

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: 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. To tokenize we convert text to lowercase, remove special characters, and split it by whitespaces.

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).

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:

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

The **max** in denominator is needed to avoid division by zero and in **log** to avoid negative values, so the formula is more stable for unexpected input

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

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

```

zaurali@zaurali:~/Documents/developer/INNO_S25/BD/final_ass_2/big_data_assignment_2
25/04/15 18:39:06 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster master
25/04/15 18:39:06 INFO MemoryStore: MapOutputTrackerMasterEndpoint stopped!
cluster master
25/04/15 18:39:06 INFO MemoryStore: MemoryStore cleared
cluster master
25/04/15 18:39:06 INFO BlockManager: BlockManager stopped
cluster master
25/04/15 18:39:06 INFO OutputCommitCoordinator: OutputCommitCoordinator stopped!
cluster master
25/04/15 18:39:06 INFO OutputCommitCoordinator: OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster master
25/04/15 18:39:06 INFO ShutdownHookManager: Shutdown hook called
cluster master
25/04/15 18:39:06 INFO ShutdownHookManager: Deleting directory /tmp/spark-ab152793-c564-48dc-b1e1-b3a3f6546b08
cluster master
25/04/15 18:39:06 INFO ShutdownHookManager: Deleting directory /tmp/spark-064503b4-a7f8-4f88-9890-559ff733a5
cluster master
Putting data to hdfs
cluster master
Data directory uploaded
cluster master
Found 100 items
cluster master
-rw-r--r-- 1 root supergroup 3284 2025-04-15 18:42 /data/10031136 A Decade in the Grave.txt
cluster master
-rw-r--r-- 1 root supergroup 529 2025-04-15 18:44 /data/10078432 A Case for the Court.txt
cluster master
-rw-r--r-- 1 root supergroup 616 2025-04-15 18:43 /data/1009975 A Different Light album.txt
cluster master
-rw-r--r-- 1 root supergroup 607 2025-04-15 18:43 /data/1010309 A Different Light book.txt
cluster master
-rw-r--r-- 1 root supergroup 591 2025-04-15 18:40 /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 18:39 /data/10223157 A Balinese Trance Seance.txt
cluster master
-rw-r--r-- 1 root supergroup 3188 2025-04-15 18:44 /data/10238605 A Book of Family Genetics.txt
cluster master
-rw-r--r-- 1 root supergroup 814 2025-04-15 18:43 /data/10238605 A Book Sinking Story.txt
cluster master
-rw-r--r-- 1 root supergroup 310 2025-04-15 18:41 /data/10254892 A Flat Men.txt
cluster master
-rw-r--r-- 1 root supergroup 800 2025-04-15 18:39 /data/10254919 A Dolls House 1973 Loevy film.txt
cluster master
-rw-r--r-- 1 root supergroup 16918 2025-04-15 18:41 /data/10254939 A House of Our Time.txt
cluster master
-rw-r--r-- 1 root supergroup 5718 2025-04-15 18:41 /data/10399316 A Flowering Tree.txt
cluster master
-rw-r--r-- 1 root supergroup 2435 2025-04-15 18:44 /data/10534791 A Black and White World.txt
cluster master
-rw-r--r-- 1 root supergroup 1138 2025-04-15 18:44 /data/10534791 A Black and White World book.txt
cluster master
-rw-r--r-- 1 root supergroup 16899 2025-04-15 18:39 /data/1067891 A Hard Days Night song.txt
cluster master
-rw-r--r-- 1 root supergroup 1098 2025-04-15 18:44 /data/1083442 A Hillbilly Tribute to ACDC.txt
cluster master
-rw-r--r-- 1 root supergroup 17000 2025-04-15 18:44 /data/1083442 A Hillbilly Tribute to Donny B.txt
cluster master
-rw-r--r-- 1 root supergroup 6764 2025-04-15 18:42 /data/10858097 A Dangerous Path.txt
cluster master
-rw-r--r-- 1 root supergroup 12157 2025-04-15 18:44 /data/10900703 A Dictionary of Canadianisms_on_Historical_Principles.txt
cluster master
-rw-r--r-- 1 root supergroup 2806 2025-04-15 18:43 /data/10917293 A Bad Spell in Yurtland.txt
cluster master
-rw-r--r-- 1 root supergroup 4450 2025-04-15 18:41 /data/10917293 A Book of Canadian Dianetics.txt
cluster master
-rw-r--r-- 1 root supergroup 923 2025-04-15 18:41 /data/11161641 A Blueprint of the World.txt
cluster master
-rw-r--r-- 1 root supergroup 2573 2025-04-15 18:43 /data/1118101 A Hanging.txt
cluster master
-rw-r--r-- 1 root supergroup 1337 2025-04-15 18:43 /data/1118101 A Hanging book.txt
cluster master
-rw-r--r-- 1 root supergroup 588 2025-04-15 18:40 /data/11315857 A Go Go Potshot album.txt
cluster master
-rw-r--r-- 1 root supergroup 333 2025-04-15 18:44 /data/11490217 A Guide to Groovy_Lovin.txt
cluster master
-rw-r--r-- 1 root supergroup 500 2025-04-15 18:43 /data/11513735 A Handful of Greenies.txt
cluster master
-rw-r--r-- 1 root supergroup 7529 2025-04-15 18:42 /data/11513735 A Handful of the West.txt
cluster master
-rw-r--r-- 1 root supergroup 1029 2025-04-15 18:39 /data/11753053 A Journal of the Plague Year album.txt
cluster master
-rw-r--r-- 1 root supergroup 597 2025-04-15 18:43 /data/11874220 A Lifetime or More.txt
cluster master
-rw-r--r-- 1 root supergroup 1137 2025-04-15 18:43 /data/11874220 A Lifetime or More book.txt
Cluster master
-rw-r--r-- 1 root supergroup 863 2025-04-15 18:42 /data/11930321 A Fragile Hope.txt
cluster master
-rw-r--r-- 1 root supergroup 6843 2025-04-15 18:40 /data/11984610 A Catalogue of Crime.txt
cluster master
-rw-r--r-- 1 root supergroup 7440 2025-04-15 18:41 /data/12132596 A Christmas Carol.txt
cluster master
-rw-r--r-- 1 root supergroup 1098 2025-04-15 18:41 /data/12132596 A Crystal Christmas.txt
cluster master
-rw-r--r-- 1 root supergroup 7681 2025-04-15 18:42 /data/12212399 A Flintstones Christmas Carol.txt

