Methodology

app.sh
The central orchestrator of the workflow. It performs the following steps in sequence:
Starts Services: Initializes Hadoop, Spark, and Cassandra via start-services.sh.
Prepares Data: Triggers prepare_data.sh to download and preprocess the dataset.
Builds Indexes: Executes index.sh to run MapReduce jobs for inverted indexing and statistical analysis.
Maintains Accessibility: Uses tail -f /dev/null to keep the container running indefinitely, enabling post-execution debugging and manual queries.

init-cassandra.sh
Ensures Cassandra is fully operational before schema creation:
Waits for Readiness: Polls Cassandra until it responds to CQL queries.
Initializes Schema: Executes cassandra-init.cql to create tables for documents, inverted indexes, and statistics.

prepare_data.sh
Manages the dataset lifecycle:
Downloads a.parquet: Uses wget to fetch the file.
Transfers to HDFS: Copies the dataset to HDFS for distributed processing.
Generates Metadata: Runs prepare_data.py to split the Parquet file into text documents and create a metadata file in HDFS (/index/data).

index.sh
Coordinates two MapReduce pipelines:
Pipeline 1 (mapper1.py + reducer1.py):
Tokenizes documents into (term, id, term frequency) tuples.
Stores results in Cassandra's indexs and documents tables.
Pipeline 2 (mapper2.py + reducer2.py):
Computes document lengths and aggregates global statistics (total documents, average length).
Populates the documents_stats table for BM25 calculations.

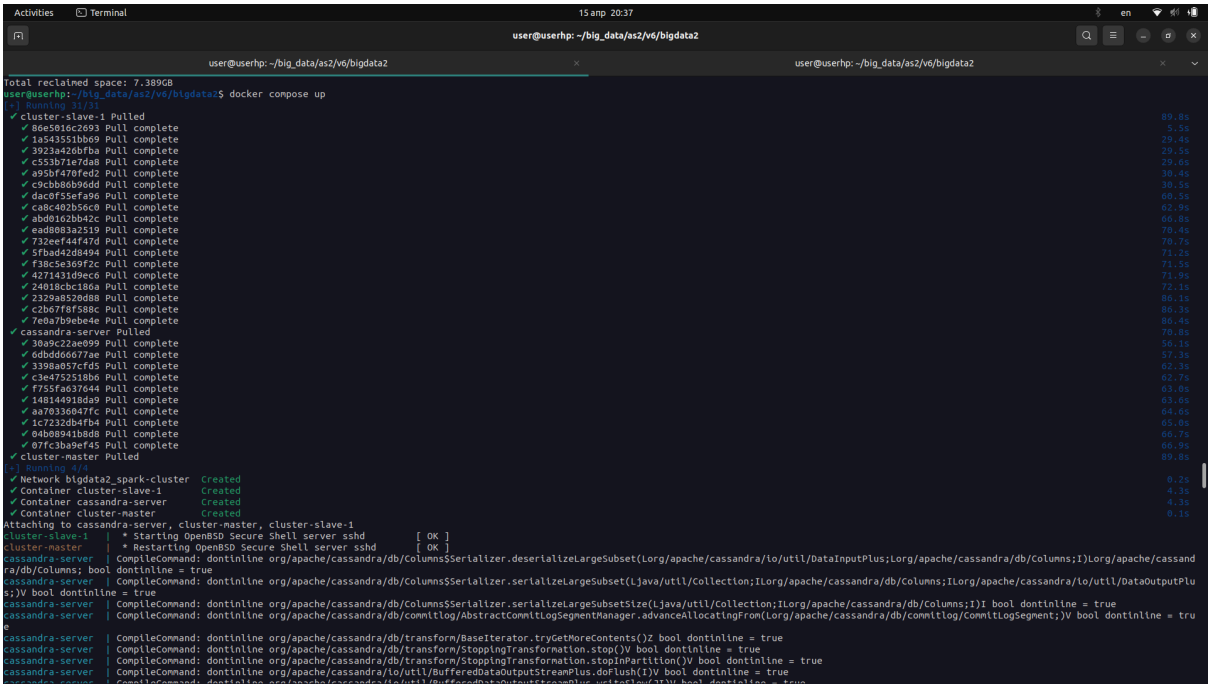
search.sh
Executes search queries using Spark:
Passes Arguments: Accepts a query string
Configures Dependencies: Includes Cassandra connector and Python environment.
Invokes query.py: Runs the ranking logic and returns the top 10 results.

