

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
```

```
Activities Terminal 15 apr 2038 user@userhp: ~/big_data/as2/v6/bigdata2
```

```
total reclaimed space: 7.389GB  
user@userhp:~/big_data/as2/v6/bigdata2$ docker compose up  
[+] Running 1/1  
✓ cluster-slave-1 Pulled 89.8s  
  8be5e1c2d93 Pull complete 5.5s  
  1a543531b469 Pull complete 29.4s  
  3923a426bfba Pull complete 29.5s  
  c553b71e7da8 Pull complete 29.6s  
  a95bf4f06ed0 Pull complete 30.4s  
  cc8cb809edd Pull complete 30.5s  
  dac0f5fefaf6 Pull complete 60.5s  
  ca8c402b5dc0 Pull complete 62.9s  
  adbd1ebbbzcc Pull complete 66.8s  
  ead8083a2519 Pull complete 70.4s  
  732eeef4fd7d Pull complete 70.7s  
  f5bde2d094 Pull complete 71.2s  
  f18C5a369f2c Pull complete 71.5s  
  4271431d9ec6 Pull complete 71.9s  
  24018bc186a Pull complete 72.1s  
  232ea82ed68 Pull complete 86.1s  
  c2b678cf588c Pull complete 86.3s  
  76da79bebade Pull complete 86.4s  
✓ cassandra-server Pulled 89.8s  
  30a9c22ae099 Pull complete 56.1s  
  6dbdd6677ae Pull complete 57.3s  
  339aa857cfd5 Pull complete 62.3s  
  c9e42251bbb Pull complete 62.7s  
  f755fa037644 Pull complete 63.0s  
  148144918da9 Pull complete 63.8s  
  a78336047fc Pull complete 84.6s  
  1c7232b4fb4 Pull complete 63.9s  
  04b08941bad8 Pull complete 66.7s  
  07fc3abdf45 Pull complete 66.9s  
✓ cluster-master Pulled 89.8s  
[+] Running 4/4  
✓ network bigdata2_spark-cluster Created 0.2s  
✓ container cluster-slave-1 Created 4.3s  
✓ container cassandra-server Created 4.3s  
✓ container cluster-master Created 0.1s  
Attaching to cassandra-server, cluster-master, cluster-slave-1  
cluster-slave-1 | * Starting OpenSSH Secure Shell server sshd [ OK ]  
cluster-master | * Restarting OpenSSH Secure Shell server sshd [ OK ]  
cassandra-server | CompilingCommand: dntline org/apache/cassandra/db/Columns$Serializer.serializeLargeSubset(Lorg/apache/cassandra/io/util/DataInputPlus;Lorg/apache/cassandra/db/Columns;)I Lorg/apache/cassand  
r/db/Columns; bool dntline = true  
cassandra-server | CompilingCommand: dntline org/apache/cassandra/db/Columns$Serializer.serializeLargeSubset([Ljava/util/Collection;Lorg/apache/cassandra/db/Columns;)I I bool dntline = true  
cassandra-server | CompilingCommand: dntline org/apache/cassandra/db/comintlog/AbstractComitLogSegmentManager.advanceLocatingFrom(Lorg/apache/cassandra/db/comintlog/ComitLogSegment;)V bool dntline = tru  
e  
cassandra-server | CompilingCommand: dntline org/apache/cassandra/db/transform/BasetTransformer.tryGetCurrentContents()J bool dntline = true  
cassandra-server | CompilingCommand: dntline org/apache/cassandra/db/transform/StoppingTransformation.stop(V) bool dntline = true  
cassandra-server | CompilingCommand: dntline org/apache/cassandra/db/transform/StoppingTransformation.stoptPartition(JV) bool dntline = true  
cassandra-server | CompilingCommand: dntline org/apache/cassandra/io/util/BufferedDataOutputStreampPlus.doFlush(IJV) bool dntline = true  
CompilingCommand: dntline org/apache/cassandra/io/util/BufferedDataOutputStreampPlus.doFlush(IJV) bool dntline = true  
[+] Running 1/1  
cluster-master | Starting ResourceManager  
cluster-master | Starting nodemanagers  
cluster-slave-2 | ssh: Could not resolve hostname cluster-slave-2: Temporary failure in name resolution  
cluster-slave-3 | ssh: Could not resolve hostname cluster-slave-3: Temporary failure in name resolution  
cluster-slave-4 | ssh: Could not resolve hostname cluster-slave-4: Temporary failure in name resolution  
cluster-slave-5 | ssh: Could not resolve hostname cluster-slave-5: Temporary failure in name resolution  
cluster-master | 400 org.apache.hadoop.hdfs.server.namenode.SecondaryNameNode  
164 org.apache.hadoop.hdfs.server namenode.NameNode  
cluster-master | 101 jdk.jcmd/run.tools.java.ps -ln  
484 org.apache.hadoop.mapreduce.v2.hs.JobHistoryServer  
cluster-master | 633 org.apache.hadoop.yarn.server.ResourceManager  
cluster-master | Configured Capacity: 265286293504 (191.19 GB)  
cluster-master | Present Capacity: 4494175692 (41.86 GB)  
cluster-master | DFS Remaining: 4494175696 (41.86 GB)  
cluster-master | DFS Used: 24576 (24 KB)  