```

Figure 1: HDFS ls /data.

```

zaurali@zaurali:~/Documents/developer/INNO_S25/BD/final_ass_2/big_data_assignment_2
25/04/15 18:39:06 /Index/data/_SUCCESS
cluster master
-rw-r--r-- 1 root supergroup 3793 2025-04-15 18:39 /data/9704239 A Contention for Honor and Riches.txt
cluster master
-rw-r--r-- 1 root supergroup 7029 2025-04-15 18:41 /data/9860012 A Dream Come True.txt
cluster master
-rw-r--r-- 1 root supergroup 1479 2025-04-15 18:41 /data/9860012 A Dream Come True song.txt
cluster master
-rw-r--r-- 1 root supergroup 1968 2025-04-15 18:40 /data/9870217 A Date with Lucy.txt
cluster master
-rw-r--r-- 1 root supergroup 5447 2025-04-15 18:43 /data/9919932 A Family Affair musical.txt
cluster master
-rw-r--r-- 1 root supergroup 619 2025-04-15 18:43 /data/9919932 A Family Affair musical book.txt
cluster master
-rw-r--r-- 1 root supergroup 896 2025-04-15 18:39 /data/9962726 A Book of Human Language.txt
cluster master
-rw-r--r-- 1 root supergroup 412 2025-04-15 18:43 /data/9982828 A Good Enough Day.txt
cluster master
Found 10 items
cluster master
-rw-r--r-- 1 root supergroup 0 2025-04-15 18:39 /Index/data/_SUCCESS
cluster master
-rw-r--r-- 1 root supergroup 3558533 2025-04-15 18:39 /Index/data/part-00000-66dbec81-05c1-43cd-9cal-2c7c4b3a34d5-c000.csv
cluster master
Deleted /tmp/index/step1
cluster master
Deleted /tmp/index/step2
cluster master
Deleted /tmp/index/step3
cluster master
Starting first MapReduce job - Term frequencies and positions...
cluster master
packageJobJar: [/tmp/streamjob11033767291884/] /tmp/streamjob11033767291884.jar tmpJobJar
cluster master
29/04/15 18:44:58.558 INFO client.YarnClientImpl: DelegationTokenRenewer: Connecting to ResourceManager as cluster-master/172.28.0.5:8882
cluster master
29/04/15 18:44:58.558 INFO client.YarnClientImpl: DelegationTokenRenewer: Connecting to ResourceManager as cluster-master/172.28.0.5:8882
cluster master
2025-04-15 18:45:05.753 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/_staging/job_1744742297326_0001
cluster master
2025-04-15 18:45:05.784 INFO mapred.FileInputFormat: Total input file(s) to process : 1
cluster master
2025-04-15 18:45:05.784 INFO mapred.FileInputFormat: Total input file(s) to process : 1
cluster master
2025-04-15 18:45:59.144 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744742297326_0001
cluster master
2025-04-15 18:45:59.144 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744742297326_0001
cluster master
2025-04-15 18:45:59.283 INFO mapreduce.ResourceUploader: Unable to find 'resource-types.xml'.
cluster master
2025-04-15 18:44:59.671 INFO impl.YarnClientImpl: Submitted application application_1744742297326_0001
cluster master
2025-04-15 18:44:59.671 INFO impl.YarnClientImpl: Application report from cluster-master/172.28.0.5:8882: appattempt_1744742297326_0001
cluster master
2025-04-15 18:44:59.712 INFO mapreduce.Job: Running job: job_1744742297326_0001
cluster master
2025-04-15 18:45:05.783 INFO mapreduce.Job: Job job_1744742297326_0001 running in uber mode : false
cluster master
2025-04-15 18:45:05.784 INFO mapreduce.Job: map 0% reduce 0%
cluster master
2025-04-15 18:45:05.784 INFO mapreduce.Job: map 100% reduce 0%
cluster master
2025-04-15 18:45:15.873 INFO mapreduce.Job: Job job_1744742297326_0001 reduce 100%
cluster master
2025-04-15 18:45:15.883 INFO mapreduce.Job: Job job_1744742297326_0001 completed successfully
cluster master
File System Counters
cluster master
FILE: Number of bytes read=11795296
cluster master
FILE: Number of bytes written=24429100
cluster master
FILE: Number of large read operations=0
cluster master
FILE: Number of write operations=0
cluster master
HDFS: Number of bytes read=11795296
cluster master
HDFS: Number of bytes written=6728946
cluster master
HDFS: Number of read operations=1
cluster master
HDFS: Number of write 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 2: HDFS ls /index/data.

```

zaurall@zaurall:~/Documents/developer/INNO_525/BD/final_ass_2/big_data_assignment_2
cluster-master 2025-04-15 18:45:05,973 INFO mapreduce.Job: Job job_1744742297326_0001 running in uber mode : false
cluster-master 2025-04-15 18:45:09,639 INFO mapreduce.Job: map 100% reduce 0%
cluster-master 2025-04-15 18:45:15,873 INFO mapreduce.Job: map 100% reduce 100%
cluster-master 2025-04-15 18:45:16,881 INFO mapreduce.Job: Job job_1744742297326_0001 completed successfully
cluster-master 2025-04-15 18:45:16,881 INFO mapreduce.Job: Counters: 54
cluster-master | File System Counters
cluster-master |   FILE: Number of bytes read=0
cluster-master |   FILE: Number of bytes written=24420106
cluster-master |   FILE: Number of read operations=0
cluster-master |   FILE: Number of large read operations=0
cluster-master |   HDFS: Number of bytes read=350221
cluster-master |   HDFS: Number of bytes written=6728946
cluster-master |   HDFS: Number of read operations=0
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
Job Counters:
cluster-master | Launched map tasks=2
cluster-master | Launched reduce tasks=1
cluster-master | Data locality: map tasks=1
cluster-master | Total time spent by all maps in occupied slots (ms)=4476
cluster-master | Total time spent by all reduces in occupied slots (ms)=3167
cluster-master | Total vcore-milliseconds taken by all map tasks=4476
cluster-master | Total vcore-milliseconds taken by all reduce tasks=3167
cluster-master | Total megabyte-milliseconds taken by all map tasks=4583424
cluster-master | Total megabyte-milliseconds taken by all reduce tasks=3243008
cluster-master | Map-Reduce Framework
cluster-master |   Map input records=1800
cluster-master |   Map output records=500022
cluster-master |   Map output bytes=10635246
cluster-master |   Map output materialized bytes=11795302
cluster-master |   Input File(s) = /tmp/index/step1
cluster-master |   Combine input records=0
cluster-master |   Combine output records=0
cluster-master |   Reduce input groups=10004
cluster-master |   Reduce shuffle bytes=11795302
cluster-master |   Reduce input records=500022
cluster-master |   Reduce output records=2500952
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)=103
cluster-master |   CPU time spent (ms)=4768
cluster-master |   Physical memory (bytes) snapshot=9277668384
cluster-master |   Virtual memory (bytes) snapshot=7779811808
cluster-master |   Total committed heap usage (bytes)=942669824
zaurall@zaurall:~/Documents/developer/INNO_525/BD/final_ass_2/big_data_assignment_2

```

Figure 3: MapReduce pipeline 1

