Methodology

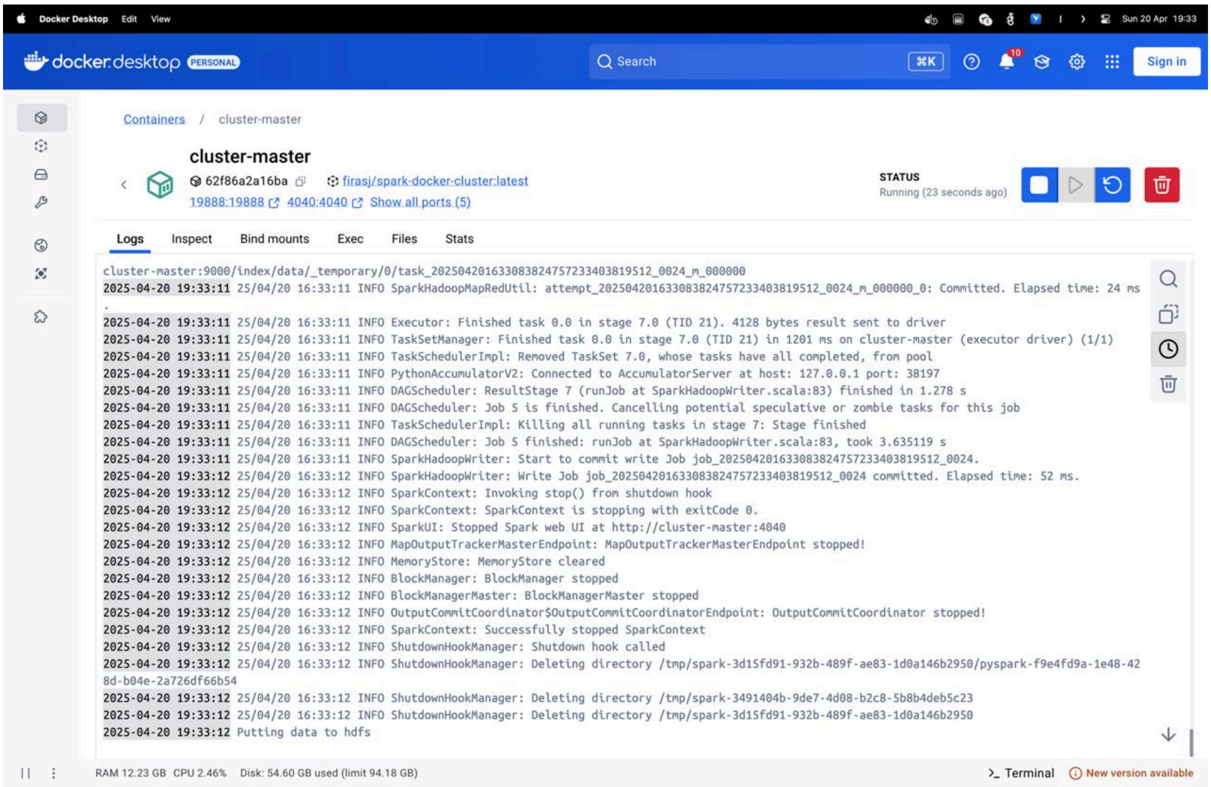
After the data preparation (prepared_data.py), the system builds the indices (index.sh). It begins by activating the virtual environment, ensures the Cassandra driver is bundled, and runs app.py to configure the database schema. This script defines four Cassandra tables: one for the inverted index to store word-document relationships with frequencies, one for storing document titles and lengths, a table for overall statistics like total document count, and another for vocabulary data including TF-IDF values used in scoring.

The main indexing logic is split between the mapper and reducer. The mapper reads through each document, breaks down the text into individual words, filters out common stopwords, and emits relevant information such as terms, document IDs, and frequencies, along with metadata like the title and document length. The reducer then gathers this data, calculates global statistics like IDF, and writes all the processed information to Cassandra in batches for efficiency.