cluster-master | DFS Used%: 0.00%  
cluster-master | Replicated blocks:  
cluster-master | Under replicated blocks: 0  
cluster-master | Blocks with corrupt replicas: 0  
cluster-master | Missing blocks: 0  
cluster-master | Missing blocks (with replication factor 1): 0  
cluster-master | Low redundancy blocks with highest priority to recover: 0  
cluster-master | Pending deletion blocks: 0  
cluster-master | Erasure Coded Block Groups:  
cluster-master | Low redundancy block groups: 0  
cluster-master | Block groups with corrupt internal blocks: 0  
cluster-master | Missing block groups: 0  
cluster-master | Low redundancy blocks with highest priority to recover: 0  
cluster-master | Pending deletion blocks: 0  
-----  
cluster-master | Live datanodes (1):  
cluster-master | Name: 172.31.0.2:9866 (cluster-slave-1.bigdata2_spark-cluster)  
cluster-master | Hostname: cluster-slave-1  
cluster-master | Decommission Status : Normal  
cluster-master | Configured Capacity: 265286293504 (191.19 GB)  
cluster-master | DFS Used: 24576 (24 KB)  
cluster-master | Non DFS Used: 149842006016 (139.55 GB)  
cluster-master | DFS Remaining: 4494175696 (41.86 GB)  
cluster-master | DFS Used%: 0.00%  
cluster-master | DFS Remaining%: 21.89%  
cluster-master | Configured Cache Capacity: 0 (0 B)  
cluster-master | Cache Used: 0 (0 B)  
cluster-master | Cache Remaining: 0 (0 B)  
cluster-master | Cache Used%: 100.00%  
cluster-master | Cache Remaining%: 0.00%  
cluster-master | Xceivers: 0  
cluster-master | Last contact: Tue Apr 15 17:15:11 GMT 2025  
cluster-master | Last Block Report: Tue Apr 15 17:15:00 GMT 2025  
cluster-master | Num of Blocks: 0  
cluster-master | Info node-sc.DFS
```

wait

```
Activities Terminal 15 apr 20:38 en
user@userhp: ~/big_data/as2/v6/bigdata2
cassandra-server INFO [Native-Transport-Requests-7] 2025-04-15 17:16:52,055 ColumnFamilyStore.java:499 - Initializing my db, terms
cassandra-server INFO [NonPeriodicTasks:1] 2025-04-15 17:16:52,067 BigFormat.java:231 - Deleting sstable: /opt/cassandra/data/data/system_schema/tables-afdfb9dbdc136088056eedc302ba09/nb-7-big
cassandra-server INFO [CompactionExecutor:1] 2025-04-15 17:16:52,068 CompactionTask.java:258 - Compacted (6c2d0470-1a1d-11f0-bc5e-6d7ddc1cae5) 4 sstables to [/opt/cassandra/data/data/system_schema/tables-afdfb9dbdc136088056eedc302ba09/nb-9-big], to level=0, 4.836KB (0.79% of original) to 3.858KB (0.79% of original) in 167ms. Read throughput = 28.867KB/s, Write throughput = -58/s. 9 total partition merged to 6. Partition merge counts were [15, 41, ]. Time spent writing keys = 49ms
cassandra-server INFO [NonPeriodicTasks:1] 2025-04-15 17:16:52,071 BigFormat.java:231 - Deleting sstable: /opt/cassandra/data/data/system_schema/tables-afdfb9dbdc136088056eedc302ba09/nb-8-big
cassandra-server INFO [NonPeriodicTasks:1] 2025-04-15 17:16:52,075 BigFormat.java:231 - Deleting sstable: /opt/cassandra/data/data/system_schema/tables-afdfb9dbdc136088056eedc302ba09/nb-5-big
cassandra-server INFO [NonPeriodicTasks:1] 2025-04-15 17:16:52,079 BigFormat.java:231 - Deleting sstable: /opt/cassandra/data/data/system_schema/tables-afdfb9dbdc136088056eedc302ba09/nb-6-big
cluster-master install a.parquet
cluster-master - 2025-04-15 17:16:52-- https://storage.googleapis.com/kaggle-data-sets/3521629/6146260/compressed/a.parquet.zip?X-Goog-Algorithm=GOOG4-RSA-SHA256&X-Goog-Credential=kaggle-com%40kaggle-1916071-iam.gserviceaccount.com%2F20250415%2Fauto%2Fstorage.googleapis.com%2Fgoog4-requests&X-Goog-Date=20250415T190154Z&X-Goog-Expires=252000&X-Goog-SignedHeaders=host&X-Goog-Signature=6ba3c109ccf0d750c683b3c7e4a56e18d40b11bb4f8aac5a66a3eb509c13f435f50bb38ed27b795870d9f725e30542c37bf340b3b9808f610af071eaddc499be707f12b064f1a7d3d13708d1d585807a083c020701ae301521ce4816767c0aad19ca5de4dc92a3330989d034ae99cfd5f7f653bed1af53356e4e27748e47f4eb593c4508f0e65d2d2d3e3a65d104c9b7892604cbaca0b2cf512c7b77b789c277b96f3476146f58936528a8a03c51e134e5b19b0ecf8a2dc2d2f0f460cf668b0964a71b39c6e4737832b8cf8a7e68e393bbcb1a6b19c8b45e3a0104e5d00b0b1b3880120f93eaa4b7de4b3c4f002b0b481
cluster-master Resolving storage.googleapis.com (storage.googleapis.com)... 64.233.165.207, 108.177.14.207, 209.85.233.207, ...
