Simple Search Engine using Hadoop, MapReduce, Cassandra and Spark RDD

Apollinariia Azarskova

April 15, 2025

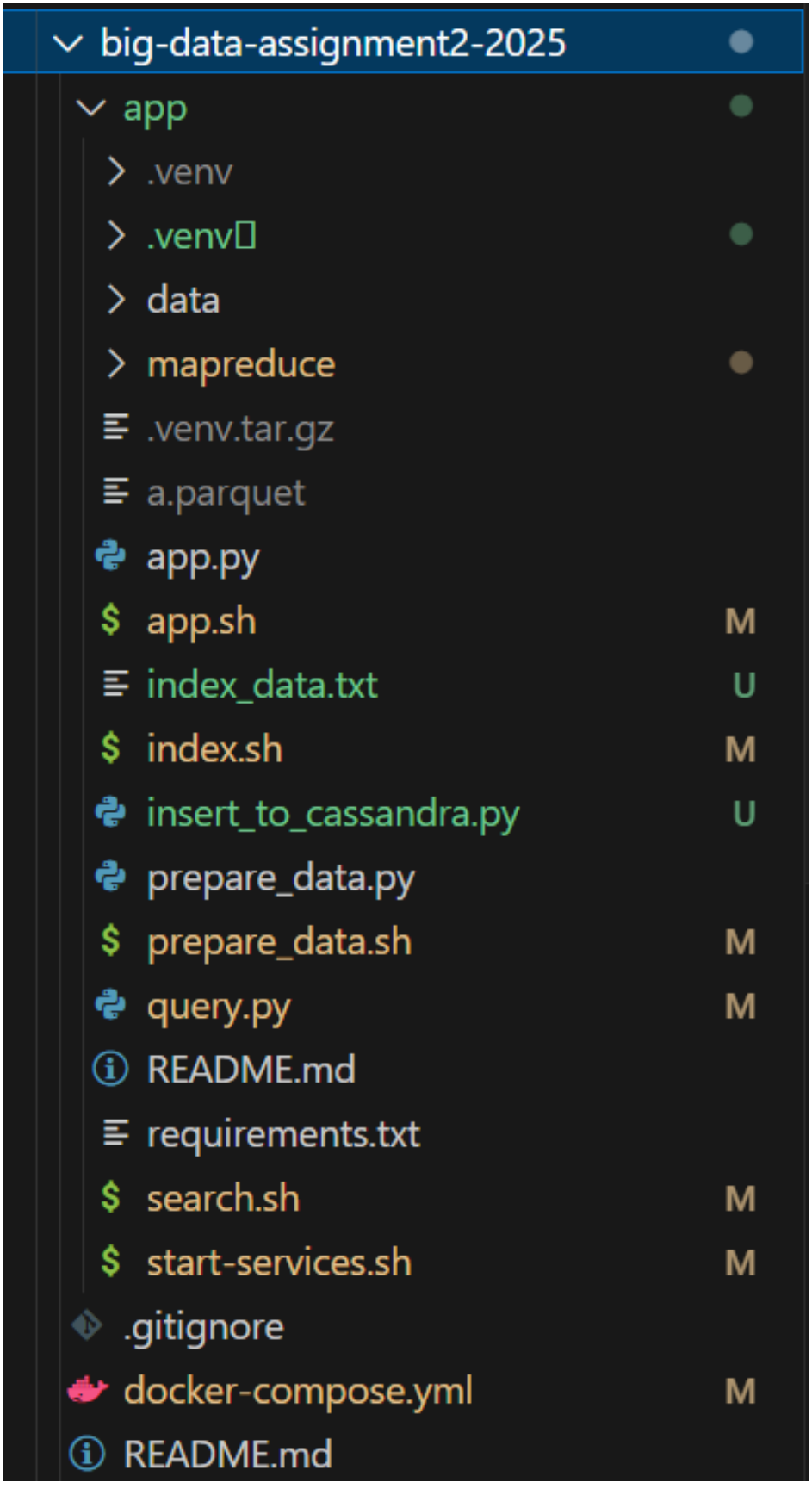
1. Objective

The goal of this project is to build a simple distributed search engine capable of retrieving relevant documents from a corpus of Wikipedia articles using the BM25 ranking algorithm. The system is implemented using a combination of Hadoop MapReduce, Apache Cassandra, and Apache Spark RDDs.