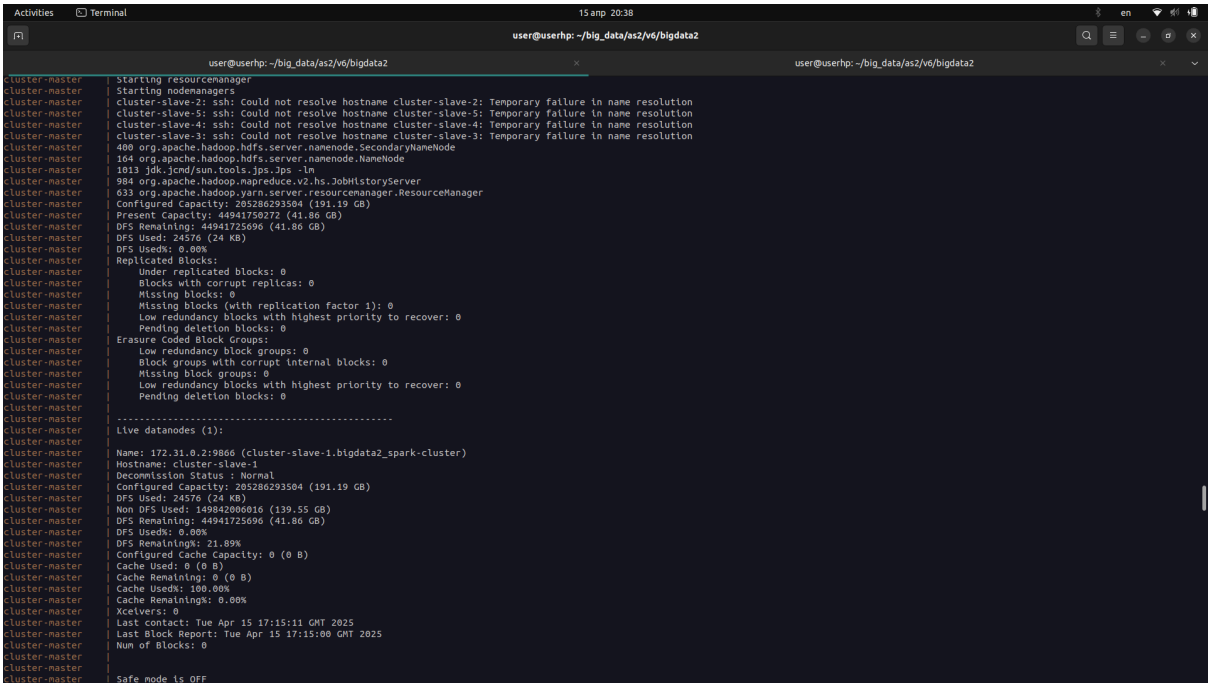
query.py
The ranking engine implements a stabilized BM25 algorithm:
Fetches Data: Retrieves term frequencies, document lengths, and global stats from Cassandra.
Computes Scores: Uses a modified BM25 formula to prevent division-by-zero errors:
python
Ranks Results: Sorts documents by relevance and returns IDs and titles.

Demonstration

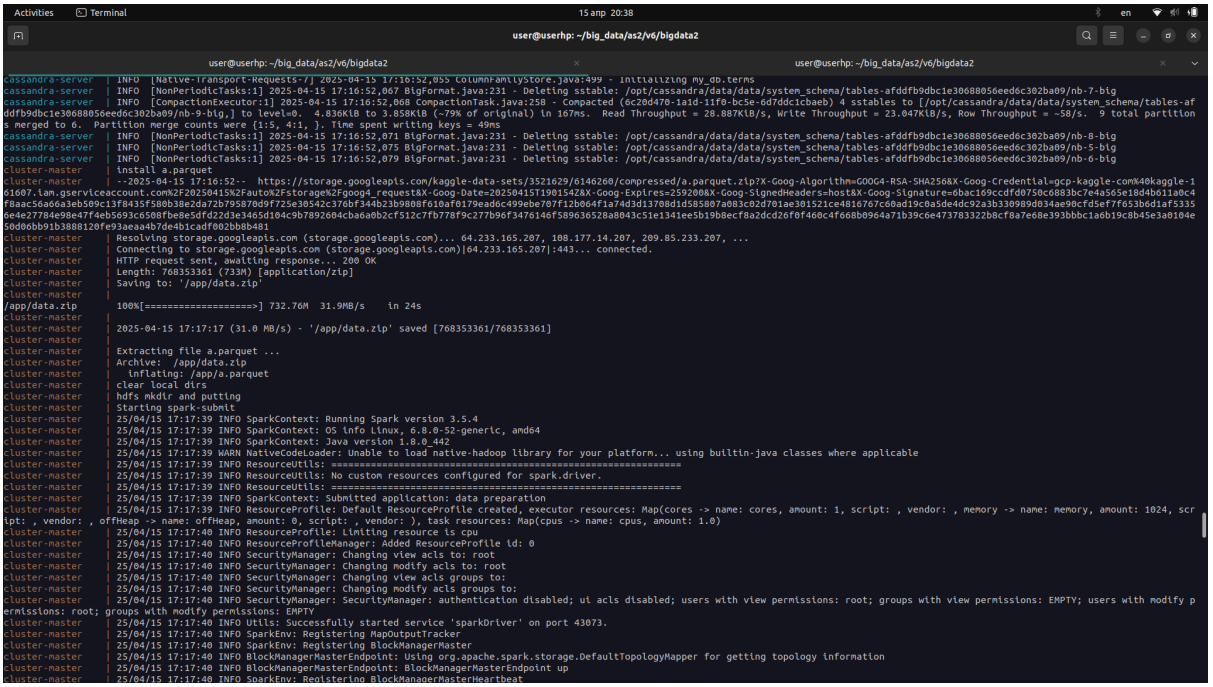
git clone https://github.com/alieAblaeva/bigdata2