cluster-master Connecting to storage.googleapis.com (storage.googleapis.com)[64.233.165.207]:443... connected.
cluster-master HTTP request sent, awaiting response... 200 OK
cluster-master Length: 76835361 (73M) [application/zip]
cluster-master Saving to: '/app/data.zip'
cluster-master 100%[=====] 732.76M 31.9MB/s in 24s
cluster-master 2025-04-15 17:17:17 (31.0 MB/s) - '/app/data.zip' saved [768353361/768353361]
cluster-master Extracting file a.parquet ...
cluster-master Archive: /app/data.zip
cluster-master Inflating: /app/a.parquet
cluster-master clear local dir
cluster-master hdfs mkdir and putting
cluster-master Starting spark-submit
cluster-master 25/04/15 17:17:39 INFO SparkContext: Running Spark version 3.5.4
cluster-master 25/04/15 17:17:39 INFO SparkContext: OS info Linux, 6.8.0-52-generic, amd64
cluster-master 25/04/15 17:17:39 INFO SparkContext: Java version 1.8.0_442
cluster-master 25/04/15 17:17:39 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
cluster-master 25/04/15 17:17:39 INFO ResourceUtils: Subsequent applications: data preparation
cluster-master 25/04/15 17:17:39 INFO ResourceUtils: No custom resources configured for spark.driver.
cluster-master 25/04/15 17:17:39 INFO ResourceUtils: =====
cluster-master 25/04/15 17:17:39 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 1024, scr
cluster-master tpt: , vendor: , offheap -> name: offheap, amount: 0, script: , vendor: ), task resources: Map(cpu -> name: cpu, amount: 1.0)
cluster-master 25/04/15 17:17:40 INFO ResourceProfile: Listing resource is cpu
cluster-master 25/04/15 17:17:40 INFO ResourceProfileManager: Added ResourceProfile id: 0
cluster-master 25/04/15 17:17:40 INFO SecurityManager: Changing view acls to: root
cluster-master 25/04/15 17:17:40 INFO SecurityManager: Changing modify acls to: root
cluster-master 25/04/15 17:17:40 INFO SecurityManager: Changing view acls groups to:
cluster-master 25/04/15 17:17:40 INFO SecurityManager: Changing modify acls groups to:
cluster-master 25/04/15 17:17:40 INFO SecurityManager: SecurityManager: authentication disabled; ut acls disabled; users with view permissions: root; groups with view permissions: EMPT; users with modify p
cluster-master ermissions: root; groups with modify permissions: EMPT;
cluster-master 25/04/15 17:17:40 INFO Utils: Successfully started service 'sparkDriver' on port 43073.
cluster-master 25/04/15 17:17:40 INFO SparkEnv: Registering MapOutputTracker
cluster-master 25/04/15 17:17:40 INFO SparkEnv: Registering BlockManagerMaster
cluster-master 25/04/15 17:17:40 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
cluster-master 25/04/15 17:17:40 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
cluster-master 25/04/15 17:17:40 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
```

a.parquet downloading

```
Activities Terminal 15 apr 20:39 en
user@userhp: ~/big_data/as2/v6/bigdata2
cluster-master 25/04/15 17:17:57 INFO DRGOScheduler: Job 5 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master 25/04/15 17:17:57 INFO TaskSchedulerImpl: Killing all running tasks in stage 8: stage finished
cluster-master 25/04/15 17:17:57 INFO FileFormatWriter: Job 5 finished: save at NativeMethodAccessorImpl.java:5: took 0.381933 s
cluster-master 25/04/15 17:17:57 INFO FileFormatWriter: Start to commit write Job 01872c04-5343-4c41-9f58-895c045145d0
cluster-master 25/04/15 17:17:57 INFO FileFormatWriter: Write Job 01872c04-5343-4c41-9f58-895c045145d0 committed. Elapsed time: 27 ms.
cluster-master 25/04/15 17:17:57 INFO FileFormatWriter: Job 5 finished: processing stats for write Job 01872c04-5343-4c41-9f58-895c045145d0.
cluster-master 25/04/15 17:17:57 INFO SparkContext: Invoking stop() from shutdown hook
cluster-master 25/04/15 17:17:57 INFO SparkContext: SparkContext is stopping with exitcode 0.
