

BigData Assignment 2 report

April 2025

Assignment goal is to create a full-text search engine that can index a collection of documents in large volume and generate appropriate ranked responses for user searches. The steps are:

Data Preparation: Spark is employed to preprocess a massive Parquet dataset by document sampling and document conversion to plain text files.

Indexing: Hadoop MapReduce programs tokenized documents, calculate term frequencies/positions, compute document frequency, and extract document metadata.

Data Persistence: Processed index data is loaded into Apache Cassandra. I created three tables: term_document, documents_info and document_frequency.

1. table `term_document` - contains term frequency in particular document and positions in symbols for each term in the document:

```
CREATE TABLE IF NOT EXISTS term_document (
  term text,
  doc_id text,
  tf int,
  positions list<int>,
  PRIMARY KEY (term, doc_id)
)
```

2. table `documents_info` - has documents id, titles and length of words in document:

```
CREATE TABLE IF NOT EXISTS documents_info (
  doc_id text PRIMARY KEY,
  title text,
  length int
)
```

3. table `document_frequency` - the same as vocabulary of terms, also include number of documents which has particular term:

```
CREATE TABLE IF NOT EXISTS document_frequency (
  term text PRIMARY KEY,
  df int
)
```

Query Processing: Query Processing: Apache Spark is used to process user queries to aggregate BM25 scores.

Containerization: Containerization: Docker Compose establishes a multi-container environment for nodes of the Spark cluster and for the Cassandra server.

1 Methodology

1.1 Data Preparation

The preparation of data is carried out in `prepare_data.py` and `prepare_data.sh`. It loads a Parquet file using Apache Spark, from which a sample of documents is obtained.

Each document title is normalized by spaces, because some of them contained `\n` or `\t` symbols that will break the division in future steps (since we divide `doc_id`, `title` and `text` using `\t` symbol in csv file).

After that we specify the filenames in the format `<doc_id>_<title>.txt`. We put the text to the file and save it in data folder in hdfs.

Additionally, we save `.tsv` file that contain rows with documents in the following format `<doc_id \t title \t text>`

1.2 Indexing using MapReduce pipelines

I have 3 pairs of MapReduce files to perform the indexing:

Job 1: Term Frequencies and Positions

Mapper: (`mapper1.py`) Tokenized the document content and output each token along with its document ID and positional index.

Reducer: (`reducer1.py`) I grouped tokens by term and document ID, aggregated their positions, and output the term, document ID, term frequency, and a comma-separated list of positions.

Job 2: Document Frequencies

Mapper: (`mapper2.py`) Extract the term and document id from Job 1 output.

Reducer: (`reducer2.py`) Aggregates unique document ids per term and outputs the document frequency.

Job 3: Document Metadata

Mapper: (`mapper3.py`) Processes each document to calculate its length (total word or token count) and outputs the metadata (document id, title, length). To tokenize we convert text to lowercase, remove special characters, and split it by whitespaces.

Reducer: (`reducer3.py`) Passes the metadata output directly.

1.3 Data Import into Apache Cassandra

In `app.py` I established a connection to the Cassandra cluster, created the keyspace and three required tables, read the MapReduce output from HDFS using shell commands and executed batch insertions into Cassandra. If a batch is too large, it falls back to individual insertions.

1.4 Query Processing and BM25 Ranking

During the query phase (`query.py`), a user's search query is handled through the following steps:

1. The query is first normalized by converting it to lowercase and splitting it into tokens using regular expressions.
2. For each token, the system retrieves the document frequency and term frequencies from Cassandra.
3. The BM25 relevance score is calculated using the formula:

$$\log \left(\max \left(1, \frac{\text{doc_count}}{\max(1, \text{df})} \right) \right)$$

$$score = idf \times \frac{tf \times (k1 + 1)}{tf + k1 \times \left(1 - b + b \times \frac{doc_length}{avg_doc_length}\right)}$$

where `idf` is the inverse document frequency, `tf` is the term frequency, and `doc_length` and `avg_doc_length` are used to adjust for document length. `k1` = 1.0 and `b` = 0.75

4. Spark RDDs are used to aggregate BM25 scores for each document.
5. Finally, the documents with the highest BM25 scores are returned, showing the document ID, title, and score for each.

1.5 Containerization with Docker Compose

The `docker-compose.yml` file orchestrates the multi-container setup:

- **Cluster Master:** Runs the Spark master node and executes the application orchestration script (`app.sh`).
- **Cluster Slaves:** Two Spark slave nodes enhance the distributed processing capabilities.
- **Cassandra Server:** Provides the distributed storage layer for the index.

All containers are connected over a dedicated network allowing seamless communication between Spark and Cassandra.

2 Demonstration

2.1 Running the System

To run the system, follow these steps:

1. Clone the repository and navigate to the root directory.
2. Ensure you have Docker and Docker Compose installed and at least 10 GB of RAM on your machine.
3. Place the Parquet file (e.g., `a.parquet`) in the `./app` folder. Also change the name of the parquet file in the following line of `app.sh`: `bash prepare_data.sh a.parquet` if you use another file
4. Start the containers by running: `docker-compose up`
5. If you want to run particular query, change it in `app.sh` in the following line: `bash search.sh "Your query"`