docker compose up

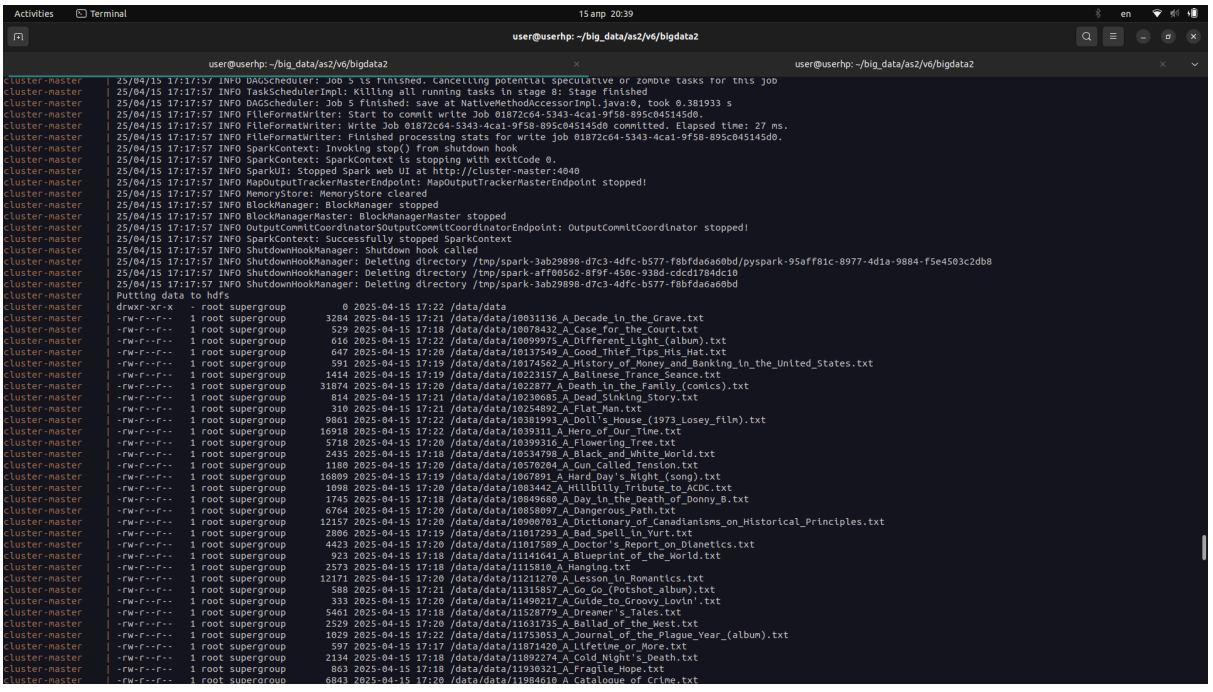


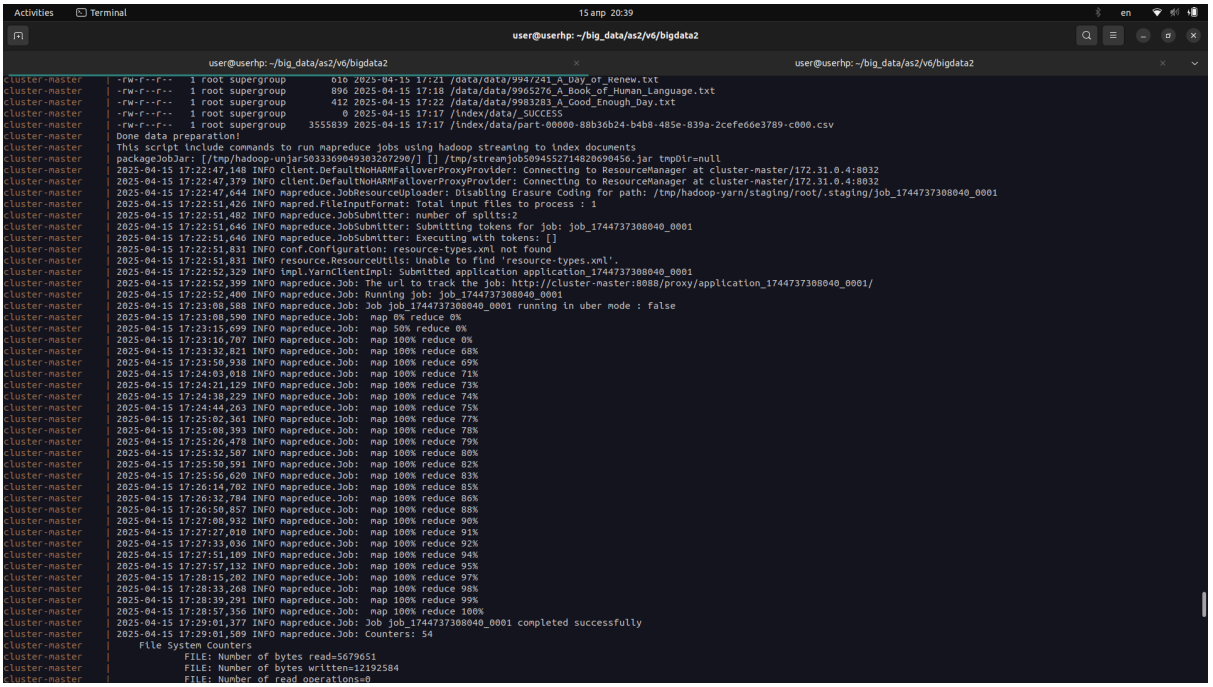


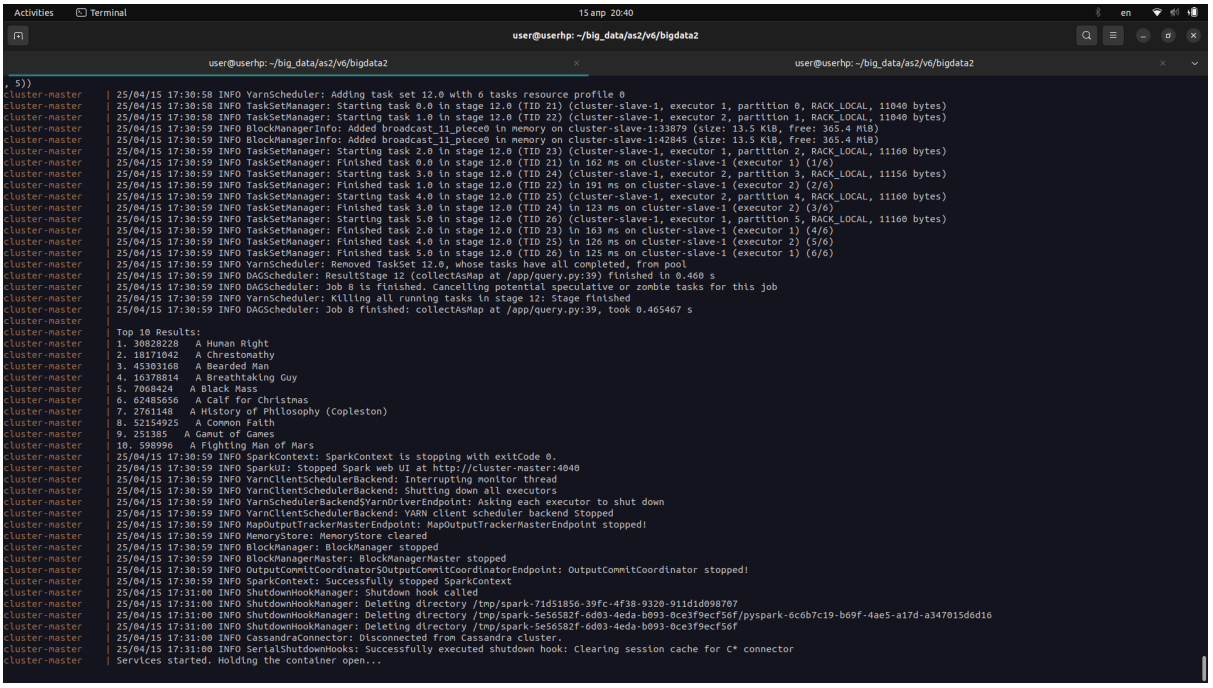
wait



a.parquet downloading



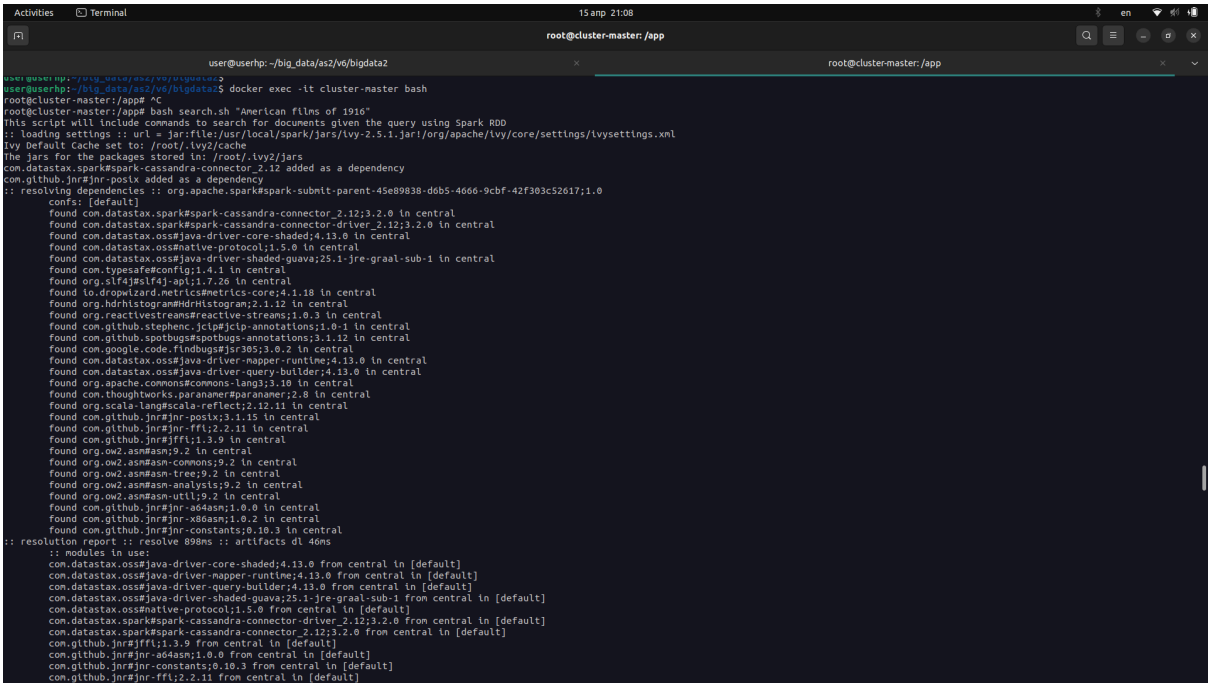

Activities
Terminal
15 апр 20:39
en
user@userhp: ~/big_data/as2/v6/bigdata2
user@userhp: ~/big_data/as2/v6/bigdata2
user@userhp: ~/big_data/as2/v6/bigdata2
cluster-master | -rw-r--r--   1 root supergroup   616 2025-04-15 17:21 /data/data/994/241_A_Day_or_Renew.txt
cluster-master | -rw-r--r--   1 root supergroup   896 2025-04-15 17:18 /data/data/9965276_A_Book_of_Human_Language.txt
cluster-master | -rw-r--r--   1 root supergroup   412 2025-04-15 17:22 /data/data/9983283_A_Good_Enough_Day.txt
cluster-master | -rw-r--r--   1 root supergroup   0 2025-04-15 17:17 /index/data/_SUCCESS
cluster-master | -rw-r--r--   1 root supergroup   3555839 2025-04-15 17:17 /index/data/part-00000-88b36b24-b4b8-485e-839a-2cefe66e3789-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 | packageJobJar: [/tmp/hadoop-unjar5033309049303267290/] [] /tmp/streamjob5094552714820690456.jar tmpDir=null
cluster-master | 2025-04-15 17:22:47,148 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.31.0.4:8032
cluster-master | 2025-04-15 17:22:47,379 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.31.0.4:8032
cluster-master | 2025-04-15 17:22:47,644 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744737308040_0001
cluster-master | 2025-04-15 17:22:51,426 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master | 2025-04-15 17:22:51,482 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master | 2025-04-15 17:22:51,646 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744737308040_0001