cluster-master 25/04/15 17:17:57 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master 25/04/15 17:17:57 INFO BlockManagerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master 25/04/15 17:17:57 INFO MemoryStore: MemoryStore cleared
cluster-master 25/04/15 17:17:57 INFO BlockManager: BlockManager stopped
cluster-master 25/04/15 17:17:57 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master 25/04/15 17:17:57 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master 25/04/15 17:17:57 INFO SparkContext: Successfully stopped SparkContext
cluster-master 25/04/15 17:17:57 INFO ShutdownHookManager: Shutdown hook called
cluster-master 25/04/15 17:17:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-3ab29898-d7c3-4dfe-b577-fbfda6a60bd/pyspark-95aff81c-8977-4d1a-9884-f5e4503c2db8
cluster-master 25/04/15 17:17:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-aff00502-8f9f-4d9c-938d-cdd1784dc01d
cluster-master 25/04/15 17:17:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-3ab29898-d7c3-4dfe-b577-fbfda6a60bd
cluster-master Putting data to hdfs
cluster-master drwxr-xr-x - root supergroup 0 2025-04-15 17:22 /data/data
cluster-master -rw-r--r-- 1 root supergroup 3284 2025-04-15 17:21 /data/data/10031136_A Decade in the Grave.txt
cluster-master -rw-r--r-- 1 root supergroup 529 2025-04-15 17:18 /data/data/10078432_A Case For the Court.txt
cluster-master -rw-r--r-- 1 root supergroup 616 2025-04-15 17:22 /data/data/10099975_A Different Light (album).txt
cluster-master -rw-r--r-- 1 root supergroup 647 2025-04-15 17:20 /data/data/10137549_A Good Thief Tips His Hat.txt
cluster-master -rw-r--r-- 1 root supergroup 591 2025-04-15 17:19 /data/data/10174562_A History of Money and Banking in the United States.txt
cluster-master -rw-r--r-- 1 root supergroup 1414 2025-04-15 17:19 /data/data/10223157_A Balinese Trance Seance.txt
cluster-master -rw-r--r-- 1 root supergroup 31874 2025-04-15 17:20 /data/data/1022877_A Death in the Family (comics).txt
cluster-master -rw-r--r-- 1 root supergroup 814 2025-04-15 17:21 /data/data/10236865_A Dead Sinking Story.txt
cluster-master -rw-r--r-- 1 root supergroup 210 2025-04-15 17:21 /data/data/10254892_A Flat Man.txt
cluster-master -rw-r--r-- 1 root supergroup 9861 2025-04-15 17:22 /data/data/10381993_A Doll's House (1973 Losey film).txt
cluster-master -rw-r--r-- 1 root supergroup 16918 2025-04-15 17:22 /data/data/1039311_A Hero of Our Time.txt
cluster-master -rw-r--r-- 1 root supergroup 5718 2025-04-15 17:20 /data/data/10399316_A Flowering Tree.txt
cluster-master -rw-r--r-- 1 root supergroup 2435 2025-04-15 17:18 /data/data/10534790_A Black and White World.txt
cluster-master -rw-r--r-- 1 root supergroup 1180 2025-04-15 17:20 /data/data/10570204_A Gun Called Tension.txt
cluster-master -rw-r--r-- 1 root supergroup 16809 2025-04-15 17:19 /data/data/1067091_A Hard Day's Night (song).txt
cluster-master -rw-r--r-- 1 root supergroup 1098 2025-04-15 17:20 /data/data/1083442_A Hillbilly Tribute to ACDC.txt
cluster-master -rw-r--r-- 1 root supergroup 1745 2025-04-15 17:18 /data/data/10849680_A Day in the Death of Donny B.txt
cluster-master -rw-r--r-- 1 root supergroup 6764 2025-04-15 17:20 /data/data/10858097_A Dangerous Path.txt
cluster-master -rw-r--r-- 1 root supergroup 82157 2025-04-15 17:20 /data/data/10890703_A Dictionary of Canadianisms on Historical Principles.txt
cluster-master -rw-r--r-- 1 root supergroup 2806 2025-04-15 17:19 /data/data/11017293_A Bad Spell in Yurt.txt
cluster-master -rw-r--r-- 1 root supergroup 4423 2025-04-15 17:20 /data/data/11017589_A Doctor's Report on Dianetics.txt
cluster-master -rw-r--r-- 1 root supergroup 923 2025-04-15 17:18 /data/data/1114641_A Blueprint of the World.txt
cluster-master -rw-r--r-- 1 root supergroup 2573 2025-04-15 17:18 /data/data/1115810_A Hanging.txt
cluster-master -rw-r--r-- 1 root supergroup 12171 2025-04-15 17:20 /data/data/1121270_A Lesson in Romanitics.txt
cluster-master -rw-r--r-- 1 root supergroup 588 2025-04-15 17:21 /data/data/11315857_A Go Go (Potsdott album).txt
cluster-master -rw-r--r-- 1 root supergroup 333 2025-04-15 17:20 /data/data/11498217_A Guide to Groovy Lovin'.txt
cluster-master -rw-r--r-- 1 root supergroup 5461 2025-04-15 17:18 /data/data/11528779_A Dreamer's Tales.txt
cluster-master -rw-r--r-- 1 root supergroup 2529 2025-04-15 17:20 /data/data/11631735_A Ballad of the West.txt
cluster-master -rw-r--r-- 1 root supergroup 1049 2025-04-15 17:22 /data/data/11735053_A Journal of the Plague Year (album).txt
cluster-master -rw-r--r-- 1 root supergroup 597 2025-04-15 17:17 /data/data/11871420_A Lifetime or More.txt
cluster-master -rw-r--r-- 1 root supergroup 2134 2025-04-15 17:18 /data/data/11892274_A Cold Night's Death.txt
cluster-master -rw-r--r-- 1 root supergroup 863 2025-04-15 17:18 /data/data/11930321_A Fragile Hope.txt