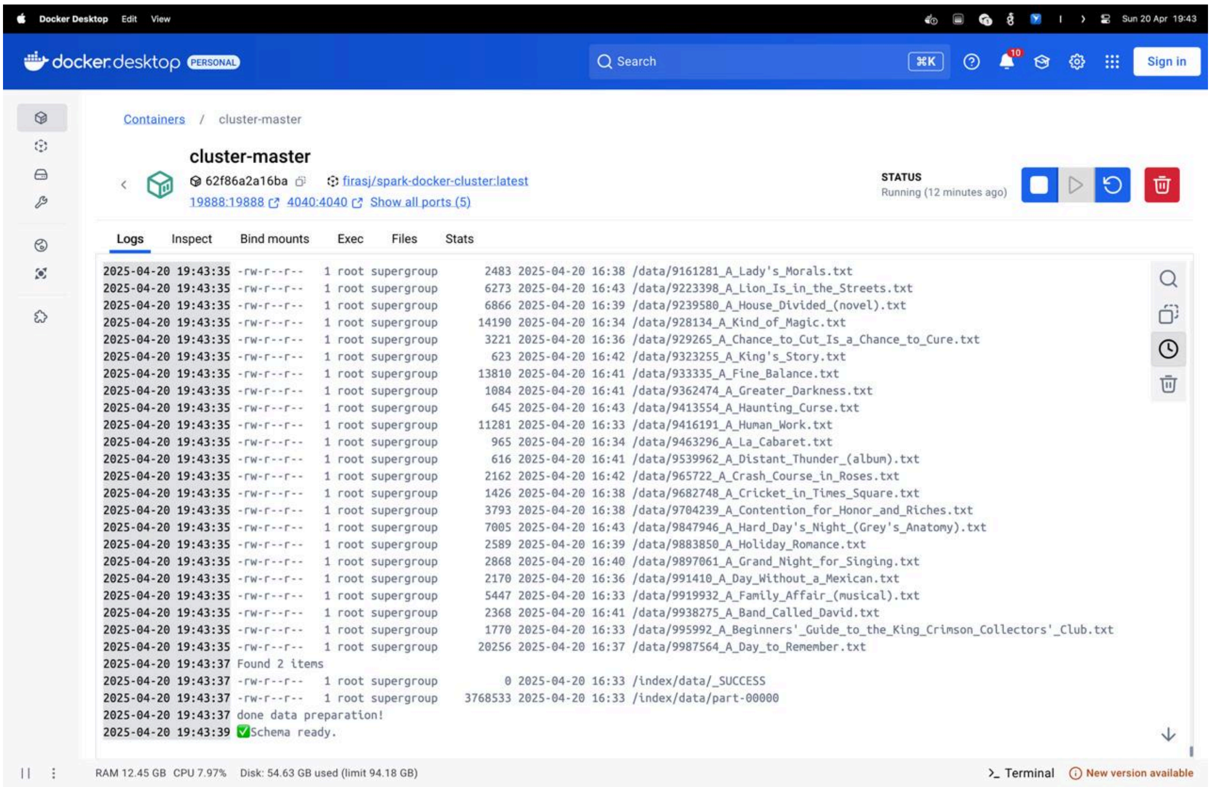
All requirements of the indexer were fully implemented, and indexing was successfully completed. I precomputed all necessary values for search result retrieval — all 1000 documents were indexed correctly. I also developed the full query pipeline, including all logic to retrieve results using the precomputed data. The logic works correctly and returns expected outputs locally. However, running the query pipeline in distributed (yarn) mode currently fails, potentially due to insufficient resources.