```

zaurall@zaurall:~/Documents/developer/INNO_525/BD/final_ass_2/big_data_assignment_2
cluster-master 2025-04-15 18:45:05,973 INFO mapreduce.Job: Physical memory (bytes) snapshot=9277668384
cluster-master 2025-04-15 18:45:05,973 INFO mapreduce.Job: Virtual memory (bytes) snapshot=7779811808
cluster-master 2025-04-15 18:45:05,973 INFO mapreduce.Job: Total committed heap usage (bytes)=942669824
cluster-master 2025-04-15 18:45:05,973 INFO mapreduce.Job: Peak Map Physical memory (bytes)=348459088
cluster-master 2025-04-15 18:45:05,973 INFO mapreduce.Job: Peak Map Virtual memory (bytes)=2128752
cluster-master 2025-04-15 18:45:05,973 INFO mapreduce.Job: Peak Reduce Physical memory (bytes)=2128752
cluster-master 2025-04-15 18:45:05,973 INFO mapreduce.Job: Peak Reduce Virtual memory (bytes)=2599866368
Shuffle Errors
cluster-master | File+IO=0
cluster-master | CONNECTION=0
cluster-master | IO ERROR=0
cluster-master | WRONGHOSTNAME=0
cluster-master | WRONG MAP=0
cluster-master | WRONG REDUCE=0
File Input Format Counters
cluster-master | Bytes Read=6728946
cluster-master | Bytes Written=6728946
cluster-master | File Output Format Counters
cluster-master | Bytes Written=6728946
2025-04-15 18:45:16,000 INFO streaming.StreamJob: Output directory: /tmp/index/step1
First MapReduce job succeeded. Sample output:
cluster-master | packagerJobJar: [/tmp/streamjob996230145529468002.jar] impDir=null
cluster-master | 2025-04-15 18:45:16,000 INFO client.RMProxy: Connecting to ResourceManager at cluster-master/172.28.0.5:8032
cluster-master | 2025-04-15 18:45:17,441 INFO client.NodemanagerHttpCallOverProxyProvider: Connecting to ResourceManager at cluster-master/172.28.0.5:8032
cluster-master | 2025-04-15 18:45:17,641 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/_staging/job_1744742297326_0002
cluster-master | 2025-04-15 18:45:17,641 INFO mapreduce.JobResourceUploader: Uploading local file to path: /tmp/hadoop-yarn/staging/root/_staging/job_1744742297326_0002
cluster-master | 2025-04-15 18:45:17,991 INFO mapreduce.JobSubmitter: number of splits=2
cluster-master | 2025-04-15 18:45:17,974 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744742297326_0002
cluster-master | 2025-04-15 18:45:17,974 INFO mapreduce.JobSubmitter: number of split tokens=2. Executing with tokens [1]
cluster-master | 2025-04-15 18:45:18,187 INFO mapreduce.JobResourceUploader: Resource Configuration: resource-types.xml not found
cluster-master | 2025-04-15 18:45:18,187 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master | 2025-04-15 18:45:18,187 INFO impl.YarnClientImpl: Submitted application application_1744742297326_0002
cluster-master | 2025-04-15 18:45:18,187 INFO impl.YarnClientImpl: Application report for application_1744742297326_0002 from cluster-master:8088/proxy/application_1744742297326_0002
cluster-master | 2025-04-15 18:45:18,198 INFO mapreduce.Job: Running job: job_1744742297326_0002
cluster-master | 2025-04-15 18:45:26,281 INFO mapreduce.Job: map 100% reduce 100%
cluster-master | 2025-04-15 18:45:30,332 INFO mapreduce.Job: map 100% reduce 0%
cluster-master | 2025-04-15 18:45:35,361 INFO mapreduce.Job: map 100% reduce 100%
cluster-master | 2025-04-15 18:45:35,372 INFO mapreduce.Job: Job job_1744742297326_0002 completed successfully
cluster-master | 2025-04-15 18:45:35,372 INFO mapreduce.Job: Counters: 54
cluster-master | File System Counters
cluster-master |   FILE: Number of bytes read=0
cluster-master |   FILE: Number of bytes written=1985
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=0
cluster-master |   HDFS: Number of large read operations=0
cluster-master |   HDFS: Number of write operations=2
zaurall@zaurall:~/Documents/developer/INNO_525/BD/final_ass_2/big_data_assignment_2

```

Figure 4: MapReduce pipeline 2

```
zaurali@zaurali:~/Documents/developer/Hadoop_S25/BD/big_data_assignment_2$ ./big_data_assignment_2
zaurali@zaurali:~/Documents/developer/Hadoop_S25/BD/big_data_assignment_2$ ./big_data_assignment_2
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)=3955
cluster-master      Total time spent by all maps reduces in occupied slots (ms)=1963
cluster-master      Total time spent by all map tasks (ms)=5955
cluster-master      Total time spent by all reduce tasks (ms)=1963
cluster-master      Total vcore-milliseconds taken by all map tasks=3955
cluster-master      Total vcore-milliseconds taken by all reduce tasks=1963
cluster-master      Total negabyte-milliseconds taken by all map tasks=404920
cluster-master      Total negabyte-milliseconds taken by all reduce tasks=201012
Map-Reduce Metrics
cluster-master      Map input records=259952
cluster-master      Map output records=259952
cluster-master      Map output bytes=1095752
cluster-master      Map output materialized bytes=4351991
cluster-master      Input split bytes=210
cluster-master      Combine input records=0
cluster-master      Combiner input bytes=0
cluster-master      Reduce input groups=40864
cluster-master      Reduce shuffle bytes=4351991
cluster-master      Reduce shuffle bytes=4351992
cluster-master      Reduce output Records=40864
cluster-master      Spilled Records=501904
cluster-master      Shuffled Maps=1
cluster-master      Failed Shuffles=0
cluster-master      Merged Map outputs=2
cluster-master      GC time elapsed (ms)=112
cluster-master      CPU time spent (ms)=120
cluster-master      Physical memory (bytes) snapshot=990597120
cluster-master      Virtual memory (bytes) snapshot=777825944
cluster-master      Total physical memory (bytes)=990597120
cluster-master      Peak Map Physical memory (bytes)=367321888
cluster-master      Peak Map Virtual memory (bytes)=2591084544
cluster-master      Max Reduce Physical memory (bytes)=256366448
cluster-master      Peak Reduce Virtual memory (bytes)=259438016
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_MAGIC=0
cluster-master          WRONG_REDUCE=0
cluster-master      File Input Format Counters
cluster-master          Bytes Read=6733842
cluster-master      File Output Format Counters
```

Figure 5: MapReduce pipeline 2