cluster-master -rw-r--r-- 1 root supergroup 6843 2025-04-15 17:20 /data/data/11984610_A Catalogue of Crime.txt
```

```
Activities Terminal 15 apr 20:39
user@userhp: ~/big_data/as2/v6/bigdata2
cluster-master ~-rw-r--r-- 1 root supergroup 616 2025-04-15 17:21 /data/data/9947241_A_Day_of_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-b4b0-485e-839a-2cfe66e3789-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-unjar583330649383267290/] [/tmp/streamjob589455271482609456.jar tmpDir=null]
cluster-master 2025-04-15 17:22:47,148 INFO client.DefaultHMRHFAOverProxyProvider: Connecting to ResourceManager at cluster-master/172.31.0.4:8032
cluster-master 2025-04-15 17:22:47,379 INFO client.DefaultHMRHFAOverProxyProvider: 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 17:22:52,399 INFO mapreduce.Job: The url to track the job: http://cluster-master:8080/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 17:23:00,588 INFO mapreduce.Job: Job job_1744737308040_0001 running in user mode : false
cluster-master 2025-04-15 17:23:00,590 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 17:23:33,821 INFO mapreduce.Job: map 100% reduce 68%
cluster-master 2025-04-15 17:23:59,938 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 17:24:21,119 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 17:25:03,361 INFO mapreduce.Job: map 100% reduce 77%
cluster-master 2025-04-15 17:25:06,393 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 17:25:55,620 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 apr 20:40
user@userhp: ~/big_data/as2/v6/bigdata2
cluster-master ~-rw-r--r-- 1 root supergroup 616 2025-04-15 17:21 /data/data/9947241_A_Day_of_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-b4b0-485e-839a-2cfe66e3789-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-unjar583330649383267290/] [/tmp/streamjob589455271482609456.jar tmpDir=null]
cluster-master 2025-04-15 17:22:47,148 INFO client.DefaultHMRHFAOverProxyProvider: Connecting to ResourceManager at cluster-master/172.31.0.4:8032
cluster-master 2025-04-15 17:22:47,379 INFO client.DefaultHMRHFAOverProxyProvider: 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 17:22:52,399 INFO mapreduce.Job: The url to track the job: http://cluster-master:8080/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 17:23:00,588 INFO mapreduce.Job: Job job_1744737308040_0001 running in user mode : false
cluster-master 2025-04-15 17:23:00,590 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 17:23:33,821 INFO mapreduce.Job: map 100% reduce 68%
cluster-master 2025-04-15 17:23:59,938 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 17:24:21,119 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 17:25:03,361 INFO mapreduce.Job: map 100% reduce 77%
cluster-master 2025-04-15 17:25:06,393 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 17:25:55,620 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
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-1:133079 (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-1:142845 (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) (3/6)
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: 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 (collectMap 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: collectMap at /app/query.py:39, took 0.465467 s
cluster-master Top 10 Results:
cluster-master 1. 30828228 A Human Right
cluster-master 2. 18171842 A Chrestomathy
cluster-master 3. 45303168 A Bearded Man
cluster-master 4. 16378814 A Breathing 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 Game 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 YarnSchedulerBackendYarnDriverEndpoint: 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-4f38-9320-911dd998707
cluster-master 25/04/15 17:31:00 INFO ShutdownHookManager: Deleting directory /tmp/spark-5e56582f-6d03-4eda-b093-0ce3f9ecf56f/pyspark-6c0bc7c19-b69f-4ae5-a17d-a347815dd616
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"

```
Activities Terminal 15 apr 21:08 root@cluster-master: /app
user@userhp: ~/big_data/as2/v6/bigdata2
user@userhp:~/big_data/as2/v6/bigdata2$ docker exec -it cluster-master bash
root@cluster-master:/app# "C
root@cluster-master:/app# bash search.sh "American films of 1916"
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-jar1/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.sparkspark-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-45e89838-d6b5-4666-9cbf-42f303c52617;1.0
confs: [default]
found com.datastax.sparkspark-cassandra-connector_2.12:3.2.0 in central
found com.datastax.sparkspark-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#jnr-ffi: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 89ms :: artifacts dl 4ms
:: 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]
com.datastax.oss#java-driver-query-builder:4.13.0 from central in [default]