Demonstration

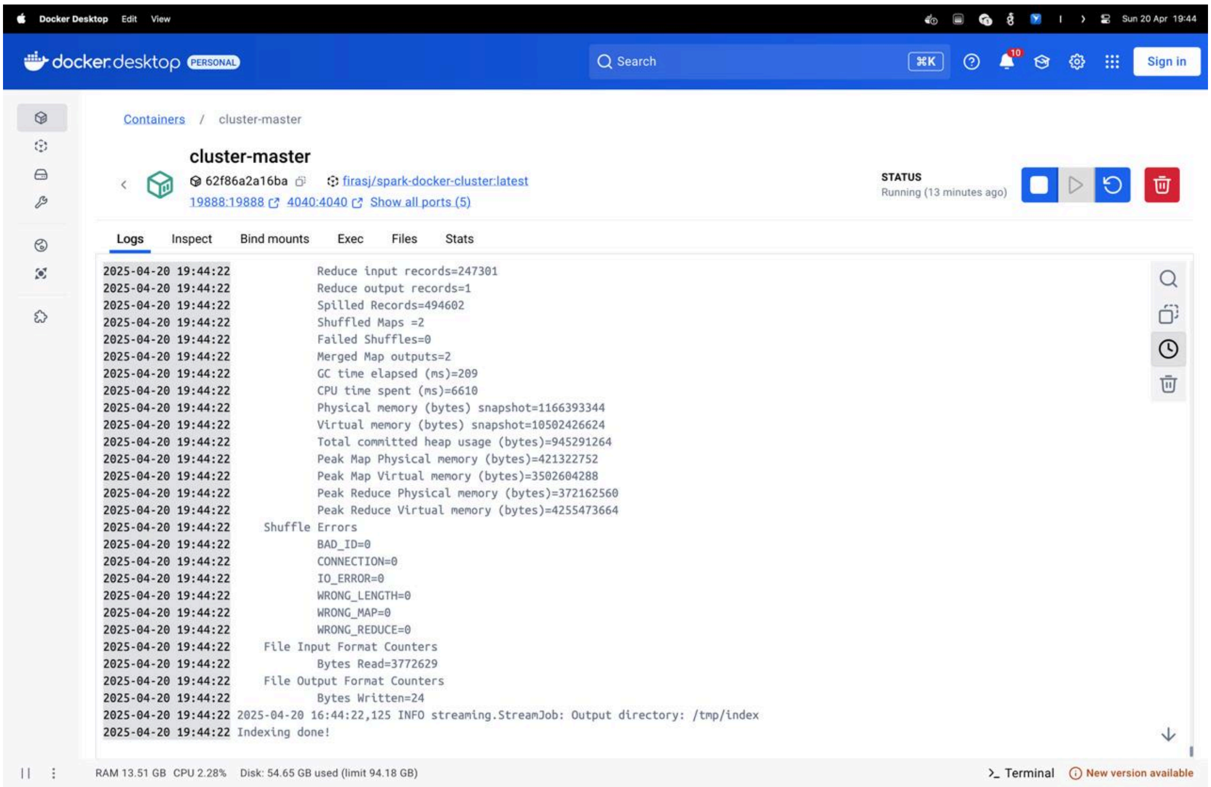
For running **docker compose up -d**



```
cluster-master:9000/index/data/_temporary/0/task_202504201633083824757233403819512_0024_m_000000
2025-04-20 19:33:11 25/04/20 16:33:11 INFO SparkHadoopMapRedUtil: attempt_202504201633083824757233403819512_0024_m_000000_0: Committed. Elapsed time: 24 ms
.
2025-04-20 19:33:11 25/04/20 16:33:11 INFO Executor: Finished task 0.0 in stage 7.0 (TID 21). 4128 bytes result sent to driver
2025-04-20 19:33:11 25/04/20 16:33:11 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID 21) in 1201 ms on cluster-master (executor driver) (1/1)
2025-04-20 19:33:11 25/04/20 16:33:11 INFO TaskSchedulerImpl: Removed TaskSet 7.0, whose tasks have all completed, from pool
2025-04-20 19:33:11 25/04/20 16:33:11 INFO PythonAccumulatorV2: Connected to AccumulatorServer at host: 127.0.0.1 port: 38197
2025-04-20 19:33:11 25/04/20 16:33:11 INFO DAGScheduler: ResultStage 7 (runJob at SparkHadoopWriter.scala:83) finished in 1.278 s
2025-04-20 19:33:11 25/04/20 16:33:11 INFO DAGScheduler: Job 5 is finished. Cancelling potential speculative or zombie tasks for this job
2025-04-20 19:33:11 25/04/20 16:33:11 INFO TaskSchedulerImpl: Killing all running tasks in stage 7: Stage finished
2025-04-20 19:33:11 25/04/20 16:33:11 INFO DAGScheduler: Job 5 finished: runJob at SparkHadoopWriter.scala:83, took 3.635119 s
2025-04-20 19:33:11 25/04/20 16:33:11 INFO SparkHadoopWriter: Start to commit write Job job_202504201633083824757233403819512_0024.
2025-04-20 19:33:12 25/04/20 16:33:12 INFO SparkHadoopWriter: Write Job job_202504201633083824757233403819512_0024 committed. Elapsed time: 52 ms.
