

Methodology

I used a.parquet file and decided to put the code to prepare data inside of the index.sh, because I was unable to succeed with your suggestion (for some reason nothing appeared in hadoop). I just do the same logic and put data in hadoop after python script that prepares data for hadoop.

In app.sh I initialize cassandra keyspace and tables inside that to store the data and call other scripts (prepare_data, index.sh, search.sh)

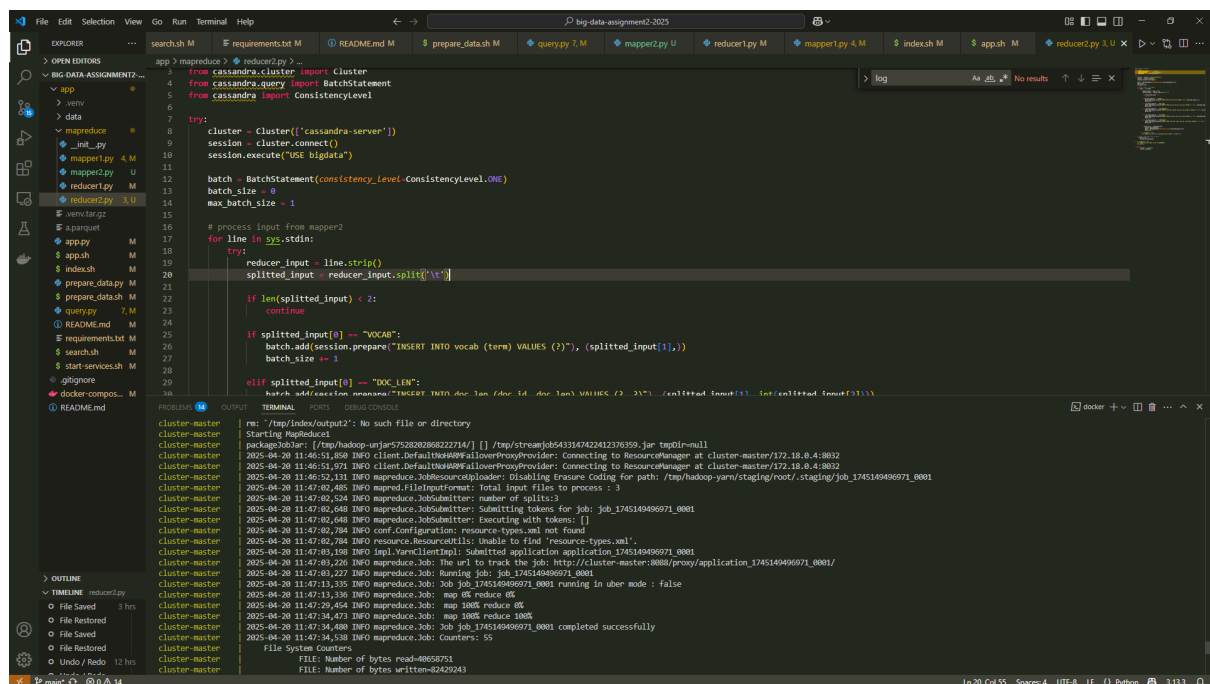
In the query.py I just read data from cassandra, calculate necessary for bm25 calculation values, call functions to calculate it and produce top documents.

mapper and reducer were done using reference from lab5 and youtube videos. I pass output from mapper1 to reducer1 (structured info about doc len, term freq, and document content). In reducer 1 it's combined together and passed further. In reducer2 I used batches and process records to avoid overloading of the system and be sure that the system will work properly.