com.datastax.oss#java-driver-shaded-guava:25.1-jre-graal-sub-1 from central in [default]
com.datastax.oss#native-protocol:1.5.0 from central in [default]
com.datastax.sparkspark-cassandra-connector-driver_2.12:3.2.0 from central in [default]
com.datastax.sparkspark-cassandra-connector_2.12:3.2.0 from central in [default]
com.github.jnr#jnr-ffi:1.3.9 from central in [default]
com.github.jnr#jnr-a64asm:1.0.0 from central in [default]
com.github.jnr#jnr-constants:0.10.3 from central in [default]
com.github.jnr#jnr-ffi:2.2.11 from central in [default]
com.github.jnr#jnr-x86asm:1.0.2 from central in [default]
com.ow2.asm#asm:9.2 from central in [default]
com.ow2.asm#asm-analysis:9.2 from central in [default]
com.ow2.asm#asm-commons:9.2 from central in [default]
com.ow2.asm#asm-tree:9.2 from central in [default]
com.ow2.asm#asm-util:9.2 from central in [default]
com.thoughtworks.paranamer#paranamer:2.8 from central in [default]
org.scala-lang#scala-reflect:2.12.11 from central in [default]
org.reactivestreams#reactive-streams:1.0.3 from central in [default]
org.hdrhistogram#hdrhistogram:2.1.12 from central in [default]
io.dropwizard.metrics#metrics-core:4.1.18 from central in [default]
org.slf4j#slf4j-api:1.7.26 from central in [default]
typesafe.config:1.4.1 from central in [default]
```

```
Activities Terminal 15 apr 21:37 user@userhp: ~/big_data/as2/v6/bigdata2
user@userhp:~/big_data/as2/v6/bigdata2$
cluster-master 25/04/15 18:36:56 INFO TaskSetManager: Starting task 0.0 in stage 12.0 (TID 23) (cluster-slave-1, executor 2, partition 0, RACK_LOCAL, 11040 bytes)
cluster-master 25/04/15 18:36:56 INFO TaskSetManager: Starting task 1.0 in stage 12.0 (TID 24) (cluster-slave-1, executor 1, partition 1, RACK_LOCAL, 11040 bytes)
cluster-master 25/04/15 18:36:56 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on cluster-slave-1:339077 (size: 13.5 KiB, free: 365.4 MiB)
cluster-master 25/04/15 18:36:56 INFO TaskSetManager: Starting task 2.0 in stage 12.0 (TID 25) (cluster-slave-1, executor 2, partition 2, RACK_LOCAL, 11160 bytes)
cluster-master 25/04/15 18:36:56 INFO TaskSetManager: Finished task 0.0 in stage 12.0 (TID 23) in 158 ms on cluster-slave-1 (executor 2) (1/7)
cluster-master 25/04/15 18:36:56 INFO TaskSetManager: Starting task 3.0 in stage 12.0 (TID 26) (cluster-slave-1, executor 1, partition 3, RACK_LOCAL, 11160 bytes)
cluster-master 25/04/15 18:36:56 INFO TaskSetManager: Finished task 1.0 in stage 12.0 (TID 24) in 165 ms on cluster-slave-1 (executor 1) (2/7)
cluster-master 25/04/15 18:36:57 INFO TaskSetManager: Starting task 4.0 in stage 12.0 (TID 27) (cluster-slave-1, executor 2, partition 4, RACK_LOCAL, 11040 bytes)
cluster-master 25/04/15 18:36:57 INFO TaskSetManager: Starting task 5.0 in stage 12.0 (TID 28) (cluster-slave-1, executor 1, partition 5, RACK_LOCAL, 11040 bytes)
cluster-master 25/04/15 18:36:57 INFO TaskSetManager: Finished task 3.0 in stage 12.0 (TID 26) in 123 ms on cluster-slave-1 (executor 1) (3/7)
cluster-master 25/04/15 18:36:57 INFO TaskSetManager: Finished task 2.0 in stage 12.0 (TID 25) in 133 ms on cluster-slave-1 (executor 2) (4/7)
cluster-master 25/04/15 18:36:57 INFO TaskSetManager: Starting task 6.0 in stage 12.0 (TID 29) (cluster-slave-1, executor 2, partition 6, RACK_LOCAL, 11036 bytes)
cluster-master 25/04/15 18:36:57 INFO TaskSetManager: Finished task 4.0 in stage 12.0 (TID 27) in 123 ms on cluster-slave-1 (executor 2) (5/7)
cluster-master 25/04/15 18:36:57 INFO TaskSetManager: Finished task 5.0 in stage 12.0 (TID 28) in 118 ms on cluster-slave-1 (executor 1) (6/7)
cluster-master 25/04/15 18:36:57 INFO TaskSetManager: Finished task 6.0 in stage 12.0 (TID 29) in 108 ms on cluster-slave-1 (executor 2) (7/7)
cluster-master 25/04/15 18:36:57 INFO YarnScheduler: Removed TaskSet 12.0, whose tasks have all completed, from pool
cluster-master 25/04/15 18:36:57 INFO DAGScheduler: ResultStage 12 (collect$map at /app/query.py:39) finished in 0.569 s
cluster-master 25/04/15 18:36:57 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master 25/04/15 18:36:57 INFO YarnScheduler: Killing all running tasks in stage 12: Stage finished
cluster-master 25/04/15 18:36:57 INFO DAGScheduler: Job 0 finished: collect$map at /app/query.py:39, took 0.517484 s
cluster-master
cluster-master Top 10 Results:
cluster-master 1. 12955622 A Day at School
cluster-master 2. 13478922 A Knight of the Range
cluster-master 3. 44853014 A Gutter Magdalene
cluster-master 4. 25993670 A Kiss for Cinderella (film)
cluster-master 5. 40474725 A Corner in Cotton
cluster-master 6. 45681561 A Fair Impostor (novel)
cluster-master 7. 45681521 A Fair Impostor
cluster-master 8. 1356924 A Child's Garden of Verses
cluster-master 9. 56880998 A J Balliol Saloon
cluster-master 10. 15670656 A Final Reckoning
cluster-master 25/04/15 18:36:57 INFO SparkContext: SparkContext is stopping with exitcode 0.