2025-04-20 19:33:12 25/04/20 16:33:12 INFO SparkContext: Invoking stop() from shutdown hook
2025-04-20 19:33:12 25/04/20 16:33:12 INFO SparkContext: SparkContext is stopping with exitCode 0.
2025-04-20 19:33:12 25/04/20 16:33:12 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
2025-04-20 19:33:12 25/04/20 16:33:12 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
2025-04-20 19:33:12 25/04/20 16:33:12 INFO MemoryStore: MemoryStore cleared
2025-04-20 19:33:12 25/04/20 16:33:12 INFO BlockManager: BlockManager stopped
2025-04-20 19:33:12 25/04/20 16:33:12 INFO BlockManagerMaster: BlockManagerMaster stopped
2025-04-20 19:33:12 25/04/20 16:33:12 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
2025-04-20 19:33:12 25/04/20 16:33:12 INFO SparkContext: Successfully stopped SparkContext
2025-04-20 19:33:12 25/04/20 16:33:12 INFO ShutdownHookManager: Shutdown hook called
2025-04-20 19:33:12 25/04/20 16:33:12 INFO ShutdownHookManager: Deleting directory /tmp/spark-3d15fd91-932b-489f-ae83-1d0a146b2950/pyspark-f9e4fd9a-1e48-42
8d-b04e-2a726df66b54
2025-04-20 19:33:12 25/04/20 16:33:12 INFO ShutdownHookManager: Deleting directory /tmp/spark-3491404b-9de7-4d08-b2c8-5b8b4deb5c23
2025-04-20 19:33:12 25/04/20 16:33:12 INFO ShutdownHookManager: Deleting directory /tmp/spark-3d15fd91-932b-489f-ae83-1d0a146b2950
2025-04-20 19:33:12 Putting data to hdfs
```

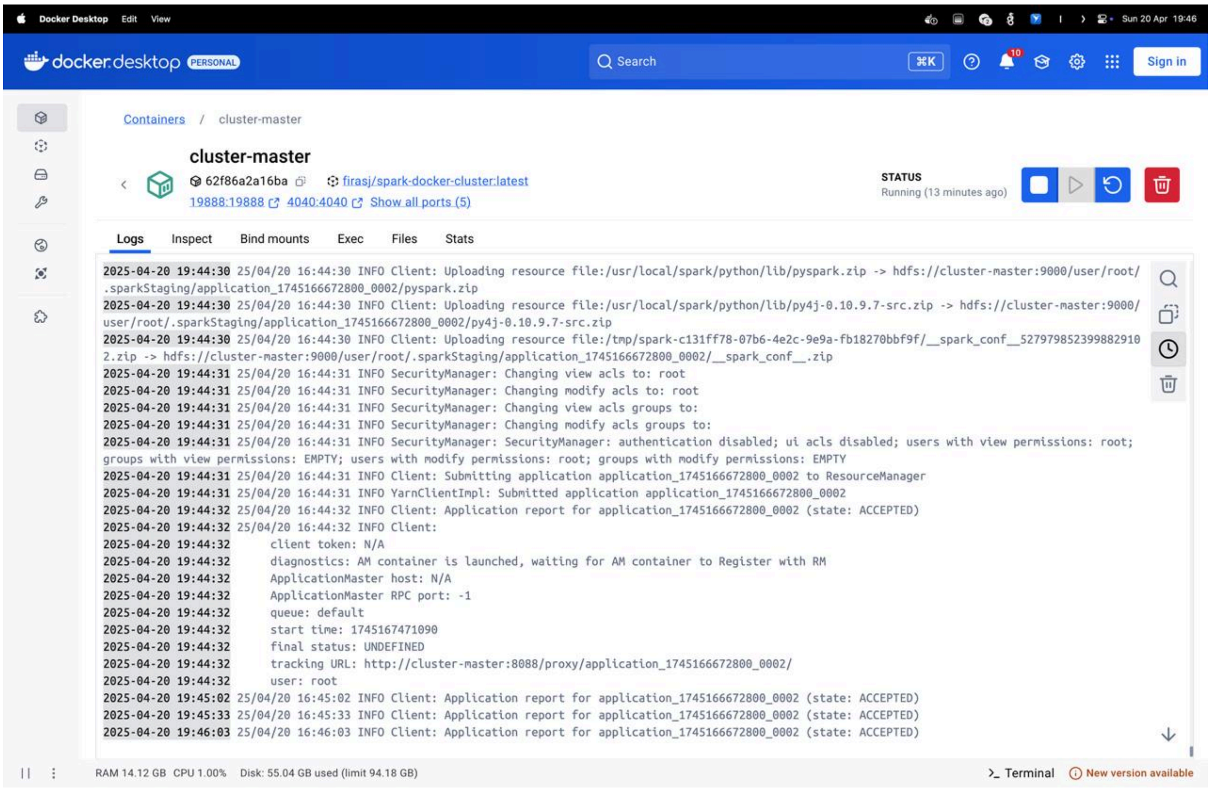


```
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       2483 2025-04-20 16:38 /data/9161281_A_Lady's_Morals.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       6273 2025-04-20 16:43 /data/9223398_A_Lion_Is_in_the_Streets.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       6866 2025-04-20 16:39 /data/9239580_A_House_Divided_(novel).txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup      14190 2025-04-20 16:34 /data/928134_A_Kind_of_Magic.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       3221 2025-04-20 16:36 /data/929265_A_Chance_to_Cut_Is_a_Chance_to_Cure.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup        623 2025-04-20 16:42 /data/9323255_A_King's_Story.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup      13810 2025-04-20 16:41 /data/933335_A_Fine_Balance.