row.doc_id, score)
```

Demonstration

steps for start

git clone https://github.com/EgorLeb/big_data_a2.git

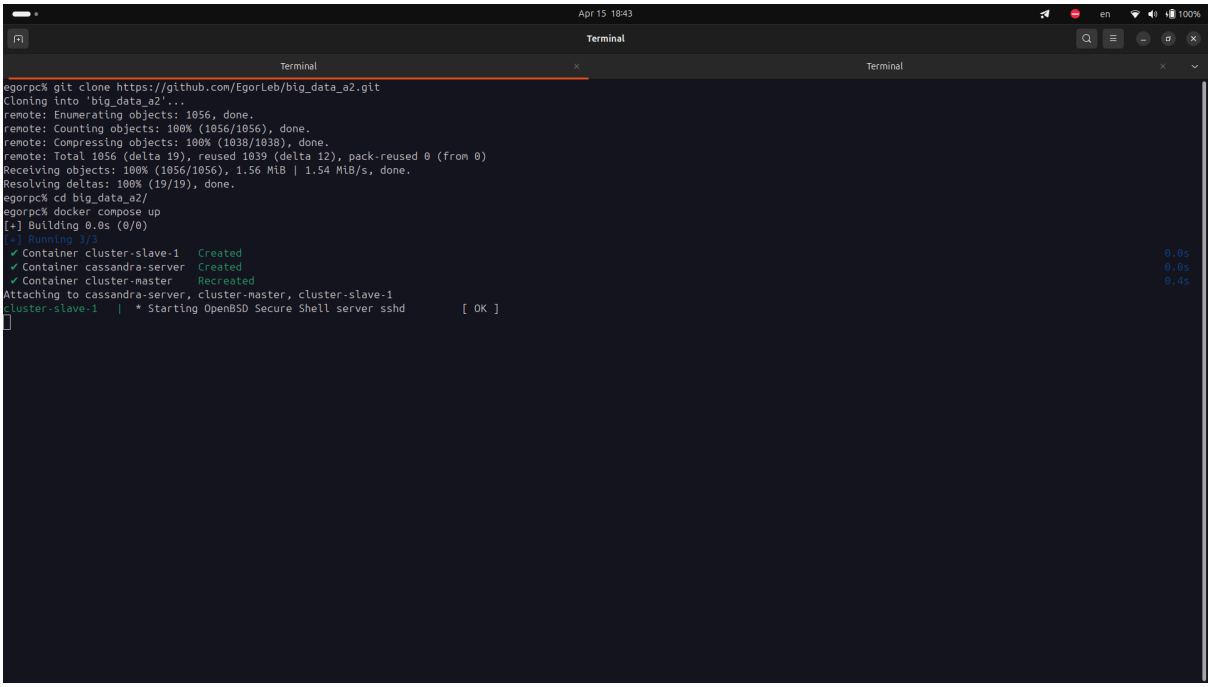
docker compose up

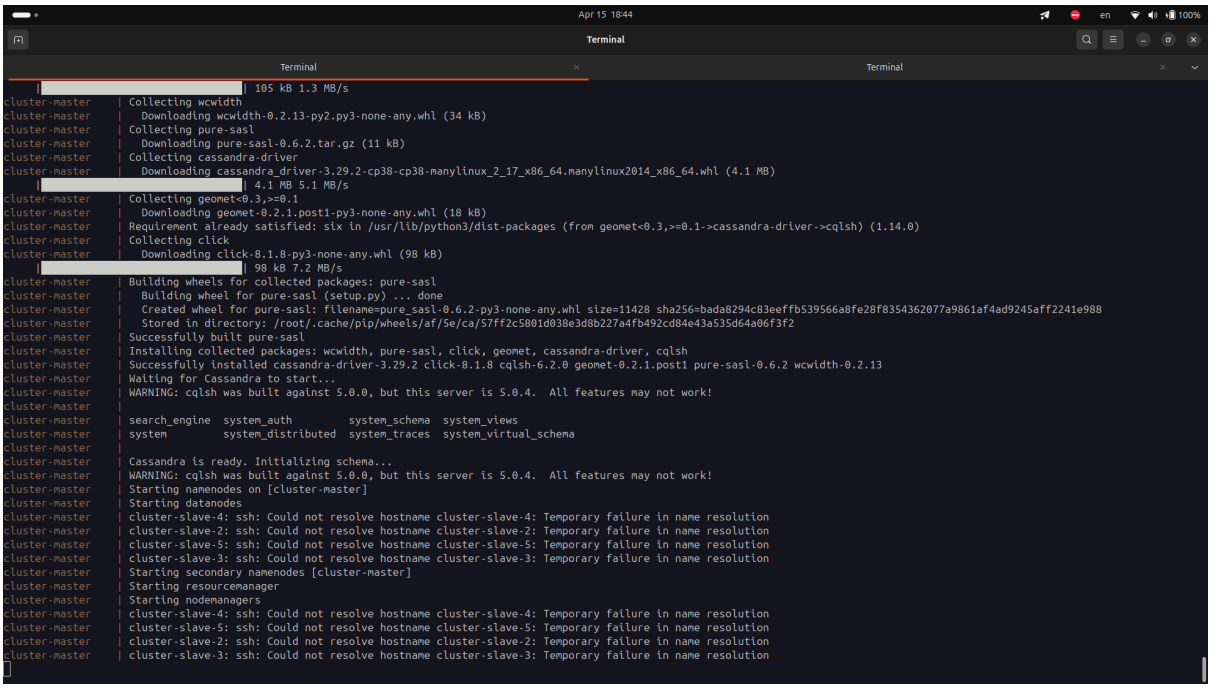
in other console

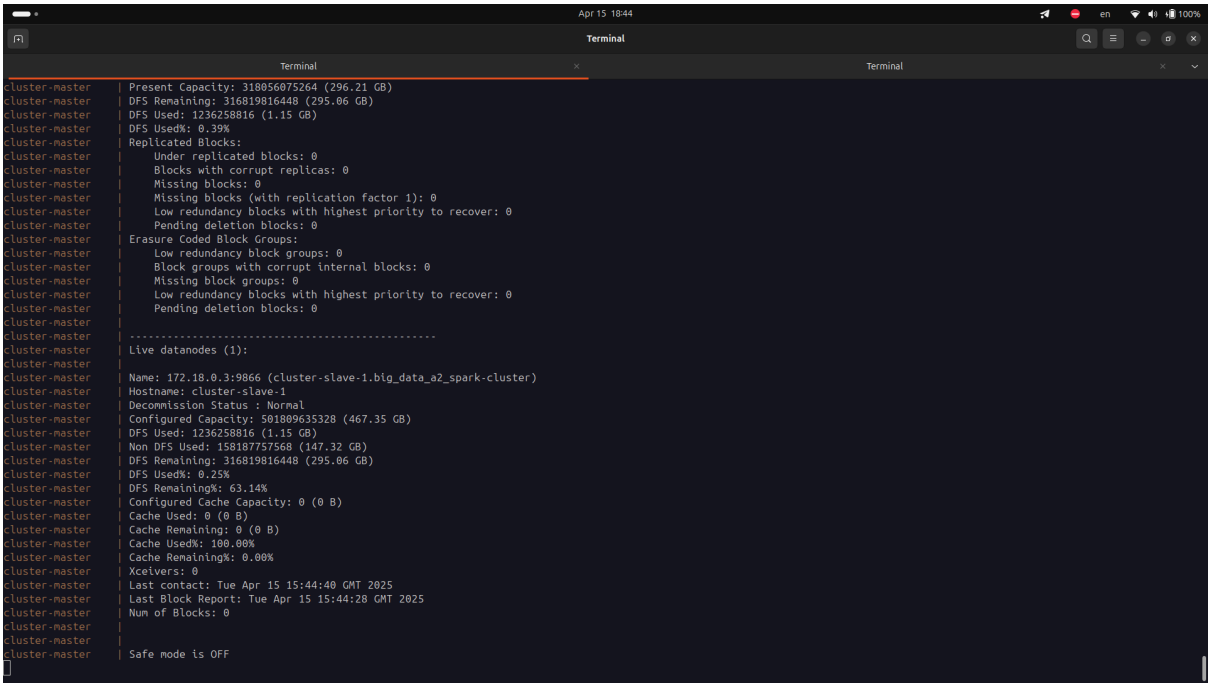
docker exec -it cluster-master /bin/bash
bash search.sh "book about the dog"
bash search.sh "chance of the corner if i will walk"

I looked at the files that came out, they contain the right words, but the meaning doesn't really fit, the problem is, I think, the number of files

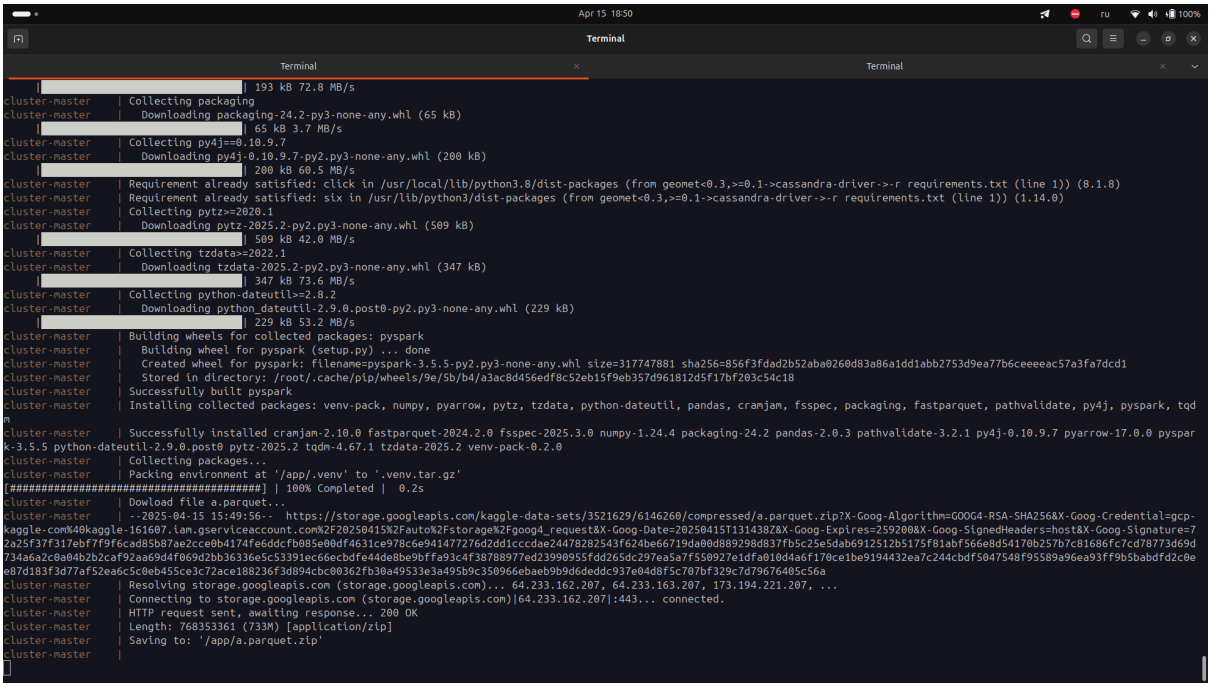
screenshots:


Apr 15 18:43
en
100%
Terminal
Terminal
Terminal
egorpc% git clone https://github.com/EgorLeb/big_data_a2.git
Cloning into 'big_data_a2'...
remote: Enumerating objects: 1056, done.
remote: Counting objects: 100% (1056/1056), done.
remote: Compressing objects: 100% (1038/1038), done.
remote: Total 1056 (delta 19), reused 1039 (delta 12), pack-reused 0 (from 0)
Receiving objects: 100% (1056/1056), 1.56 MiB | 1.54 MiB/s, done.
Resolving deltas: 100% (19/19), done.
egorpc% cd big_data_a2/
egorpc% docker compose up
[+] Building 0.0s (0/0)
[+] Running 3/3
✔ Container cluster-slave-1 Created 0.0s
✔ Container cassandra-server Created 0.0s
✔ Container cluster-master Recreated 0.4s
Attaching to cassandra-server, cluster-master, cluster-slave-1
cluster-slave-1 | * Starting OpenBSD Secure Shell server sshd [OK]

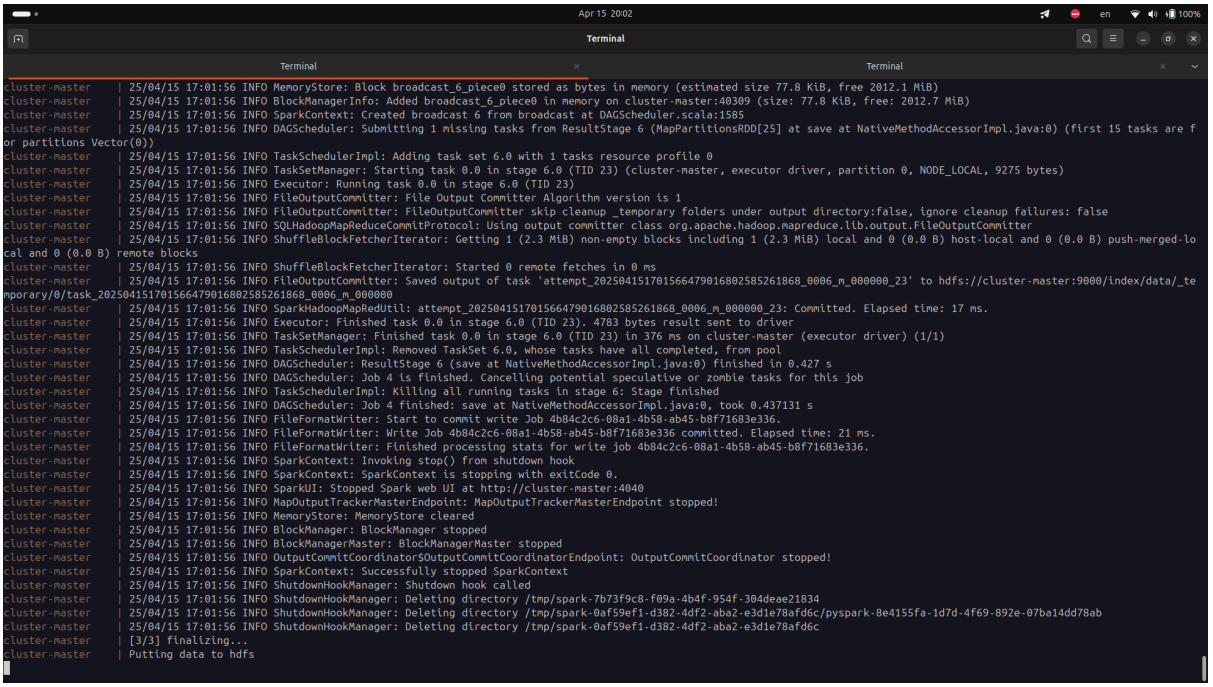

Apr 15 18:44
en
100%
Terminal
Terminal
Terminal
| 105 kB 1.3 MB/s
cluster-master | Collecting wcwidth
cluster-master | Downloading wcwidth-0.2.13-py2.py3-none-any.whl (34 kB)
cluster-master | Collecting pure-sasl
cluster-master | Downloading pure-sasl-0.6.2.tar.gz (11 kB)
cluster-master | Collecting cassandra-driver
cluster-master | Downloading cassandra_driver-3.29.2-cp38-cp38-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (4.1 MB)
| 4.1 MB 5.1 MB/s
cluster-master | Collecting geomet<0.3,>=0.1
cluster-master | Downloading geomet-0.2.1.post1-py3-none-any.whl (18 kB)
cluster-master | Requirement already satisfied: six in /usr/lib/python3/dist-packages (from geomet<0.3,>=0.1->cassandra-driver->cqlsh) (1.14.0)
cluster-master | Collecting click
cluster-master | Downloading click-8.1.8-py3-none-any.whl (98 kB)
| 98 kB 7.2 MB/s
cluster-master | Building wheels for collected packages: pure-sasl
cluster-master | Building wheel for pure-sasl (setup.py) ... done