2. Architecture Overview

The system consists of the following components:

- **Data Preparation:** Transform Wikipedia dataset from parquet format to plain text.
- **Indexer (MapReduce):** Tokenizes text and builds an inverted index.
- **Storage:** Index and document statistics are stored in Cassandra.
- **Ranker (Spark):** Processes search queries and ranks documents using BM25.


big-data-assignment2-2025
app
.venv
.venv
data
mapreduce
.venv.tar.gz
a.parquet
app.py
app.sh M
index_data.txt U
index.sh M
insert_to_cassandra.py U
prepare_data.py
prepare_data.sh M
query.py M
README.md
requirements.txt
search.sh M
start-services.sh M
.gitignore
docker-compose.yml M
README.md

3. Data Preparation

- Used a subset of the Wikipedia dump `a.parquet`.
- Selected 1000 documents and converted each to a `.txt` file.
- Combined into a single tab-separated file `index_data.txt`.
- Uploaded to HDFS under `/index/data`.

```
(.venv) root@cluster-master:/app# bash prepare_data.sh
25/04/15 21:49:45 INFO SparkContext: Running Spark version 3.5.4
25/04/15 21:49:45 INFO SparkContext: OS info Linux, 5.15.146.1-microsoft-standard-WSL2, amd64
25/04/15 21:49:45 INFO SparkContext: Java version 1.8.0_442
25/04/15 21:49:45 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
25/04/15 21:49:46 INFO ResourceUtils: ==============================================================
25/04/15 21:49:46 INFO ResourceUtils: No custom resources configured for spark.driver.
25/04/15 21:49:46 INFO ResourceUtils: ==============================================================
25/04/15 21:49:46 INFO SparkContext: Submitted application: data preparation
25/04/15 21:49:46 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount:
1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
25/04/15 21:49:46 INFO ResourceProfile: Limiting resource is cpu
25/04/15 21:49:46 INFO ResourceProfileManager: Added ResourceProfile id: 0
25/04/15 21:49:46 INFO SecurityManager: Changing view acls to: root
25/04/15 21:49:46 INFO SecurityManager: Changing modify acls to: root
25/04/15 21:49:46 INFO SecurityManager: Changing view acls groups to:
25/04/15 21:49:46 INFO SecurityManager: Changing modify acls groups to:
25/04/15 21:49:46 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
25/04/15 21:49:46 INFO Utils: Successfully started service 'sparkDriver' on port 44575.
25/04/15 21:49:46 INFO SparkEnv: Registering MapOutputTracker
25/04/15 21:49:46 INFO SparkEnv: Registering BlockManagerMaster
25/04/15 21:49:46 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
25/04/15 21:49:46 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
25/04/15 21:49:46 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
25/04/15 21:49:46 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-b9402ad0-41fd-45bb-a6fc-fa80c78d3409
25/04/15 21:49:46 INFO MemoryStore: MemoryStore started with capacity 366.3 MiB
25/04/15 21:49:46 INFO SparkEnv: Registering OutputCommitCoordinator
25/04/15 21:49:46 INFO JettyUtils: Start Jetty 0.0.0.0:4040 for SparkUI
25/04/15 21:49:46 INFO Utils: Successfully started service 'SparkUI' on port 4040.
25/04/15 21:49:46 INFO Executor: Starting executor ID driver on host cluster-master
25/04/15 21:49:46 INFO Executor: OS info Linux, 5.15.146.1-microsoft-standard-WSL2, amd64
25/04/15 21:49:46 INFO Executor: Java version 1.8.0_442
25/04/15 21:49:46 INFO Executor: Starting executor with user classpath (userClassPathFirst = false): ''
25/04/15 21:49:46 INFO Executor: Created or updated repl class loader org.apache.spark.util.MutableURLClassLoader@7fe743c for default.
25/04/15 21:49:46 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 39977.
25/04/15 21:49:46 INFO NettyBlockTransferService: Server created on cluster-master:39977
25/04/15 21:49:46 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
25/04/15 21:49:46 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, cluster-master, 39977, None)
25/04/15 21:49:46 INFO BlockManagerMasterEndpoint: Registering block manager cluster-master:39977 with 366.3 MiB RAM, BlockManagerId(driver, cluster-master, 39977, None)
25/04/15 21:49:46 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, cluster-master, 39977, None)
25/04/15 21:49:46 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, cluster-master, 39977, None)
25/04/15 21:49:47 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir.
25/04/15 21:49:47 INFO SharedState: Warehouse path is 'file:/app/spark-warehouse'.
25/04/15 21:49:48 INFO InMemoryFileIndex: It took 51 ms to list leaf files for 1 paths.
25/04/15 21:49:48 INFO SparkContext: Starting job: parquet at NativeMethodAccessorImpl.java:0
25/04/15 21:49:48 INFO DAGScheduler: Got job 0 (parquet at NativeMethodAccessorImpl.java:0) with 1 output partitions
25/04/15 21:49:48 INFO DAGScheduler: Final stage: ResultStage 0 (parquet at NativeMethodAccessorImpl.java:0)
25/04/15 21:49:48 INFO DAGScheduler: Parents of final stage: List()
25/04/15 21:49:48 INFO DAGScheduler: Missing parents: List()
25/04/15 21:49:48 INFO DAGScheduler: Submitting ResultStage 0 (MapPartitionsRDD[1] at parquet at NativeMethodAccessorImpl.java:0), which has no missing parents
25/04/15 21:49:48 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 104.1 KiB, free 366.2 MiB)
25/04/15 21:49:48 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 37.8 KiB, free 366.2 MiB)
```