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       1084 2025-04-20 16:41 /data/9362474_A_Greater_Darkness.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup        645 2025-04-20 16:43 /data/9413554_A_Haunting_Curse.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup      11281 2025-04-20 16:33 /data/9416191_A_Human_Work.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup        965 2025-04-20 16:34 /data/9463296_A_La_Cabaret.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup        616 2025-04-20 16:41 /data/9539962_A_Distant_Thunder_(album).txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       2162 2025-04-20 16:42 /data/965722_A_Crash_Course_in_Roses.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       1426 2025-04-20 16:38 /data/9682748_A_Cricket_in_Times_Square.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       3793 2025-04-20 16:38 /data/9704239_A_Contention_for_Honor_and_Riches.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       7005 2025-04-20 16:43 /data/9847946_A_Hard_Day's_Night_(Grey's_Anatomy).txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       2589 2025-04-20 16:39 /data/9883850_A_Holiday_Romance.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       2868 2025-04-20 16:40 /data/9897061_A_Grand_Night_for_Singing.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       2170 2025-04-20 16:36 /data/991410_A_Day_Without_a_Mexican.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       5447 2025-04-20 16:33 /data/9919932_A_Family_Affair_(musical).txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       2368 2025-04-20 16:41 /data/9938275_A_Band_Called_David.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup       1770 2025-04-20 16:33 /data/995992_A_Beginners'_Guide_to_the_King_Crimson_Collectors'_Club.txt
2025-04-20 19:43:35 -rw-r--r--   1 root supergroup      20256 2025-04-20 16:37 /data/9987564_A_Day_to_Remember.txt
2025-04-20 19:43:37 Found 2 items
2025-04-20 19:43:37 -rw-r--r--   1 root supergroup          0 2025-04-20 16:33 /index/data/_SUCCESS
2025-04-20 19:43:37 -rw-r--r--   1 root supergroup    3768533 2025-04-20 16:33 /index/data/part-00000
2025-04-20 19:43:37 done data preparation!
2025-04-20 19:43:39 ✅Schema ready.