cluster-master | Created wheel for pure-sasl: filename=pure_sasl-0.6.2-py3-none-any.whl size=11428 sha256=bada8294c83eeffb539566a8fe28f8354362077a9861af4ad9245aff2241e988
cluster-master | Stored in directory: /root/.cache/pip/wheels/af/5e/ca/57ff2c5801d038e3d8b227a4fb492cd84e43a535d64a06f3f2
cluster-master | Successfully built pure-sasl
cluster-master | Installing collected packages: wcwidth, pure-sasl, click, geomet, cassandra-driver, cqlsh
cluster-master | Successfully installed cassandra-driver-3.29.2 click-8.1.8 cqlsh-6.2.0 geomet-0.2.1.post1 pure-sasl-0.6.2 wcwidth-0.2.13
cluster-master | Waiting for Cassandra to start...
cluster-master | WARNING: cqlsh was built against 5.0.0, but this server is 5.0.4. All features may not work!
cluster-master |
cluster-master | search_engine system_auth system_schema system_views
cluster-master | system system_distributed system_traces system_virtual_schema
cluster-master |
cluster-master | Cassandra is ready. Initializing schema...
cluster-master | WARNING: cqlsh was built against 5.0.0, but this server is 5.0.4. All features may not work!
cluster-master | Starting namenodes on [cluster-master]
cluster-master | Starting datanodes
cluster-master | cluster-slave-4: ssh: Could not resolve hostname cluster-slave-4: Temporary failure in name resolution
cluster-master | cluster-slave-2: ssh: Could not resolve hostname cluster-slave-2: Temporary failure in name resolution
cluster-master | cluster-slave-5: ssh: Could not resolve hostname cluster-slave-5: Temporary failure in name resolution
cluster-master | cluster-slave-3: ssh: Could not resolve hostname cluster-slave-3: Temporary failure in name resolution
cluster-master | Starting secondary namenodes [cluster-master]
cluster-master | Starting resourcemanager
cluster-master | Starting nodemanagers
cluster-master | cluster-slave-4: ssh: Could not resolve hostname cluster-slave-4: Temporary failure in name resolution
cluster-master | cluster-slave-5: ssh: Could not resolve hostname cluster-slave-5: Temporary failure in name resolution
cluster-master | cluster-slave-2: ssh: Could not resolve hostname cluster-slave-2: Temporary failure in name resolution
cluster-master | cluster-slave-3: ssh: Could not resolve hostname cluster-slave-3: Temporary failure in name resolution

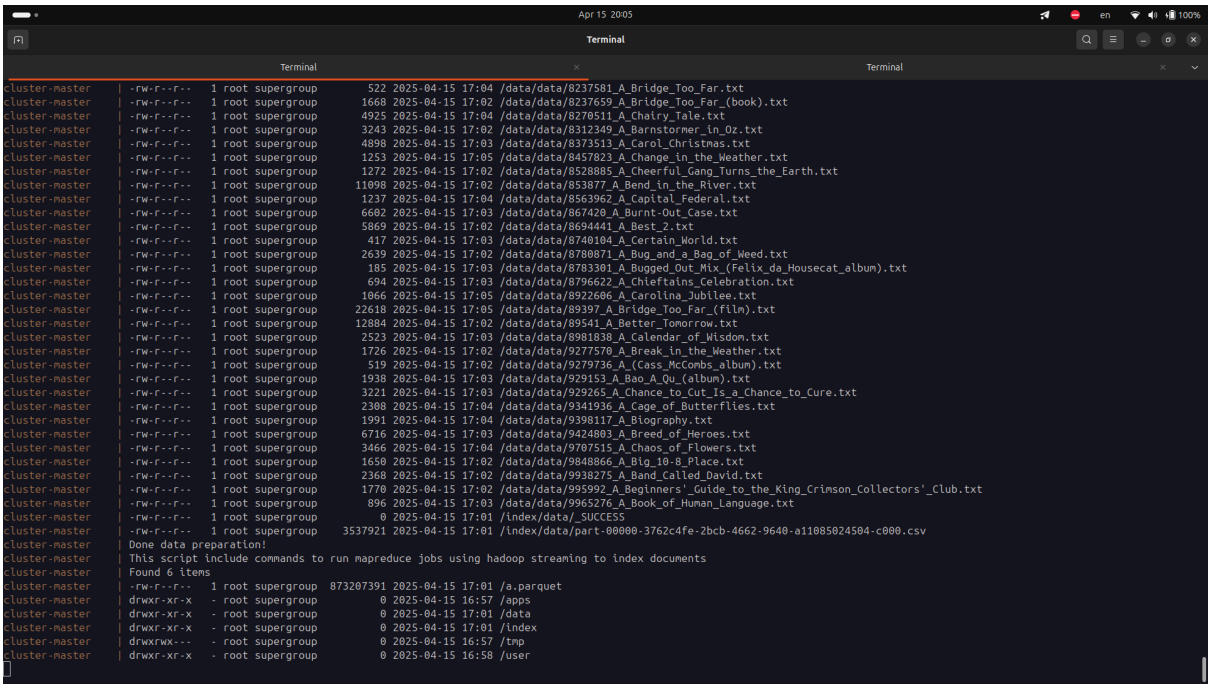


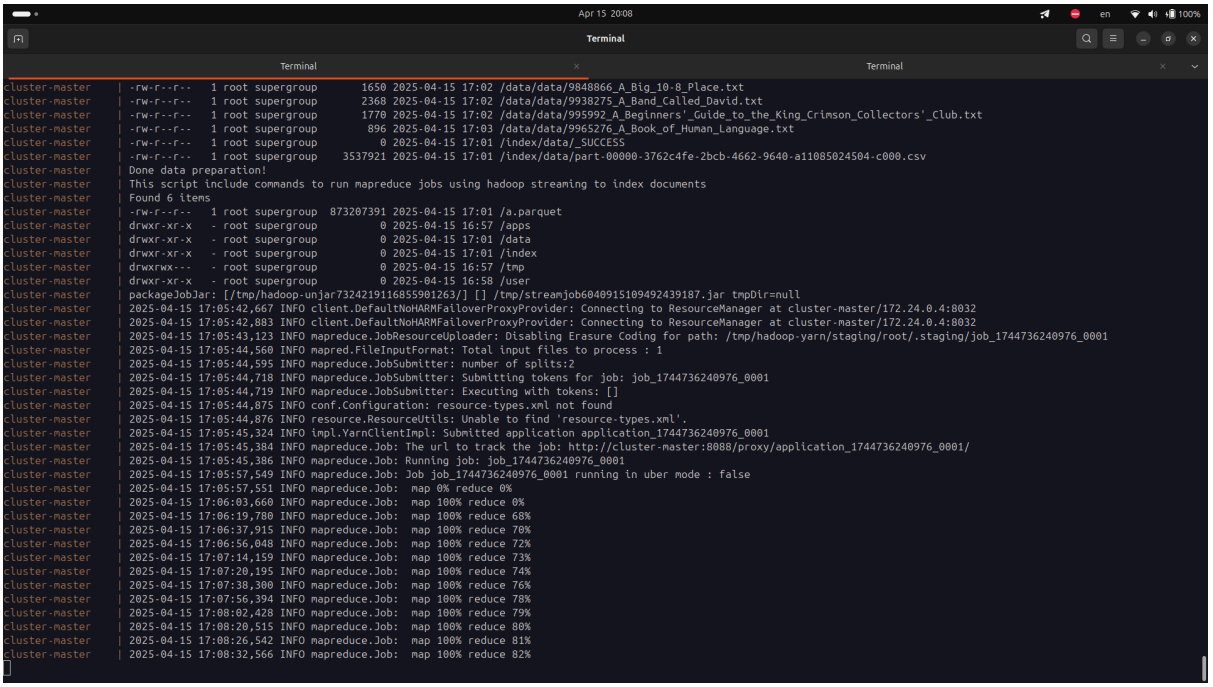
here need to wait

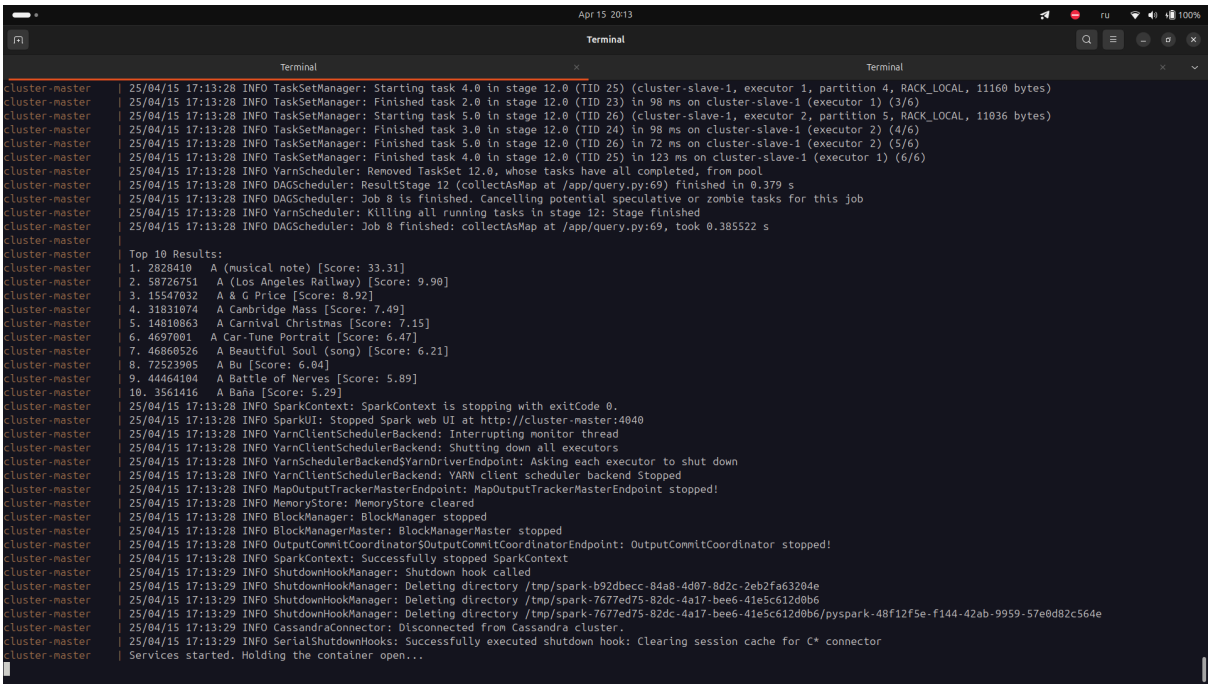


a.parquet installing - wait

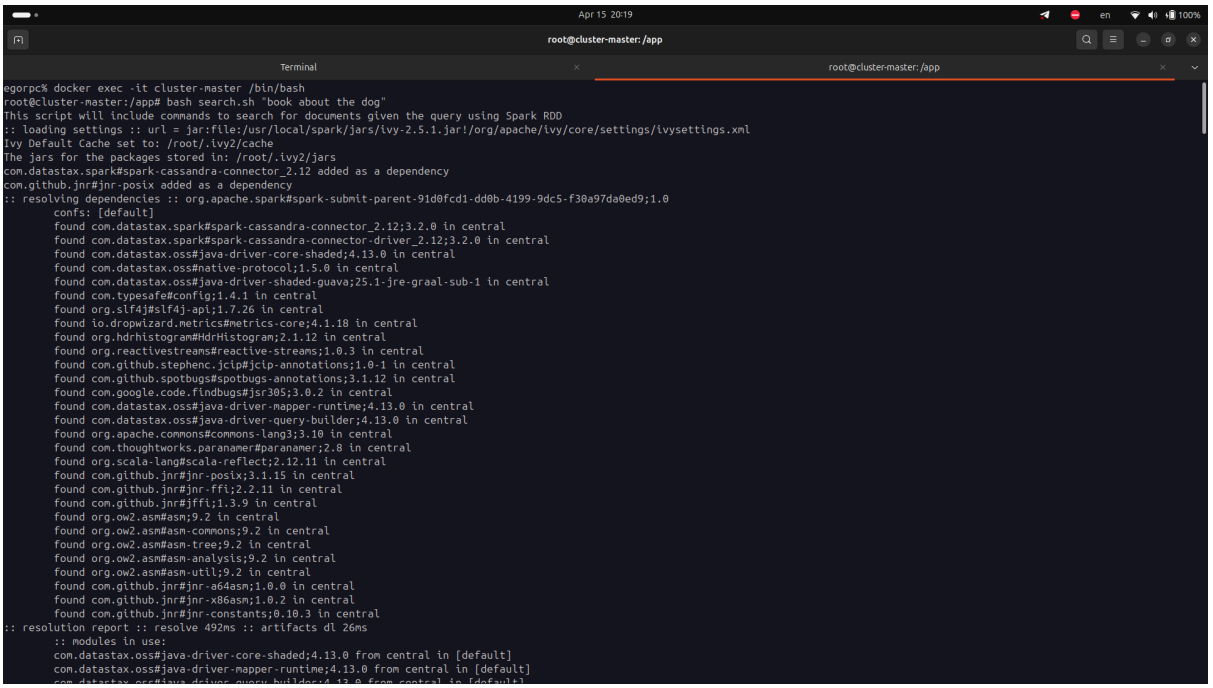

Apr 15 20:02
en
100%
Terminal
