

# IBD Assignment 2 Report

*Anna Gromova, DS-01*

*GitHub link: <https://github.com/anngrosha/hadoop-search>*

## Methodology

For this project, I implemented a simple search engine using Hadoop MapReduce, Cassandra, and Spark RDD. The system indexes documents and allows searching using either BM25 ranking or vector similarity search.

## Data Preparation

I started by processing the Wikipedia dataset in Parquet format. The `prepare_data.py` script reads the Parquet file using PySpark, selects 1000 random documents, and saves each document as a separate text file. Each file is named using the document ID and title (with spaces replaced by underscores). The script also creates an input file for MapReduce in HDFS with tab-separated document ID, title, and content.

## MapReduce Pipeline

I implemented two MapReduce jobs for indexing:

1. **Term Frequency Pipeline** (`mapper1.py`, `reducer1.py`):
  - The mapper tokenizes each document's text and emits (word, doc\_id, term\_count, doc\_length) tuples
  - It also emits document metadata (!META! records)
  - The reducer collects all documents for each word and outputs them
2. **Document Frequency Pipeline** (`mapper2.py`, `reducer2.py`):
  - The mapper takes the term frequency output and emits (word, 1) for each document containing the word
  - The reducer counts how many documents contain each word (document frequency)

## Cassandra Storage

The `app.py` script handles Cassandra operations:

- Creates keyspace and tables for document metadata, term index, vocabulary, and corpus statistics
- Loads data from MapReduce output into Cassandra
- Creates vector indexes when vector search is enabled

Tables include:

- `document_metadata`: Stores document IDs, titles, lengths and vector representations
- `term_index`: Stores term frequencies per document
- `vocabulary`: Stores document frequencies for all terms
- `corpus_stats`: Stores average document length and total document count

## Search Implementation

The `query.py` script provides two search methods.

The **vector search** implementation works by:

1. Creating a vocabulary from all unique terms in the corpus
2. Representing each document as a vector where:
  - o Each dimension corresponds to a vocabulary term
  - o The value is the TF-IDF weight for that term in the document
3. Normalizing vectors to unit length
4. Representing queries the same way using the same vocabulary
5. Calculating cosine similarity between query and document vectors

The vectors are stored in Cassandra using the `VECTOR` type and indexed for efficient similarity search. This allows finding documents similar to a query even when they don't contain exact term matches.

The **BM25 ranking** considers:

- Term frequency in each document
- Document frequency of each term
- Document lengths and average document length
- The BM25 formula parameters ( $k_1=1.2$ ,  $b=0.75$ )

This provides better ranking than simple TF-IDF by accounting for document length normalization and term frequency saturation.

Both return 10 most relevant document in the end.

## Optimizations

Key optimizations include using Cassandra's built-in vector indexing for fast similarity search, caching document titles during search to reduce database queries, batch processing of documents during vector creation, and proper connection handling and error recovery for Cassandra operations

The system demonstrates how different big data technologies can work together to build a functional search engine, with Hadoop for batch processing, Cassandra for storage, and Spark for query processing.

## Demonstration

### Running the Project

To run the project:

1. Place the `a.parquet` file in the app folder
2. Run `docker-compose up` (or `docker compose up`)

The system will:

- Start Hadoop, Spark and Cassandra services
- Process the Parquet file and create text documents
- Run the MapReduce indexing pipeline
- Load data into Cassandra
- Perform a sample search for "dog food"

### Changing Queries

To change the search query:

1. Edit `app.sh` and modify the `bash search.sh` command
2. For example: `bash search.sh "computer science"`

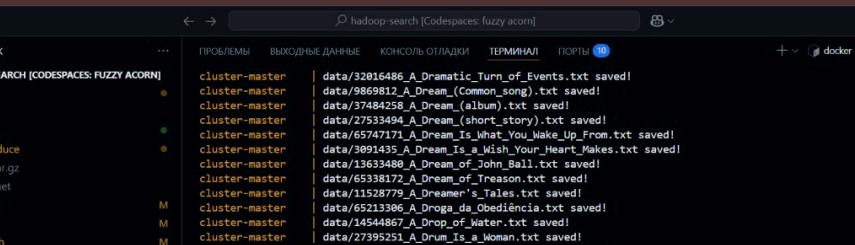
For vector search:

