BigData Assignment 2 report

Nikita Kurkulskiu

April 14, 2025

Methodology

gitlink: https://github.com/dadagust/big-data-assignment2-2025

1. Data Preparation

Begin by collecting a sample of plain text documents. Each document is stored in the `data/` folder and pushed to the Hadoop Distributed File System (HDFS). The documents follow a naming format `<doc_id>_<doc_title>.txt`, allowing us to parse `doc_id` in MapReduce jobs.

2. First MapReduce Pipeline (Inverted Index)

Mapper: For each document, the mapper reads lines of text and extracts its `doc_id` from the filename, then tokenizes the lines (using basic regex) into terms. For each term, the mapper emits `(term, doc_id, 1)`.

Reducer: The reducer aggregates partial term counts for each `(term, doc_id)` pair to produce the final term frequency (TF). Concurrently, it maintains the total number of terms per document (`doc_length`), so that each document's length is inserted into the `doc_length` table. The reducer stores:

- `(term, doc_id, freq)` tuples in the `inverted_index` table.
- `(doc_id, length)` in the `doc_length` table.

Both tables reside in Cassandra database under a dedicated keyspace (e.g., `search_keyspace`).

3. Second MapReduce Pipeline (Term Statistics)

Mapper: To compute document frequency (DF), each mapper outputs `(term, doc_id)` but ensures each term in a document is emitted only once. The mapper thus relies on a set-based approach to avoid duplicates within the same document.

Reducer: This reducer receives `(term, doc_id)` pairs and accumulates a set of unique documents for every term. It also looks up total document count and sums of lengths from the `doc_length` table to compute:

- Document frequency (`df`) for each term,
- Total number of documents (`n`),
- The average document length (`avg_doc_length`).

These are stored in a new Cassandra table `statistics(term, df, n, avg_doc_length)`, enabling BM25 computations later.

4. Retrieval (PySpark & BM25)

For query processing, developed a `query.py` Spark application that:

1. Reads a user query from the command line,
2. Loads `inverted_index`, `doc_length`, and `statistics` tables from Cassandra,
3. Prepares an RDD from the in-memory data (in a simple implementation) and filters for terms present in the user query,
4. Computes BM25 for each `(term, doc_id)` using aggregated statistics:

$$\text{BM25}(d, q) = \sum_{t \in q} \ln\!\left(\frac{n - \text{df}(t) + 0.5}{\text{df}(t) + 0.5}\right) \times \frac{\text{freq}(t, d)\,(k+1)}{\text{freq}(t, d) + k\left(1 - b + b \cdot \frac{|d|}{\overline{|d|}}\right)}$$

5. Ranks documents by their BM25 score and returns the top 10 matches.

Hence, the retrieval is orchestrated via Spark RDD operations, preserving distributed parallelism and upholding the assignment's requirement of avoiding single-machine frameworks.

5. Deployment on YARN

Both indexing pipelines (`index.sh`) and the BM25 retrieval job (`search.sh`) are packaged for deployment on a multi-node Hadoop cluster.

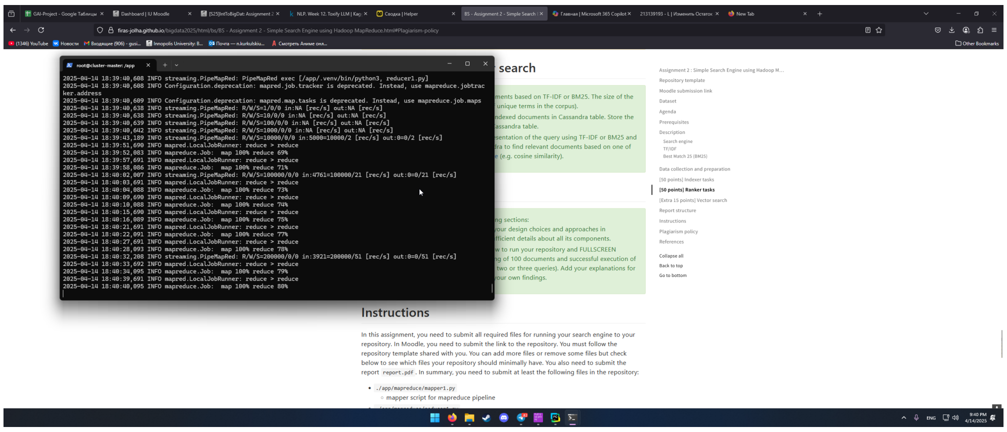
- The indexing pipelines run via Hadoop Streaming MapReduce (in either local or distributed mode).

