

Big Data Search Engine

using Hadoop MapReduce

Amine Trabelsi

Submitted for Big Data
Assignment 2

a.trabelsi@innopolis.university
group DS-01
Innopolis University
2025

Contents

Methodology	2
MapReduce Indexing Pipeline	2
Query Processing with BM25 Ranking	2
Demonstration	4
Running the code	4
Indexing Documents	4
Cassandra Inverted Index	5
Search Results	6
Observations	8

Methodology

MapReduce Indexing Pipeline

The search engine relies on a Hadoop MapReduce pipeline to tokenize documents and build an inverted index stored in Cassandra.

Mapper (`mapper1.py`)

Each line of the input dataset (stored in HDFS) is expected to follow this tab-separated format:

```
<doc_id>\t<title>\t<text>
```

The mapper performs the following operations:

- Splits each line into `doc_id`, `title`, and `text`.
- Tokenizes the text into lowercase alphanumeric terms using a regular expression (`(\w+)`).
- Emits a key-value pair for each term occurrence in the format:

```
<term>\t<doc_id>\t1
```

This effectively prepares the data for counting how many times each term appears in each document.

Reducer (`reducer1.py`)

The reducer receives all emitted term-document pairs sorted by key, and:

- Aggregates the total frequency (`tf`) of each term per document.
- Inserts the final result into a Cassandra table called `inverted_index`, with the schema:

```
(term TEXT, doc_id TEXT, tf INT,  
PRIMARY KEY (term, doc_id))
```

- All Cassandra table creation and insertion is handled directly within the reducer.
- The reducer also logs inserted terms to standard error for debugging purposes.

Query Processing with BM25 Ranking

After the inverted index is stored in Cassandra, a PySpark application (`query.py`) is used to process user queries and retrieve the top 10 most relevant documents using the BM25 scoring algorithm.

Workflow:

1. The user provides a query via command-line input.
2. The PySpark application connects to Cassandra and retrieves relevant entries from the `inverted_index` table using a `SELECT` query with all query terms.
3. Because only `(term, doc_id, tf)` data is available, the script:
 - Estimates document lengths by summing term frequencies (TFs) per document.
 - Assigns placeholder document titles (e.g., "Document_<doc_id>").
 - Uses static placeholder values for inverse document frequency (IDF) and average document length.
4. Using the PySpark RDD API:
 - The BM25 score is computed for each document-term pair using the formula:

$$\text{BM25} = \text{idf} * (\text{tf} * (\text{k1} + 1)) / (\text{tf} + \text{k1} * (1 - \text{b} + \text{b} * (\text{dl} / \text{avgdl})))$$

- Scores are aggregated per document.
 - The top 10 documents are selected based on the total BM25 score.
5. The final output displays the document IDs, their generated titles, and BM25 scores in descending order.

Execution: The PySpark application is launched in a distributed fashion on the Hadoop YARN cluster via the `search.sh` script:

```
bash search.sh "search engine indexing"
```

Demonstration

Running the container

To run the code it is sufficient to run:

```
docker-compose up --build
```

this will build three images: cluster-master, cluster-slave, and cassandra-server