1. Edit `app.sh` to include `--vector` flag:

```
bash index.sh --vector  
bash search.sh "<query>" --vector
```

### Screenshots

1. Data Files Creation:

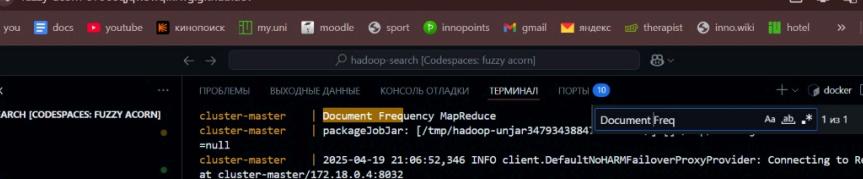


```
hadoop-search [Codespaces: fuzzy acorn]
cluster-master  data/32016486_A_Dramatic_Turn_of_Events.txt saved!
cluster-master  data/39869812_A_Dream_(Common_song).txt saved!
cluster-master  data/37484258_A_Dream_(album).txt saved!
cluster-master  data/27533494_A_Dream_(short_story).txt saved!
cluster-master  data/65747171_A_Dream_Is_What_You_Wake_Up_From.txt saved!
cluster-master  data/3091435_A_Dream_Is_a_Wish_Your_Heart_Makes.txt saved!
cluster-master  data/13633480_A_Dream_of_John_Ball.txt saved!
cluster-master  data/65338172_A_Dream_of_Treason.txt saved!
cluster-master  data/11528779_A_Dreamer_s_Tales.txt saved!
cluster-master  data/65123306_A_Draga_da_Obediencia.txt saved!
cluster-master  data/14544867_A_Drop_of_Water.txt saved!
cluster-master  data/27395251_A_Drum_Is_a_Woman.txt saved!
cluster-master  data/65714859_A_Drummer_Boy_Christmas.txt saved!
cluster-master  data/74058962_A_Drummer_Boy_Christmas_Tour.txt saved!
cluster-master  data/68255267_A_Duke_and_No_Duke.txt saved!
cluster-master  data/13958799_A_Dying_Colonialism.txt saved!
cluster-master  data/5003381_A_Dying_Light_in_Corduba.txt saved!
cluster-master  data/36902574_A_Donde.txt saved!
cluster-master  data/28798362_A_Encomeada_A_Pobra_de_Trives.txt saved!
cluster-master  data/32222497_A_Ergo.txt saved!
cluster-master  data/13985505_A_Escrava_Isaura.txt saved!
cluster-master  data/64688311_A_European_Requiem.txt saved!
cluster-master  data/32101003_A_Face_in_the_Crowd_(Michael_Martin_Murphy_and_Holly_Dunn_song).txt saved!
cluster-master  data/36788181_A_Face_to_Die_in_the_Crowd_(novella).txt saved!
cluster-master  data/3651016_A_Face_to_Fix_for_Trix.txt saved!
cluster-master  data/45681521_A_Face_impostor_.txt saved!
cluster-master  data/45681561_A_Fair_Impostor_(novel).txt saved!
cluster-master  data/41086718_A_Fair_to_Remember_(Modern_Family).txt saved!
cluster-master  data/8702341_A_Fairly_Honourable_Defeat.txt saved!
```