- The `search.sh` script launches a `spark-submit` job that connects to the Cassandra cluster and processes queries across the YARN resource manager.

This architecture ensures a scalable approach: indexing leverages batch-oriented MapReduce, while real-time queries rely on Spark's low-latency RDD transformations to rank and retrieve documents.

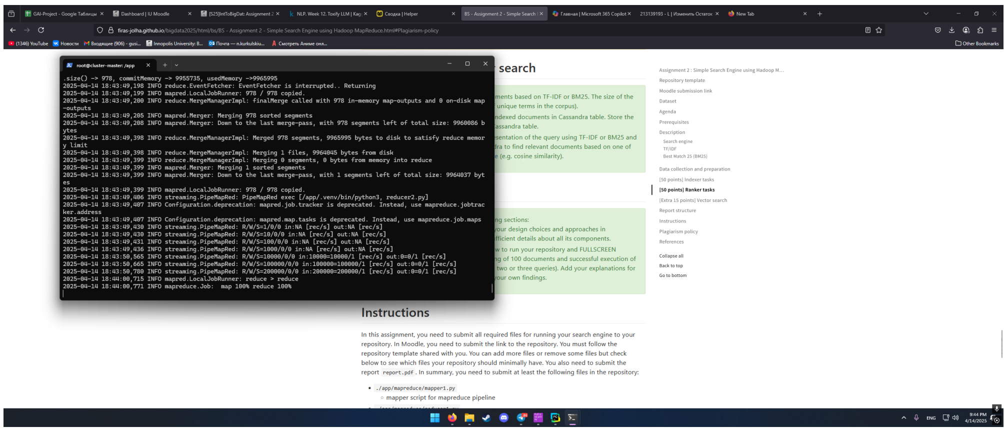
Demonstration

to run the repository enter the cluster-master node and run bash app.sh



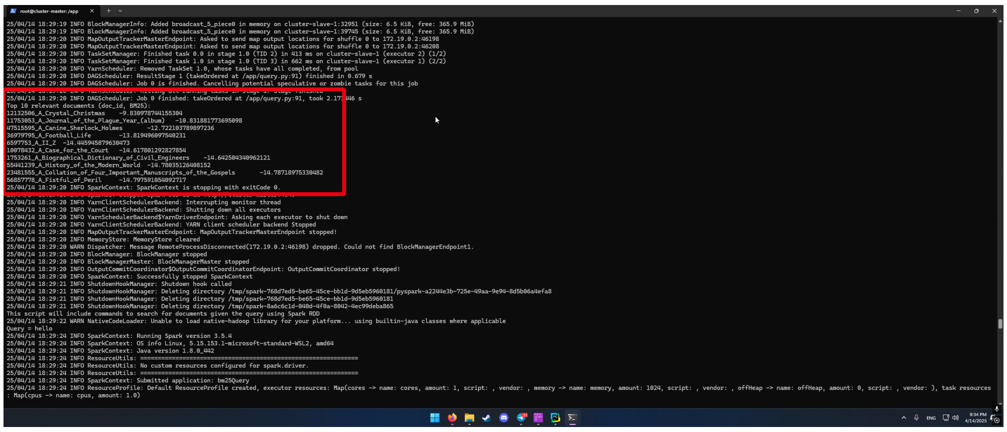
```
2025-04-14 18:39:40,608 INFO streaming.PipeMapRed: PipeMapRed exec [/app/.venv/bin/python3, reducer1.py]
2025-04-14 18:39:40,608 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtrac
ker.address
2025-04-14 18:39:40,609 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2025-04-14 18:39:40,638 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
2025-04-14 18:39:40,638 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
2025-04-14 18:39:40,639 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
2025-04-14 18:39:40,642 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
2025-04-14 18:39:43,189 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:5000=10000/2 [rec/s] out:0=0/2 [rec/s]
2025-04-14 18:39:51,690 INFO mapred.LocalJobRunner: reduce > reduce
2025-04-14 18:39:52,083 INFO mapreduce.Job:  map 100% reduce 69%
2025-04-14 18:39:57,691 INFO mapred.LocalJobRunner: reduce > reduce
2025-04-14 18:39:58,086 INFO mapreduce.Job:  map 100% reduce 71%
2025-04-14 18:40:02,007 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:4761=100000/21 [rec/s] out:0=0/21 [rec/s]
2025-04-14 18:40:03,691 INFO mapred.LocalJobRunner: reduce > reduce
2025-04-14 18:40:04,088 INFO mapreduce.Job:  map 100% reduce 73%
2025-04-14 18:40:09,690 INFO mapred.LocalJobRunner: reduce > reduce
2025-04-14 18:40:10,088 INFO mapreduce.Job:  map 100% reduce 74%
2025-04-14 18:40:15,690 INFO mapred.LocalJobRunner: reduce > reduce
2025-04-14 18:40:16,089 INFO mapreduce.Job:  map 100% reduce 75%
2025-04-14 18:40:21,691 INFO mapred.LocalJobRunner: reduce > reduce
2025-04-14 18:40:22,091 INFO mapreduce.Job:  map 100% reduce 77%
2025-04-14 18:40:27,691 INFO mapred.LocalJobRunner: reduce > reduce
2025-04-14 18:40:28,093 INFO mapreduce.Job:  map 100% reduce 78%
2025-04-14 18:40:32,208 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:3921=200000/51 [rec/s] out:0=0/51 [rec/s]
2025-04-14 18:40:33,692 INFO mapred.LocalJobRunner: reduce > reduce