2.2 Screenshots

```
zauri@zauri: ~/Documents/developer/NN05_525/BO/assignment_2
cluster-master 25/04/15 14:26:52 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master 25/04/15 14:26:52 INFO OutputCommitCoordinatorSubOutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master 25/04/15 14:26:52 INFO SparkContext: Successfully stopped SparkContext
cluster-master 25/04/15 14:26:52 INFO ShutdownHookManager: Shutdown hook called
cluster-master 25/04/15 14:26:52 INFO ShutdownHookManager: Deleting directory /tmp/spark-88055fcf-011a-41ab-8750-4d57707dc4a2
cluster-master 25/04/15 14:26:52 INFO ShutdownHookManager: Deleting directory /tmp/spark-b4622547-82c6-4f71-a47e-bee6d49ca3ee
cluster-master 25/04/15 14:26:52 INFO ShutdownHookManager: Deleting directory /tmp/spark-b4622547-82c6-4f71-a47e-bee6d49ca3ee/pyspark-3c74c915-aa90-458f-af73-eta28a7bb8ab
cluster-master Preparing data.py completed
cluster-master Putting data to hdfs
cluster-master Data directory uploaded
cluster-master Found 1000 items
cluster-master -rw-r--r-- 1 root supergroup 3284 2025-04-15 14:30 /data/1003136 A Decade in the Grave.txt
cluster-master -rw-r--r-- 1 root supergroup 529 2025-04-15 14:32 /data/1007432 A Case for the Court.txt
cluster-master -rw-r--r-- 1 root supergroup 616 2025-04-15 14:31 /data/10099975 A Different Light_album.txt
cluster-master -rw-r--r-- 1 root supergroup 647 2025-04-15 14:28 /data/10137449 A Good Thief Tips His Hat.txt
cluster-master -rw-r--r-- 1 root supergroup 591 2025-04-15 14:28 /data/10174262 A History of Money and Banking in the United States.txt
cluster-master -rw-r--r-- 1 root supergroup 1414 2025-04-15 14:27 /data/10223157 A Balinese Trance Seance.txt
cluster-master -rw-r--r-- 1 root supergroup 31874 2025-04-15 14:30 /data/1022877 A Death in the Family_comics.txt
cluster-master -rw-r--r-- 1 root supergroup 814 2025-04-15 14:31 /data/10238885 A Dead Sinking Story.txt
cluster-master -rw-r--r-- 1 root supergroup 310 2025-04-15 14:29 /data/10254092 A Flat Man.txt
cluster-master -rw-r--r-- 1 root supergroup 9861 2025-04-15 14:27 /data/10303093 A Dolls House 1973 Losey_film.txt
cluster-master -rw-r--r-- 1 root supergroup 16918 2025-04-15 14:27 /data/1039311 A Hero of Our Time.txt
cluster-master -rw-r--r-- 1 root supergroup 5718 2025-04-15 14:28 /data/10399316 A Flowering Tree.txt
cluster-master -rw-r--r-- 1 root supergroup 2435 2025-04-15 14:32 /data/10534798 A Black and White World.txt
cluster-master -rw-r--r-- 1 root supergroup 1180 2025-04-15 14:31 /data/10570284 A Gun Called Tension.txt
cluster-master -rw-r--r-- 1 root supergroup 10800 2025-04-15 14:26 /data/1067603 A Hard Days Night_song.txt
cluster-master -rw-r--r-- 1 root supergroup 1098 2025-04-15 14:32 /data/1083442 A Hillbilly Tribute to ACDC.txt
cluster-master -rw-r--r-- 1 root supergroup 1745 2025-04-15 14:27 /data/10849600 A Day in the Death of Donny B.txt
cluster-master -rw-r--r-- 1 root supergroup 6764 2025-04-15 14:29 /data/10850997 A Dangerous Path.txt
cluster-master -rw-r--r-- 1 root supergroup 12157 2025-04-15 14:32 /data/10900703 A Dictionary of Canadianisms on Historical Principles.txt
cluster-master -rw-r--r-- 1 root supergroup 2866 2025-04-15 14:31 /data/11017392 A Bird and Spell in Yurt.txt
cluster-master -rw-r--r-- 1 root supergroup 4423 2025-04-15 14:32 /data/11017589 A Doctors Report on Diabetics.txt
cluster-master -rw-r--r-- 1 root supergroup 923 2025-04-15 14:29 /data/11141641 A Blueprint of the World.txt
cluster-master -rw-r--r-- 1 root supergroup 2572 2025-04-15 14:31 /data/1115818 A Hanging.txt
cluster-master -rw-r--r-- 1 root supergroup 12171 2025-04-15 14:32 /data/11211270 A Lesson in Romanticism.txt
cluster-master -rw-r--r-- 1 root supergroup 362 2025-04-15 14:28 /data/1135557 A Go on Potshot_album.txt
cluster-master -rw-r--r-- 1 root supergroup 333 2025-04-15 14:32 /data/11490217 A Guide to Groovy Lovin.txt
cluster-master -rw-r--r-- 1 root supergroup 5461 2025-04-15 14:31 /data/11526770 A Dreamers Tales.txt
cluster-master -rw-r--r-- 1 root supergroup 2529 2025-04-15 14:30 /data/11631125 A Ballad of the West.txt
cluster-master -rw-r--r-- 1 root supergroup 1029 2025-04-15 14:27 /data/11753853 A Journal of the Plague Year_album.txt
cluster-master -rw-r--r-- 1 root supergroup 597 2025-04-15 14:31 /data/11871420 A Lifetime or More.txt
cluster-master -rw-r--r-- 1 root supergroup 2134 2025-04-15 14:27 /data/11892774 A Cold Nights Death.txt
cluster-master -rw-r--r-- 1 root supergroup 863 2025-04-15 14:30 /data/11930321 A Fragile Hope.txt
cluster-master -rw-r--r-- 1 root supergroup 6843 2025-04-15 14:30 /data/11984610 A Catalogue of Crime.txt
cluster-master -rw-r--r-- 1 root supergroup 7441 2025-04-15 14:30 /data/12000397 A King and No King.txt
cluster-master -rw-r--r-- 1 root supergroup 1698 2025-04-15 14:29 /data/12132586 A Crystal Christmas.txt
cluster-master -rw-r--r-- 1 root supergroup 7681 2025-04-15 14:30 /data/12212199 A Firststones Christmas Carol.txt
cluster-master -rw-r--r-- 1 root supergroup 758 2025-04-15 14:29 /data/1240312 A Giant Alien Force More Violent_Sick Than Anything You Can Imagine.txt
cluster-master -rw-r--r-- 1 root supergroup 364 2025-04-15 14:27 /data/12459639 A Day at the Races_video.txt
cluster-master -rw-r--r-- 1 root supergroup 6324 2025-04-15 14:27 /data/12621170 A Fool in Love.txt
cluster-master -rw-r--r-- 1 root supergroup 10068 2025-04-15 14:30 /data/12666064 A Bird in the House.txt
```

Figure 1: HDFS ls /data.

```
zauri@zauri: ~/Documents/developer/NN05_525/BO/assignment_2
cluster-master -rw-r--r-- 1 root supergroup 3578 2025-04-15 14:27 /data/99580821 A Gunshot to the Head of Treason.txt
cluster-master -rw-r--r-- 1 root supergroup 1874 2025-04-15 14:31 /data/9146522 A Field Guide to the Birds of Hawaii and the Tropical Pacific.txt
cluster-master -rw-r--r-- 1 root supergroup 2483 2025-04-15 14:27 /data/9161281 A Ladys Morals.txt
cluster-master -rw-r--r-- 1 root supergroup 3328 2025-04-15 14:32 /data/921686 A Lie of the Mind.txt
cluster-master -rw-r--r-- 1 root supergroup 1938 2025-04-15 14:30 /data/929153 A Bao A Qu_album.txt
cluster-master -rw-r--r-- 1 root supergroup 3221 2025-04-15 14:30 /data/929265 A Chance to Cut Is a Chance to Cure.txt
cluster-master -rw-r--r-- 1 root supergroup 645 2025-04-15 14:31 /data/9415554 A Hanging Comes.txt
cluster-master -rw-r--r-- 1 root supergroup 3431 2025-04-15 14:29 /data/961187 A Hangover You Dont Deserve.txt
cluster-master -rw-r--r-- 1 root supergroup 3767 2025-04-15 14:27 /data/9704239 A Confession for Honor and Riches.txt
cluster-master -rw-r--r-- 1 root supergroup 7035 2025-04-15 14:31 /data/9847940 A Hard Days Night Greys Anatomy.txt
cluster-master -rw-r--r-- 1 root supergroup 1478 2025-04-15 14:29 /data/9869812 A Dream Common_song.txt
cluster-master -rw-r--r-- 1 root supergroup 1960 2025-04-15 14:27 /data/9870217 A Date with Luyu.txt
cluster-master -rw-r--r-- 1 root supergroup 5447 2025-04-15 14:31 /data/9919932 A Family Affair musical.txt
cluster-master -rw-r--r-- 1 root supergroup 616 2025-04-15 14:30 /data/9947241 A Day of Renewal.txt
cluster-master -rw-r--r-- 1 root supergroup 806 2025-04-15 14:27 /data/9965216 A Book of Human Language.txt
cluster-master -rw-r--r-- 1 root supergroup 412 2025-04-15 14:31 /data/9983283 A Good Enough Day.txt
cluster-master -rw-r--r-- 1 root supergroup 0 2025-04-15 14:26 /index/data/_SUCCESS
cluster-master -rw-r--r-- 1 root supergroup 355833 2025-04-15 14:26 /index/data/part-00000-lad64673-2285-4ec7-a289-45741f08df37-c000.csv
cluster-master Done data preparation
cluster-master Deleted /tmp/index/step1
cluster-master Deleted /tmp/index/step2
cluster-master Deleted /tmp/index/step3
cluster-master /index/data
cluster-master Starting first MapReduce job - Term frequencies and positions...
cluster-master packageJobJar: [/tmp/hadoop-unjar1701122783650861694/] [] /tmp/streamjob657212939900292612.jar tmpDir=null
cluster-master 2025-04-15 14:33:12.541 INFO client.DefaultHadoopFileProxyProvider: Connecting to ResourceManager at cluster-master/172.28.0.3:8032
cluster-master 2025-04-15 14:33:12.715 INFO client.DefaultHadoopFileProxyProvider: Connecting to ResourceManager at cluster-master/172.28.0.3:8032
cluster-master 2025-04-15 14:33:12.944 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.1744727167128_0001
cluster-master 2025-04-15 14:33:12.921 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master 2025-04-15 14:33:13.384 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master 2025-04-15 14:33:13.421 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744727167128_0001
cluster-master 2025-04-15 14:33:13.421 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master 2025-04-15 14:33:13.592 INFO conf.Configuration: resource-types.xml not found
cluster-master 2025-04-15 14:33:13.592 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'
cluster-master 2025-04-15 14:33:13.980 INFO impl.YarnClientImpl: Submitted application application_1744727167128_0001
cluster-master 2025-04-15 14:33:14.021 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744727167128_0001/
cluster-master 2025-04-15 14:33:14.022 INFO mapreduce.Job: Running job: job_1744727167128_0001
cluster-master 2025-04-15 14:33:20.139 INFO mapreduce.Job: Job job_1744727167128_0001 running in uber mode : false
cluster-master 2025-04-15 14:33:20.141 INFO mapreduce.Job: map 0% reduce 0%
cluster-master 2025-04-15 14:33:25.226 INFO mapreduce.Job: map 100% reduce 0%
cluster-master 2025-04-15 14:33:31.263 INFO mapreduce.Job: map 100% reduce 100%
cluster-master 2025-04-15 14:33:31.271 INFO mapreduce.Job: Job job_1744727167128_0001 completed successfully
cluster-master 2025-04-15 14:33:31.351 INFO mapreduce.Job: Counters: 54
cluster-master File System Counters
cluster-master FILE: Number of bytes read=11795296
cluster-master FILE: Number of bytes written=24420103
cluster-master FILE: Number of read operations=0
cluster-master FILE: Number of large read operations=0
cluster-master FILE: Number of write operations=0
cluster-master HDFS: Number of bytes read=3560221
```

