

Assignment 2 - Simple Search Engine using Hadoop MapReduce

Rufina Gafiatullina, r.gafiatullina@innopolis.university
Innopolis University

1 Methodology

This search engine performs indexing, ranking and retrieval of top 10 text documents for query. The system is designed using distributed technologies to work with large amounts of data.

Full code you can find in GitHub: [big-data-assignment2-2025](#)

1.1 Overall Architecture

The overall pipeline of the search engine is following:

1. **Data Collection:** A collection of text documents is loaded into the HDFS for indexing.
2. **Indexing (MapReduce):** Words are extracted from documents and normalized (e.g., lowercased). MapReduce is used to compute the components for BM25 score.
3. **Storage (Cassandra):** The results from previous step are stored in Cassandra.
4. **Query Processing (Spark):** Using a query, Spark loads from Cassandra needed information for query terms, calculates BM25 score for documents, and retrieves 10 documents with high score.

1.2 Cassandra

The schema in Cassandra is effective for searching of term frequencies, document statistics, and metadata of the whole document collection.

To have more precise and categorized searches in the future, it is useful to specify the type of collection to which a document belongs (e.g., scientific articles, news reports, or fiction texts). For this purpose, each document is associated with a **corpus**, allowing for faster and more targeted searches. However, for now all documents are stored in a default corpus labeled "whole_corpus".

1.2.1 Keyspace

```
1 CREATE KEYSPACE IF NOT EXISTS index_keyspace
2 WITH replication = {'class': 'SimpleStrategy', 'replication_factor': 1};
```

This defines the keyspace used for indexing process.

1.2.2 doc_frequency_of_term table

This table stores the number of documents containing the specific term within a specific corpus.

```

1 CREATE TABLE IF NOT EXISTS doc_frequency_of_term (
2     term TEXT,
3     corpus_name TEXT,
4     doc_frequency INT,
5     PRIMARY KEY (corpus_name, term)
6 );

```

The data is partitioned by `corpus_name`, allowing you to quickly find data for a particular corpus.

Within each partition, data is clustered by `term`, allowing efficient querying of term frequency within a corpus.

1.2.3 term_frequency_in_doc table

This table stores the frequency of each term in a particular document that has the document ID and title.

```

1 CREATE TABLE IF NOT EXISTS term_frequency_in_doc (
2     term TEXT,
3     corpus_name TEXT,
4     doc_id INT,
5     doc_title TEXT,
6     term_frequency INT,
7     PRIMARY KEY ((term, corpus_name), doc_id, doc_title)
8 );

```

Due to partitioning by `(term, corpus_name)`, term frequencies for the same term within a specific corpus are stored together in the same partition, which optimizes reads and writes for specific terms in a corpus.

Also data is clustering by `doc_id, doc_title`. This allows efficient retrieval of term frequencies for a given document.

1.2.4 doc_info table

This table stores metadata about each document including its length, which is used for calculating BM25 score.

```

1 CREATE TABLE IF NOT EXISTS doc_info (
2     doc_id INT,
3     doc_title TEXT,
4     doc_length INT,
5     PRIMARY KEY (doc_id, doc_title)
6 );

```

1.2.5 corpus_info table

This table stores metadata about each corpus, including total number of documents and total length. Now this table contains only one corpus information for `whole_corpus`.

```

1 CREATE TABLE IF NOT EXISTS corpus_info (
2     corpus_name TEXT PRIMARY KEY,
3     doc_n INT,
4     total_doc_length INT
5 );

```

1.3 Document Indexing with Hadoop MapReduce

The indexing process contains a two-stage MapReduce pipeline. By default, documents in `/index/data` in HDFS are indexed. But it is possible to give as an argument a local document, which is first converted to the required format, then copied to the HDFS `/local_files`, and indexed.

1.3.1 MapReduce1

- **Mapper 1:** The mapper takes the format (`<doc_id> <doc_title> <doc_text>`) separated by tab. It takes the text of the document, extract just the words (alphanumeric and apostrophes), obtain unique words for document, and maps 1 to each word in the set.
- **Reducer 1:** The reducer aggregates same words to obtain number of documents that have this word. After that it works with Cassandra and inserts (if term does not exists in table) or updates (if we index new document and tables in Cassandra already store information about previous indexing) values to `_doc_frequency_of_term` table.