cluster-master 25/04/15 18:36:57 INFO SparkUI: stopped spark web UI at http://cluster-master:4040
cluster-master 25/04/15 18:36:57 INFO YarnClientSchedulerBackend: Interrupting monitor thread
cluster-master 25/04/15 18:36:57 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster-master 25/04/15 18:36:57 INFO YarnClientSchedulerBackend: Asking each executor to shut down
cluster-master 25/04/15 18:36:57 INFO YarnClientSchedulerBackend: YARN client scheduler backend stopped
cluster-master 25/04/15 18:36:57 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master 25/04/15 18:36:57 INFO MemoryStore: MemoryStore cleared
cluster-master 25/04/15 18:36:57 INFO BlockManager: BlockManager stopped
cluster-master 25/04/15 18:36:57 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master 25/04/15 18:36:57 INFO OutputCommitCoordinatorSubOutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master 25/04/15 18:36:57 INFO SparkContext: Successfully stopped SparkContext
cluster-master 25/04/15 18:36:57 INFO ShutdownHookManager: Shutdown hook called
cluster-master 25/04/15 18:36:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-100abcec-390c-41e3-9a2b-329753f62fff/pyspark-54e258ae-94b0-47aa-a305-b030396c8738
cluster-master 25/04/15 18:36:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-7a8ba29a-c8e2-ac2c-b6a4-26af876ed319
cluster-master 25/04/15 18:36:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-100abcec-390c-41e3-9a2b-329753f62fff
cluster-master 25/04/15 18:36:58 INFO CassandraConnector: Disconnected from Cassandra cluster.
cluster-master 25/04/15 18:36:58 INFO SerialShutdownHooks: Successfully executed shutdown hook: Clearing session cache for C* connector
cluster-master Services started. Holding the container open...
```

the first doc contains this line! its work!

bash search.sh "films about the future"

```
Activities Terminal 15 apr 21:42 en root@cluster-master: /app
user@userhp: ~/big_data/as2/v6/bigdata2$ docker exec -it cluster-master bash
root@cluster-master: /app# bash search.sh "films about the future"
This script will include commands to search for documents given the query using Spark RDD
:: loading settings :: url = jar:file:/usr/local/spark/jars/Livy-2.5.1-jar1/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
con.datastax.sparkspark-cassandra-connector_2.12 added as a dependency
con.github.jnr#jnr-posix added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-f58aadd3-8a64-40b0-b14f-b610b978cdd3;1.0
  confs: [default]
  found con.datastax.sparkspark-cassandra-connector_2.12;3.2.0 in central
  found con.datastax.sparkspark-cassandra-connector-driver_2.12;3.2.0 in central
  found con.datastax.oss#java-driver-core-shaded;4.13.0 in central
  found con.datastax.oss#native-protocol;1.5.0 in central
  found con.datastax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 in central
  found con.typesafeconfig;1.4.1 in central
  found org.slf4j#slf4j-api;1.7.26 in central
  found io.dropwizard.metrics#metrics-core;1.18 in central
  found org.hdrhistogram#HdrHistogram;2.1.12 in central
  found org.reactivestreams#reactive-streams;1.0.3 in central
  found con.github.stephenc.jcip#jcip-annotations;1.0-1 in central
  found con.github.spotbugs#spotbugs-annotations;1.12 in central
  found con.google.code.findbugs#jsr305;3.0.2 in central
  found con.datastax.oss#java-driver-mapper-runtime;4.13.0 in central
  found con.datastax.oss#java-driver-query-builder;4.13.0 in central
  found org.apache.commons#commons-lang3;3.10 in central
  found con.thoughtworks.paranamer#paranamer;2.8 in central
  found org.scala-lang#scala-reflect;2.12.11 in central
  found con.github.jnr#jnr-posix;3.1.15 in central
  found con.github.jnr#jnr-ffi;2.2.11 in central
  found con.github.jnr#jnr-ffi;1.3.9 in central
  found org.ow2.asm#asm;5.2 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 con.github.jnr#jnr-a64asm;1.0.0 in central
  found con.github.jnr#jnr-x86asm;1.0.2 in central
  found con.github.jnr#jnr-constants;0.16.3 in central
  found con.github.jnr#jnr-posix;3.1.15 in central
  resolution report :: resolve 822ms :: artifacts dl 37ms
  :: modules in use:
    con.datastax.oss#java-driver-core-shaded;4.13.0 from central in [default]
    con.datastax.oss#java-driver-mapper-runtime;4.13.0 from central in [default]
    con.datastax.oss#java-driver-query-builder;4.13.0 from central in [default]
    con.datastax.oss#native-protocol;1.5.0 from central in [default]
    con.datastax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 from central in [default]
    con.datastax.sparkspark-cassandra-connector-driver_2.12;3.2.0 from central in [default]
    con.datastax.sparkspark-cassandra-connector_2.12;3.2.0 from central in [default]
    con.github.jnr#jnr-ffi;1.3.9 from central in [default]
    con.github.jnr#jnr-a64asm;1.0.0 from central in [default]
    con.github.jnr#jnr-constants;0.16.3 from central in [default]