Figure 2: HDFS ls /index/data.

```
zaur@zaur: ~/Documents/developer/INNO_S25/BO/assignment_2
cluster-master 2025-04-15 14:33:31.273 INFO mapreduce.Job: Job job_1744727167128_0001 completed successfully
cluster-master 2025-04-15 14:33:31.351 INFO mapreduce.Job: Counters: 54
cluster-master File System Counters
cluster-master FILE: Number of bytes read=11795296
cluster-master FILE: Number of bytes written=24420103
cluster-master FILE: Number of read operations=0
cluster-master FILE: Number of large read operations=0
cluster-master FILE: Number of write operations=4
cluster-master HDFS: Number of bytes read=3560221
cluster-master HDFS: Number of bytes written=672046
cluster-master HDFS: Number of read operations=11
cluster-master HDFS: Number of large read operations=0
cluster-master HDFS: Number of write operations=2
cluster-master HDFS: Number of bytes read erasure-coded=0
cluster-master Job Counters
cluster-master Launched map tasks=2
cluster-master Launched reduce tasks=1
cluster-master Data-local map tasks=2
cluster-master Total time spent by all maps in occupied slots (ms)=4735
cluster-master Total time spent by all reduces in occupied slots (ms)=3441
cluster-master Total time spent by all map tasks (ms)=4735
cluster-master Total time spent by all reduce tasks (ms)=3441
cluster-master Total vcore-milliseconds taken by all map tasks=4735
cluster-master Total vcore-milliseconds taken by all reduce tasks=3441
cluster-master Total megabyte-milliseconds taken by all map tasks=4848640
cluster-master Total megabyte-milliseconds taken by all reduce tasks=3523384
cluster-master Map-Reduce Framework
cluster-master Map input records=1000
cluster-master Map output records=50022
cluster-master Map output bytes=10635246
cluster-master Map output materialized bytes=11795302
cluster-master Input split bytes=292
cluster-master Combine input records=0
cluster-master Combine output records=0
cluster-master Reduce input groups=40864
cluster-master Reduce shuffle bytes=1195302
cluster-master Reduce input records=50022
cluster-master Reduce output records=25952
cluster-master Spilled Records=100044
cluster-master Shuffled Maps=2
cluster-master Failed Shuffles=0
cluster-master Merged Map outputs=2
cluster-master GC time elapsed (ms)=125
cluster-master CPU time spent (ms)=5230
cluster-master Physical memory (bytes) snapshot=924459008
cluster-master Virtual memory (bytes) snapshot=7788888064
cluster-master Total committed heap usage (bytes)=950534144
cluster-master Peak Map Physical memory (bytes)=348119840
cluster-master Peak Map Virtual memory (bytes)=2593759232
cluster-master Peak Reduce Physical memory (bytes)=269590416
cluster-master Peak Reduce Virtual memory (bytes)=2603798528
```

Figure 3: MapReduce pipeline 1

```
zaur@zaur: ~/Documents/developer/INNO_S25/BO/assignment_2
cluster-master Peak Map Virtual memory (bytes)=2593759232
cluster-master Peak Reduce Physical memory (bytes)=269590416
cluster-master Peak Reduce Virtual memory (bytes)=2603798528
cluster-master Shuffle
cluster-master BAD ID=0
cluster-master CONNECTION=0
cluster-master IO Error=0
cluster-master WRONG LENGTH=0
cluster-master WRONG MAP=0
cluster-master WRONG REDUCE=0
cluster-master File Input Format Counters
cluster-master Bytes Read=355929
cluster-master File Output Format Counters
cluster-master Bytes Written=672046
cluster-master 2025-04-15 14:33:31.351 INFO streaming.StreamJob: Output directory: /tmp/index/step1
cluster-master First MapReduce job succeeded. Sample output:
cluster-master Starting second MapReduce job: Document frequencies...
cluster-master packageJobJar: [/tmp/hadoop-unjar/7012723283428450021/] [] /tmp/streamjob1698010171080770865.jar tmpDir=null
cluster-master 2025-04-15 14:33:32.924 INFO client.DefaultHARFaiLowerProxyProvider: Connecting to ResourceManager at cluster-master/172.28.0.5:8032
cluster-master 2025-04-15 14:33:33.207 INFO client.DefaultHARFaiLowerProxyProvider: Connecting to ResourceManager at cluster-master/172.28.0.5:8032
cluster-master 2025-04-15 14:33:33.499 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/staging/job_1744727167128_0002
cluster-master 2025-04-15 14:33:33.774 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master 2025-04-15 14:33:33.820 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master 2025-04-15 14:33:33.938 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744727167128_0002
cluster-master 2025-04-15 14:33:33.938 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master 2025-04-15 14:33:34.089 INFO conf.Configuration: resource.types.xml not found
cluster-master 2025-04-15 14:33:34.089 INFO resource.ResourceUtils: Unable to find 'resource.types.xml'
cluster-master 2025-04-15 14:33:34.102 INFO Impl.YarnClientImpl: Submitted application application_1744727167128_0002
cluster-master 2025-04-15 14:33:34.234 INFO mapreduce.Job: The url to track the job: http://cluster-master:8080/proxy/application_1744727167128_0002/
cluster-master 2025-04-15 14:33:34.240 INFO mapreduce.Job: Running job: job_1744727167128_0002
cluster-master 2025-04-15 14:33:41.315 INFO mapreduce.Job: Job job_1744727167128_0002 running in user mode : false
cluster-master 2025-04-15 14:33:41.316 INFO mapreduce.Job: map 0% reduce 0%
cluster-master 2025-04-15 14:33:45.365 INFO mapreduce.Job: map 100% reduce 0%
cluster-master 2025-04-15 14:33:51.411 INFO mapreduce.Job: map 100% reduce 100%
cluster-master 2025-04-15 14:33:51.419 INFO mapreduce.Job: Job job_1744727167128_0002 completed successfully
cluster-master 2025-04-15 14:33:51.499 INFO mapreduce.Job: Counters: 54
cluster-master File System Counters
cluster-master FILE: Number of bytes read=4351985
cluster-master FILE: Number of bytes written=9533496
cluster-master FILE: Number of read operations=0
cluster-master FILE: Number of large read operations=0
cluster-master FILE: Number of write operations=0
cluster-master HDFS: Number of bytes read=673252
cluster-master HDFS: Number of bytes written=417954
cluster-master HDFS: Number of read operations=11
cluster-master HDFS: Number of large read operations=0
cluster-master HDFS: Number of write operations=2
cluster-master HDFS: Number of bytes read erasure-coded=0
cluster-master Job Counters
cluster-master Launched map tasks=2
cluster-master Launched reduce tasks=1
```