Also, I modify docker-compose file and provide port 9042 to connect to it

Demonstration

to run the system just run docker compose up --build
1st query: best football player



```
File Edit Selection View Go Run Terminal Help
big-data-assignment2-2025
EXPLORER
  app
  data
  mapreduce
  _init_.py
  mapper1.py 4 M
  mapper2.py U
  reducer1.py M
  reducer2.py 3.4 U
  venv
  venv.tar.gz
  a.parquet
  app.py M
  app.sh M
  index.sh M
  prepare_data.py M
  prepare_data.sh M
  query.py 7.7 M
  requirements.txt M
  README.md M
  ignore
  docker-compose... M
  README.md
search.sh M requirements.txt M README.md M prepare_data.sh M query.py 7.7 M mapper1.py 4 M reducer1.py M reducer2.py 3.4 U index.sh M app.sh M
app > mapreduce > reducer2.py >
1 from cassandra.cluster import Cluster
2 from cassandra.query import BatchStatement
3 from cassandra import ConsistencyLevel
4
5 try:
6     cluster = Cluster(['cassandra-server'])
7     session = cluster.connect()
8     session.execute("USE bigdata")
9
10    batch = BatchStatement(consistency_level=ConsistencyLevel.ONE)
11    batch_size = 0
12    max_batch_size = 1
13
14    # process input from mapper2
15    for line in sys.stdin:
16        try:
17            reducer_input = line.strip()
18            splitted_input = reducer_input.split("\t")
19
20            if len(splitted_input) < 2:
21                continue
22
23            if splitted_input[0] == "VOCAB":
24                batch.add(session.prepare("INSERT INTO vocab (term) VALUES (?)"), (splitted_input[1],))
25                batch_size += 1
26
27            elif splitted_input[0] == "DOC_LEN":
28                batch.add(session.prepare("INSERT INTO doc_len (doc_id, doc_len) VALUES (?, ?)"), (splitted_input[1], splitted_input[2]))
29                batch_size += 1
30
31            if batch_size == max_batch_size:
32                batch.execute()
33                batch_size = 0
34
35    batch.execute()
36
37    session.close()
38 except Exception as e:
39    print(f"Error: {e}")
40
```

```
cluster-master cluster-master
rm: /tmp/index/output2: No such file or directory
Starting MapReduce1
package.json: [/tmp/hadoop-usrj57520203080222716/] [/tmp/strawjob4333494292012370393_jar.tgz]-null
2025-04-20 11:46:51,850 INFO client.DefaultHadoopClient: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-20 11:46:51,971 INFO client.DefaultHadoopClient: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-20 11:46:52,131 INFO mapreduce.JobResourceUploader: Disabling fsnative coding for path: /tmp/hadoop-usrj57520203080222716/staging/job_1745149496971_0001
2025-04-20 11:47:02,485 INFO mapreduce.JobResourceUploader: total input files to process = 3
2025-04-20 11:47:02,524 INFO mapreduce.JobSubmitter: number of splits=3
2025-04-20 11:47:02,648 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1745149496971_0001
2025-04-20 11:47:02,648 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-20 11:47:02,784 INFO conf.Configuration: resource-types.xml not found
2025-04-20 11:47:02,784 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'
2025-04-20 11:47:03,188 INFO impl.VersionClientImpl: Submitted application application_1745149496971_0001
2025-04-20 11:47:03,226 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1745149496971_0001/
2025-04-20 11:47:03,227 INFO mapreduce.Job: Running Job: job_1745149496971_0001
2025-04-20 11:47:11,335 INFO mapreduce.Job: Job job_1745149496971_0001 running in uber mode : false
2025-04-20 11:47:13,336 INFO mapreduce.Job: map 0% reduce 0%
2025-04-20 11:47:29,454 INFO mapreduce.Job: map 100% reduce 0%
2025-04-20 11:47:34,472 INFO mapreduce.Job: map 100% reduce 100%
2025-04-20 11:47:34,480 INFO mapreduce.Job: Job job_1745149496971_0001 completed successfully
2025-04-20 11:47:34,538 INFO mapreduce.Job: Counters: 55
File System Counters
FILE: Number of bytes read=40628751
FILE: Number of bytes written=62420243
Ln 20, Col 55 Spaces: 4 UTF-8 LF (i) Python 3.13.3
```

The screenshot displays a JupyterLab interface with the following components:

- Top Bar:** Standard application menus (File, Edit, Selection, View, Go, Run, Terminal, Help) and a breadcrumb path: `big-data-assignment-2025`.
- Left Panel:**
 - EXPLORER:** Shows a file tree with folders like `app`, `src`, and `data`. The `index.sh` file is selected.
 - OPEN EDITORS:** Lists open files including `index.sh`, `prepare_data.sh`, `mapper1.py`, `mapper2.py`, `mapper3.py`, `reducer1.py`, `reducer2.py`, `reducer3.py`, `app.py`, and `app.sh`.
 - OUTLINE:** Shows a tree view of the project structure.
 - TERMINAL:** Contains a list of file save actions.
- Center Panel:** The code editor displays a shell script named `index.sh`. The script defines a function `main` that iterates over files in a directory, splits them into chunks, and processes them using `mapreduce.py`. It also includes a `def create_doc(row):` function for generating document data.
- Right Panel:**
 - TERMINAL:** Shows the output of the `index.sh` script, including logs for `mapreduce.py` and `mapreduce.sh`. It details the execution of `mapreduce.py` with various flags and the output of `mapreduce.sh`, which shows the progress of the MapReduce job.
 - FILE EXPLORER:** Displays the file structure of the project, including `app`, `src`, and `data` folders.