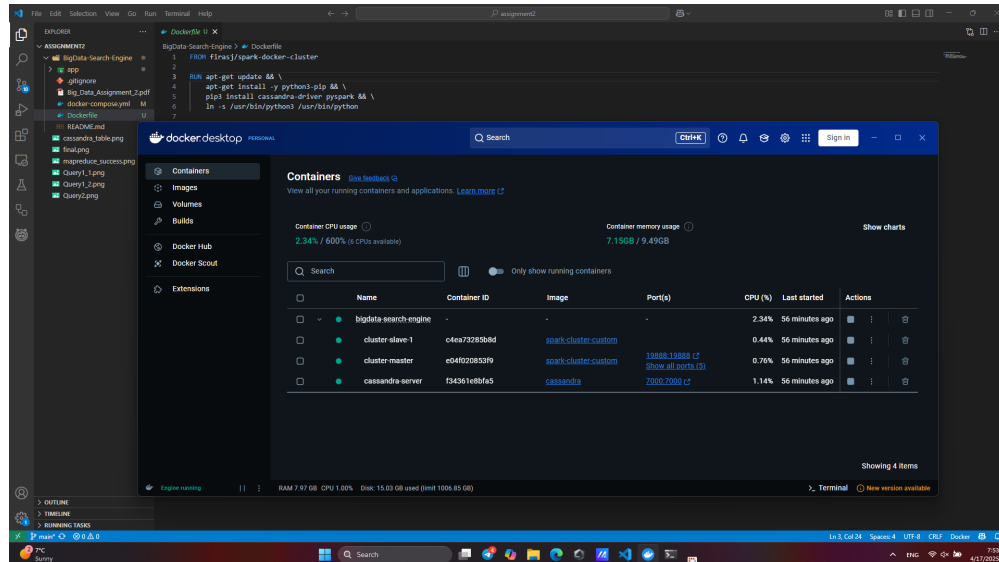


Figure 1: Running docker container

Indexing Documents with Hadoop MapReduce

The following screenshot shows the successful execution of the MapReduce pipeline over 993 documents stored in HDFS. The mapper and reducer jobs were executed in a fully distributed environment using Hadoop YARN.

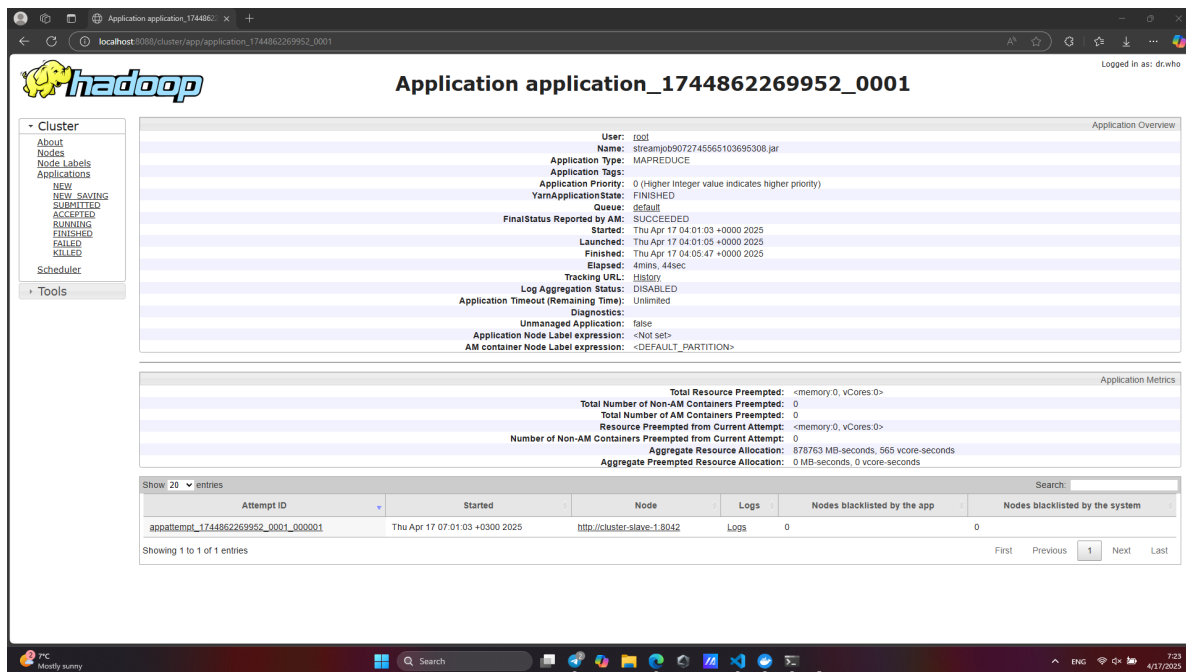


Figure 2: Hadoop YARN UI: MapReduce job completed successfully

Inverted Index Stored in Cassandra

After the reducer finishes, the term frequencies are stored in the `inverted_index` table in the Cassandra database. Below is a screenshot of a sample query run using `cqlsh`.