4. Indexer: MapReduce Pipeline

- **mapper1.py:** Emits each term with document ID and frequency 1.
- **reducer1.py:** Aggregates term frequencies and stores data in Cassandra.
- Also stores document length in a separate table.

```
(.venv) root@cluster-master:/app# bash index.sh
MapReduce index initiation
Deleted /tmp/index
2025-04-15 21:51:29,132 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/app/mapreduce/mapper1.py, /app/mapreduce/reducer1.py, /tmp/hadoop-unjar2523606184664720847/] [] /tmp/streamjob5740925844440951431.jar tmpDir=null
2025-04-15 21:51:29,554 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.20.0.4:8032
2025-04-15 21:51:29,710 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.20.0.4:8032
2025-04-15 21:51:29,832 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744739338664_0015
2025-04-15 21:51:30,174 INFO mapred.FileInputFormat: Total input files to process : 1
2025-04-15 21:51:30,220 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-15 21:51:30,308 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744739338664_0015
2025-04-15 21:51:30,308 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-15 21:51:30,410 INFO conf.Configuration: resource-types.xml not found
2025-04-15 21:51:30,410 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-15 21:51:30,464 INFO impl.YarnClientImpl: Submitted application application_1744739338664_0015
2025-04-15 21:51:30,502 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744739338664_0015/
2025-04-15 21:51:30,504 INFO mapreduce.Job: Running job: job_1744739338664_0015
2025-04-15 21:51:35,565 INFO mapreduce.Job: Job job_1744739338664_0015 running in uber mode : false
2025-04-15 21:51:35,566 INFO mapreduce.Job:  map 0% reduce 0%
2025-04-15 21:51:38,601 INFO mapreduce.Job:  map 100% reduce 0%
2025-04-15 21:51:42,621 INFO mapreduce.Job:  map 100% reduce 100%
2025-04-15 21:51:43,633 INFO mapreduce.Job: Job job_1744739338664_0015 completed successfully
2025-04-15 21:51:43,679 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=10612894
                FILE: Number of bytes written=22055236
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3545044
                HDFS: Number of bytes written=4384500
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=3377
                Total time spent by all reduces in occupied slots (ms)=1727
                Total time spent by all map tasks (ms)=3377
                Total time spent by all reduce tasks (ms)=1727
                Total vcore-milliseconds taken by all map tasks=3377
                Total vcore-milliseconds taken by all reduce tasks=1727
                Total megabyte-milliseconds taken by all map tasks=3458048
                Total megabyte-milliseconds taken by all reduce tasks=1768448
        Map-Reduce Framework
                Map input records=1000
                Map output records=581022
                Map output bytes=9450844
                Map output materialized bytes=10612900
                Input split bytes=178
```