2025-04-14 18:40:34,095 INFO mapreduce.Job:  map 100% reduce 79%
2025-04-14 18:40:39,691 INFO mapred.LocalJobRunner: reduce > reduce
2025-04-14 18:40:40,095 INFO mapreduce.Job:  map 100% reduce 80%
```

Fig. 1. Cluster running MapReduce and Spark jobs



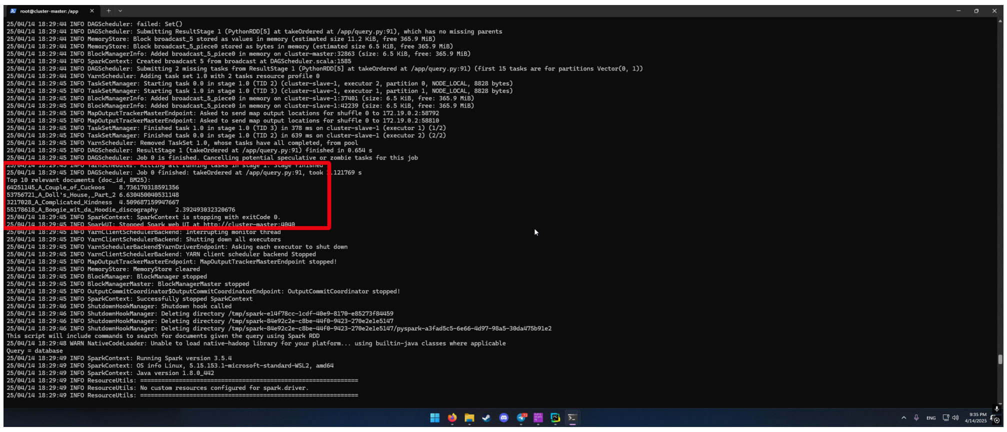
```
.size() -> 978, commitMemory -> 9955735, usedMemory ->9965995
2025-04-14 18:43:49,198 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2025-04-14 18:43:49,199 INFO mapred.LocalJobRunner: 978 / 978 copied.
2025-04-14 18:43:49,200 INFO reduce.MergeManagerImpl: finalMerge called with 978 in-memory map-outputs and 0 on-disk map
-outputs
2025-04-14 18:43:49,205 INFO mapred.Merger: Merging 978 sorted segments
2025-04-14 18:43:49,208 INFO mapred.Merger: Down to the last merge-pass, with 978 segments left of total size: 9960086 b
ytes
2025-04-14 18:43:49,398 INFO reduce.MergeManagerImpl: Merged 978 segments, 9965995 bytes to disk to satisfy reduce memor
y limit
2025-04-14 18:43:49,398 INFO reduce.MergeManagerImpl: Merging 1 files, 9964045 bytes from disk
2025-04-14 18:43:49,399 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2025-04-14 18:43:49,399 INFO mapred.Merger: Merging 1 sorted segments
2025-04-14 18:43:49,399 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 9964037 byt
es
2025-04-14 18:43:49,399 INFO mapred.LocalJobRunner: 978 / 978 copied.
2025-04-14 18:43:49,406 INFO streaming.PipeMapRed: PipeMapRed exec [/app/.venv/bin/python3, reducer2.py]
2025-04-14 18:43:49,407 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtrac
ker.address
2025-04-14 18:43:49,407 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2025-04-14 18:43:49,430 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
2025-04-14 18:43:49,430 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
2025-04-14 18:43:49,431 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
2025-04-14 18:43:49,436 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
2025-04-14 18:43:50,565 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:10000=10000/1 [rec/s] out:0=0/1 [rec/s]
2025-04-14 18:43:50,665 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:100000=100000/1 [rec/s] out:0=0/1 [rec/s]
2025-04-14 18:43:50,780 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:200000=200000/1 [rec/s] out:0=0/1 [rec/s]
2025-04-14 18:44:00,715 INFO mapred.LocalJobRunner: reduce > reduce
2025-04-14 18:44:00,771 INFO mapreduce.Job:  map 100% reduce 100%
```