    con.github.jnr#jnr-ffi;2.2.11 from central in [default]
    con.github.jnr#jnr-posix;3.1.15 from central in [default]
    con.github.jnr#jnr-x86asm;1.0.2 from central in [default]
```

```
Activities Terminal 15 apr 21:42 en root@cluster-master: /app
user@userhp: ~/big_data/as2/v6/bigdata2$
25/04/15 18:41:52 INFO YarnScheduler: Adding task set 12.0 with 7 tasks resource profile 0
25/04/15 18:41:52 INFO TaskSetManager: Starting task 0.0 in stage 12.0 (TID 23) (cluster-slave-1, executor 2, partition 0, RACK_LOCAL, 11160 bytes)
25/04/15 18:41:52 INFO TaskSetManager: Starting task 1.0 in stage 12.0 (TID 24) (cluster-slave-1, executor 1, partition 1, RACK_LOCAL, 10920 bytes)
25/04/15 18:41:52 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on cluster-slave-1:33545 (size: 13.5 KiB, free: 365.5 MiB)
25/04/15 18:41:52 INFO BlockManagerInfo: Added broadcast_12_piece0 in memory on cluster-slave-1:33922 (size: 13.5 KiB, free: 365.5 MiB)
25/04/15 18:41:52 INFO TaskSetManager: Starting task 2.0 in stage 12.0 (TID 25) (cluster-slave-1, executor 1, partition 2, RACK_LOCAL, 11160 bytes)
25/04/15 18:41:52 INFO TaskSetManager: Finished task 1.0 in stage 12.0 (TID 24) in 125 ms on cluster-slave-1 (executor 1) (1/7)
25/04/15 18:41:52 INFO TaskSetManager: Starting task 3.0 in stage 12.0 (TID 26) (cluster-slave-1, executor 2, partition 3, RACK_LOCAL, 11040 bytes)
25/04/15 18:41:52 INFO TaskSetManager: Finished task 0.0 in stage 12.0 (TID 23) in 140 ms on cluster-slave-1 (executor 2) (2/7)
25/04/15 18:41:52 INFO TaskSetManager: Starting task 4.0 in stage 12.0 (TID 27) (cluster-slave-1, executor 2, partition 4, RACK_LOCAL, 11036 bytes)
25/04/15 18:41:52 INFO TaskSetManager: Finished task 3.0 in stage 12.0 (TID 26) in 110 ms on cluster-slave-1 (executor 2) (3/7)
25/04/15 18:41:52 INFO TaskSetManager: Starting task 5.0 in stage 12.0 (TID 28) (cluster-slave-1, executor 1, partition 5, RACK_LOCAL, 11040 bytes)
25/04/15 18:41:52 INFO TaskSetManager: Finished task 2.0 in stage 12.0 (TID 25) in 134 ms on cluster-slave-1 (executor 1) (4/7)
25/04/15 18:41:52 INFO TaskSetManager: Starting task 6.0 in stage 12.0 (TID 29) (cluster-slave-1, executor 2, partition 6, RACK_LOCAL, 11160 bytes)
25/04/15 18:41:52 INFO TaskSetManager: Finished task 4.0 in stage 12.0 (TID 27) in 94 ms on cluster-slave-1 (executor 2) (5/7)
25/04/15 18:41:52 INFO TaskSetManager: Finished task 5.0 in stage 12.0 (TID 28) in 90 ms on cluster-slave-1 (executor 1) (6/7)
25/04/15 18:41:52 INFO TaskSetManager: Finished task 6.0 in stage 12.0 (TID 29) in 115 ms on cluster-slave-1 (executor 2) (7/7)
25/04/15 18:41:52 INFO YarnScheduler: Removed TaskSet 12.0, whose tasks have all completed, from pool.
25/04/15 18:41:52 INFO DAGScheduler: ResultStage 12 (collectTaskMap at /app/query.py:39) finished in 0.457 s
25/04/15 18:41:52 INFO DAGScheduler: Job 8 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/15 18:41:52 INFO YarnScheduler: Killing all running tasks in stage 12: Stage finished
25/04/15 18:41:52 INFO DAGScheduler: Job 8 finished: collectTaskMap at /app/query.py:39, took 0.465573 s

Top 10 Results:
1. 7868751 A Christmas Carol (2006 film)
2. 6622869 A Is for Atom
3. 51573838 A Happy Day of Jlnsa Maeng
4. 15287294 A Day at the Zoo
5. 37670240 A Date with the Falcon
6. 48920674 A House Built on Water
7. 39388324 A Field in England
8. 26995523 A Defeated Peace
9. 19749366 A Hundred Things Keep Me Up at Night
10. 10381993 A Doll's House (1973 Losey film)
25/04/15 18:41:52 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/15 18:41:52 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/15 18:41:52 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/15 18:41:53 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/15 18:41:53 INFO YarnClientSchedulerBackend: YARN client scheduler backend stopped
25/04/15 18:41:53 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/15 18:41:53 INFO MemoryStore: MemoryStore cleared
25/04/15 18:41:53 INFO BlockManager: BlockManager stopped
25/04/15 18:41:53 INFO BlockManagerMaster: BlockManagerMaster stopped
25/04/15 18:41:53 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/15 18:41:53 INFO SparkContext: Successfully stopped SparkContext
25/04/15 18:41:53 INFO ShutdownHookManager: Shutdown hook called
25/04/15 18:41:53 INFO ShutdownHookManager: Deleting directory /tmp/spark-9ba181c0-f549-4d99-a7e4-4eb7bb3e58e8
25/04/15 18:41:53 INFO ShutdownHookManager: Deleting directory /tmp/spark-9ba181c0-f549-4d99-a7e4-4eb7bb3e58e8/pyspark-0eac0157-23dc-4df9-9bb5-1258c21e5cfa
25/04/15 18:41:53 INFO ShutdownHookManager: Deleting directory /tmp/spark-358b7f09-9779-4e2e-b021-baxe93a80237
25/04/15 18:41:54 INFO CassandraConnector: Disconnected from Cassandra cluster.
25/04/15 18:41:54 INFO SerialShutdownHooks: Successfully executed shutdown hook: Clearing session cache for C* connector
root@cluster-master: /app#
```

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.