1.3.2 MapReduce2

- **Mapper 2:** The mapper takes the format (`<doc_id> <doc_title> <doc_text>`) separated by tab. It takes the text of the document, extract just the words (alphanumeric and apostrophes), and prints `doc_id`, `doc_title`, and `word` with assigned 1.
- **Reducer 2:** The reducer computes term frequency per document, document metadata, and corpus information. After that it works with Cassandra and inserts (if term does not exists in table) or updates (if we index new document and tables in Cassandra already store information about previous indexing) values to `term_frequency_in_doc`, `doc_info`, and `corpus_info` tables.

1.3.3 Initializing Cassandra Schema

Before indexing, the Cassandra schema is initialized.

```

1 cqlsh cassandra-server -f /app/cassandra/schema.cql

```

This ensures that all necessary tables present in Cassandra before inserting index data.

1.4 Query Processing and Ranking with Spark RDD

1.4.1 Query Preprocessing

Each input query is tokenized into unique lowercase terms:

```
1 query_terms = list(set([word.lower() for word in query[1].split()]))
```

1.4.2 Retrieving Needed Info for Cassandra

Using Spark, I load information for each unique term in a given query from the `doc_frequency_of_term` table from Cassandra. If none of the terms exist in the corpus, the program immediately writes an empty result.

For terms that exist, I collect information about their frequency of appearance in documents. The corresponding rows from the `term_frequency_in_doc` table, which contains the frequencies of terms for each document, are also loaded. Then using the id and title of the documents, I extract information about each document from the table `doc_info`. For these documents, I will count the score. In addition, the corpus information is extracted from the `corpus_info` table in Cassandra.

1.4.3 BM25 Score Calculation

The BM25 formula is following:

$$\text{BM25}(t, d) = \log \left(\frac{N}{\text{df}(t)} \right) \cdot \frac{(k_1 + 1) \cdot \text{tf}(t, d)}{k_1 \cdot \left(1 - b + b \cdot \frac{\text{dl}}{\text{dl}_{\text{avg}}} \right) + \text{tf}(t, d)}$$

Where:

- N is the total number of documents,
- $\text{df}(t)$ is the document frequency of query term t ,
- $\text{tf}(t, d)$ is the frequency of term t in document d ,
- dl is the length of document d ,
- dl_{avg} is the average document length across the corpus,
- k_1 and b are hyperparameters (set to 1 and 0.75 respectively).

1.4.4 Ranking and Output

After computing BM25 scores for each term-document pair, scores are aggregated per document using summation and sorted in descending order. The top 10 results go to a CSV file:

```
1 ranked_docs = joined_df.groupBy("doc_id", "doc_title").sum("bm25")
2 ranked_docs.orderBy(col("sum(bm25)").desc()).limit(10)
```

1.5 Limitations

This system has following limitations:

- **Simple Tokenization:** The current system uses simple whitespace-based tokenization without support for stemming or stop word removal.
- **No query tokenization:** The system does not remove punctuation marks and other non-alphabetic characters in the query, which may affect the results.
- **Limitation to a single enclosure:** Although the corpus concept is implemented, there is no input mechanism to easily specify or switch between corpora during searching or indexing.

1.6 Summary

Pipeline Step	Script Name	Input	Output
Data collection and preparation	prepare_data.sh	a.parquet	/index/data and /data in HDFS
Indexing tasks	index.sh	Raw .txt file in local system or /index/data directory in HDFS	Records in Cassandra tables
Ranking with BM25	search.sh	User query, Cassandra tables	Top 10 ranked documents written to CSV in /bm25_output in HDFS

2 Demonstration

2.1 Instructions to Run

2.1.1 Main Pipeline

To run search engine pipeline, clone the repository and run Docker:

```

1 git clone https://github.com/Rufina2323/big-data-assignment2-2025.git
2 cd big-data-assignment2-2025
3 docker compose up

```

This will find documents for "football game" query searching in /index/data documents.