```
zaurali@zaurali:~/Documents/develop/NINO_S25/BD$ final_ass_2/big_data_assignment_2
zaurali@zaurali:~/Documents/develop/NINO_S25/BD$ final_ass_2/big_data_assignment_2

cluster:master      File Input Format Counters
cluster:master          Bytes Read=6733042
cluster:master      File Output Format Counters
cluster:master          Bytes Written=905398
cluster:master      Map Reducer Counters
cluster:master          map tasks=1
cluster:master          Launched map tasks=1
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)=494
cluster:master      Total time spent by all maps in occupied slots (ms)=1562
cluster:master      Total time spent by all map tasks (ms)=3494
cluster:master      Total time spent by all reduce tasks (ms)=1562
cluster:master      Total vcore-milliseconds taken by all map tasks=304
cluster:master      Total vcore-milliseconds taken by all reduce tasks=3162
cluster:master      Total megabyte-milliseconds taken by all map tasks=3577856
cluster:master      Total megabyte-milliseconds taken by all reduce tasks=1599488
cluster:master      Map-Reduce Framework
cluster:master          map tasks=1
cluster:master          reduce tasks=1
cluster:master          Launched map tasks=1
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)=494
cluster:master      Total time spent by all maps in occupied slots (ms)=1562
cluster:master      Total time spent by all map tasks (ms)=3494
cluster:master      Total time spent by all reduce tasks (ms)=1562
cluster:master      Total vcore-milliseconds taken by all map tasks=304
cluster:master      Total vcore-milliseconds taken by all reduce tasks=3162
cluster:master      Total megabyte-milliseconds taken by all map tasks=3577856
cluster:master      Total megabyte-milliseconds taken by all reduce tasks=1599488

2025-04-15 18:45:35.434 INFO Streaming.StreamJob: Output directory: /tmp/index/step2
cluster:master      Second MapReduce job succeeded!
cluster:master      Starting third MapReduce job - document metadata...
cluster:master      Job ID: job_1744742297326_0003 [mapred://streamjob3948244683491796.jar:tmpDir=null]
cluster:master      2025-04-15 18:45:36.832 INFO client.DefaultNotAMasterExceptionProvider: Connecting to ResourceManager at cluster-master/172.28.0.5:8082
cluster:master      2025-04-15 18:45:36.985 INFO client.DefaultNotAMasterFallbackProxyProvider: Connecting to ResourceManager at cluster-master/172.28.0.5:8082
cluster:master      2025-04-15 18:45:37.139 INFO client.RMProxy: Connecting to ResourceManager at cluster-master/172.28.0.5:8082
cluster:master      2025-04-15 18:45:37.531 INFO mapred.FileInputFormat: Total input files to process : 1
cluster:master      2025-04-15 18:45:37.594 INFO mapred.JobSubmitter: number of splits:1
cluster:master      2025-04-15 18:45:37.608 INFO mapred.JobSubmitter: Starting job: job_1744742297326_0003
cluster:master      2025-04-15 18:45:37.608 INFO conf.Configuration: resource-types.xml not found
cluster:master      2025-04-15 18:45:37.608 INFO conf.Configuration: unable to find resource types.xml
cluster:master      2025-04-15 18:45:37.608 INFO mapred.JobResourceRequests: unable to find resource types.xml
cluster:master      2025-04-15 18:45:37.764 INFO mapred.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744742297326_0003/
cluster:master      2025-04-15 18:45:37.764 INFO mapred.Job: Running job: job_1744742297326_0003
cluster:master      2025-04-15 18:45:45.792 INFO mapred.Job: Map 0% reduce 0%
cluster:master      2025-04-15 18:45:49.838 INFO mapred.Job: map 100% reduce 0%
cluster:master      2025-04-15 18:45:52.866 INFO mapred.Job: map 100% reduce 100%
cluster:master      2025-04-15 18:45:52.925 INFO mapred.Job: Job ID: job_1744742297326_0003 completed successfully
cluster:master      2025-04-15 18:45:52.925 INFO mapred.Job: Counters: 54

cluster:master      File System Counters
cluster:master          FILE: Number of bytes read=29405
cluster:master          FILE: Number of bytes written=905398
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          FILE: Number of bytes read=3560221
cluster:master          HDFS: Number of bytes read=3560221
cluster:master          HDFS: Number of bytes written=3560220
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)=494
cluster:master      Total time spent by all maps in occupied slots (ms)=1562
cluster:master      Total time spent by all map tasks (ms)=3494
cluster:master      Total time spent by all reduce tasks (ms)=1562
cluster:master      Total vcore-milliseconds taken by all map tasks=304
cluster:master      Total vcore-milliseconds taken by all reduce tasks=3162
cluster:master      Total megabyte-milliseconds taken by all map tasks=3577856
cluster:master      Total megabyte-milliseconds taken by all reduce tasks=1599488
```

Figure 6: MapReduce pipeline 3