## 2. Term Frequency Pipeline:

```
cluster-master | Term Frequency MapReduce
cluster-master | packageJobJar: [/tmp/hadoop-unjar8063546783]
null
cluster-master | 2025-04-19 21:06:27,945 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8082
cluster-master | 2025-04-19 21:06:28,103 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8082
cluster-master | 2025-04-19 21:06:28,387 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/_staging/job_1745896711350_0001
cluster-master | 2025-04-19 21:06:29,416 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master | 2025-04-19 21:06:30,255 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master | 2025-04-19 21:06:30,851 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1745896711350_0001
cluster-master | 2025-04-19 21:06:30,851 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master | 2025-04-19 21:06:31,016 INFO conf.Configuration: resource-types.xml not found
cluster-master | 2025-04-19 21:06:31,016 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master | 2025-04-19 21:06:31,412 INFO impl.YarnClientImpl: Submitted application application_1745896711350_0001
cluster-master | 2025-04-19 21:06:31,448 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1745896711350_0001
cluster-master | 2025-04-19 21:06:31,444 INFO mapreduce.Job: Running job: job_1745896711350_0001
cluster-master | 2025-04-19 21:06:38,535 INFO mapreduce.Job: Job job_1745896711350_0001 running in uber mode : false
cluster-master | 2025-04-19 21:06:38,536 INFO mapreduce.Job: map 0% reduce 0%
cluster-master | 2025-04-19 21:06:44,607 INFO mapreduce.Job: map 100% reduce 0%
cluster-master | 2025-04-19 21:06:49,629 INFO mapreduce.Job: map 100% reduce 100%
cluster-master | 2025-04-19 21:06:56,640 INFO mapreduce.Job: Job job_1745896711350_0001 completed successfully
cluster-master | 2025-04-19 21:06:56,706 INFO mapreduce.Job: Counters: 54
cluster-master | File System Counters
cluster-master | FILE: Number of bytes read=6016538
cluster-master | FILE: Number of bytes written=12862599
cluster-master | FILE: Number of read operations=20
```

### 3. Document Frequency Pipeline:

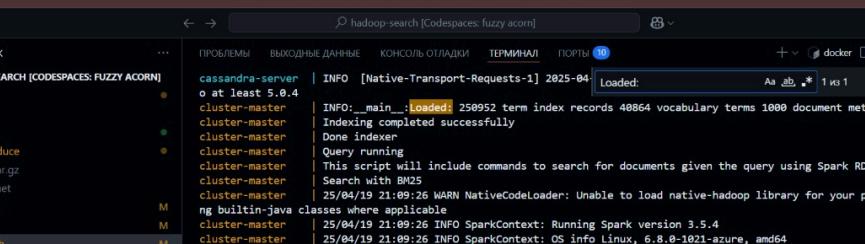


The screenshot shows a browser window with several tabs open. The active tab is 'fuzzy-acorn-97966cjqw5wqfxrwg.github.dev'. The page displays a file tree for a 'HADOOP-SEARCH [CODESPACES: FUZZY ACORN]' project. The tree includes files like 'app', 'venv', 'data', 'mapreduce', 'F. venv.tar.gz', 'E. a.parquet', 'app.py', '\$ app.sh', '\$ indexsh', 'prepare\_data.py', '\$ prepare\_data.sh', 'query.py', 'README.md', 'requirements.txt', 'searchsh', 'start-services.sh', 'vector\_indexer.py', '.gitignore', 'docker-compose.yml', 'get-docker.sh', 'СТРУКТУРА', and 'ВРЕМЕННАЯ ШКАЛА'. The 'cluster-master' log file is expanded, showing logs related to Frequency MapReduce, packageJobJar, and various INFO messages from the mapreduce and cluster-master components. The terminal tab is also visible at the top.