cluster-master | 2025-04-15 17:22:51,646 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master | 2025-04-15 17:22:51,831 INFO conf.Configuration: resource-types.xml not found
cluster-master | 2025-04-15 17:22:51,831 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master | 2025-04-15 17:22:52,329 INFO impl.YarnClientImpl: Submitted application application_1744737308040_0001
cluster-master | 2025-04-15 17:22:52,399 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744737308040_0001/
cluster-master | 2025-04-15 17:22:52,400 INFO mapreduce.Job: Running job: job_1744737308040_0001
cluster-master | 2025-04-15 17:23:08,588 INFO mapreduce.Job: Job job_1744737308040_0001 running in uber mode : false
cluster-master | 2025-04-15 17:23:08,590 INFO mapreduce.Job: map 0% reduce 0%
cluster-master | 2025-04-15 17:23:15,699 INFO mapreduce.Job: map 50% reduce 0%
cluster-master | 2025-04-15 17:23:16,707 INFO mapreduce.Job: map 100% reduce 0%
cluster-master | 2025-04-15 17:23:32,821 INFO mapreduce.Job: map 100% reduce 68%
cluster-master | 2025-04-15 17:23:50,938 INFO mapreduce.Job: map 100% reduce 69%
cluster-master | 2025-04-15 17:24:03,018 INFO mapreduce.Job: map 100% reduce 71%
cluster-master | 2025-04-15 17:24:21,129 INFO mapreduce.Job: map 100% reduce 73%
cluster-master | 2025-04-15 17:24:38,229 INFO mapreduce.Job: map 100% reduce 74%
cluster-master | 2025-04-15 17:24:44,263 INFO mapreduce.Job: map 100% reduce 75%
cluster-master | 2025-04-15 17:25:02,361 INFO mapreduce.Job: map 100% reduce 77%
cluster-master | 2025-04-15 17:25:08,393 INFO mapreduce.Job: map 100% reduce 78%