Fig. 2. Cluster ended running MapReduce and Spark jobs



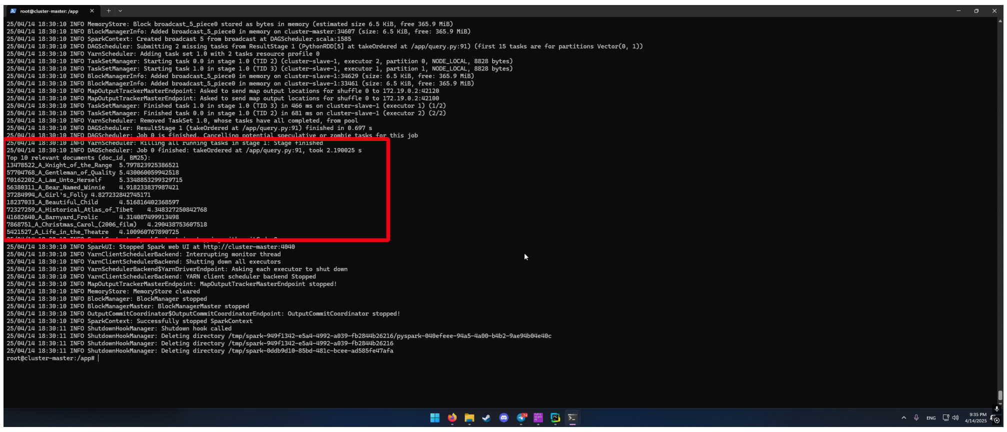
```
25/04/14 18:29:19 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on cluster-slave-1:32951 (size: 6.5 KiB, free: 365.9 MiB)
25/04/14 18:29:19 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on cluster-slave-1:39745 (size: 6.5 KiB, free: 365.9 MiB)
25/04/14 18:29:20 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to 172.19.0.2:46198
25/04/14 18:29:20 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to 172.19.0.2:46208
25/04/14 18:29:20 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2) in 413 ms on cluster-slave-1 (executor 2) (1/2)
25/04/14 18:29:20 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (TID 3) in 662 ms on cluster-slave-1 (executor 1) (2/2)
25/04/14 18:29:20 INFO YarnScheduler: Removed TaskSet 1.0, whose tasks have all completed, from pool
25/04/14 18:29:20 INFO DAGScheduler: ResultStage 1 (takeOrdered at /app/query.py:91) finished in 0.679 s
25/04/14 18:29:20 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/14 18:29:20 INFO YarnScheduler: Killing all running tasks in stage 1: Stage finished
25/04/14 18:29:20 INFO DAGScheduler: Job 0 finished: takeOrdered at /app/query.py:91, took 2.17[illegible]446 s
Top 10 relevant documents (doc_id, BM25):
12132506_A_Crystal_Christmas    -9.830978744155304
11753053_A_Journal_of_the_Plague_Year_(album)   -10.831881773695098
47515595_A_Canine_Sherlock_Holmes       -12.722103789897236
36979795_A_Football_Life        -13.819496097540231
6597753_A_II_Z  -14.445945879630473
10078432_A_Case_for_the_Court   -14.617801292827854
1753261_A_Biographical_Dictionary_of_Civil_Engineers    -14.642504340962121
55441239_A_History_of_the_Modern_World  -14.78035126408152
23481555_A_Collation_of_Four_Important_Manuscripts_of_the_Gospels       -14.78718975330482
56857778_A_Fistful_of_Peril     -14.797591054092717
25/04/14 18:29:20 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/14 18:29:20 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/14 18:29:20 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/14 18:29:20 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/14 18:29:20 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