Terminal
Terminal
cluster-master | 25/04/15 17:01:56 INFO MemoryStore: Block broadcast_6_piece0 stored as bytes in memory (estimated size 77.8 KiB, free 2012.1 MiB)
cluster-master | 25/04/15 17:01:56 INFO BlockManagerInfo: Added broadcast_6_piece0 in memory on cluster-master:40309 (size: 77.8 KiB, free: 2012.7 MiB)
cluster-master | 25/04/15 17:01:56 INFO SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:1585
cluster-master | 25/04/15 17:01:56 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 6 (MapPartitionsRDD[25] at save at NativeMethodAccessorImpl.java:0) (first 15 tasks are for partitions Vector(0))
cluster-master | 25/04/15 17:01:56 INFO TaskSchedulerImpl: Adding task set 6.0 with 1 tasks resource profile 0
cluster-master | 25/04/15 17:01:56 INFO TaskSetManager: Starting task 0.0 in stage 6.0 (TID 23) (cluster-master, executor driver, partition 0, NODE_LOCAL, 9275 bytes)
cluster-master | 25/04/15 17:01:56 INFO Executor: Running task 0.0 in stage 6.0 (TID 23)
cluster-master | 25/04/15 17:01:56 INFO FileOutputCommitter: File Output Committer Algorithm version is 1
cluster-master | 25/04/15 17:01:56 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
cluster-master | 25/04/15 17:01:56 INFO SQLHadoopMapReduceCommitProtocol: Using output committer class org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
cluster-master | 25/04/15 17:01:56 INFO ShuffleBlockFetcherIterator: Getting 1 (2.3 MiB) non-empty blocks including 1 (2.3 MiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
cluster-master | 25/04/15 17:01:56 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
cluster-master | 25/04/15 17:01:56 INFO FileOutputCommitter: Saved output of task 'attempt_202504151701566479016802585261868_0006_m_000000_23' to hdfs://cluster-master:9000/index/data/_temporary/0/task_202504151701566479016802585261868_0006_m_000000
cluster-master | 25/04/15 17:01:56 INFO SparkHadoopMapRedUtil: attempt_202504151701566479016802585261868_0006_m_000000_23: Committed. Elapsed time: 17 ms.
cluster-master | 25/04/15 17:01:56 INFO Executor: Finished task 0.0 in stage 6.0 (TID 23). 4783 bytes result sent to driver
cluster-master | 25/04/15 17:01:56 INFO TaskSetManager: Finished task 0.0 in stage 6.0 (TID 23) in 376 ms on cluster-master (executor driver) (1/1)
cluster-master | 25/04/15 17:01:56 INFO TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool
cluster-master | 25/04/15 17:01:56 INFO DAGScheduler: ResultStage 6 (save at NativeMethodAccessorImpl.java:0) finished in 0.427 s