cluster-master | 2025-04-15 17:25:26,478 INFO mapreduce.Job: map 100% reduce 79%
cluster-master | 2025-04-15 17:25:32,507 INFO mapreduce.Job: map 100% reduce 80%
cluster-master | 2025-04-15 17:25:50,591 INFO mapreduce.Job: map 100% reduce 82%
cluster-master | 2025-04-15 17:25:56,620 INFO mapreduce.Job: map 100% reduce 83%
cluster-master | 2025-04-15 17:26:14,702 INFO mapreduce.Job: map 100% reduce 85%
cluster-master | 2025-04-15 17:26:32,784 INFO mapreduce.Job: map 100% reduce 86%
cluster-master | 2025-04-15 17:26:50,857 INFO mapreduce.Job: map 100% reduce 88%
cluster-master | 2025-04-15 17:27:08,932 INFO mapreduce.Job: map 100% reduce 90%
cluster-master | 2025-04-15 17:27:27,010 INFO mapreduce.Job: map 100% reduce 91%
cluster-master | 2025-04-15 17:27:33,036 INFO mapreduce.Job: map 100% reduce 92%
cluster-master | 2025-04-15 17:27:51,109 INFO mapreduce.Job: map 100% reduce 94%
cluster-master | 2025-04-15 17:27:57,132 INFO mapreduce.Job: map 100% reduce 95%
cluster-master | 2025-04-15 17:28:15,202 INFO mapreduce.Job: map 100% reduce 97%
cluster-master | 2025-04-15 17:28:33,268 INFO mapreduce.Job: map 100% reduce 98%
cluster-master | 2025-04-15 17:28:39,291 INFO mapreduce.Job: map 100% reduce 99%
cluster-master | 2025-04-15 17:28:57,356 INFO mapreduce.Job: map 100% reduce 100%
cluster-master | 2025-04-15 17:29:01,377 INFO mapreduce.Job: Job job_1744737308040_0001 completed successfully
cluster-master | 2025-04-15 17:29:01,509 INFO mapreduce.Job: Counters: 54
cluster-master | File System Counters
cluster-master | FILE: Number of bytes read=5679651
cluster-master | FILE: Number of bytes written=12192584
cluster-master | FILE: Number of read operations=0