## 4. Loading Data to Cassandra:

```
cluster-master | WRONG_MAP=0
cluster-master | WRONG_REDUCE=0
File Input Format Counters
    Bytes Read=5599724
File Output Format Counters
    Bytes Written=417954
2025-04-19 21:07:12,226 INFO streaming.StreamJob: Output directory: /tmp/document_frequencies
cluster-master | loading data into Cassandra
cluster-master | WARNING:cassandra.cluster:Cluster._init__ called with contact_points specified, but no load_balancing
_policy. In the next major version, this will raise an error; please specify a load-balancing policy. (contact_points = ['
cassandra-server'], lbp = None)
cluster-master | WARNING:cassandra.connection:An authentication challenge was not sent, this is suspicious because the
driver expects authentication (configured authenticator = PlainTextAuthenticator)
cluster-master | INFO:cassandra.policies:Using datacenter 'dc1' for DCAwareRoundRobinPolicy (via host '172.18.0.2:9042'
); if incorrect, please specify a local_dc to the constructor, or limit contact points to local cluster nodes
cluster-master | WARNING:cassandra.connection:An authentication challenge was not sent, this is suspicious because the
driver expects authentication (configured authenticator = PlainTextAuthenticator)
cluster-master | INFO:_main_:Schema created successfully
cluster-master | WARNING:cassandra.cluster:Cluster._init__ called with contact_points specified, but no load balancing
_policy. In the next major version, this will raise an error; please specify a load-balancing policy. (contact_points = ['
cassandra-server'], lbp = None)
cluster-master | WARNING:cassandra.connection:An authentication challenge was not sent, this is suspicious because the
driver expects authentication (configured authenticator = PlainTextAuthenticator)
cluster-master | INFO:cassandra.policies:Using datacenter 'dc1' for DCAwareRoundRobinPolicy (via host '172.18.0.2:9042'
); if incorrect, please specify a local_dc to the constructor, or limit contact points to local cluster nodes
cluster-master | WARNING:cassandra.connection:An authentication challenge was not sent, this is suspicious because the
driver expects authentication (configured authenticator = PlainTextAuthenticator)
cluster-master | INFO [Native-Transport-Requests-1] 2025-04-19 21:07:14,024 QueryProcessor.java:654 - Fully upgraded t
o at least 5.0.4
```

## 5. Cassandra Success Message:



The screenshot shows a browser window with multiple tabs open. The active tab is titled "indexsh - hadoop-search [Codespace: fuzzy acorn]". The content area displays a terminal session with the following command and output:

```
java -jar hadoop-search-0.0.1-SNAPSHOT.jar
```

```
INFO [Native-Transport-Requests-1] 2025-04-25 09:26:30.110 +0000 Loaded: /app/hadoop-search.jar
```

```
INFO [main] 2025-04-25 09:26:30.110 +0000 main :Loaded: 250952 term index records 40864 vocabulary terms 1000 document metadata entries
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 indexing completed successfully
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 Done indexer
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 Query running
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 This script will include commands to search for documents given the query using Spark RDD
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 Search with BM25
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:26 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin Java classes where applicable
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:26 INFO SparkContext: Running Spark version 3.5.4
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:26 INFO SparkContext: OS info Linux, 6.8.0-1021-azure, amd64
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:27 INFO ResourceUtils: =====
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:27 INFO ResourceUtils: No custom resources configured for spark.driver.
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:27 INFO ResourceUtils: =====
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:27 INFO SparkContext: Submitted application: SearchEngineQuery
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:27 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores => name: cores, amount: 1, script: <script>, vendor: <memory> -> name: memory, amount: 1024, script: <script>, vendor: <offheap> -> name: offheap, amount: 8, script: <script>, vendor: <script>), task resources: Map(cpus => name: cpus, amount: 1.0)
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:27 INFO ResourceProfile: Limiting resource is cpus at 1 tasks per executor
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:27 INFO ResourceProfileManager: AddedResourceProfile id: 0
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:27 INFO SecurityManager: Changing view acls to: root
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:27 INFO SecurityManager: Changing modify acls to: root
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:27 INFO SecurityManager: Changing view acls groups to:
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:27 INFO SecurityManager: Changing modify acls groups to:
```

```
INFO [cluster-master] 2025-04-25 09:26:30.110 +0000 25/04/19 21:09:27 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
```

## 6. Top 10 Documents:

The screenshot shows a browser window with multiple tabs open. The active tab is titled "indexsh - hadoop-search [Codespace]". The page content displays a file tree under "HADOOP-SEARCH [CODESPACES: FUZZY ACORN]" and a log output from a "cluster-master" node. The log shows the application starting and performing a search for "B25" documents. The search results are listed as follows:

Doc ID	Title	Score
200010	A_Boy_and_His_Dog	29.2825
896285	A_Grand_Day_Out	22.4138
35739084	A_Kitty_Bobo_Show	20.9332
32395144	A_Dog's_Purpose	18.5525
43781676	A_Dog_Called_Ego	17.9945
34652846	A_Dog_in_a_Drawer	17.5384
17488265	A_Boy_and_His_Dog_(1946_film)	17.4178
33568949	A_Little_Bit_of_Cucumber	17.2177
48516568	A_Dog_Named_Gucci	16.9565

The browser interface includes a search bar ("hadoop-search [Codespaces: fuzzy acorn]"), a top navigation bar with tabs like "ПРОВОДНИК", "ПРОБЛЕМЫ", "ВЫХОДНЫЕ ДАННЫЕ", "КОНСОЛЬ ОТЛАДКИ", "ТЕРМИНАЛ", and "ПОРТЫ", and a sidebar with various icons representing different tools and services.

## 7. Vector Creation:

```
cluster-master | INFO:cassandra.policies:Using datacenter 'dc1' for DCAwareRoundRobinPolicy (via host '172.18.0.2:9042'); if incorrect, please specify a local_dc to the constructor, or limit contact points to local cluster nodes
cluster-master | WARNING:cassandra.connection:An authentication challenge was not sent, this is suspicious because the driver expects authentication (configured authenticator = PlainTextAuthenticator)
cassandra-server | INFO [Native-Transport-Requests-1] 2025-04-19 21:38:50,185 QueryProcessor.java:654 - Fully upgraded to at least 5.0.4
cluster-master | WARNING:cassandra.protocol:Server warning: 'USE <keyspace>' with prepared statements is considered to be an anti-pattern due to ambiguity in non-qualified table names. Please consider removing instances of 'Session#setKeyspace(<keyspace>)', 'Session#execute("USE <keyspace>")' and 'cluster.newSession(<keyspace>) from your code, and always use fully qualified table name (e.g. <keyspace>.<table>). Keyspace used: search_engine, statement keyspace: search_engine, statement id: f1ff9c06ddebccae5ddf78bd5d8efb49a
cluster-master | INFO:_main_:Loaded: 258952 term index records 40864 vocabulary terms 1000 document metadata entries
cluster-master | Creating vector representations
cluster-master | WARNING:cassandra.cluster:Cluster._init_ called with contact_points specified, but no load_balancing_policy. In the next major version, this will raise an error; please specify a load-balancing policy. (contact_points = ['cassandra-server'], lbp = None)
cluster-master | WARNING:cassandra.connection:An authentication challenge was not sent, this is suspicious because the driver expects authentication (configured authenticator = PlainTextAuthenticator)
cluster-master | INFO:cassandra.policies:Using datacenter 'dc1' for DCAwareRoundRobinPolicy (via host '172.18.0.2:9042'); if incorrect, please specify a local_dc to the constructor, or limit contact points to local cluster nodes
cluster-master | WARNING:cassandra.connection:An authentication challenge was not sent, this is suspicious because the driver expects authentication (configured authenticator = PlainTextAuthenticator)
cluster-master | WARNING:cassandra.cluster:Cluster._init_ called with contact_points specified, but no load_balancing_policy. In the next major version, this will raise an error; please specify a load-balancing policy. (contact_points = ['cassandra-server'], lbp = None)
cluster-master | WARNING:cassandra.connection:An authentication challenge was not sent, this is suspicious because the driver expects authentication (configured authenticator = PlainTextAuthenticator)
```

## Conclusion

The search engine worked correctly and retrieved 10 most relevant documents for each query. As seen in the screenshots, BM25 returned good results by analyzing term frequencies and document similarities. The system successfully combined Hadoop, Cassandra and Spark to process and search through the data efficiently.