```



```
2025-04-20 19:44:22             Reduce input records=247301
2025-04-20 19:44:22             Reduce output records=1
2025-04-20 19:44:22             Spilled Records=494602
2025-04-20 19:44:22             Shuffled Maps =2
2025-04-20 19:44:22             Failed Shuffles=0
2025-04-20 19:44:22             Merged Map outputs=2
2025-04-20 19:44:22             GC time elapsed (ms)=209
2025-04-20 19:44:22             CPU time spent (ms)=6610
2025-04-20 19:44:22             Physical memory (bytes) snapshot=1166393344
2025-04-20 19:44:22             Virtual memory (bytes) snapshot=10502426624
2025-04-20 19:44:22             Total committed heap usage (bytes)=945291264
2025-04-20 19:44:22             Peak Map Physical memory (bytes)=421322752
2025-04-20 19:44:22             Peak Map Virtual memory (bytes)=3502604288
2025-04-20 19:44:22             Peak Reduce Physical memory (bytes)=372162560
2025-04-20 19:44:22             Peak Reduce Virtual memory (bytes)=4255473664
2025-04-20 19:44:22     Shuffle Errors
2025-04-20 19:44:22             BAD_ID=0
2025-04-20 19:44:22             CONNECTION=0
2025-04-20 19:44:22             IO_ERROR=0
2025-04-20 19:44:22             WRONG_LENGTH=0
2025-04-20 19:44:22             WRONG_MAP=0
2025-04-20 19:44:22             WRONG_REDUCE=0
2025-04-20 19:44:22     File Input Format Counters
2025-04-20 19:44:22             Bytes Read=3772629
2025-04-20 19:44:22     File Output Format Counters
2025-04-20 19:44:22             Bytes Written=24
2025-04-20 19:44:22 2025-04-20 16:44:22,125 INFO streaming.StreamJob: Output directory: /tmp/index
2025-04-20 19:44:22 Indexing done!
```

Fig 1, 2, 3. Indexing the documents



```
2025-04-20 19:44:30 25/04/20 16:44:30 INFO Client: Uploading resource file:/usr/local/spark/python/lib/pyspark.zip -> hdfs://cluster-master:9000/user/root/
.sparkStaging/application_1745166672800_0002/pyspark.zip
2025-04-20 19:44:30 25/04/20 16:44:30 INFO Client: Uploading resource file:/usr/local/spark/python/lib/py4j-0.10.9.7-src.zip -> hdfs://cluster-master:9000/
user/root/.sparkStaging/application_1745166672800_0002/py4j-0.10.9.7-src.zip
2025-04-20 19:44:30 25/04/20 16:44:30 INFO Client: Uploading resource file:/tmp/spark-c131ff78-07b6-4e2c-9e9a-fb18270bbf9f/__spark_conf__527979852399882910
2.zip -> hdfs://cluster-master:9000/user/root/.sparkStaging/application_1745166672800_0002/__spark_conf__.zip
2025-04-20 19:44:31 25/04/20 16:44:31 INFO SecurityManager: Changing view acls to: root
2025-04-20 19:44:31 25/04/20 16:44:31 INFO SecurityManager: Changing modify acls to: root
2025-04-20 19:44:31 25/04/20 16:44:31 INFO SecurityManager: Changing view acls groups to:
2025-04-20 19:44:31 25/04/20 16:44:31 INFO SecurityManager: Changing modify acls groups to:
2025-04-20 19:44:31 25/04/20 16:44:31 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: root;
groups with view permissions: EMPTY; users with modify permissions: root; groups with modify permissions: EMPTY
2025-04-20 19:44:31 25/04/20 16:44:31 INFO Client: Submitting application application_1745166672800_0002 to ResourceManager
2025-04-20 19:44:31 25/04/20 16:44:31 INFO YarnClientImpl: Submitted application application_1745166672800_0002
2025-04-20 19:44:32 25/04/20 16:44:32 INFO Client: Application report for application_1745166672800_0002 (state: ACCEPTED)
2025-04-20 19:44:32 25/04/20 16:44:32 INFO Client:
2025-04-20 19:44:32      client token: N/A
2025-04-20 19:44:32      diagnostics: AM container is launched, waiting for AM container to Register with RM
2025-04-20 19:44:32      ApplicationMaster host: N/A
2025-04-20 19:44:32      ApplicationMaster RPC port: -1
2025-04-20 19:44:32      queue: default
2025-04-20 19:44:32      start time: 1745167471090
2025-04-20 19:44:32      final status: UNDEFINED
2025-04-20 19:44:32      tracking URL: http://cluster-master:8088/proxy/application_1745166672800_0002/
2025-04-20 19:44:32      user: root
2025-04-20 19:45:02 25/04/20 16:45:02 INFO Client: Application report for application_1745166672800_0002 (state: ACCEPTED)
2025-04-20 19:45:33 25/04/20 16:45:33 INFO Client: Application report for application_1745166672800_0002 (state: ACCEPTED)
2025-04-20 19:46:03 25/04/20 16:46:03 INFO Client: Application report for application_1745166672800_0002 (state: ACCEPTED)
```

Fig 4. Attempt to run query in distributed (yarn) mode