25/04/14 18:29:20 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
25/04/14 18:29:20 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/14 18:29:20 INFO MemoryStore: MemoryStore cleared
25/04/14 18:29:20 WARN Dispatcher: Message RemoteProcessDisconnected(172.19.0.2:46198) dropped. Could not find BlockManagerEndpoint1.
25/04/14 18:29:20 INFO BlockManager: BlockManager stopped
25/04/14 18:29:20 INFO BlockManagerMaster: BlockManagerMaster stopped
25/04/14 18:29:20 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/14 18:29:20 INFO SparkContext: Successfully stopped SparkContext
25/04/14 18:29:21 INFO ShutdownHookManager: Shutdown hook called
25/04/14 18:29:21 INFO ShutdownHookManager: Deleting directory /tmp/spark-768d7ed5-be65-45ce-bb1d-9d5eb5960181/pyspark-a2244e3b-725e-49aa-9e94-8d5b06a4efa8
25/04/14 18:29:21 INFO ShutdownHookManager: Deleting directory /tmp/spark-768d7ed5-be65-45ce-bb1d-9d5eb5960181
25/04/14 18:29:21 INFO ShutdownHookManager: Deleting directory /tmp/spark-8a6c6c1d-040d-4f0a-8042-4ec99deba865
This script will include commands to search for documents given the query using Spark RDD
25/04/14 18:29:22 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Query = hello
25/04/14 18:29:24 INFO SparkContext: Running Spark version 3.5.4
25/04/14 18:29:24 INFO SparkContext: OS info Linux, 5.15.153.1-microsoft-standard-WSL2, amd64
25/04/14 18:29:24 INFO SparkContext: Java version 1.8.0_442
25/04/14 18:29:24 INFO ResourceUtils: ==============================================================
25/04/14 18:29:24 INFO ResourceUtils: No custom resources configured for spark.driver.
25/04/14 18:29:24 INFO ResourceUtils: ==============================================================
25/04/14 18:29:24 INFO SparkContext: Submitted application: bm25Query
25/04/14 18:29:24 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources
: Map(cpus -> name: cpus, amount: 1.0)
```