```
zaurali@zaurali: ~/Documents/developer/NINO_S25/B0Final_ss_2/big_data_assignment_2 | zaurali@zaurali: ~/Documents/developer/NINO_S25/B0Final_ss_2/big_data_assignment_2
cluster-master | Total megabyte milliseconds taken by all map tasks=3577856
cluster-master | Total megabyte milliseconds taken by all reduce tasks=1599488
cluster-master | Map-Reduce Framework
cluster-master | Map input records=1000
cluster-master | Map output records=1800
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 | Reduce input records=1000
cluster-master | Reduce shuffle bytes=37951
cluster-master | Reduce input records=1000
cluster-master | Reduce shuffle bytes=10000
cluster-master | Spilled Records=2000
cluster-master | Shuffled Maps =2
cluster-master | Failed Maps =0
cluster-master | Merged Map outputs=2
cluster-master | GC time elapsed (ms)=110
cluster-master | CPU time spent (ms)=154
cluster-master | Physical memory snapshot=85725936
cluster-master | Virtual memory (bytes) snapshot=774777344
cluster-master | Total committed heap usage (bytes)=188628288
cluster-master | Peak heap committed memory (bytes)=188628288
cluster-master | Peak Max Virtual memory (bytes)=259118928
cluster-master | Peak Reduce Physical memory (bytes)=252423968
cluster-master | Peak Reduce Virtual memory (bytes)=2594991088
cluster-master | Shuffle bytes=0
cluster-master | BAD ID=0
cluster-master | CONNECTION=0
cluster-master | IS_EOF=0
cluster-master | WRONG_LENGTH=0
cluster-master | WRONG_MAP=0
cluster-master | WRONG_HEARTBEAT=0
cluster-master | File Input Format Counters
cluster-master | Bytes Read=355929
cluster-master | File Output Format Counters
cluster-master | Bytes Written=355929
cluster-master | 2025-04-15 18:45:52.025 INFO streaming.StreamJob: Output directory: /tmp/index/step3
cluster-master | Importing data into Cassandra...
cluster-master | 
cluster-master | 
cluster-master | [2025-04-15 18:45:53.610] -> WARNING: Downgrading core protocol version from 66 to 67 for 172.28.0.2:9042. To avoid this, it is best practice to explicitly set Cluster(protocol_version) to the desired version. http://datastax.github.io/python-driver/api/cassandra/cluster.html#core_protocol_version
cluster-master | [2025-04-15 18:45:53.610] -> INFO: Using core protocol version 67 for host '172.28.0.2:9042'. Note that this 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#core_protocol_version
cluster-master | [2025-04-15 18:45:53.704] -> INFO: Using datacenter='datacenter1' for DCAwareRoundRobinPolicy (via host '172.28.0.2:9042'); if incorrect, please specify a local_dc to the cluster
cluster-master | [2025-04-15 18:45:53.787] -> INFO: Successfully connected to Cassandra cluster
cluster-master | [2025-04-15 18:45:53.806] -> INFO: Successfully created keyspace and tables
cluster-master | [2025-04-15 18:45:53.896] -> INFO: Reading term_index data from /tmp/index/step1
```

Figure 7: MapReduce pipeline 3

```
zaural@zaural:~/Documents/developer/NINO_S25/BD$ ./big_data_assignment_2
zaural@zaural:~/Documents/developer/NINO_S25/BD$ ./big_data_assignment_2

cluster-master [2025-04-15 18:46:53.707] : INFO : Successfully fully connected to Cassandra cluster
cluster-master [2025-04-15 18:46:53.806] : INFO : Successfully created keyspaces and tables
cluster-master [2025-04-15 18:46:53.806] : INFO : Reading term index data from /tmp/index/step1
[WARN] [Native-Transport-Requests-2] 2025-04-15 18:46:08,858 BatchStatement.java:362 - Batch for [search_engine.term.document] is of size 5.020KB, exceeding specified threshold of 5.000KB by 20.
cluster-master [2025-04-15 18:46:08,861] : WARNING - Server warning: Batch for [search_engine.term.document] is of size 5140, exceeding specified threshold of 5120 by 20.
[WARN] [Native-Transport-Requests-2] 2025-04-15 18:46:14,912 BatchStatement.java:362 - Batch for [search_engine.term.document] is of size 5.099KB, exceeding specified threshold of 5.000KB by 99.
cassandra-server [2025-04-15 18:46:14,914] : WARNING - Server warning: Batch for [search_engine.term.document] is of size 5220, exceeding specified threshold of 5120 by 100.
[WARN] [Native-Transport-Requests-1] 2025-04-15 18:46:17,188 BatchStatement.java:362 - Batch for [search_engine.term.document] is of size 6.699KB, exceeding specified threshold of 5.000KB by 399.
cluster-master [2025-04-15 18:46:17,187] : INFO : Successfully imported 1000 document metadata entries
[WARN] [Native-Transport-Requests-1] 2025-04-15 18:46:17,187 BatchStatement.java:362 - Batch for [search_engine.term.document] is of size 6660, exceeding specified threshold of 5120 by 1740.
cassandra-server [2025-04-15 18:46:17,187] : WARNING - Server warning: Batch for [search_engine.term.document] is of size 5.008KB by 2508.
[WARN] [Native-Transport-Requests-1] 2025-04-15 18:46:39,733 BatchStatement.java:362 - Batch for [search_engine.term.document] is of size 5.244KB, exceeding specified threshold of 5.000KB by 244.
cluster-master [2025-04-15 18:46:39,734] : 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 18:46:42,584] : INFO : Successfully imported 250952 term index entries
cluster-master [2025-04-15 18:46:42,585] : INFO : Reading document frequency data from /tmp/index/step2
cluster-master [2025-04-15 18:46:42,585] : INFO : Processing document frequency data from /tmp/index/step2
cluster-master [2025-04-15 18:46:48,034] : INFO : Importing document metadata from Mapreduce output
cluster-master [2025-04-15 18:46:51,321] : INFO : Successfully imported 1000 document metadata entries
cluster-master [2025-04-15 18:46:51,321] : INFO : Successfully imported index data into Cassandra
cluster-master [2025-04-15 18:46:51,321] : INFO : Successfully imported index data into Cassandra
cluster-master [2025-04-15 18:46:51,321] : INFO : Searching for: how to build nuclear bomb
cluster-master [2025-04-15 18:46:52,928] : WARNING - Server warning: No heap memory available to load native-hadoop library for your platform... using builtin java classes where applicable
cluster-master [2025-04-15 18:46:53,308] : INFO : Processing query; how to build nuclear bomb?
cluster-master [2025-04-15 18:46:53,308] : [cassandra.cluster]:WARNING: cluster, init : called with contact points not 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}], bbf = None)
cluster-master [2025-04-15 18:46:53,308] : [cassandra.cluster]:INFO: Upgrading 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.Cluster.protocol_version
cluster-master [2025-04-15 18:46:53,310] : [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.Cluster.protocol_version
cluster-master [2025-04-15 18:46:53,310] : [cassandra.cluster]:INFO: Upgrading core protocol version from 65 to 66 for 172.28.0.2:9042; if incorrect, please specify a local dc in the constructor, or limit contact points to local cluster nodes
cluster-master [2025-04-15 18:46:53,310] : [search-engine]:INFO: Successfully connected to Cassandra cluster
cluster-master [2025-04-15 18:46:53,310] : [spark]:INFO: sparkContext: OS Linux, 6.8.0-51-generic, amd64
cluster-master [2025-04-15 18:46:53,310] : [spark]:INFO: sparkContext: Java version 1.8.0_422
cluster-master [2025-04-15 18:46:53,310] : [spark]:INFO: sparkContext: sparkHome /usr/local/spark
cluster-master [2025-04-15 18:46:53,310] : [spark]:INFO: ResourcesUtil: No custom resources configured for spark.driver
cluster-master [2025-04-15 18:46:53,310] : [spark]:INFO: ResourcesUtil: spark driver application has been selected to build nuclear bomb
cluster-master [2025-04-15 18:46:53,310] : [spark]:INFO: ResourcesUtil: Default ResourceProfile created: executor resources: Map[cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memroy, amount: 1024, script: , vendor: , ] ; fileSystem resources: Map[cpu -> name: cores, amount: 1, script: , vendor: , task resources: Map[cpu -> name: cpus, amount: 1.0]
cluster-master [2025-04-15 18:46:53,310] : [spark]:INFO: ResourceProfile: Limiting resource is cpus at 1 tasks per executor
cluster-master [2025-04-15 18:46:53,310] : [spark]:INFO: ResourceProfile: spark.executor.instances is set to: 1
cluster-master [2025-04-15 18:46:53,310] : [spark]:INFO: SecurityManager: Changing view acls to: root
cluster-master [2025-04-15 18:46:53,310] : [spark]:INFO: SecurityManager: Changing modify acls to: root
cluster-master [2025-04-15 18:46:53,310] : [spark]:INFO: SecurityManager: Changing delete acls to: root
cluster-master [2025-04-15 18:46:53,310] : [spark]:INFO: SecurityManager: Changing modify acls groups to:
cluster-master [2025-04-15 18:46:53,310] : [spark]:INFO: SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: root; groups with view permissions: EMPTY
; users with modify permissions: root; groups with modify permissions: EMPTY
```

Figure 8: import index data to Cassandra