cluster-master | 25/04/15 17:01:56 INFO DAGScheduler: Job 4 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master | 25/04/15 17:01:56 INFO TaskSchedulerImpl: Killing all running tasks in stage 6: Stage finished
cluster-master | 25/04/15 17:01:56 INFO DAGScheduler: Job 4 finished: save at NativeMethodAccessorImpl.java:0, took 0.437131 s
cluster-master | 25/04/15 17:01:56 INFO FileFormatWriter: Start to commit write Job 4b84c2c6-08a1-4b58-ab45-b8f71683e336.
cluster-master | 25/04/15 17:01:56 INFO FileFormatWriter: Write Job 4b84c2c6-08a1-4b58-ab45-b8f71683e336 committed. Elapsed time: 21 ms.
cluster-master | 25/04/15 17:01:56 INFO FileFormatWriter: Finished processing stats for write job 4b84c2c6-08a1-4b58-ab45-b8f71683e336.
cluster-master | 25/04/15 17:01:56 INFO SparkContext: Invoking stop() from shutdown hook
cluster-master | 25/04/15 17:01:56 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master | 25/04/15 17:01:56 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master | 25/04/15 17:01:56 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master | 25/04/15 17:01:56 INFO MemoryStore: MemoryStore cleared
cluster-master | 25/04/15 17:01:56 INFO BlockManager: BlockManager stopped
cluster-master | 25/04/15 17:01:56 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master | 25/04/15 17:01:56 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master | 25/04/15 17:01:56 INFO SparkContext: Successfully stopped SparkContext
cluster-master | 25/04/15 17:01:56 INFO ShutdownHookManager: Shutdown hook called
cluster-master | 25/04/15 17:01:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-7b73f9c8-f09a-4b4f-954f-304deae21834
cluster-master | 25/04/15 17:01:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-0af59ef1-d382-4df2-aba2-e3d1e78afd6c/pyspark-8e4155fa-1d7d-4f69-892e-07ba14dd78ab
cluster-master | 25/04/15 17:01:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-0af59ef1-d382-4df2-aba2-e3d1e78afd6c
cluster-master | [3/3] finalizing...
cluster-master | Putting data to hdfs


Apr 15 20:05
en
100%
Terminal
Terminal
Terminal
cluster-master | -rw-r--r--   1 root supergroup   522 2025-04-15 17:04 /data/data/8237581_A_Bridge_Too_Far.txt
cluster-master | -rw-r--r--   1 root supergroup   1668 2025-04-15 17:02 /data/data/8237659_A_Bridge_Too_Far_(book).txt
cluster-master | -rw-r--r--   1 root supergroup   4925 2025-04-15 17:04 /data/data/8270511_A_Chairy_Tale.txt