Figure 4: MapReduce pipeline 2

```
zaural@zaural: ~/Documents/developer/HNO_S25/BO/assignment_2
cluster-master Job Counters
cluster-master   Launched map tasks=2
cluster-master   Launched reduce tasks=1
cluster-master   Data-local map tasks=2
cluster-master   Total time spent by all maps in occupied slots (ms)=4829
cluster-master   Total time spent by all reduces in occupied slots (ms)=2593
cluster-master   Total time spent by all map tasks (ms)=4829
cluster-master   Total time spent by all reduce tasks (ms)=2593
cluster-master   Total vcore-milliseconds taken by all map tasks=4829
cluster-master   Total vcore-milliseconds taken by all reduce tasks=2593
cluster-master   Total megabyte-milliseconds taken by all map tasks=4125696
cluster-master   Total megabyte-milliseconds taken by all reduce tasks=2363072
cluster-master Map-Reduce Framework
cluster-master   Map input records=259952
cluster-master   Map output records=259952
cluster-master   Map output bytes=3850075
cluster-master   Map output materialized bytes=4351991
cluster-master   Input split bytes=210
cluster-master   Combine input records=0
cluster-master   Combine output records=0
cluster-master   Reduce input groups=40864
cluster-master   Reduce shuffle bytes=4351991
cluster-master   Reduce input records=259952
cluster-master   Reduce output records=40864
cluster-master   Spilled Records=501994
cluster-master   Shuffled Maps=2
cluster-master   Failed Shuffles=0
cluster-master   Merged Map outputs=2
cluster-master   GC time elapsed (ms)=126
cluster-master   CPU time spent (ms)=5048
cluster-master   Physical memory (bytes) snapshot=801774592
cluster-master   Virtual memory (bytes) snapshot=7781584896
cluster-master   Total committed heap usage (bytes)=91833984
cluster-master   Peak Map Physical memory (bytes)=309075968
cluster-master   Peak Map Virtual memory (bytes)=2591170560
cluster-master   Peak Reduce Physical memory (bytes)=2265366880
cluster-master   Peak Reduce Virtual memory (bytes)=2608484864
cluster-master Shuffle Errors
cluster-master   BAD_ID=0
cluster-master   CONNECTION=0
cluster-master   IO_ERROR=0
cluster-master   WRONG_LENGTH=0
cluster-master   WRONG_MAP=0
cluster-master   WRONG_REQUIE=0
cluster-master File Input Format Counters
cluster-master   Bytes Read=33842
cluster-master File Output Format Counters
cluster-master   Bytes Written=417954
cluster-master 2025-04-15 14:33:51.409 INFO streaming.StreamJob: Output directory: /tmp/index/step2
cluster-master Second MapReduce job succeeded!
cluster-master Starting third MapReduce job - Document metadata...
```

Figure 5: MapReduce pipeline 2

```
zaural@zaural: ~/Documents/developer/HNO_S25/BO/assignment_2
cluster-master 2025-04-15 14:33:51.409 INFO streaming.StreamJob: Output directory: /tmp/index/step2
cluster-master Second MapReduce job succeeded!
cluster-master Starting third MapReduce job - Document metadata...
cluster-master packageJobJar: [/tmp/hadoop-unjar/7830221207778410/] [] /tmp/streamjob3092319681109552562.jar tmpDir=null
cluster-master 2025-04-15 14:33:53.264 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.28.0.5:8032
cluster-master 2025-04-15 14:33:53.440 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.28.0.5:8032
cluster-master 2025-04-15 14:33:53.632 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/staging/job_1744727167128_0003
cluster-master 2025-04-15 14:33:53.864 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master 2025-04-15 14:33:53.910 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master 2025-04-15 14:33:54.069 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744727167128_0003
cluster-master 2025-04-15 14:33:54.069 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master 2025-04-15 14:33:54.215 INFO conf.Configuration: resource-types.xml not found
cluster-master 2025-04-15 14:33:54.261 INFO impl.YarnClientImpl: Submitted application application_1744727167128_0003
cluster-master 2025-04-15 14:33:54.289 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744727167128_0003/
cluster-master 2025-04-15 14:33:54.290 INFO mapreduce.Job: Running job: job_1744727167128_0003
cluster-master 2025-04-15 14:34:01.383 INFO mapreduce.Job: Job job_1744727167128_0003 running in uber mode : false
cluster-master 2025-04-15 14:34:01.384 INFO mapreduce.Job: map 0% reduce 0%
cluster-master 2025-04-15 14:34:06.439 INFO mapreduce.Job: map 100% reduce 0%
cluster-master 2025-04-15 14:34:09.464 INFO mapreduce.Job: map 100% reduce 100%
cluster-master 2025-04-15 14:34:09.472 INFO mapreduce.Job: Job job_1744727167128_0003 completed successfully
cluster-master 2025-04-15 14:34:09.533 INFO mapreduce.Job: Counters: 54
cluster-master File System Counters
cluster-master   FILE: Number of bytes read=37945
cluster-master   FILE: Number of bytes written=905401
cluster-master   FILE: Number of read operations=0
cluster-master   FILE: Number of large read operations=0
cluster-master   FILE: Number of write operations=0
cluster-master   HDFS: Number of bytes read=3560221
cluster-master   HDFS: Number of bytes written=23939
cluster-master   HDFS: Number of read operations=11
cluster-master   HDFS: Number of large read operations=0
cluster-master   HDFS: Number of write operations=2
cluster-master   HDFS: Number of bytes read erasure-coded=0
cluster-master Job Counters
cluster-master   Launched map tasks=2
cluster-master   Launched reduce tasks=1
cluster-master   Data-local map tasks=2
cluster-master   Total time spent by all maps in occupied slots (ms)=3572
cluster-master   Total time spent by all reduces in occupied slots (ms)=1622
cluster-master   Total time spent by all map tasks (ms)=3572
cluster-master   Total time spent by all reduce tasks (ms)=1622
cluster-master   Total vcore-milliseconds taken by all map tasks=3572
cluster-master   Total vcore-milliseconds taken by all reduce tasks=1622
cluster-master   Total megabyte-milliseconds taken by all map tasks=3657728
cluster-master   Total megabyte-milliseconds taken by all reduce tasks=1668928
cluster-master Map-Reduce Framework
cluster-master   Map input records=1000
cluster-master   Map output records=1000
cluster-master   Map output bytes=35930
cluster-master   Map output materialized bytes=37951
```