```

zaurall@zaurall:~/Documents/developer/INNO_525/BD/final_ass_2/big_data_assignment_2$ docker exec -it cassandra-server cqlsh
Connected to cluster at 127.0.0.1:9042
Cassandra 3.2.0 | Cassandra 3.0.4 | CQL spec 3.4.7 | Native protocol V5
Use HELP for help.
cqlsh> use search_engine ;
cqlsh> select count(*) from document_frequency;
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 document_frequency ;
count
-----
40864
(1 rows)
Warnings :
Aggregation query used without partition key
cqlsh> search_engine> select count(*) from documents_info ;
count
-----
1000
(1 rows)
Warnings :
Aggregation query used without partition key
cqlsh> search_engine> select count(*) from term_document ;
count
-----
250952
(1 rows)
Warnings :
Aggregation query used without partition key
cqlsh> search_engine> select * from document_frequency limit 10 ;
term          | df
-----+
dobson       | 1
sain         | 1
bessus       | 1
ix           | 4
await        | 2
libertad     | 1
-----+
250952
(1 rows)

```

Figure 9: Cassandra tables content

```

zaurall@zaurall:~/Documents/developer/INNO_525/BD/final_ass_2/big_data_assignment_2$ docker exec -it cassandra-server cqlsh
Connected to cluster at 127.0.0.1:9042
Cassandra 3.2.0 | Cassandra 3.0.4 | CQL spec 3.4.7 | Native protocol V5
Use HELP for help.
cqlsh> search_engine> select * from document_frequency limit 10 ;
term          | df
-----+
dobson       | 1
sain         | 1
bessus       | 1
ix           | 4
await        | 2
libertad     | 1
-----+
504908      | 39
previous     | 7
maclean      | 1
gibbs        | 6
(10 rows)
cqlsh> search_engine> select * from documents_info limit 10 ;
doc_id | length | title
-----+
1023665   | 115   | A Dead Sinking Story
27568194  | 567   | A Hero Ain't Nothin' but a Sandwich (film)
39710446  | 330   | A Little Bit of Luck
32320009  | 300   | A Change Is Gonna Come (Upstairs, Downstairs)
51794980  | 306   | A Family Secret (Upstairs, Downstairs)
56880098  | 564   | A J Balliol Salmon
29798396  | 63    | A Laundromat, A Girl Lives
37196774  | 1675  | A Case of Conscience
504908   | 1334  | A Case of Dusty
7765742   | 660   | A Girl Called Dusty
(10 rows)
cqlsh> search_engine> select * from term_document limit 10 ;
term          | doc_id | positions          | tf
-----+
dobson       | 14634600 | [1840, 1339, 1108, 1040] | 4
sain         | 14444655 | [1840, 1339, 1108, 1029] | 3
bessus       | 12000397 | [1617, 1006, 1029] | 3
ix           | 19789591 | [542] | 1
ix           | 19789591 | [1] | 1
ix           | 67078438 | [64, 159] | 2
ix           | 73922368 | [413] | 1
ix           | 4136530  | [130, 167] | 1
await        | 4136530  | [167] | 1
libertad     | 47812009 | [20, 41] | 2
(10 rows)
cqlsh> search_engine> 
```

Figure 10: Cassandra tables content