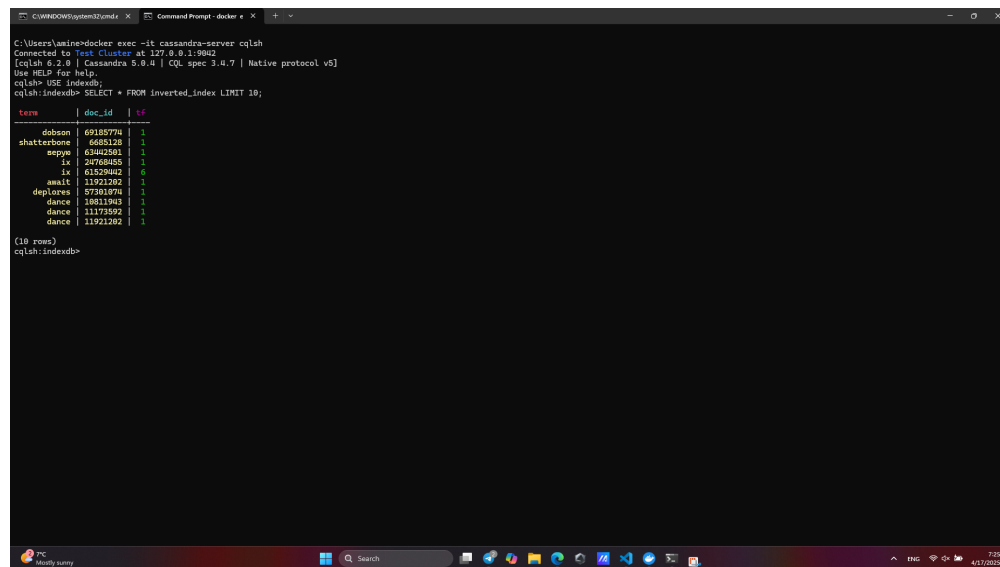
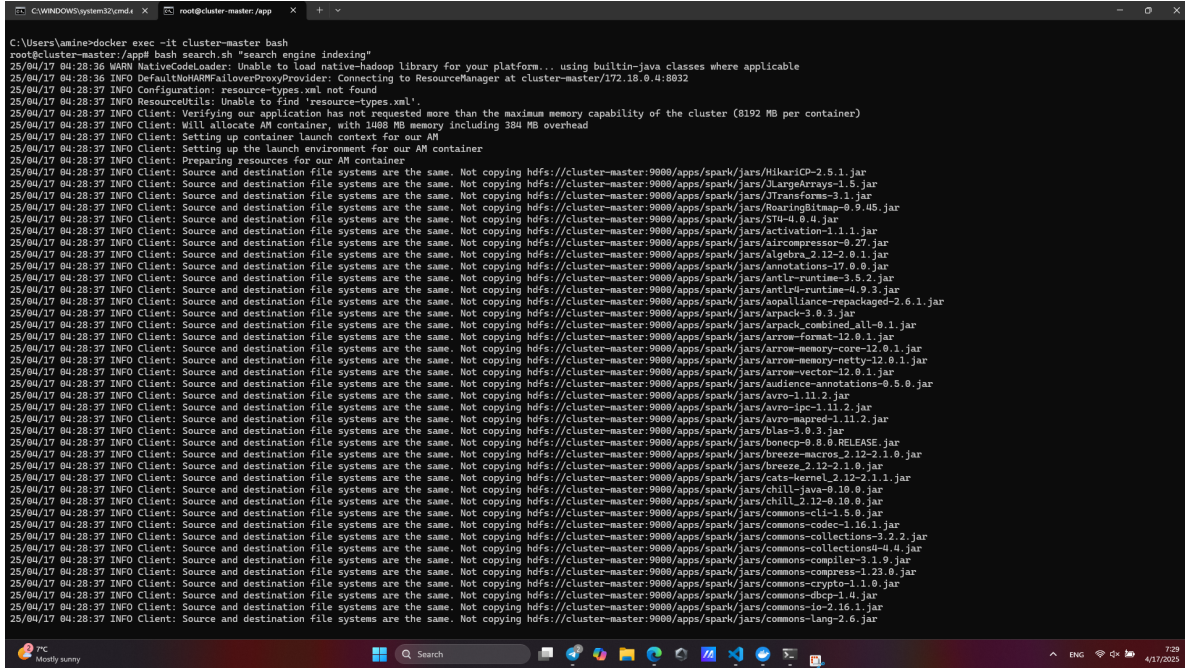


Figure 3: Sample rows from inverted_index table in Cassandra

Search Results for Example Queries

The following screenshots show the results of running the BM25-based search engine using `query.py` and retrieving the top 10 most relevant documents for each query.