Figure 6: MapReduce pipeline 3


```
zaural@zaural: ~/Documents/developer/INNO_S25/BO/assignment_2
cluster-master Total vcore-milliseconds taken by all map tasks=3572
cluster-master Total vcore-milliseconds taken by all reduce tasks=1622
cluster-master Total megabyte-milliseconds taken by all map tasks=3657728
cluster-master Total megabyte-milliseconds taken by all reduce tasks=1660928
cluster-master Map-Reduce Framework
cluster-master Map input records=1000
cluster-master Map output records=1000
cluster-master Map output bytes=35939
cluster-master Map output materialized bytes=37951
cluster-master Input split bytes=292
cluster-master Combine input records=0
cluster-master Combine output records=0
cluster-master Reduce input groups=1000
cluster-master Reduce shuffle bytes=37951
cluster-master Reduce input records=1000
cluster-master Reduce output records=1000
cluster-master Spilled Records=2000
cluster-master Shuffled Maps=2
cluster-master Failed Shuffles=0
cluster-master Merged Map outputs=2
cluster-master GC time elapsed (ms)=101
cluster-master CPU time spent (ms)=1540
cluster-master Physical memory (bytes) snapshot=806666248
cluster-master Virtual memory (bytes) snapshot=775707248
cluster-master Total committed heap usage (bytes)=810840784
cluster-master Peak Map Physical memory (bytes)=103770440
cluster-master Peak Map Virtual memory (bytes)=2501285248
cluster-master Peak Reduce Physical memory (bytes)=199847936
cluster-master Peak Reduce Virtual memory (bytes)=250359336
cluster-master Shuffle Errors
cluster-master BAD ID=0
cluster-master CONNECTION=0
cluster-master IO ERROR=0
cluster-master WRONG LENGTH=0
cluster-master WRONG MAP=0
cluster-master WRONG REDUCE=0
cluster-master File Input Format Counters
cluster-master Bytes Read=3559929
cluster-master File Output Format Counters
cluster-master Bytes Written=35939
cluster-master 2025-04-15 14:34:09.533 INFO streaming.StreamJob: Output directory: /tmp/index/step3
cluster-master Importing data into Cassandra...
cluster-master 2025-04-15 14:34:10.091 WARNING - Cluster._init_ called with contact points specified, but no load_balancing policy. In the next major version, this will raise an error; please specify a load_balancing policy. (contact points = ['cassandra-server'], lbp = None)
cluster-master 2025-04-15 14:34:10.127 WARNING - Downgrading core protocol version from 66 to 65 for 172.28.0.2:9042. To avoid this, it is best practice to explicitly set Cluster(protocol version) to the version supported by your cluster. http://datastax.github.io/python-driver/api/cassandra/cluster.html#cassandra.cluster.Cluster.protocol_version
cluster-master 2025-04-15 14:34:10.130 WARNING - Downgrading core protocol version from 65 to 5 for 172.28.0.2:9042. To avoid this, it is best practice to explicitly set Cluster(protocol version) to the version supported by your cluster. http://datastax.github.io/python-driver/api/cassandra/cluster.html#cassandra.cluster.Cluster.protocol_version
cluster-master 2025-04-15 14:34:10.214 INFO - Using datacenter 'datacenter1' for DCAwareRoundRobinPolicy (via host '172.28.0.2:9042'); if incorrect, please specify a local_dc to the constructor, or limit contact points to local cluster nodes
cluster-master 2025-04-15 14:34:10.285 INFO - Successfully connected to Cassandra cluster
cluster-master 2025-04-15 14:34:10.285 INFO - Successfully connected to Cassandra cluster
```

Figure 7: MapReduce pipeline 3

```
zaural@zaural: ~/Documents/developer/INNO_S25/BO/assignment_2
cluster-master File Input Format Counters
cluster-master Bytes Read=3559929
cluster-master File Output Format Counters
cluster-master Bytes Written=35939
cluster-master 2025-04-15 14:34:09.533 INFO streaming.StreamJob: Output directory: /tmp/index/step3
cluster-master Importing data into Cassandra...
cluster-master 2025-04-15 14:34:10.091 WARNING - Cluster._init_ called with contact points specified, but no load_balancing policy. In the next major version, this will raise an error; please specify a load_balancing policy. (contact points = ['cassandra-server'], lbp = None)
cluster-master 2025-04-15 14:34:10.127 WARNING - Downgrading core protocol version from 66 to 65 for 172.28.0.2:9042. To avoid this, it is best practice to explicitly set Cluster(protocol version) to the version supported by your cluster. http://datastax.github.io/python-driver/api/cassandra/cluster.html#cassandra.cluster.Cluster.protocol_version
cluster-master 2025-04-15 14:34:10.130 WARNING - Downgrading core protocol version from 65 to 5 for 172.28.0.2:9042. To avoid this, it is best practice to explicitly set Cluster(protocol version) to the version supported by your cluster. http://datastax.github.io/python-driver/api/cassandra/cluster.html#cassandra.cluster.Cluster.protocol_version
cluster-master 2025-04-15 14:34:10.214 INFO - Using datacenter 'datacenter1' for DCAwareRoundRobinPolicy (via host '172.28.0.2:9042'); if incorrect, please specify a local_dc to the constructor, or limit contact points to local cluster nodes
cluster-master 2025-04-15 14:34:10.285 INFO - Successfully connected to Cassandra cluster
cluster-master 2025-04-15 14:34:10.285 INFO - Successfully connected to Cassandra cluster
cluster-master 2025-04-15 14:34:10.388 INFO - Successfully created keyspace and tables
cluster-master 2025-04-15 14:34:10.398 INFO - Reading term index data from /tmp/index/step1
cluster-master WARN [Native-Transport-Requests-2] 2025-04-15 14:34:27.601 BatchStatement.java:362 - Batch for [search.engine.term.document] is of size 5.020KiB, exceeding specified threshold of 5.000KiB by 20B.
cluster-master 2025-04-15 14:34:27.602 WARNING - Server warning: Batch for [search.engine.term.document] is of size 5140, exceeding specified threshold of 5120 by 20.
cluster-master WARN [Native-Transport-Requests-1] 2025-04-15 14:34:31.064 BatchStatement.java:362 - Batch for [search.engine.term.document] is of size 5.098KiB, exceeding specified threshold of 5.000KiB by 100B.
cluster-master 2025-04-15 14:34:31.065 WARNING - Server warning: Batch for [search.engine.term.document] is of size 5220, exceeding specified threshold of 5120 by 109.
cluster-master WARN [Native-Transport-Requests-2] 2025-04-15 14:34:33.020 BatchStatement.java:362 - Batch for [search.engine.term.document] is of size 6.699KiB, exceeding specified threshold of 5.000KiB by 1.699KiB.
cluster-master 2025-04-15 14:34:33.021 WARNING - Server warning: Batch for [search.engine.term.document] is of size 6860, exceeding specified threshold of 5120 by 1740.
cluster-master WARN [Native-Transport-Requests-3] 2025-04-15 14:34:59.081 BatchStatement.java:362 - Batch for [search.engine.term.document] is of size 5.244KiB, exceeding specified threshold of 5.000KiB by 240B.
cluster-master 2025-04-15 14:34:59.082 WARNING - Server warning: Batch for [search.engine.term.document] is of size 5370, exceeding specified threshold of 5120 by 250.
cluster-master 2025-04-15 14:35:00.523 INFO - Successfully imported 250922 term index entries
cluster-master 2025-04-15 14:35:02.525 INFO - Reading document frequency data from /tmp/index/step2
cluster-master 2025-04-15 14:35:08.346 INFO - Successfully imported 48864 document frequency entries
cluster-master 2025-04-15 14:35:08.346 INFO - Importing document metadata from MapReduce output
cluster-master 2025-04-15 14:35:11.242 INFO - Successfully imported 1000 document metadata entries
cluster-master 2025-04-15 14:35:11.244 INFO - Successfully imported index data into Cassandra
cluster-master Successfully imported index data into Cassandra
cluster-master Searching for: this is a query!
cluster-master 2025-04-15 14:35:13.585 [search-engine] INFO: Processing query: this is a query!
cluster-master 2025-04-15 14:35:13.587 [cassandra.cluster] WARNING: Cluster._init_ called with contact points specified, but no load_balancing policy. In the next major version, this will raise an error; please specify a load_balancing policy. (contact points = ['cassandra-server'], lbp = None)
cluster-master 2025-04-15 14:35:13.592 [cassandra.cluster] WARNING: Downgrading core protocol version from 66 to 65 for 172.28.0.2:9042. To avoid this, it is best practice to explicitly set Cluster(protocol version) to the version supported by your cluster. http://datastax.github.io/python-driver/api/cassandra/cluster.html#cassandra.cluster.Cluster.protocol_version
cluster-master 2025-04-15 14:35:13.592 [cassandra.cluster] WARNING: Downgrading core protocol version from 65 to 5 for 172.28.0.2:9042. To avoid this, it is best practice to explicitly set Cluster(protocol version) to the version supported by your cluster. http://datastax.github.io/python-driver/api/cassandra/cluster.html#cassandra.cluster.Cluster.protocol_version
cluster-master 2025-04-15 14:35:13.602 [cassandra.cluster] INFO: Using datacenter 'datacenter1' for DCAwareRoundRobinPolicy (via host '172.28.0.2:9042'); if incorrect, please specify a local_dc to the constructor, or limit contact points to local cluster nodes
cluster-master 2025-04-15 14:35:13.609 [search-engine] INFO: Successfully connected to Cassandra cluster
cluster-master 25/04/15 14:35:13 INFO SparkContext: Running Spark version 3.5.4
cluster-master 25/04/15 14:35:13 INFO SparkContext: OS info Linux, 6.8.0-51-generic, amd64
cluster-master 25/04/15 14:35:13 INFO SparkContext: Java version 1.8.0_442
```