2.1.2 Indexing your File

If you want to index your file, before you run docker, uncomment the last line in app.sh (it is needed so that the master node does not exit after finishing executing the script) and comment index.sh and search.sh executing:

```

1 # Run the indexer
2 # bash index.sh
3
4 # Run the ranker
5 # bash search.sh "football game"
6
7 tail -f /dev/null

```

Then write in master node following command, where you should specify path to your file that you want to index (for example, /data/sample.txt):

```
1 bash index.sh /data/sample.txt
```

2.1.3 Searching your Query

If you want to search your query, before you run docker, uncomment the last line in app.sh (it is needed so that the master node does not exit after finishing executing the script) and comment search.sh executing:

```

1 # Run the ranker
2 # bash search.sh "football game"
3
4 tail -f /dev/null

```

Then write in master node following command, where you should specify your query:

```
1 bash search.sh "query"
```

2.2 Examples

I run only this part of app.sh to show then that code for indexing and searching work correctly:

```

1 #!/bin/bash
2 # Start ssh server
3 service ssh restart
4
5 # Starting the services
6 bash start-services.sh
7
8 # Creating a virtual environment
9 python3 -m venv .venv
10 source .venv/bin/activate
11
12 # Install any packages
13 pip install -r requirements.txt
14
15 # Package the virtual env.

```

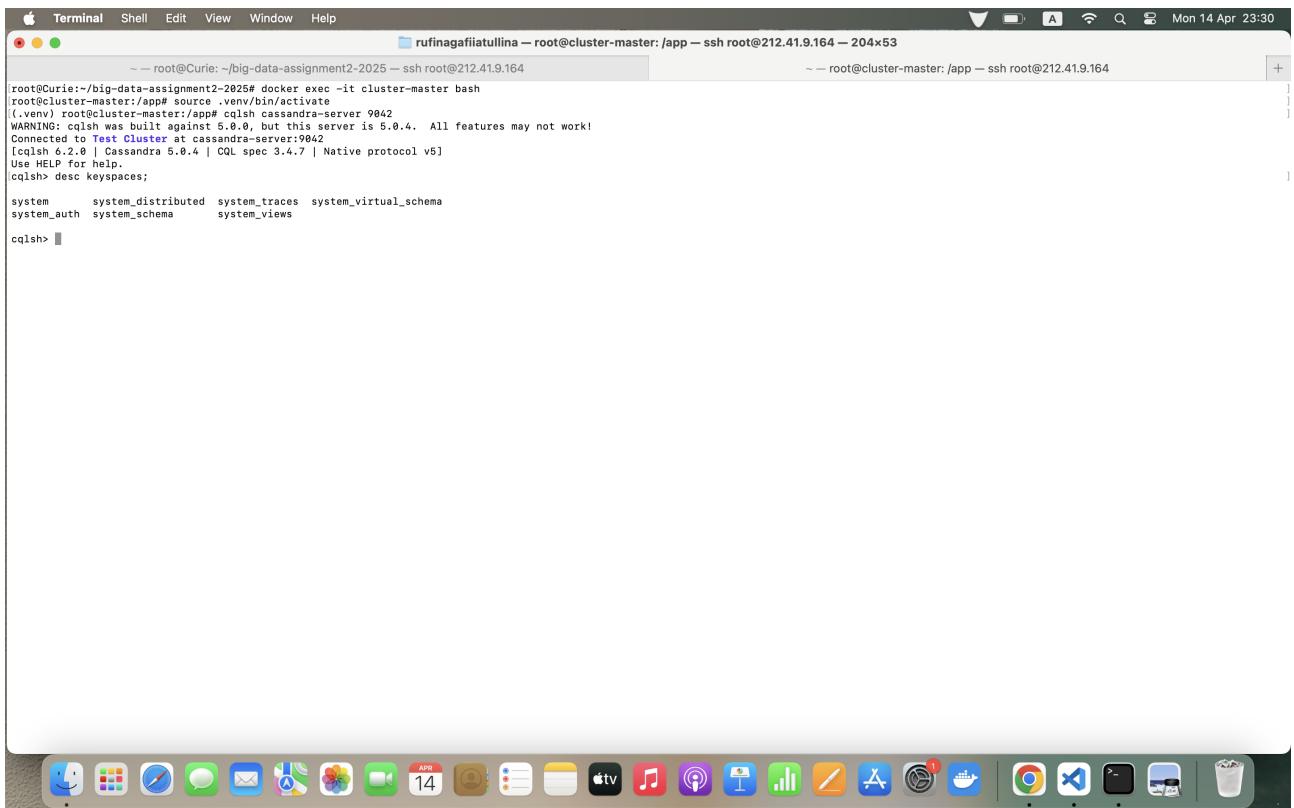
```

16 venv-pack -o .venv.tar.gz
17
18 # Collect data
19 bash prepare_data.sh