cluster-master | -rw-r--r--   1 root supergroup   3243 2025-04-15 17:02 /data/data/8312349_A_Barnstormer_in_Oz.txt
cluster-master | -rw-r--r--   1 root supergroup   4898 2025-04-15 17:03 /data/data/8373513_A_Carol_Christmas.txt
cluster-master | -rw-r--r--   1 root supergroup   1253 2025-04-15 17:05 /data/data/8457823_A_Change_in_the_Weather.txt
cluster-master | -rw-r--r--   1 root supergroup   1272 2025-04-15 17:02 /data/data/8528885_A_Cheerful_Gang_Turns_the_Earth.txt
cluster-master | -rw-r--r--   1 root supergroup   11098 2025-04-15 17:02 /data/data/853877_A_Bend_in_the_River.txt
cluster-master | -rw-r--r--   1 root supergroup   1237 2025-04-15 17:04 /data/data/8563962_A_Capital_Federal.txt
cluster-master | -rw-r--r--   1 root supergroup   6602 2025-04-15 17:03 /data/data/867420_A_Burnt-Out_Case.txt
cluster-master | -rw-r--r--   1 root supergroup   5869 2025-04-15 17:02 /data/data/8694441_A_Best_2.txt
cluster-master | -rw-r--r--   1 root supergroup   417 2025-04-15 17:03 /data/data/8740104_A_Certain_World.txt
cluster-master | -rw-r--r--   1 root supergroup   2639 2025-04-15 17:02 /data/data/8780871_A_Bug_and_a_Bag_of_Weed.txt
cluster-master | -rw-r--r--   1 root supergroup   185 2025-04-15 17:03 /data/data/8783301_A_Bugged_Out_Mix_(Felix_da_Housecat_album).txt
cluster-master | -rw-r--r--   1 root supergroup   694 2025-04-15 17:03 /data/data/8796622_A_Chieftains_Celebration.txt
cluster-master | -rw-r--r--   1 root supergroup   1066 2025-04-15 17:05 /data/data/8922606_A_Carolina_Jubilee.txt
cluster-master | -rw-r--r--   1 root supergroup   22618 2025-04-15 17:05 /data/data/89397_A_Bridge_Too_Far_(film).txt
cluster-master | -rw-r--r--   1 root supergroup   12884 2025-04-15 17:02 /data/data/89541_A_Better_Tomorrow.txt
cluster-master | -rw-r--r--   1 root supergroup   2523 2025-04-15 17:03 /data/data/8981838_A_Calendar_of_Wisdom.txt
cluster-master | -rw-r--r--   1 root supergroup   1726 2025-04-15 17:02 /data/data/9277570_A_Break_in_the_Weather.txt
cluster-master | -rw-r--r--   1 root supergroup   519 2025-04-15 17:02 /data/data/9279736_A_(Cass_McCombs_album).txt
cluster-master | -rw-r--r--   1 root supergroup   1938 2025-04-15 17:03 /data/data/929153_A_Bao_A_Qu_(album).txt
cluster-master | -rw-r--r--   1 root supergroup   3221 2025-04-15 17:03 /data/data/929265_A_Chance_to_Cut_Is_a_Chance_to_Cure.txt
cluster-master | -rw-r--r--   1 root supergroup   2308 2025-04-15 17:04 /data/data/9341936_A_Cage_of_Butterflies.txt
cluster-master | -rw-r--r--   1 root supergroup   1991 2025-04-15 17:04 /data/data/9398117_A_Biography.txt