Figure 8: import index data to Cassandra

```

zaurall@zaurall:~/Documents/developer/INNO_S25/BD/assignment_2
Connected to test cluster at 127.0.0.1:9042
[cqlsh 6.2.0 | Cassandra 5.0.4 | CQL spec 3.4.7 | Native protocol v5]
Use HELP for help.
cqlsh> use search_engine ;
cqlsh:search_engine> select count(*) FROM
document_frequency      search_engine,      system_auth,      system_schema,      system_views,      term_document
documents_info           system,          system_distributed,  system_traces,      system_virtual_schema,
cqlsh:search_engine> select count(*) FROM documents_info;
count
-----
0
(1 rows)
Warnings :
Aggregation query used without partition key
cqlsh:search_engine> select count(*) FROM documents_info;
count
-----
0
(1 rows)
Warnings :
Aggregation query used without partition key
cqlsh:search_engine> select count(*) FROM document_frequency documents_info;
cqlsh:search_engine> select count(*) FROM document_frequency ;
count
-----
0
(1 rows)
Warnings :
Aggregation query used without partition key
cqlsh:search_engine> select count(*) FROM term_document ;
count
-----
20435
(1 rows)

```

Figure 9: Cassandra tables content

```

(1 rows)
Warnings :
Aggregation query used without partition key
cqlsh:search_engine> select * FROM term_document limit 10;
term | doc_id | positions | tf
-----|-----|-----|-----
564b0 | 617705 | [1856] | 1
483 | 8227618 | [485] | 1
517 | 322487 | [79] | 1
517 | 53278497 | [197] | 1
517 | 978826 | [112] | 1
1843 | 1398342 | [239] | 1
1843 | 38842440 | [114, 123, 18] | 3
4311t | 28319546 | [986] | 1
4311t | 28862620 | [1588] | 1
255 | 185393 | [16887] | 1
(10 rows)
cqlsh:search_engine> select * FROM term_document where term = 'this' limit 10;
term | doc_id | positions | tf
-----|-----|-----|-----
(0 rows)
cqlsh:search_engine> select * FROM term_document where term = 'is' limit 10;
term | doc_id | positions | tf
-----|-----|-----|-----
(0 rows)
cqlsh:search_engine> select * FROM term_document where term = 'a' limit 10;
term | doc_id | positions | tf
-----|-----|-----|-----
(0 rows)
cqlsh:search_engine> select * FROM term_document where term = '255' limit 10;
term | doc_id | positions | tf
-----|-----|-----|-----
255 | 105281 | [14987] | 1
255 | 1467895 | [19381] | 1
255 | 29401892 | [6771] | 1
255 | 9385747 | [4227] | 1
(4 rows)
cqlsh:search_engine> select * FROM term_document limit 100;

```

Figure 10: Cassandra tables content


```

zauri@zauri: ~/Documents/developer/INNO_S25/BD/assignment_2
zauri@zauri: ~/Documents/developer/INNO_S25/BD/assignment_2
zauri@zauri: ~/Documents/developer/INNO_S25/BD/assignment_2
1964 | 3599665 | [49] | 1
1964 | 35184664 | [189] | 1
1964 | 37803108 | [159] | 1
1964 | 40896793 | [187] | 1
1964 | 40842526 | [232] | 1
1964 | 41080200 | [138] | 1
1964 | 43495241 | [1265] | 1
1964 | 4403 | [2164] | 1
1964 | 46534688 | [106] | 1
1964 | 49994999 | [397, 349, 373, 389, 381, 405, 365, 357] | 8

(100 rows)
cqlsh:search_engine>
cqlsh:search_engine> select count(*) FROM term_document ;

count
-----
19435

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:search_engine> zauri@zauri:~/Documents/developer/INNO_S25/BD/assignment_2 docker exec -it cassandra-server cqlsh
Connected to test cluster at 127.0.0.1:9042
[cqlsh 0.2.0 | Cassandra 5.0.4 | CQL spec 3.4.7 | Native protocol v5]
Use HELP for help.
cqlsh> use search_engine ;
cqlsh:search_engine> select * FROM
document_frequency search_engine system_auth system schema system views term_document
documents_info system system distributed system traces system virtual schema
cqlsh:search_engine> select * FROM document_frequency limit 10;

term | df
-----
dobson | 1
sain | 1
bessus | 1
ix | 4
sweet | 2
libertad | 1
dance | 39
previews | 7
maclean | 1
gibbs | 6

(10 rows)
cqlsh:search_engine> select count(*) FROM document_frequency;

count
-----
40864

(1 rows)

```

Figure 13: Cassandra tables content

```

zauri@zauri: ~/Documents/developer/INNO_S25/BD/assignment_2
zauri@zauri: ~/Documents/developer/INNO_S25/BD/assignment_2
zauri@zauri: ~/Documents/developer/INNO_S25/BD/assignment_2
dance | 39
previews | 7
maclean | 1
gibbs | 6

(10 rows)
cqlsh:search_engine> select count(*) FROM document_frequency;

count
-----
40864

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:search_engine> select * FROM documents_info limit 10;

doc_id | length | title
-----
18228685 | 115 | A Dead Sinking Story
27588194 | 567 | A Hero Ain't Nothin' but a Sandwich (film)
39710446 | 330 | A Little Bit of Luck
31294093 | 338 | A Change Is Gonna Come (Jack McBratt album)
51794980 | 306 | A Family Secret (Upstairs, Downstairs)
36880998 | 564 | A J Balliol Salmon
28794362 | 61 | A Encarnado, A Pobra de Trives
1719074 | 1075 | A Case of Conscience
584808 | 1334 | A Case of Identity
7745742 | 660 | A Girl Called Dusty

(10 rows)
cqlsh:search_engine> select count(*) FROM documents_info;

count
-----
1080

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:search_engine> select * FROM ;
document_frequency search_engine system_auth system schema system views term_document
documents_info system system distributed system traces system virtual schema
cqlsh:search_engine> select * FROM term_document LIMIT 10;

term | doc_id | positions | tf
-----

```

Figure 14: Cassandra tables content

```

38294693 | 338 | A Change Is Gonna Come (Jack Mc Huff album)
31794980 | 366 | A Family Secret (Upstairs, Downstairs)
56880980 | 564 | A J Ballico Salmon
28798362 | 63 | A Encomenda, A Pobra de Trives
3719074 | 1675 | A Case of Conscience
584888 | 1334 | A Case of Identity
7765742 | 660 | A Girl Called Dusty

(10 rows)
cqlsh:search_engine select count(*) FROM documents_info;
count
-----
1090
(1 rows)

Warnings :
Aggregation query used without partition key
cqlsh:search_engine select * FROM ;
document_frequency search_engine system_auth system_schema system_views term_document
documents_info system system_traces system_virtual_schema
cqlsh:search_engine select * FROM term_document LIMIT 10;

term | doc_id | positions | tf
-----|-----|-----|-----
dobson | 13631486 | [60] | 1
sain | 1448455 | [1840, 1339, 1180, 1540] | 4
bessus | 12006397 | [1817, 1090, 1620] | 3
ix | 10789501 | [54] | 1
ix | 32497421 | [34] | 1
ix | 67078438 | [64, 150] | 2
ix | 72922368 | [43] | 1
awalt | 41675801 | [1367] | 1
awalt | 4136530 | [672] | 1
Libertad | 47612098 | [20, 41] | 2

(10 rows)
cqlsh:search_engine select count(*) FROM term_document;
count
-----
250952
(1 rows)

Warnings :
Aggregation query used without partition key
cqlsh:search_engine