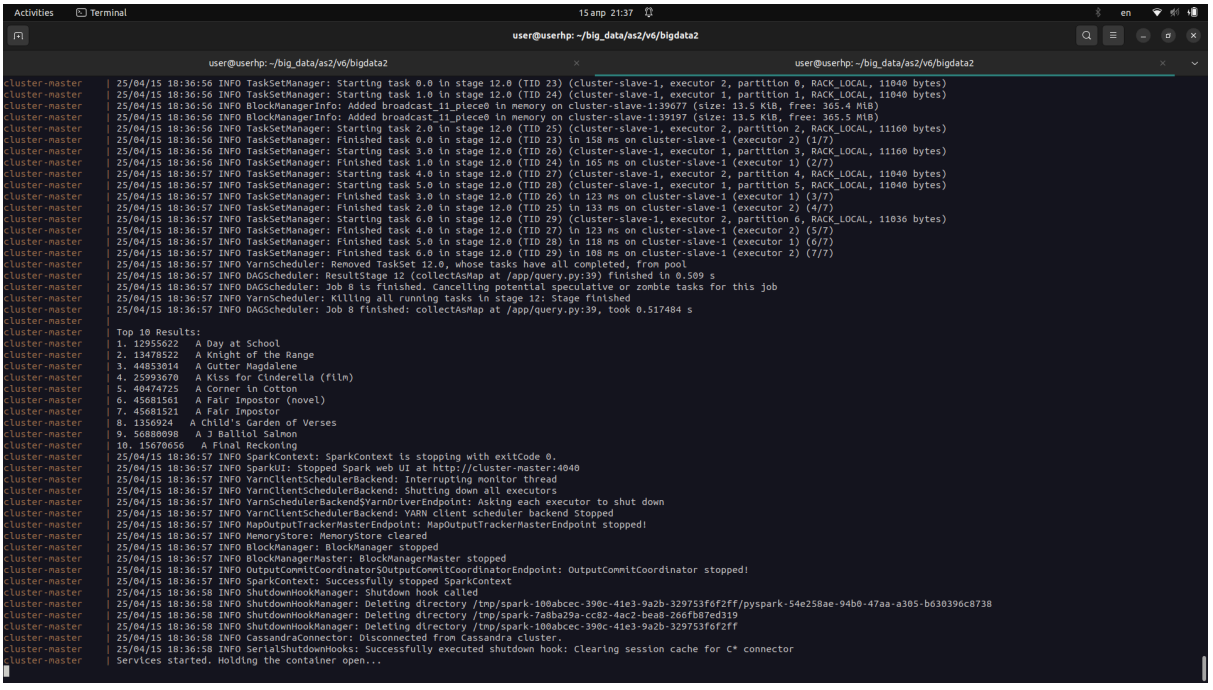

Activities
Terminal
15 апр 20:40
en
user@userhp: ~/big_data/as2/v6/bigdata2
user@userhp: ~/big_data/as2/v6/bigdata2
user@userhp: ~/big_data/as2/v6/bigdata2
, 5))
cluster-master | 25/04/15 17:30:58 INFO YarnScheduler: Adding task set 12.0 with 6 tasks resource profile 0
cluster-master | 25/04/15 17:30:58 INFO TaskSetManager: Starting task 0.0 in stage 12.0 (TID 21) (cluster-slave-1, executor 1, partition 0, RACK_LOCAL, 11040 bytes)
cluster-master | 25/04/15 17:30:58 INFO TaskSetManager: Starting task 1.0 in stage 12.0 (TID 22) (cluster-slave-1, executor 2, partition 1, RACK_LOCAL, 11040 bytes)
cluster-master | 25/04/15 17:30:59 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on cluster-slave-1:33879 (size: 13.5 KiB, free: 365.4 MiB)
cluster-master | 25/04/15 17:30:59 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on cluster-slave-1:42845 (size: 13.5 KiB, free: 365.4 MiB)
cluster-master | 25/04/15 17:30:59 INFO TaskSetManager: Starting task 2.0 in stage 12.0 (TID 23) (cluster-slave-1, executor 1, partition 2, RACK_LOCAL, 11160 bytes)
cluster-master | 25/04/15 17:30:59 INFO TaskSetManager: Finished task 0.0 in stage 12.0 (TID 21) in 162 ms on cluster-slave-1 (executor 1) (1/6)
cluster-master | 25/04/15 17:30:59 INFO TaskSetManager: Starting task 3.0 in stage 12.0 (TID 24) (cluster-slave-1, executor 2, partition 3, RACK_LOCAL, 11156 bytes)
cluster-master | 25/04/15 17:30:59 INFO TaskSetManager: Finished task 1.0 in stage 12.0 (TID 22) in 191 ms on cluster-slave-1 (executor 2) (2/6)
cluster-master | 25/04/15 17:30:59 INFO TaskSetManager: Starting task 4.0 in stage 12.0 (TID 25) (cluster-slave-1, executor 2, partition 4, RACK_LOCAL, 11160 bytes)
cluster-master | 25/04/15 17:30:59 INFO TaskSetManager: Finished task 3.0 in stage 12.0 (TID 24) in 123 ms on cluster-slave-1 (executor 2) (3/6)
cluster-master | 25/04/15 17:30:59 INFO TaskSetManager: Starting task 5.0 in stage 12.0 (TID 26) (cluster-slave-1, executor 1, partition 5, RACK_LOCAL, 11160 bytes)
cluster-master | 25/04/15 17:30:59 INFO TaskSetManager: Finished task 2.0 in stage 12.0 (TID 23) in 163 ms on cluster-slave-1 (executor 1) (4/6)
cluster-master | 25/04/15 17:30:59 INFO TaskSetManager: Finished task 4.0 in stage 12.0 (TID 25) in 126 ms on cluster-slave-1 (executor 2) (5/6)
cluster-master | 25/04/15 17:30:59 INFO TaskSetManager: Finished task 5.0 in stage 12.0 (TID 26) in 125 ms on cluster-slave-1 (executor 1) (6/6)
cluster-master | 25/04/15 17:30:59 INFO YarnScheduler: Removed TaskSet 12.0, whose tasks have all completed, from pool
cluster-master | 25/04/15 17:30:59 INFO DAGScheduler: ResultStage 12 (collectAsMap at /app/query.py:39) finished in 0.460 s