cluster-master | -rw-r--r--   1 root supergroup   6716 2025-04-15 17:03 /data/data/9424803_A_Breed_of_Heroes.txt
cluster-master | -rw-r--r--   1 root supergroup   3466 2025-04-15 17:04 /data/data/9707515_A_Chaos_of_Flowers.txt
cluster-master | -rw-r--r--   1 root supergroup   1650 2025-04-15 17:02 /data/data/9848866_A_Big_10-8_Place.txt
cluster-master | -rw-r--r--   1 root supergroup   2368 2025-04-15 17:02 /data/data/9938275_A_Band_Called_David.txt
cluster-master | -rw-r--r--   1 root supergroup   1770 2025-04-15 17:02 /data/data/995992_A_Beginners'_Guide_to_the_King_Crimson_Collectors'_Club.txt
cluster-master | -rw-r--r--   1 root supergroup   896 2025-04-15 17:03 /data/data/9965276_A_Book_of_Human_Language.txt
cluster-master | -rw-r--r--   1 root supergroup   0 2025-04-15 17:01 /index/data/_SUCCESS
cluster-master | -rw-r--r--   1 root supergroup   3537921 2025-04-15 17:01 /index/data/part-00000-3762c4fe-2bcb-4662-9640-a11085024504-c000.csv
cluster-master | Done data preparation!
cluster-master | This script include commands to run mapreduce jobs using hadoop streaming to index documents
cluster-master | Found 6 items
cluster-master | -rw-r--r--   1 root supergroup 873207391 2025-04-15 17:01 /a.parquet
cluster-master | drwxr-xr-x   - root supergroup   0 2025-04-15 16:57 /apps
cluster-master | drwxr-xr-x   - root supergroup   0 2025-04-15 17:01 /data
cluster-master | drwxr-xr-x   - root supergroup   0 2025-04-15 17:01 /index
cluster-master | drwxrwx---   - root supergroup   0 2025-04-15 16:57 /tmp
cluster-master | drwxr-xr-x   - root supergroup   0 2025-04-15 16:58 /user

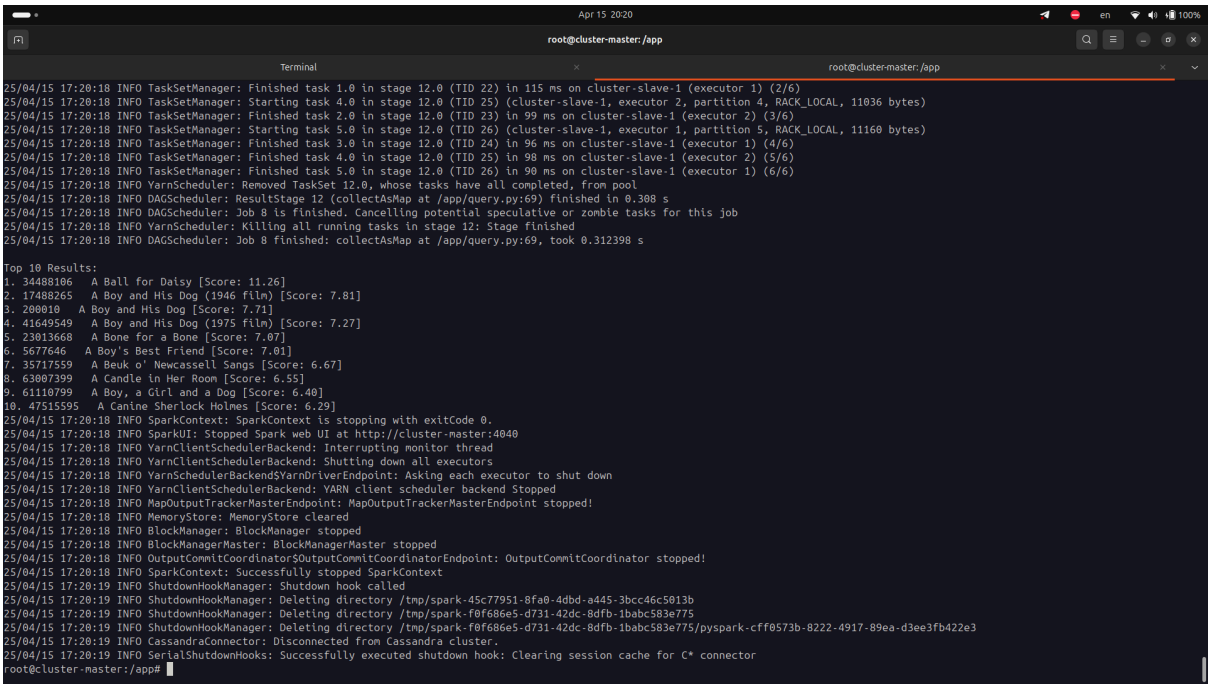

Apr 15 20:08
en
100%
Terminal
Terminal
Terminal
cluster-master | -rw-r--r-- 1 root supergroup 1650 2025-04-15 17:02 /data/data/9848866_A_Big_10-8_Place.txt
cluster-master | -rw-r--r-- 1 root supergroup 2368 2025-04-15 17:02 /data/data/9938275_A_Band_Called_David.txt
cluster-master | -rw-r--r-- 1 root supergroup 1770 2025-04-15 17:02 /data/data/995992_A_Beginners'_Guide_to_the_King_Crimson_Collectors'_Club.txt
cluster-master | -rw-r--r-- 1 root supergroup 896 2025-04-15 17:03 /data/data/9965276_A_Book_of_Human_Language.txt
cluster-master | -rw-r--r-- 1 root supergroup 0 2025-04-15 17:01 /index/data/_SUCCESS
cluster-master | -rw-r--r-- 1 root supergroup 3537921 2025-04-15 17:01 /index/data/part-00000-3762c4fe-2bcb-4662-9640-a11085024504-c000.csv
cluster-master | Done data preparation!