```

Figure 15: Cassandra tables content

```

cluster-master 2025/04/15 14:35:41 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (TID 2) in 628 ms on cluster-slave-2 (executor 1) (1/2)
cluster-master 2025/04/15 14:35:41 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (TID 3) in 627 ms on cluster-slave-1 (executor 2) (2/2)
cluster-master 2025/04/15 14:35:41 INFO YarnScheduler: Removed TaskSet 1.0, whose tasks have all completed, from pool
cluster-master 2025/04/15 14:35:41 INFO DAGScheduler: ResultStage 1 (takeOrdered at /app/query.py:247) finished in 0.647 s
cluster-master 2025/04/15 14:35:41 INFO DAGScheduler: Job 0 is finished, Cancelling potential speculative or zombie tasks for this job
cluster-master 2025/04/15 14:35:41 INFO YarnScheduler: Killing all running tasks in stage 1: Stage finished
cluster-master 2025/04/15 14:35:41 INFO DAGScheduler: Job 0 finished: takeOrdered at /app/query.py:247, took 2.339473 s

Search Results for: this is a query!
=====
1. ID: 47515595 Score: 8.9019 Title: A Canine Sherlock Holmes
2. ID: 38628228 Score: 1.6101 Title: A Human Right
3. ID: 18171842 Score: 1.5942 Title: A Chrestomathy
4. ID: 7868424 Score: 1.5864 Title: A Black Mass
5. ID: 57278016 Score: 1.5804 Title: A Book of American Martyrs
6. ID: 41801556 Score: 1.5638 Title: A (The Walking Dead)
7. ID: 45393168 Score: 1.5618 Title: A Bearded Man
8. ID: 16378814 Score: 1.5409 Title: A Breathtaking Guy
9. ID: 62485656 Score: 1.5363 Title: A Galf for Christmas
10. ID: 2761148 Score: 1.5357 Title: A History of Philosophy (Copleston)

25/04/15 14:35:41 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/15 14:35:41 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/15 14:35:41 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/15 14:35:41 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/15 14:35:41 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
25/04/15 14:35:41 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/15 14:35:41 INFO MemoryStore: MemoryStore cleared
25/04/15 14:35:41 INFO BlockManager: BlockManager stopped
25/04/15 14:35:41 INFO OutputCommitCoordinatorOutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/15 14:35:41 INFO SparkContext: Successfully stopped SparkContext
25/04/15 14:35:42 INFO ShutdownHookManager: Shutdown hook called
25/04/15 14:35:42 INFO ShutdownHookManager: Deleting directory /tmp/spark-654f8b99-1e35-4c7a-9af3-5b61d354dc2c
25/04/15 14:35:42 INFO ShutdownHookManager: Deleting directory /tmp/spark-654f8b99-1e35-4c7a-9af3-5b61d354dc2c/pyspark-b3be686f-b8e2-4407-a9b3-95891d63f34d
cluster-master exited with code 0
cassandra-server WARN [Native-Transport-Requests-1] 2025-04-15 14:37:38.402 SelectStatement.java:557 - Aggregation query used without partition key
aggregation type: AGGREGATE EVERYTHING
INFO [ReadStage-2] 2025-04-15 14:37:38.906 MonitoringTask.java:95 - Scheduling monitoring task with report interval of 5000 ms, max operations 50
INFO [ScheduledTasks-1] 2025-04-15 14:37:43.912 NoSpamLogger.java:104 - Some operations were slow, details available at debug level (debug.log)
cassandra-server WARN [Native-Transport-Requests-1] 2025-04-15 14:38:00.358 SelectStatement.java:557 - Aggregation query used without partition key
cassandra-server WARN [Native-Transport-Requests-1] 2025-04-15 14:38:00.358 NoSpamLogger.java:107 - Aggregation query used without partition key on table search_engine.documents_info, ag
aggregation type: AGGREGATE EVERYTHING
cassandra-server WARN [Native-Transport-Requests-1] 2025-04-15 14:38:25.958 SelectStatement.java:557 - Aggregation query used without partition key
aggregation type: AGGREGATE EVERYTHING
cassandra-server WARN [Native-Transport-Requests-1] 2025-04-15 14:38:25.959 NoSpamLogger.java:107 - Aggregation query used without partition key on table search_engine.term_document, ag
aggregation type: AGGREGATE EVERYTHING

```

Figure 16: Query 1 results

```
zaur@zaur: ~/Documents/developer/INNO_S25/BO/assignment_2
cluster-master 25/04/15 14:46:18 INFO MemoryStore: Block broadcast 1 piece0 stored as bytes in memory (estimated size 6.5 KiB, free 366.3 MiB)
cluster-master 25/04/15 14:46:18 INFO BlockManagerInfo: Added broadcast 1 piece0 in memory on cluster-master:45727 (size: 6.5 KiB, free: 366.3 MiB)
cluster-master 25/04/15 14:46:18 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1585
cluster-master 25/04/15 14:46:18 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 1 (PythonRDD[5] at takeOrdered at /app/query.py:247) (first 15 tasks are for partitions V
ector(0, 1))
cluster-master 25/04/15 14:46:18 INFO YarnScheduler: Adding task set 1.0 with 2 tasks resource profile 0
cluster-master 25/04/15 14:46:18 INFO TaskSetManager: Starting task 1.0 in stage 1.0 (TID 3) (cluster-slave-2, executor 2, partition 1, NODE LOCAL, 8828 bytes)
cluster-master 25/04/15 14:46:18 INFO BlockManagerInfo: Added broadcast 1 piece0 in memory on cluster-slave-1:41865 (size: 6.5 KiB, free: 366.3 MiB)
cluster-master 25/04/15 14:46:18 INFO BlockManagerInfo: Added broadcast 1 piece0 in memory on cluster-slave-2:37869 (size: 6.5 KiB, free: 366.3 MiB)
cluster-master 25/04/15 14:46:18 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to 172.28.0.2:54924
cluster-master 25/04/15 14:46:18 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to 172.28.0.4:35356
cluster-master 25/04/15 14:46:18 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2) in 555 ms on cluster-slave-1 (executor 1) (1/2)
cluster-master 25/04/15 14:46:18 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (TID 3) in 676 ms on cluster-slave-2 (executor 2) (2/2)
cluster-master 25/04/15 14:46:18 INFO YarnScheduler: Removed TaskSet 1.0, whose tasks have all completed, from pool
cluster-master 25/04/15 14:46:19 INFO DAGScheduler: ResultStage 1 (takeOrdered at /app/query.py:247) finished in 0.697 s
cluster-master 25/04/15 14:46:19 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master 25/04/15 14:46:19 INFO YarnScheduler: Killing all running tasks in stage 1: stage finished
cluster-master 25/04/15 14:46:19 INFO DAGScheduler: Job 0 finished: takeOrdered at /app/query.py:247, took 2.423933 s
cluster-master
cluster-master Search Results for: how to gain muscles?
cluster-master =====
cluster-master 1. ID: 63546862 Score: 6.660 Title: A Life of Sin
cluster-master 2. ID: 55590139 Score: 6.2130 Title: A Day in the Life of the Dummies
cluster-master 3. ID: 48474725 Score: 5.7168 Title: A Corner in Cotton
cluster-master 4. ID: 11528779 Score: 5.7043 Title: A Dreamer's Tales
cluster-master 5. ID: 37676249 Score: 5.5010 Title: A Date with the Falcon
cluster-master 6. ID: 45434976 Score: 5.0801 Title: A Dainty Politician
cluster-master 7. ID: 7609581 Score: 4.4212 Title: A Bold Stroke for a Wife
cluster-master 8. ID: 45474532 Score: 4.2741 Title: A Closer Look (Steve Harley & Cockney Rebel album)
cluster-master 9. ID: 559184 Score: 4.1914 Title: A Hairdresser's Experience in High Life
cluster-master 10. ID: 6602969 Score: 3.7933 Title: A Is for Atom
cluster-master 25/04/15 14:46:19 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master 25/04/15 14:46:19 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master 25/04/15 14:46:19 INFO YarnClientSchedulerBackend: Interrupting monitor thread
cluster-master 25/04/15 14:46:19 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster-master 25/04/15 14:46:19 INFO YarnSchedulerBackendsYarnDriverEndpoint: Asking each executor to shut down
cluster-master 25/04/15 14:46:19 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
cluster-master 25/04/15 14:46:19 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master 25/04/15 14:46:19 INFO MemoryStore: MemoryStore cleared
cluster-master 25/04/15 14:46:19 INFO BlockManager: BlockManager stopped
cluster-master 25/04/15 14:46:19 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master 25/04/15 14:46:19 INFO OutputCommitCoordinatorOutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master 25/04/15 14:46:19 INFO SparkContext: Successfully stopped SparkContext
cluster-master 25/04/15 14:46:20 INFO ShutdownHookManager: Shutdown hook called
cluster-master 25/04/15 14:46:20 INFO ShutdownHookManager: Deleting directory /tmp/spark-a0d785d0-3619-42b8-b034-aeb555b91679
cluster-master 25/04/15 14:46:20 INFO ShutdownHookManager: Deleting directory /tmp/spark-a0d785d0-3619-42b8-b034-aeb555b91679/pyspark-bef449c-23f2-4a01-bf1d-3e0cb2a72fce
cluster-master 25/04/15 14:46:20 INFO ShutdownHookManager: Deleting directory /tmp/spark-54e0abf9-d2ab-45ee-bc2f-2748b527a85
cluster-master 25/04/15 14:46:20 INFO ShutdownHookManager: Deleting directory /tmp/spark-54e0abf9-d2ab-45ee-bc2f-2748b527a85
cluster-master cluster-master exited with code 0
cluster-master
cluster-master Enable Watch
```