```

zaurali@zaurali:~/Documents/developer/INNO_S25/BD/final_ass_2/big_data_assignment_2
25/04/15 18:59:58 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 11.2 KiB, free: 366.3 MiB)
cluster master
25/04/15 18:59:58 INFO BlockManager: Added broadcast_1_piece0 in memory on cluster-slave-1 (size: 6.5 KiB, free: 366.3 MiB)
cluster master
25/04/15 18:59:58 INFO BlockManager: Added broadcast_1_piece0 in memory on cluster-slave-1 (size: 6.5 KiB, free: 366.3 MiB)
cluster master
25/04/15 18:59:58 INFO DAGScheduler: Created broadcast 1 from broadcast at DAGScheduler.scala:1585
cluster master
25/04/15 18:59:58 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 18:59:58 INFO YarnScheduler: Adding task set 1.0 with 2 tasks resource profile 0
cluster master
25/04/15 18:59:58 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 2) (cluster-slave-2, executor 1, partition 0, NODE LOCAL, 8828 bytes)
cluster master
25/04/15 18:59:58 INFO BlockManager: Added broadcast_1_piece0 in memory on cluster-slave-1 (executor 2, partition 1, NODE LOCAL, 8828 bytes)
cluster master
25/04/15 18:59:58 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on cluster-slave-1:246357 (size: 6.5 KiB, free: 366.3 MiB)
cluster master
25/04/15 18:59:58 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on cluster-slave-1:24847 (size: 6.5 KiB, free: 366.3 MiB)
cluster master
25/04/15 18:59:58 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to 172.28.0.4:56022
cluster master
25/04/15 18:59:59 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2) in 771 ms on cluster-slave-2 (executor 1) (1/2)
cluster master
25/04/15 18:59:59 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (TID 3) in 786 ms on cluster-slave-1 (executor 2) (2/2)
cluster master
25/04/15 18:59:59 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
cluster master
25/04/15 18:59:59 INFO DAGScheduler: Killing all running tasks in stage 1: Stage finished
cluster master
25/04/15 18:59:59 INFO DAGScheduler: Job 0 is finished: takeOrdered at /app/query.py:247, took 2.610519 s
cluster master
Search Results for: how to gain weight
=====
1. ID: 63546062 Score: 6.6600 Title: A Lift in Sin
cluster master
2. ID: 55598139 Score: 6.2130 Title: A Day in the Life of the Dummies
cluster master
3. ID: 11528779 Score: 5.9403 Title: A Day in the Cotton
cluster master
4. ID: 11528779 Score: 5.7043 Title: A Dreamer's Tales
cluster master
5. ID: 37676240 Score: 5.3910 Title: A Date with the Falcon
cluster master
6. ID: 11528779 Score: 5.3890 Title: A Day in the Cotton
cluster master
7. ID: 7090541 Score: 4.4212 Title: A Gold 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: 11528779 Score: 3.7933 Title: A Day in the Cotton
cluster master
10. ID: 6682969 Score: 3.7933 Title: A Hairdresser's Experience in High Life
cluster master
25/04/15 18:59:59 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster master
25/04/15 18:59:59 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster master
25/04/15 18:59:59 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster master
25/04/15 18:59:59 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
cluster master
25/04/15 18:59:59 INFO YarnClientSchedulerBackend: client scheduler backend Stopped
cluster master
25/04/15 18:59:59 INFO MemoryStore: MemoryStore cleared
cluster master
25/04/15 18:59:59 INFO BlockManager: BlockManager stopped
cluster master
25/04/15 18:59:59 INFO BlockManager: BlockManager stopped
cluster master
25/04/15 18:59:59 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster master
25/04/15 18:59:59 INFO SparkContext: Successfully stopped SparkContext
cluster master
25/04/15 19:00:00 INFO ShutdownHookManager: Deleting directory /tmp/spark-bfc83a98-7a97-44b6-89af-0f752b263bfe/pyspark-4d390c81-786a-40ba-b08c-28712684a184
cluster master
25/04/15 19:00:00 INFO ShutdownHookManager: Deleting directory /tmp/spark-a0bc218-6465-4cd4-95e4-87611288e04
cluster master
25/04/15 19:00:00 INFO ShutdownHookManager: Deleting directory /tmp/spark-bfc83a98-7a97-44b6-89af-0f752b263bfe
cluster master exited with code 0

```

Figure 11: Query 1 results

```

zaurali@zaurali:~/Documents/developer/INNO_S25/BD/final_ass_2/big_data_assignment_2
25/04/15 19:07:43 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 19:07:43 INFO BlockManager: Added broadcast_1_piece0 in memory on cluster-slave-1 (size: 6.5 KiB, free: 366.3 MiB)
cluster master
25/04/15 19:07:43 INFO DAGScheduler: Created broadcast 1 from broadcast at DAGScheduler.scala:1585
cluster master
25/04/15 19:07:43 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 19:07:43 INFO YarnScheduler: Adding task set 1.0 with 2 tasks resource profile 0
cluster master
25/04/15 19:07:43 INFO TaskSetManager: Starting task 1.0 in stage 1.0 (TID 3) (cluster-slave-1, executor 1, partition 0, NODE LOCAL, 8828 bytes)
cluster master
25/04/15 19:07:43 INFO BlockManager: Added broadcast_1_piece0 in memory on cluster-slave-1 (executor 2, partition 1, NODE LOCAL, 8828 bytes)
cluster master
25/04/15 19:07:44 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to 172.28.0.3:60128
cluster master
25/04/15 19:07:44 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (TID 3) in 211 ms on cluster-slave-1 (executor 2) (1/2)
cluster master
25/04/15 19:07:44 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2) in 795 ms on cluster-slave-1 (executor 1) (2/2)
cluster master
25/04/15 19:07:44 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
cluster master
25/04/15 19:07:44 INFO DAGScheduler: Killing all running tasks in stage 1: Stage finished
cluster master
25/04/15 19:07:44 INFO DAGScheduler: Job 0 finished: takeOrdered at /app/query.py:247, took 2.559577 s
cluster master
Search Results for: this is a query!
=====
1. ID: 47515595 Score: 8.0019 Title: A Canine Sherlock Holmes
cluster master
2. ID: 30628228 Score: 1.6101 Title: A Human Right
cluster master
3. ID: 18171042 Score: 1.5942 Title: A Chrestomathy
cluster master
4. ID: 10000000 Score: 1.5880 Title: A Day in the Cotton
cluster master
5. ID: 57279810 Score: 1.5804 Title: A Book of American Martyrs
cluster master
6. ID: 41801556 Score: 1.5638 Title: A (The Walking Dead)
cluster master
7. ID: 45350000 Score: 1.5600 Title: A Bearded Man
cluster master
8. ID: 33788184 Score: 1.5480 Title: A Dangerous Guy
cluster master
9. ID: 62485656 Score: 1.5363 Title: A Calf for Christmas
cluster master
10. ID: 2761148 Score: 1.5351 Title: History of Philosophy (Copelston)
cluster master
25/04/15 19:07:44 INFO YarnClientSchedulerBackend: Stopping SparkContext with exitCode 0.
cluster master
25/04/15 19:07:44 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster master
25/04/15 19:07:44 INFO YarnClientSchedulerBackend: Interrupting monitor thread
cluster master
25/04/15 19:07:44 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster master
25/04/15 19:07:44 INFO MapOutputTrackerMasterEndpoint: Asked each executor to shut down
cluster master
25/04/15 19:07:44 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
cluster master
25/04/15 19:07:44 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster master
25/04/15 19:07:44 INFO BlockManager: BlockManager stopped
cluster master
25/04/15 19:07:44 INFO BlockManagerMaster: BlockManager stopped
cluster master
25/04/15 19:07:44 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster master
25/04/15 19:07:45 INFO SparkContext: Successfully stopped SparkContext
cluster master
25/04/15 19:07:45 INFO ShutdownHookManager: Shutdown hook called
cluster master
25/04/15 19:07:45 INFO ShutdownHookManager: Deleting directory /tmp/spark-e95ch134-3c68-4ce1-8875-421db8e9a1f5
cluster master
25/04/15 19:07:45 INFO ShutdownHookManager: Deleting directory /tmp/spark-e110855d-052b-4655-a57a-5d897d23487e
cluster master
25/04/15 19:07:45 INFO ShutdownHookManager: Deleting directory /tmp/spark-e110855d-052b-4655-a57a-5d897d23487e
cluster master
25/04/15 19:07:45 INFO ShutdownHookManager: Deleting directory /tmp/spark-5318fa2a-be70-4b90-a432-390ec6725d06
cluster master exited with code 0

```

Figure 12: Query 2 results