cluster-master | This script include commands to run mapreduce jobs using hadoop streaming to index documents
cluster-master | Found 6 items
cluster-master | -rw-r--r-- 1 root supergroup 873207391 2025-04-15 17:01 /a.parquet
cluster-master | drwxr-xr-x - root supergroup 0 2025-04-15 16:57 /apps
cluster-master | drwxr-xr-x - root supergroup 0 2025-04-15 17:01 /data
cluster-master | drwxr-xr-x - root supergroup 0 2025-04-15 17:01 /index
cluster-master | drwxrwx--- - root supergroup 0 2025-04-15 16:57 /tmp
cluster-master | drwxr-xr-x - root supergroup 0 2025-04-15 16:58 /user
cluster-master | packageJobJar: [/tmp/hadoop-unjar7324219116855901263/] [] /tmp/streamjob6040915109492439187.jar tmpDir=null
cluster-master | 2025-04-15 17:05:42,667 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.24.0.4:8032
cluster-master | 2025-04-15 17:05:42,883 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.24.0.4:8032
cluster-master | 2025-04-15 17:05:43,123 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744736240976_0001
cluster-master | 2025-04-15 17:05:44,560 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master | 2025-04-15 17:05:44,595 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master | 2025-04-15 17:05:44,718 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744736240976_0001
cluster-master | 2025-04-15 17:05:44,719 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master | 2025-04-15 17:05:44,875 INFO conf.Configuration: resource-types.xml not found
cluster-master | 2025-04-15 17:05:44,876 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master | 2025-04-15 17:05:45,324 INFO impl.YarnClientImpl: Submitted application application_1744736240976_0001
cluster-master | 2025-04-15 17:05:45,384 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744736240976_0001/
cluster-master | 2025-04-15 17:05:45,386 INFO mapreduce.Job: Running job: job_1744736240976_0001
cluster-master | 2025-04-15 17:05:57,549 INFO mapreduce.Job: Job job_1744736240976_0001 running in uber mode : false
cluster-master | 2025-04-15 17:05:57,551 INFO mapreduce.Job: map 0% reduce 0%
cluster-master | 2025-04-15 17:06:03,660 INFO mapreduce.Job: map 100% reduce 0%
cluster-master | 2025-04-15 17:06:19,780 INFO mapreduce.Job: map 100% reduce 68%
cluster-master | 2025-04-15 17:06:37,915 INFO mapreduce.Job: map 100% reduce 70%
cluster-master | 2025-04-15 17:06:56,048 INFO mapreduce.Job: map 100% reduce 72%
cluster-master | 2025-04-15 17:07:14,159 INFO mapreduce.Job: map 100% reduce 73%
cluster-master | 2025-04-15 17:07:20,195 INFO mapreduce.Job: map 100% reduce 74%
cluster-master | 2025-04-15 17:07:38,300 INFO mapreduce.Job: map 100% reduce 76%
cluster-master | 2025-04-15 17:07:56,394 INFO mapreduce.Job: map 100% reduce 78%
cluster-master | 2025-04-15 17:08:02,428 INFO mapreduce.Job: map 100% reduce 79%
cluster-master | 2025-04-15 17:08:20,515 INFO mapreduce.Job: map 100% reduce 80%
cluster-master | 2025-04-15 17:08:26,542 INFO mapreduce.Job: map 100% reduce 81%
cluster-master | 2025-04-15 17:08:32,566 INFO mapreduce.Job: map 100% reduce 82%


Apr 15 20:13
ru
100%
Terminal
Terminal
Terminal
cluster-master | 25/04/15 17:13:28 INFO TaskSetManager: Starting task 4.0 in stage 12.0 (TID 25) (cluster-slave-1, executor 1, partition 4, RACK_LOCAL, 11160 bytes)
cluster-master | 25/04/15 17:13:28 INFO TaskSetManager: Finished task 2.0 in stage 12.0 (TID 23) in 98 ms on cluster-slave-1 (executor 1) (3/6)
cluster-master | 25/04/15 17:13:28 INFO TaskSetManager: Starting task 5.0 in stage 12.0 (TID 26) (cluster-slave-1, executor 2, partition 5, RACK_LOCAL, 11036 bytes)
cluster-master | 25/04/15 17:13:28 INFO TaskSetManager: Finished task 3.0 in stage 12.0 (TID 24) in 98 ms on cluster-slave-1 (executor 2) (4/6)
cluster-master | 25/04/15 17:13:28 INFO TaskSetManager: Finished task 5.0 in stage 12.0 (TID 26) in 72 ms on cluster-slave-1 (executor 2) (5/6)
cluster-master | 25/04/15 17:13:28 INFO TaskSetManager: Finished task 4.0 in stage 12.0 (TID 25) in 123 ms on cluster-slave-1 (executor 1) (6/6)
cluster-master | 25/04/15 17:13:28 INFO YarnScheduler: Removed TaskSet 12.0, whose tasks have all completed, from pool
cluster-master | 25/04/15 17:13:28 INFO DAGScheduler: ResultStage 12 (collectAsMap at /app/query.py:69) finished in 0.379 s
cluster-master | 25/04/15 17:13:28 INFO DAGScheduler: Job 8 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master | 25/04/15 17:13:28 INFO YarnScheduler: Killing all running tasks in stage 12: Stage finished
cluster-master | 25/04/15 17:13:28 INFO DAGScheduler: Job 8 finished: collectAsMap at /app/query.py:69, took 0.385522 s
cluster-master |
cluster-master | Top 10 Results:
cluster-master | 1. 2828410 A (musical note) [Score: 33.31]
cluster-master | 2. 58726751 A (Los Angeles Railway) [Score: 9.90]
cluster-master | 3. 15547032 A & G Price [Score: 8.92]
cluster-master | 4. 31831074 A Cambridge Mass [Score: 7.49]
cluster-master | 5. 14810863 A Carnival Christmas [Score: 7.15]
cluster-master | 6. 4697001 A Car-Tune Portrait [Score: 6.47]
cluster-master | 7. 46860526 A Beautiful Soul (song) [Score: 6.21]
cluster-master | 8. 72523905 A Bu [Score: 6.04]
cluster-master | 9. 44464104 A Battle of Nerves [Score: 5.89]
cluster-master | 10. 3561416 A Baña [Score: 5.29]
cluster-master | 25/04/15 17:13:28 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master | 25/04/15 17:13:28 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master | 25/04/15 17:13:28 INFO YarnClientSchedulerBackend: Interrupting monitor thread
cluster-master | 25/04/15 17:13:28 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster-master | 25/04/15 17:13:28 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
cluster-master | 25/04/15 17:13:28 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
cluster-master | 25/04/15 17:13:28 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master | 25/04/15 17:13:28 INFO MemoryStore: MemoryStore cleared
cluster-master | 25/04/15 17:13:28 INFO BlockManager: BlockManager stopped
cluster-master | 25/04/15 17:13:28 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master | 25/04/15 17:13:28 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master | 25/04/15 17:13:28 INFO SparkContext: Successfully stopped SparkContext
cluster-master | 25/04/15 17:13:29 INFO ShutdownHookManager: Shutdown hook called
cluster-master | 25/04/15 17:13:29 INFO ShutdownHookManager: Deleting directory /tmp/spark-b92dbecc-84a8-4d07-8d2c-2eb2fa63204e
cluster-master | 25/04/15 17:13:29 INFO ShutdownHookManager: Deleting directory /tmp/spark-7677ed75-82dc-4a17-bee6-41e5c612d0b6
cluster-master | 25/04/15 17:13:29 INFO ShutdownHookManager: Deleting directory /tmp/spark-7677ed75-82dc-4a17-bee6-41e5c612d0b6/pyspark-48f12f5e-f144-42ab-9959-57e0d82c564e
cluster-master | 25/04/15 17:13:29 INFO CassandraConnector: Disconnected from Cassandra cluster.
cluster-master | 25/04/15 17:13:29 INFO SerialShutdownHooks: Successfully executed shutdown hook: Clearing session cache for C* connector
cluster-master | Services started. Holding the container open...