Query 1: search engine indexing



```
C:\Users\andine-docker> exec -it cluster-master bash
root@cluster-master:/app# bash search.sh "search engine indexing"
25/04/17 04:28:36 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
25/04/17 04:28:37 INFO Configuration: resource-types.xml not found
25/04/17 04:28:37 INFO Configuration: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
25/04/17 04:28:37 INFO ResourceUtils: Unable to find 'resource-types.xml'.
25/04/17 04:28:37 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (8192 MB per container)
25/04/17 04:28:37 INFO Client: Will allocate AM container, with 1088 MB memory including 384 MB overhead
25/04/17 04:28:37 INFO Client: Setting up container launch context for our AM
25/04/17 04:28:37 INFO Client: Setting up the launch environment for our AM container
25/04/17 04:28:37 INFO Client: Preparing resources for our AM container
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/HikariCP-2.5.1.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/3LargeArrays-1.5.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/JTransforms-3.1.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/STU-4.0.4.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/activation-1.1.1.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/aircompressor-0.27.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/algebra_2.12-2.0.1.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/annotations-17.0.9.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/antlr-runtime-3.5.2.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/antlr4-runtime-4.9.3.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/aopalliance-repackaged-2.6.1.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/arrow-combined-all-0.1.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/arrow-format-12.0.1.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/arrow-memory-core-12.0.1.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/arrow-memory-netty-12.0.1.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/arrow-vector-12.0.1.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/audience-annotations-0.5.0.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/avro-1.11.2.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/avro-lpc-1.11.2.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/avro-mapred-1.11.2.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/bias-3.0.3.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/bonecp-0.8.0.RELEASE.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/breeze-macros-2.12-2.1.0.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/breeze-2.12-2.1.0.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/cats-kernel-2.12-2.1.1.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/chill-java-0.10.0.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/chill-2.12-0.10.0.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/commons-cli-1.5.0.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/commons-codec-1.16.1.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/commons-collections-3.2.2.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/commons-collections4-4.4.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/commons-compilr-3.1.9.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/commons-compress-1.23.0.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/commons-crypto-1.1.0.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/commons-dbc-1.4.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/commons-io-2.16.1.jar
25/04/17 04:28:37 INFO Client: Source and destination file systems are the same. Not copying hdfs://cluster-master:9800/apps/spark/jars/commons-lang-2.6.jar
```