Figure 17: Query 2 results

```
zaur@zaur: ~/Documents/developer/INNO_S25/BO/assignment_2
cluster-master 25/04/15 14:50:41 INFO MemoryStore: Block broadcast 1 piece0 stored as bytes in memory (estimated size 6.5 KiB, free 366.3 MiB)
cluster-master 25/04/15 14:50:41 INFO BlockManagerInfo: Added broadcast 1 piece0 in memory on cluster-master:36439 (size: 6.5 KiB, free: 366.3 MiB)
cluster-master 25/04/15 14:50:41 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1585
cluster-master 25/04/15 14:50:41 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 1 (PythonRDD[5] at takeOrdered at /app/query.py:247) (first 15 tasks are for partitions V
ector(0, 1))
cluster-master 25/04/15 14:50:41 INFO YarnScheduler: Adding task set 1.0 with 2 tasks resource profile 0
cluster-master 25/04/15 14:50:41 INFO TaskSetManager: Starting task 1.0 in stage 1.0 (TID 3) (cluster-slave-2, executor 2, partition 0, NODE LOCAL, 8828 bytes)
cluster-master 25/04/15 14:50:41 INFO BlockManagerInfo: Added broadcast 1 piece0 in memory on cluster-slave-2:41999 (size: 6.5 KiB, free: 366.3 MiB)
cluster-master 25/04/15 14:50:41 INFO BlockManagerInfo: Added broadcast 1 piece0 in memory on cluster-slave-1:46397 (size: 6.5 KiB, free: 366.3 MiB)
cluster-master 25/04/15 14:50:42 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to 172.28.0.5:37400
cluster-master 25/04/15 14:50:42 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to 172.28.0.2:40344
cluster-master 25/04/15 14:50:42 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2) in 845 ms on cluster-slave-2 (executor 2) (1/2)
cluster-master 25/04/15 14:50:42 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (TID 3) in 848 ms on cluster-slave-1 (executor 1) (2/2)
cluster-master 25/04/15 14:50:42 INFO YarnScheduler: Removed TaskSet 1.0, whose tasks have all completed, from pool
cluster-master 25/04/15 14:50:42 INFO DAGScheduler: ResultStage 1 (takeOrdered at /app/query.py:247) finished in 0.869 s
cluster-master 25/04/15 14:50:42 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master 25/04/15 14:50:42 INFO YarnScheduler: Killing all running tasks in stage 1: stage finished
cluster-master 25/04/15 14:50:42 INFO DAGScheduler: Job 0 finished: takeOrdered at /app/query.py:247, took 2.753829 s
cluster-master
cluster-master Search Results for: how to build nuclear bomb?
cluster-master =====
cluster-master 1. ID: 21691223 Score: 13.6283 Title: A Bomb Was Stolen
cluster-master 2. ID: 6602969 Score: 12.6151 Title: A Is for Atom
cluster-master 3. ID: 267889 Score: 8.2229 Title: A Gift Upon the Shore
cluster-master 4. ID: 33758502 Score: 7.9549 Title: A God Somewhere
cluster-master 5. ID: 18185446 Score: 7.4401 Title: A Frames (band)
cluster-master 6. ID: 280019 Score: 7.3438 Title: A Boy and His Dog
cluster-master 7. ID: 38752487 Score: 6.6167 Title: A Glimmer of Hope
cluster-master 8. ID: 32016408 Score: 6.1902 Title: A Dramatic Turn of Events
cluster-master 9. ID: 67227 Score: 6.0101 Title: A Brier History of Time
cluster-master 10. ID: 1022877 Score: 5.7333 Title: A Death in the Family (comics)
cluster-master 25/04/15 14:50:42 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master 25/04/15 14:50:42 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master 25/04/15 14:50:42 INFO YarnClientSchedulerBackend: Interrupting monitor thread
cluster-master 25/04/15 14:50:42 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster-master 25/04/15 14:50:42 INFO YarnSchedulerBackendsYarnDriverEndpoint: Asking each executor to shut down
cluster-master 25/04/15 14:50:42 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
cluster-master 25/04/15 14:50:42 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master 25/04/15 14:50:42 INFO MemoryStore: MemoryStore cleared
cluster-master 25/04/15 14:50:42 INFO BlockManager: BlockManager stopped
cluster-master 25/04/15 14:50:42 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master 25/04/15 14:50:42 INFO OutputCommitCoordinatorOutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master 25/04/15 14:50:42 INFO SparkContext: Successfully stopped SparkContext
cluster-master 25/04/15 14:50:43 INFO ShutdownHookManager: Shutdown hook called
cluster-master 25/04/15 14:50:43 INFO ShutdownHookManager: Deleting directory /tmp/spark-f6b9542e-e655-4a87-9086-4218ee9f53ab/pyspark-2695aa13-12f8-45ab-9f11-2cee639a0c66
cluster-master 25/04/15 14:50:43 INFO ShutdownHookManager: Deleting directory /tmp/spark-1610f723-0394-4748-a2e0-89c8eecd5d1
cluster-master 25/04/15 14:50:43 INFO ShutdownHookManager: Deleting directory /tmp/spark-f6b9542e-e655-4a87-9086-4218ee9f53ab
cluster-master 25/04/15 14:50:43 INFO ShutdownHookManager: Deleting directory /tmp/spark-f6b9542e-e655-4a87-9086-4218ee9f53ab
cluster-master cluster-master exited with code 0
cluster-master
cluster-master Enable Watch
```

Figure 18: Query 3 results