cluster-master | 25/04/15 17:30:59 INFO DAGScheduler: Job 8 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master | 25/04/15 17:30:59 INFO YarnScheduler: Killing all running tasks in stage 12: Stage finished
cluster-master | 25/04/15 17:30:59 INFO DAGScheduler: Job 8 finished: collectAsMap at /app/query.py:39, took 0.465467 s
cluster-master |
cluster-master | Top 10 Results:
cluster-master | 1. 30828228   A Human Right
cluster-master | 2. 18171042   A Chrestomathy
cluster-master | 3. 45303168   A Bearded Man
cluster-master | 4. 16378814   A Breathtaking Guy
cluster-master | 5. 7068424   A Black Mass
cluster-master | 6. 62485656   A Calf for Christmas
cluster-master | 7. 2761148   A History of Philosophy (Copleston)
cluster-master | 8. 52154925   A Common Faith
cluster-master | 9. 251385   A Gamut of Games
cluster-master | 10. 598996   A Fighting Man of Mars
cluster-master | 25/04/15 17:30:59 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master | 25/04/15 17:30:59 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master | 25/04/15 17:30:59 INFO YarnClientSchedulerBackend: Interrupting monitor thread
cluster-master | 25/04/15 17:30:59 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster-master | 25/04/15 17:30:59 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
cluster-master | 25/04/15 17:30:59 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
cluster-master | 25/04/15 17:30:59 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master | 25/04/15 17:30:59 INFO MemoryStore: MemoryStore cleared
cluster-master | 25/04/15 17:30:59 INFO BlockManager: BlockManager stopped
cluster-master | 25/04/15 17:30:59 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master | 25/04/15 17:30:59 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master | 25/04/15 17:30:59 INFO SparkContext: Successfully stopped SparkContext
cluster-master | 25/04/15 17:31:00 INFO ShutdownHookManager: Shutdown hook called
cluster-master | 25/04/15 17:31:00 INFO ShutdownHookManager: Deleting directory /tmp/spark-71d51856-39fc-4f38-9320-911d1d098707
cluster-master | 25/04/15 17:31:00 INFO ShutdownHookManager: Deleting directory /tmp/spark-5e56582f-6d03-4eda-b093-0ce3f9ecf56f/pyspark-6c6b7c19-b69f-4ae5-a17d-a347015d6d16
cluster-master | 25/04/15 17:31:00 INFO ShutdownHookManager: Deleting directory /tmp/spark-5e56582f-6d03-4eda-b093-0ce3f9ecf56f
cluster-master | 25/04/15 17:31:00 INFO CassandraConnector: Disconnected from Cassandra cluster.
cluster-master | 25/04/15 17:31:00 INFO SerialShutdownHooks: Successfully executed shutdown hook: Clearing session cache for C* connector
cluster-master | Services started. Holding the container open...