Fig. 3. top 10 for query "this is a query!"



```
25/04/14 18:29:44 INFO DAGScheduler: failed: Set()
25/04/14 18:29:44 INFO DAGScheduler: Submitting ResultStage 1 (PythonRDD[5] at takeOrdered at /app/query.py:91), which has no missing parents
25/04/14 18:29:44 INFO MemoryStore: Block broadcast_5 stored as values in memory (estimated size 11.2 KiB, free 365.9 MiB)
25/04/14 18:29:44 INFO MemoryStore: Block broadcast_5_piece0 stored as bytes in memory (estimated size 6.5 KiB, free 365.9 MiB)
25/04/14 18:29:44 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on cluster-master:32863 (size: 6.5 KiB, free: 365.9 MiB)
25/04/14 18:29:44 INFO SparkContext: Created broadcast 5 from broadcast at DAGScheduler.scala:1585
25/04/14 18:29:44 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 1 (PythonRDD[5] at takeOrdered at /app/query.py:91) (first 15 tasks are for partitions Vector(0, 1))
25/04/14 18:29:44 INFO YarnScheduler: Adding task set 1.0 with 2 tasks resource profile 0
25/04/14 18:29:45 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 2) (cluster-slave-1, executor 2, partition 0, NODE_LOCAL, 8828 bytes)
25/04/14 18:29:45 INFO TaskSetManager: Starting task 1.0 in stage 1.0 (TID 3) (cluster-slave-1, executor 1, partition 1, NODE_LOCAL, 8828 bytes)
25/04/14 18:29:45 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on cluster-slave-1:37401 (size: 6.5 KiB, free: 365.9 MiB)
25/04/14 18:29:45 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on cluster-slave-1:42239 (size: 6.5 KiB, free: 365.9 MiB)
25/04/14 18:29:45 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to 172.19.0.2:58792
25/04/14 18:29:45 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to 172.19.0.2:58810
25/04/14 18:29:45 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (TID 3) in 378 ms on cluster-slave-1 (executor 1) (1/2)
25/04/14 18:29:45 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2) in 639 ms on cluster-slave-1 (executor 2) (2/2)
25/04/14 18:29:45 INFO YarnScheduler: Removed TaskSet 1.0, whose tasks have all completed, from pool
25/04/14 18:29:45 INFO DAGScheduler: ResultStage 1 (takeOrdered at /app/query.py:91) finished in 0.654 s
25/04/14 18:29:45 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/14 18:29:45 INFO YarnScheduler: Killing all running tasks in stage 1: Stage finished
25/04/14 18:29:45 INFO DAGScheduler: Job 0 finished: takeOrdered at /app/query.py:91, took [illegible].121769 s
Top 10 relevant documents (doc_id, BM25):
64251145_A_Couple_of_Cuckoos    8.736170318591356
53756721_A_Doll's_House,_Part_2 6.630450040531148
3217028_A_Complicated_Kindness  4.509687159947667
55178618_A_Boogie_wit_da_Hoodie_discography     2.392493032320676
25/04/14 18:29:45 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/14 18:29:45 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/14 18:29:45 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/14 18:29:45 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/14 18:29:45 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
25/04/14 18:29:45 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
25/04/14 18:29:45 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/14 18:29:45 INFO MemoryStore: MemoryStore cleared
25/04/14 18:29:45 INFO BlockManager: BlockManager stopped
25/04/14 18:29:45 INFO BlockManagerMaster: BlockManagerMaster stopped
25/04/14 18:29:45 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/14 18:29:45 INFO SparkContext: Successfully stopped SparkContext
25/04/14 18:29:46 INFO ShutdownHookManager: Shutdown hook called
25/04/14 18:29:46 INFO ShutdownHookManager: Deleting directory /tmp/spark-e14f78cc-1cdf-40e9-8170-e85273f84459
25/04/14 18:29:46 INFO ShutdownHookManager: Deleting directory /tmp/spark-84e92c2e-c8be-44f0-9423-270e2e1e5147
25/04/14 18:29:46 INFO ShutdownHookManager: Deleting directory /tmp/spark-84e92c2e-c8be-44f0-9423-270e2e1e5147/pyspark-a3fad5c5-6e66-4d97-98a5-30da475b91e2
This script will include commands to search for documents given the query using Spark RDD
25/04/14 18:29:48 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Query = database
25/04/14 18:29:49 INFO SparkContext: Running Spark version 3.5.4
25/04/14 18:29:49 INFO SparkContext: OS info Linux, 5.15.153.1-microsoft-standard-WSL2, amd64
25/04/14 18:29:49 INFO SparkContext: Java version 1.8.0_442
25/04/14 18:29:49 INFO ResourceUtils: ==============================================================
25/04/14 18:29:49 INFO ResourceUtils: No custom resources configured for spark.driver.
25/04/14 18:29:49 INFO ResourceUtils: ==============================================================
```