docker exec -it cluster-master /bin/bash
bash search.sh "book about the dog"




```
egorpc% docker exec -it cluster-master /bin/bash
root@cluster-master:/app# bash search.sh "book about the dog"
This script will include commands to search for documents given the query using Spark RDD
:: loading settings :: url = jar:file:/usr/local/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
com.datastax.spark#spark-cassandra-connector_2.12 added as a dependency
com.github.jnr#jnr-posix added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-91d0fcd1-dd0b-4199-9dc5-f30a97da0ed9;1.0
        confs: [default]
        found com.datastax.spark#spark-cassandra-connector_2.12;3.2.0 in central
        found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.2.0 in central
        found com.datastax.oss#java-driver-core-shaded;4.13.0 in central
        found com.datastax.oss#native-protocol;1.5.0 in central
        found com.datastax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 in central
        found com.typesafe#config;1.4.1 in central
        found org.slf4j#slf4j-api;1.7.26 in central
        found io.dropwizard.metrics#metrics-core;4.1.18 in central
        found org.hdrhistogram#HdrHistogram;2.1.12 in central
        found org.reactivestreams#reactive-streams;1.0.3 in central
        found com.github.stephenc.jcip#jcip-annotations;1.0-1 in central
        found com.github.spotbugs#spotbugs-annotations;3.1.12 in central
        found com.google.code.findbugs#jsr305;3.0.2 in central
        found com.datastax.oss#java-driver-mapper-runtime;4.13.0 in central
        found com.datastax.oss#java-driver-query-builder;4.13.0 in central
        found org.apache.commons#commons-lang3;3.10 in central
        found com.thoughtworks.paranamer#paranamer;2.8 in central
        found org.scala-lang#scala-reflect;2.12.11 in central
        found com.github.jnr#jnr-posix;3.1.15 in central
        found com.github.jnr#jnr-ffi;2.2.11 in central
        found com.github.jnr#jffi;1.3.9 in central
        found org.ow2.asm#asm;9.2 in central
        found org.ow2.asm#asm-commons;9.2 in central
        found org.ow2.asm#asm-tree;9.2 in central
        found org.ow2.asm#asm-analysis;9.2 in central
        found org.ow2.asm#asm-util;9.2 in central
        found com.github.jnr#jnr-a64asm;1.0.0 in central
        found com.github.jnr#jnr-x86asm;1.0.2 in central
        found com.github.jnr#jnr-constants;0.10.3 in central
:: resolution report :: resolve 492ms :: artifacts dl 26ms
        :: modules in use:
        com.datastax.oss#java-driver-core-shaded;4.13.0 from central in [default]
        com.datastax.oss#java-driver-mapper-runtime;4.13.0 from central in [default]
```




```
25/04/15 17:20:18 INFO TaskSetManager: Finished task 1.0 in stage 12.0 (TID 22) in 115 ms on cluster-slave-1 (executor 1) (2/6)
25/04/15 17:20:18 INFO TaskSetManager: Starting task 4.0 in stage 12.0 (TID 25) (cluster-slave-1, executor 2, partition 4, RACK_LOCAL, 11036 bytes)
25/04/15 17:20:18 INFO TaskSetManager: Finished task 2.0 in stage 12.0 (TID 23) in 99 ms on cluster-slave-1 (executor 2) (3/6)
25/04/15 17:20:18 INFO TaskSetManager: Starting task 5.0 in stage 12.0 (TID 26) (cluster-slave-1, executor 1, partition 5, RACK_LOCAL, 11160 bytes)
25/04/15 17:20:18 INFO TaskSetManager: Finished task 3.0 in stage 12.0 (TID 24) in 96 ms on cluster-slave-1 (executor 1) (4/6)
25/04/15 17:20:18 INFO TaskSetManager: Finished task 4.0 in stage 12.0 (TID 25) in 98 ms on cluster-slave-1 (executor 2) (5/6)
25/04/15 17:20:18 INFO TaskSetManager: Finished task 5.0 in stage 12.0 (TID 26) in 90 ms on cluster-slave-1 (executor 1) (6/6)
25/04/15 17:20:18 INFO YarnScheduler: Removed TaskSet 12.0, whose tasks have all completed, from pool
25/04/15 17:20:18 INFO DAGScheduler: ResultStage 12 (collectAsMap at /app/query.py:69) finished in 0.308 s
25/04/15 17:20:18 INFO DAGScheduler: Job 8 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/15 17:20:18 INFO YarnScheduler: Killing all running tasks in stage 12: Stage finished
25/04/15 17:20:18 INFO DAGScheduler: Job 8 finished: collectAsMap at /app/query.py:69, took 0.312398 s

Top 10 Results:
1. 34488106   A Ball for Daisy [Score: 11.26]
2. 17488265   A Boy and His Dog (1946 film) [Score: 7.81]
3. 200010   A Boy and His Dog [Score: 7.71]
4. 41649549   A Boy and His Dog (1975 film) [Score: 7.27]
5. 23013668   A Bone for a Bone [Score: 7.07]
6. 5677646   A Boy's Best Friend [Score: 7.01]
7. 35717559   A Beuk o' Newcassell Sangs [Score: 6.67]
8. 63007399   A Candle in Her Room [Score: 6.55]
9. 61110799   A Boy, a Girl and a Dog [Score: 6.40]
10. 47515595   A Canine Sherlock Holmes [Score: 6.29]
25/04/15 17:20:18 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/15 17:20:18 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/15 17:20:18 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/15 17:20:18 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/15 17:20:18 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
25/04/15 17:20:18 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
25/04/15 17:20:18 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/15 17:20:18 INFO MemoryStore: MemoryStore cleared
25/04/15 17:20:18 INFO BlockManager: BlockManager stopped
25/04/15 17:20:18 INFO BlockManagerMaster: BlockManagerMaster stopped
25/04/15 17:20:18 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/15 17:20:18 INFO SparkContext: Successfully stopped SparkContext
25/04/15 17:20:19 INFO ShutdownHookManager: Shutdown hook called
25/04/15 17:20:19 INFO ShutdownHookManager: Deleting directory /tmp/spark-45c77951-8fa0-4dbd-a445-3bcc46c5013b
25/04/15 17:20:19 INFO ShutdownHookManager: Deleting directory /tmp/spark-f0f686e5-d731-42dc-8dfb-1babc583e775
25/04/15 17:20:19 INFO ShutdownHookManager: Deleting directory /tmp/spark-f0f686e5-d731-42dc-8dfb-1babc583e775/pyspark-cff0573b-8222-4917-89ea-d3ee3fb422e3
25/04/15 17:20:19 INFO CassandraConnector: Disconnected from Cassandra cluster.
25/04/15 17:20:19 INFO SerialShutdownHooks: Successfully executed shutdown hook: Clearing session cache for C* connector
root@cluster-master:/app#
```

bash search.sh "chance of the corner if i will walk"