[illegible]

2nd query: Chat gpt transformer

```
1 #!/bin/bash
2
3 source .venv/bin/activate
4
5 # Python of the driver ((app/.venv/bin/python))
6 export PYSARK_DRIVER_PYTHON=$(which python)
7
8 unset PYSARK_PYTHON
9
10 # DOWNLOAD a.parquet or any parquet file before you run this
11
12 hdfs dfs -put -f a.parquet / && \
13 spark-submit prepare_data.py && \
14 echo "Putting data to hdfs" && \
15 hdfs dfs -put data / && \
16 hdfs dfs -ls /data && \
17 echo "done data preparation!"
```

Top 5 results for: 'Chat gpt transformer'

- 1.Document ID: 1039316 A Flowering Tree
Relevance score: 6.5079442/7841418
Beginning of the file: "A Flowering Tree is an opera in two acts composed by John Adams with libretto by Adams and Peter Se
- 2.Document ID: 45602582 A Little Bit of Fluff (1919 File)
Relevance score: 6.2812342/3591015
Beginning of the file: "A Little Bit of Fluff is a 1919 British silent comedy film directed by Karel Foss and Geoffrey H.
- 3.Document ID: 52124393 A Different American Dream
Relevance score: 5.6340570/8994095
Beginning of the file: "A Different American Dream is a 2016 American documentary film by Simon Brook and Jane I. Wells. It
- 4.Document ID: 51418778 A Double-Dyed Deceiver
Relevance score: 4.9744757/85798316
Beginning of the file: "A Double-Dyed Deceiver is a lostThe Library of Congress/FIAF American Silent Feature Film Survivor C
- 5.Document ID: 47941374 A Christmas Horror Story
Relevance score: 4.9267173/3137286
Beginning of the file: "A Christmas Horror Story is a 2015 Canadian anthology horror film directed by Grant Harvey, Steven

25/04/20 12:03:22 INFO SparkContext: SparkContext is stopping with exitcode 0.

We can see that it is the case that nothing related to this is found in documents. but. at least we obtain the output somehow related to the query

3rd query: Interesting producer movie

```
51 CREATE TABLE IF NOT EXISTS term_freq (
52   term_freq int,
53   PRIMARY KEY (term, doc_id)
54 );
55
56 EOF
57
58 cqlsh cassandra-server -f /tmp/create_cass_keyspace_and_tables.cql
59
60 # Collect data
61 bash prepare_data.sh
62 # Run the Indexer
63 bash index.sh
64 # Run the ranker
65 bash search.sh
66 # Run the ranker
67
```

Top 5 results for: 'Interesting producer movie'

- 1.Document ID: 3145259 A Fifth of Beethoven
Relevance score: 6.8216357/80342705
Beginning of the file: "\"A Fifth of Beethoven\" is a disco instrumental recorded by Walter Murphy and the Big Apple Band,
- 2.Document ID: 34923571 A Friend of Mine (2006 film)
Relevance score: 6.4548487/48464595
Beginning of the file: "A Friend of Mine (Ein Freund von mir) is a 2006 German comedy-drama film written and directed by Se
- 3.Document ID: 4136930 A Bay of Blood
Relevance score: 6.3868061/699522
Beginning of the file: "A Bay of Blood (Italian: Ecologia del delitto, lit. 'Ecology of Crime'), later retitled Reazione a
- 4.Document ID: 42889699 A Kitty, Bobo Show
Relevance score: 6.1489018/283196115
Beginning of the file: "A Kitty Bobo Show is an American animated pilot created by Kevin Kaliner and Meghan Dunn, and prod
- 5.Document ID: 54334687 A Better Tomorrow 2018
Relevance score: 5.9442022/44877094
Beginning of the file: "A Better Tomorrow 2018 (), is a Chinese action film directed by Ding Sheng and starring Wang Kai, M

Ln 67, Col 43 (26 selected) Spaces: 4 UTF-8 LF () Shell Script