Fig. 4. top 10 for query "hello"



```
25/04/14 18:30:10 INFO MemoryStore: Block broadcast_5_piece0 stored as bytes in memory (estimated size 6.5 KiB, free 365.9 MiB)
25/04/14 18:30:10 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on cluster-master:34607 (size: 6.5 KiB, free: 365.9 MiB)
25/04/14 18:30:10 INFO SparkContext: Created broadcast 5 from broadcast at DAGScheduler.scala:1585
25/04/14 18:30:10 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 1 (PythonRDD[5] at takeOrdered at /app/query.py:91) (first 15 tasks are for partitions Vector(0, 1))
25/04/14 18:30:10 INFO YarnScheduler: Adding task set 1.0 with 2 tasks resource profile 0
25/04/14 18:30:10 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 2) (cluster-slave-1, executor 2, partition 0, NODE_LOCAL, 8828 bytes)
25/04/14 18:30:10 INFO TaskSetManager: Starting task 1.0 in stage 1.0 (TID 3) (cluster-slave-1, executor 1, partition 1, NODE_LOCAL, 8828 bytes)
25/04/14 18:30:10 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on cluster-slave-1:34629 (size: 6.5 KiB, free: 365.9 MiB)
25/04/14 18:30:10 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on cluster-slave-1:33461 (size: 6.5 KiB, free: 365.9 MiB)
25/04/14 18:30:10 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to 172.19.0.2:42120
25/04/14 18:30:10 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to 172.19.0.2:42100
25/04/14 18:30:10 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (TID 3) in 466 ms on cluster-slave-1 (executor 1) (1/2)
25/04/14 18:30:10 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2) in 681 ms on cluster-slave-1 (executor 2) (2/2)
25/04/14 18:30:10 INFO YarnScheduler: Removed TaskSet 1.0, whose tasks have all completed, from pool
25/04/14 18:30:10 INFO DAGScheduler: ResultStage 1 (takeOrdered at /app/query.py:91) finished in 0.697 s
25/04/14 18:30:10 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/14 18:30:10 INFO YarnScheduler: Killing all running tasks in stage 1: Stage finished
25/04/14 18:30:10 INFO DAGScheduler: Job 0 finished: takeOrdered at /app/query.py:91, took 2.190025 s
Top 10 relevant documents (doc_id, BM25):
13478522_A_Knight_of_the_Range  5.797823925386521
57704768_A_Gentleman_of_Quality 5.430060059942518
70162202_A_Law_Unto_Herself     5.3348853299329715
56380311_A_Bear_Named_Winnie    4.918233837987421
37284994_A_Girl's_Folly 4.827232842745171
18237033_A_Beautiful_Child      4.516816402368597
72327259_A_Historical_Atlas_of_Tibet    4.348327250842768
41682640_A_Barnyard_Frolic      4.314087499913498
7868751_A_Christmas_Carol_(2006_film)   4.290438753607518
5421527_A_Life_in_the_Theatre   4.100960767890725
25/04/14 18:30:10 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/14 18:30:10 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/14 18:30:10 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/14 18:30:10 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/14 18:30:10 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
25/04/14 18:30:10 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
25/04/14 18:30:10 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/14 18:30:10 INFO MemoryStore: MemoryStore cleared
25/04/14 18:30:10 INFO BlockManager: BlockManager stopped
25/04/14 18:30:10 INFO BlockManagerMaster: BlockManagerMaster stopped
25/04/14 18:30:10 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/14 18:30:10 INFO SparkContext: Successfully stopped SparkContext
25/04/14 18:30:11 INFO ShutdownHookManager: Shutdown hook called
25/04/14 18:30:11 INFO ShutdownHookManager: Deleting directory /tmp/spark-949f1342-e5a4-4992-a039-fb2844b26216/pyspark-040efeee-94a5-4a00-b4b2-9ae94b04e40c
25/04/14 18:30:11 INFO ShutdownHookManager: Deleting directory /tmp/spark-949f1342-e5a4-4992-a039-fb2844b26216
25/04/14 18:30:11 INFO ShutdownHookManager: Deleting directory /tmp/spark-0ddb9d10-85bd-481c-bcee-ad585fe47afa
root@cluster-master:/app# |
```

Fig. 5. top 10 for query "database"