Figure 4: Bash results for query: "search engine indexing"

```
C:\WINDOWS\system32\cmd.exe root@cluster-master: /app
25/04/17 04:28:37 INFO Client: Uploading resource file:/app/query.py -> hdfs://cluster-master:9000/user/root/.sparkStaging/application_1744862269952_0003/query.py
25/04/17 04:28:37 INFO Client: Uploading resource file:/usr/local/spark/python/lib/pyspark.zip -> hdfs://cluster-master:9000/user/root/.sparkStaging/application_1744862269952_0003/pyspark.zip
25/04/17 04:28:37 INFO Client: Uploading resource file:/usr/local/spark/python/lib/py4j-0.10.9.7-src.zip -> hdfs://cluster-master:9000/user/root/.sparkStaging/application_1744862269952_0003/py4j-0.10.9.7-src.zip
25/04/17 04:28:38 INFO Client: Uploading resource file:/tmp/spark-a3ee7391-f472-438b-94ad-368be58636af/_spark_conf_274563503199913489.zip -> hdfs://cluster-master:9000/user/root/.sparkStaging/application_1744862269952_0003/_spark_conf.zip
25/04/17 04:28:38 INFO SecurityManager: Changing view acls to: root
25/04/17 04:28:38 INFO SecurityManager: Changing modify acls to: root
25/04/17 04:28:38 INFO SecurityManager: Changing view acls groups to:
25/04/17 04:28:38 INFO SecurityManager: Changing modify acls groups to:
25/04/17 04:28:38 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
25/04/17 04:28:38 INFO Client: Submitting application application_1744862269952_0003 to ResourceManager
25/04/17 04:28:38 INFO YarnClientImpl: Submitted application application_1744862269952_0003
25/04/17 04:28:39 INFO Client: Application report for application_1744862269952_0003 (state: ACCEPTED)
25/04/17 04:28:39 INFO Client:
client token: N/A
diagnostics: AM container is launched, waiting for AM container to Register with RM
ApplicationMaster host: N/A
ApplicationMaster RPC port: -1
queue: default
start time: 1744864118263
final status: UNDEFINED
tracking URL: http://cluster-master:8088/proxy/application_1744862269952_0003/
user: root
25/04/17 04:28:43 INFO Client: Application report for application_1744862269952_0003 (state: RUNNING)
25/04/17 04:28:43 INFO Client:
client token: N/A
diagnostics: N/A
ApplicationMaster host: cluster-slave-1
ApplicationMaster RPC port: 43115
queue: default
start time: 1744864118263
final status: UNDEFINED
tracking URL: http://cluster-master:8088/proxy/application_1744862269952_0003/
user: root
25/04/17 04:28:52 INFO Client: Application report for application_1744862269952_0003 (state: FINISHED)
25/04/17 04:28:52 INFO Client:
client token: N/A
diagnostics: N/A
ApplicationMaster host: cluster-slave-1
ApplicationMaster RPC port: 43115
queue: default
start time: 1744864118263
final status: SUCCEEDED
tracking URL: http://cluster-master:8088/proxy/application_1744862269952_0003/
user: root
25/04/17 04:28:52 INFO ShutdownHookManager: Shutdown hook called
25/04/17 04:28:52 INFO ShutdownHookManager: Deleting directory /tmp/spark-a3ee7391-f472-438b-94ad-368be58636af
25/04/17 04:28:52 INFO ShutdownHookManager: Deleting directory /tmp/spark-62a817c6-ea34-4bcd-92a8-e15a7f23a1e4
root@cluster-master: /app
```

Figure 5: Continuation of bash results for query: "search engine indexing"

Query 2: natural language processing

```
C:\WINDOWS\system32\cmd.exe root@cluster-master: /app
client token: N/A
diagnostics: AM container is launched, waiting for AM container to Register with RM
ApplicationMaster host: N/A
ApplicationMaster RPC port: -1
queue: default
start time: 1744864118263
final status: UNDEFINED
tracking URL: http://cluster-master:8088/proxy/application_1744862269952_0003/
user: root
25/04/17 04:28:43 INFO Client: Application report for application_1744862269952_0003 (state: RUNNING)
25/04/17 04:28:43 INFO Client:
client token: N/A
diagnostics: N/A
ApplicationMaster host: cluster-slave-1
ApplicationMaster RPC port: 43115
queue: default
start time: 1744864118263
final status: UNDEFINED
tracking URL: http://cluster-master:8088/proxy/application_1744862269952_0003/
user: root
25/04/17 04:28:52 INFO Client: Application report for application_1744862269952_0003 (state: FINISHED)
25/04/17 04:28:52 INFO Client:
client token: N/A
diagnostics: N/A
ApplicationMaster host: cluster-slave-1
ApplicationMaster RPC port: 43115
queue: default
start time: 1744864118263
final status: SUCCEEDED
tracking URL: http://cluster-master:8088/proxy/application_1744862269952_0003/
user: root
25/04/17 04:28:52 INFO ShutdownHookManager: Shutdown hook called
25/04/17 04:28:52 INFO ShutdownHookManager: Deleting directory /tmp/spark-a3ee7391-f472-438b-94ad-368be58636af
25/04/17 04:28:52 INFO ShutdownHookManager: Deleting directory /tmp/spark-62a817c6-ea34-4bcd-92a8-e15a7f23a1e4
root@cluster-master: /app python3 query.py "natural language processing"
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/04/17 04:32:00 WARN NativeCodeLoader: Unable to load native-heapoop library for your platform... using builtin-java classes where applicable

Top 10 documents:
6641931 Document_6641931 BM25 Score: 5.367
31527744 Document_31527744 BM25 Score: 5.367
37906536 Document_37906536 BM25 Score: 5.367
16522251 Document_16522251 BM25 Score: 5.367
1184662 Document_1184662 BM25 Score: 5.367
24047175 Document_24047175 BM25 Score: 5.367
59302664 Document_59302664 BM25 Score: 5.367
67662091 Document_67662091 BM25 Score: 3.3
2296501 Document_2296501 BM25 Score: 3.128
59570947 Document_59570947 BM25 Score: 3.128
root@cluster-master: /app
```