result

in other console

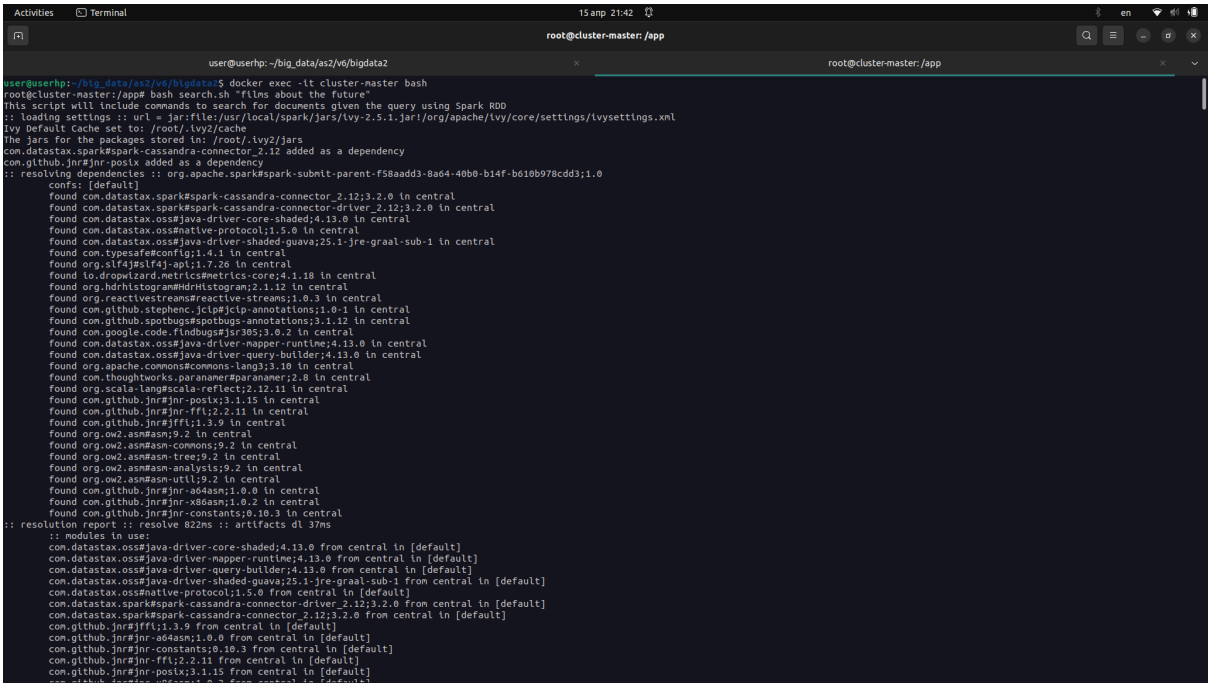
docker exec -it cluster-master bash
bash search.sh "American films of 1916"

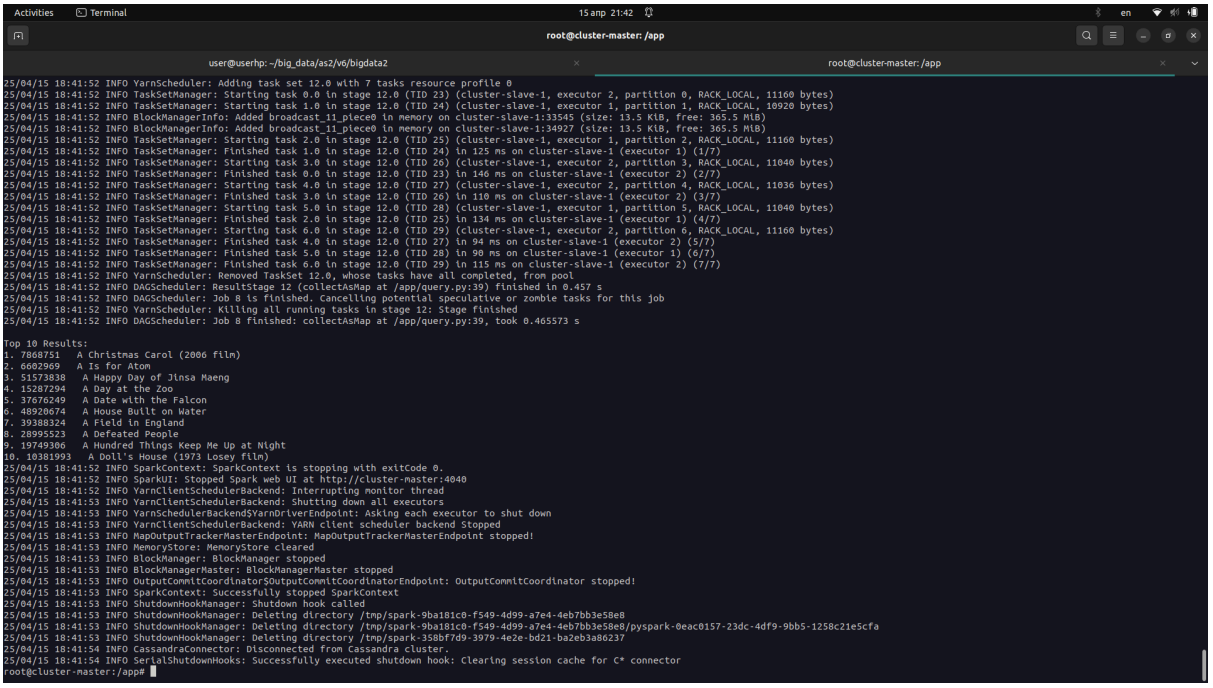




the first doc contains this line! its work!

bash search.sh "films about the future"





it can be seen that the files contain information about the query in descending order, I think this is a really primitive but working algorithm, it will show the result better if it has more files.

As it is, we can see that he did an excellent job with the first test, and he found relevant files for the second.