5. Data Model in Cassandra

- `inverted_index(term TEXT, doc_id INT, freq INT)`
- `doc_stats(doc_id INT, doc_len INT)`

```
cqlsh> USE search;
cqlsh:search>
cqlsh:search> SELECT * FROM inverted_index LIMIT 5;

 term    | doc_id    | freq
---------+-----------+------
 dobson  | 13633480  |    1
   sain  | 14404655  |    4
 bessus  | 12000397  |    3
     ix  | 19789501  |    1
     ix  | 32497421  |    1

(5 rows)
cqlsh:search>
cqlsh:search> SELECT * FROM doc_stats LIMIT 5;

 doc_id    | doc_len
-----------+---------
 65747171  |     177
  4887308  |     375
 65604188  |     219
 47595311  |     918
 23837773  |     253

(5 rows)
```

6. Ranker: BM25 with Spark RDD

- Query is parsed and tokenized.

- For each query term, fetch posting lists and document stats.
- Compute BM25 score:

$$BM25(q,d) = \sum_{i \in q} IDF(i) \cdot \frac{f_{i,d} \cdot (k+1)}{f_{i,d} + k \cdot (1 - b + b \cdot \frac{L_d}{avgdl})}$$

- Top 10 documents are returned along with titles and scores.

```
root@cluster-master:/app# bash search.sh "money history"
This script will include commands to search for documents given the query using Spark RDD
25/04/15 22:02:15 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Top 10 relevant documents for query: money
Doc ID: 53037866 | Title: A Fool and His Money | Score: 5.8917
Doc ID: 454136 | Title: A Collection of Great Dance Songs | Score: 5.5073
Doc ID: 10174562 | Title: A History of Money and Banking in the United States | Score: 5.4582
Doc ID: 18397636 | Title: A Fool and His Money (1925 film) | Score: 5.2822
Doc ID: 44853014 | Title: A Gutter Magdalene | Score: 5.2326
Doc ID: 21090146 | Title: A Gigster's Life for Me | Score: 5.007
Doc ID: 53181236 | Title: A Gentleman Friend | Score: 4.7658
Doc ID: 34589053 | Title: A Fugitive from the Past | Score: 4.2291
Doc ID: 45390106 | Title: A Desperate Crime | Score: 3.9753
Doc ID: 8768022 | Title: A Drama in Livonia | Score: 3.9569
25/04/15 22:02:18 INFO ShutdownHookManager: Shutdown hook called
25/04/15 22:02:18 INFO ShutdownHookManager: Deleting directory /tmp/spark-97ba21be-e490-4090-a902-226a9883e7cd
```

7. Conclusion

- Successfully implemented a distributed search engine with MapReduce and Spark.
- BM25 allows for high-quality ranking of documents.
- System is scalable and can be extended with vector search, relevance feedback, or frontend integration.