Figure 6: BM25 scores for query: "natural language processing"

Dashboard

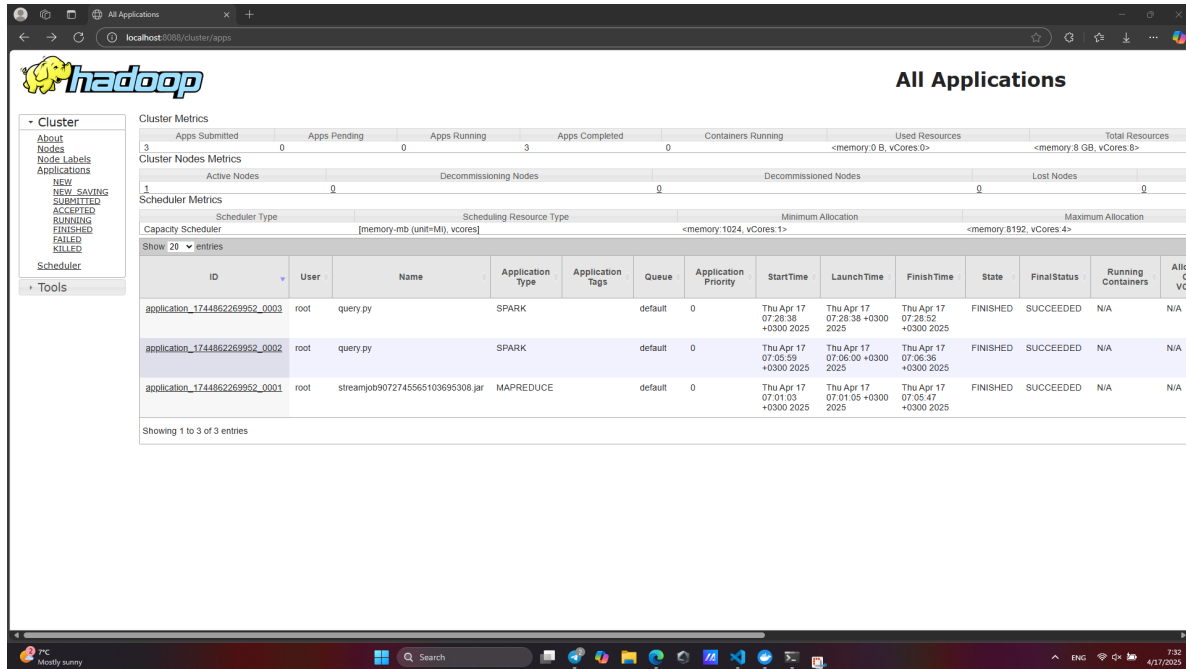


Figure 7: Successful Execution of Applications

Observations

- For the container to work, I had to keep increasing resources. I ended up with a setup of 10GB of ram allocated for docker and 6 CPU cores.
- I have faced an issue with pip3 missing, and cassandra-driver and pyspark not installed. I fixed this by adding a dockerfile with installation commands in the image.

```
FROM firasj/spark-docker-cluster
```

```
RUN apt-get update && \
    apt-get install -y python3-pip && \
    pip3 install cassandra-driver pyspark && \
    ln -s /usr/bin/python3 /usr/bin/python
```