```

zaurall@zaurall:~/Documents/developer/INNO_525/BD/final_ass_2/big_data_assignment_2
zaurall@zaurall:~/Documents/developer/INNO_525/BD/final_ass_2/big_data_assignment_2
cluster-master 25/04/15 18:47:21 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (ID 2) in 758 ms on cluster-slave-1 (executor 1) (1/2)
cluster-master 25/04/15 18:47:21 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (ID 2) in 758 ms on cluster-slave-2 (executor 2) (2/2)
cluster-master 25/04/15 18:47:21 INFO YarnScheduler: Removed TaskSet 1.0, whose tasks have all completed, from pool
cluster-master 25/04/15 18:47:21 INFO DAGScheduler: ResultStage 3 (takeOrdered at /app/query.py:247) finished in 0.788 s
cluster-master 25/04/15 18:47:21 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master 25/04/15 18:47:21 INFO DAGScheduler: Cancelling 0 speculative or all running tasks in stage 1: Stage finished
cluster-master 25/04/15 18:47:21 INFO DAGScheduler: Job 0 finished: takeOrdered at /app/query.py:247, took 2.615235 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: 1022877 Score: 10.0000 Title: A Bomb Was Stolen
cluster-master 3. ID: 2670853 Score: 8.2294 Title: A God Upon the Shore
cluster-master 4. ID: 33758502 Score: 7.9549 Title: A God Somewhere
cluster-master 5. ID: 18385446 Score: 7.4481 Title: A Frames (band)
cluster-master 6. ID: 3016486 Score: 7.0000 Title: A God's Plan to Stop
cluster-master 7. ID: 38752487 Score: 6.6167 Title: A Glimmer of Hope
cluster-master 8. ID: 32016486 Score: 6.1989 Title: A Dramatic Turn of Events
cluster-master 9. ID: 1022877 Score: 5.7933 Title: A Death in the Family (comics)
cluster-master 10. ID: 1022877 Score: 5.7933 Title: A Death in the Family (comics)
cluster-master 25/04/15 18:47:21 INFO SparkContext: SparkContext is stopping with exitcode 0
cluster-master 25/04/15 18:47:21 INFO SparkContext: Stopped Spark with url http://zaurall:4040
cluster-master 25/04/15 18:47:21 INFO SparkContext: Stopped spark://zaurall:4040@zaurall:4040
cluster-master 25/04/15 18:47:21 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster-master 25/04/15 18:47:21 INFO YarnSchedulerBackendYarnDriverEndpoint: Asking each executor to shut down
cluster-master 25/04/15 18:47:21 INFO YarnSchedulerBackendYarnDriverEndpoint: All client scheduler backend stopped
cluster-master 25/04/15 18:47:21 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master 25/04/15 18:47:21 INFO MemoryStore: MemoryStore cleared
cluster-master 25/04/15 18:47:21 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master 25/04/15 18:47:21 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master 25/04/15 18:47:21 INFO SparkContext: Successfully stopped SparkContext
cluster-master 25/04/15 18:47:22 INFO ShutdownHookManager: Deleting directory /tmp/spark-34fb9ffd-012b-4577-8ba7-347421dee965/pyspark-b0601216-11af-479b-858f-60ef219ed1ca
cluster-master 25/04/15 18:47:22 INFO ShutdownHookManager: Deleting directory /tmp/spark-ed078dd-4000-468d-83f0-521dc1305434
cluster master exited with code 0
cassandra-server  WARN [Native-Transport-Requests-1] 2025-04-15 18:48:20,278 SelectStatement.java:557 - Aggregation query used without partition key
cassandra-server  WARN [Native-Transport-Requests-1] 2025-04-15 18:48:20,278 NoSpareLogger.java:107 - Aggregation query used without partition key on table search.engine.document.frequency_aggregation type: AGGREGATE EVERYTHING
cassandra-server  INFO [ReadStage-2] 2025-04-15 18:48:20,781 MonitoringTask.java:95 Scheduling monitoring task with report interval of 5000 ms, max operations 50
cassandra-server  INFO [ScheduledTasks] 2025-04-15 18:48:20,781 NoSpareLogger.java:107 Some operations were slow, details available at tracing level (debug.log)
cassandra-server  WARN [Native-Transport-Requests-1] 2025-04-15 18:48:20,998 SelectStatement.java:557 - Aggregation query used without partition key
cassandra-server  WARN [Native-Transport-Requests-1] 2025-04-15 18:48:28,000 NoSpareLogger.java:107 - Aggregation query used without partition key on table search.engine.documents.info, aggregation type: AGGREGATE EVERYTHING
cassandra-server  WARN [Native-Transport-Requests-1] 2025-04-15 18:48:32,529 SelectStatement.java:557 - Aggregation query used without partition key
cassandra-server  WARN [Native-Transport-Requests-1] 2025-04-15 18:48:32,530 NoSpareLogger.java:107 - Aggregation query used without partition key on table search.engine.term_document, aggregation type: AGGREGATE EVERYTHING

```

Figure 13: Query 3 results