```

2.2.1 Indexing /index/data

Fig.1 shows that Cassandra does not have specified keyspace "index_keyspace".



```

Terminal Shell Edit View Window Help rufinagafiatullina — root@cluster-master: /app — ssh root@212.41.9.164 — 204x53
~ — root@Curie: ~/big-data-assignment2-2025$ docker exec -it cluster-master bash
root@Curie:~/big-data-assignment2-2025$ ssh root@212.41.9.164
root@cluster-master:/app# source .venv/bin/activate
(.venv) root@cluster-master:/app# cqlsh cassandra-server 9042
WARNING: cqlsh was built against 5.0.0, but this server is 5.0.4. All features may not work!
Connected to Test Cluster at cassandra-server:9042
[cqlsh 6.2.0 | Cassandra 5.0.4 | CQL spec 3.4.7 | Native protocol v5]
Use HELP for help.
cqlsh> desc keyspaces;
system      system_distributed  system_traces  system_virtual_schema
system_auth  system_schema     system_views
cqlsh>

```

Figure 1: Checking Cassandra keyspaces

Fig.2 shows that HDFS has needed directory /index/data. Then I execute index.sh without arguments. All index.sh outputs and logs you can see in fig.2, fig.3, fig.4, fig.5.

```

rufinagiafiatullina -- root@cluster-master: /app -- ssh root@212.41.9.164 -- 204x52
-- root@cluster-master: /app# hdfs dfs -ls /
Found 6 items
-rw-r--r-- 1 root supergroup 873207391 2025-04-14 20:26 /a.parquet
drwxr-xr-x 2 root supergroup 0 2025-04-14 20:24 /apps
drwxr-xr-x 2 root supergroup 0 2025-04-14 20:29 /data
drwxr-xr-x 2 root supergroup 0 2025-04-14 20:26 /index
drwxrwx--- 2 root supergroup 0 2025-04-14 20:24 /tmp
drwxr-xr-x 2 root supergroup 0 2025-04-14 20:26 /user
root@cluster-master:/app# bash index.sh
Setting up Cassandra schema...
WARNING: casslib was built against 5.0.0, but this server is 5.0.4. All features may not work!
Running MapReduce job...
rm: '/tmp/Index/output': No such file or directory
Running!
packageJobJar: [/tmp/hadoop-unjar1808623829219541174/] [] /tmp/streamjob2948749227226569314.jar tmpDir=null
2025-04-14 20:32:43,173 INFO client.DefaultToHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-14 20:32:43,282 INFO client.DefaultToHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-14 20:32:43,422 INFO mapred.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744662282546_0001
2025-04-14 20:32:43,441 INFO mapred.JobResourceUploader: Total input files to process : 1
2025-04-14 20:32:43,441 INFO mapred.JobResourceUploader: JobSubmitter: number of splits:2
2025-04-14 20:32:44,199 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744662282546_0001
2025-04-14 20:32:44,208 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-14 20:32:44,308 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-14 20:32:44,481 INFO impl.YarnClientImpl: Submitted application application_1744662282546_0001
2025-04-14 20:32:44,508 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744662282546_0001/
2025-04-14 20:32:44,609 INFO mapreduce.Job: Running job: job_1744662282546_0001 running in uber mode : false
2025-04-14 20:32:44,628 INFO mapreduce.Job: map 100% reduce 0%
2025-04-14 20:32:44,659 INFO mapreduce.Job: map 100% reduce 0%
2025-04-14 20:32:44,688 INFO mapreduce.Job: map 100% reduce 100%
2025-04-14 20:32:44,708 INFO mapreduce.Job: map 100% reduce 100%
2025-04-14 20:32:44,882 INFO mapreduce.Job: job_1744662282546_0001 completed successfully
2025-04-14 20:33:46,935 INFO mapreduce.Job: Counters: 5
File System Counters
FILE: Number of bytes read=2743046
FILE: Number of bytes written=6319383
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=3564573
HDFS: Number of bytes written=0
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)=4565
Total time spent by all reduces in occupied slots (ms)=43646

```

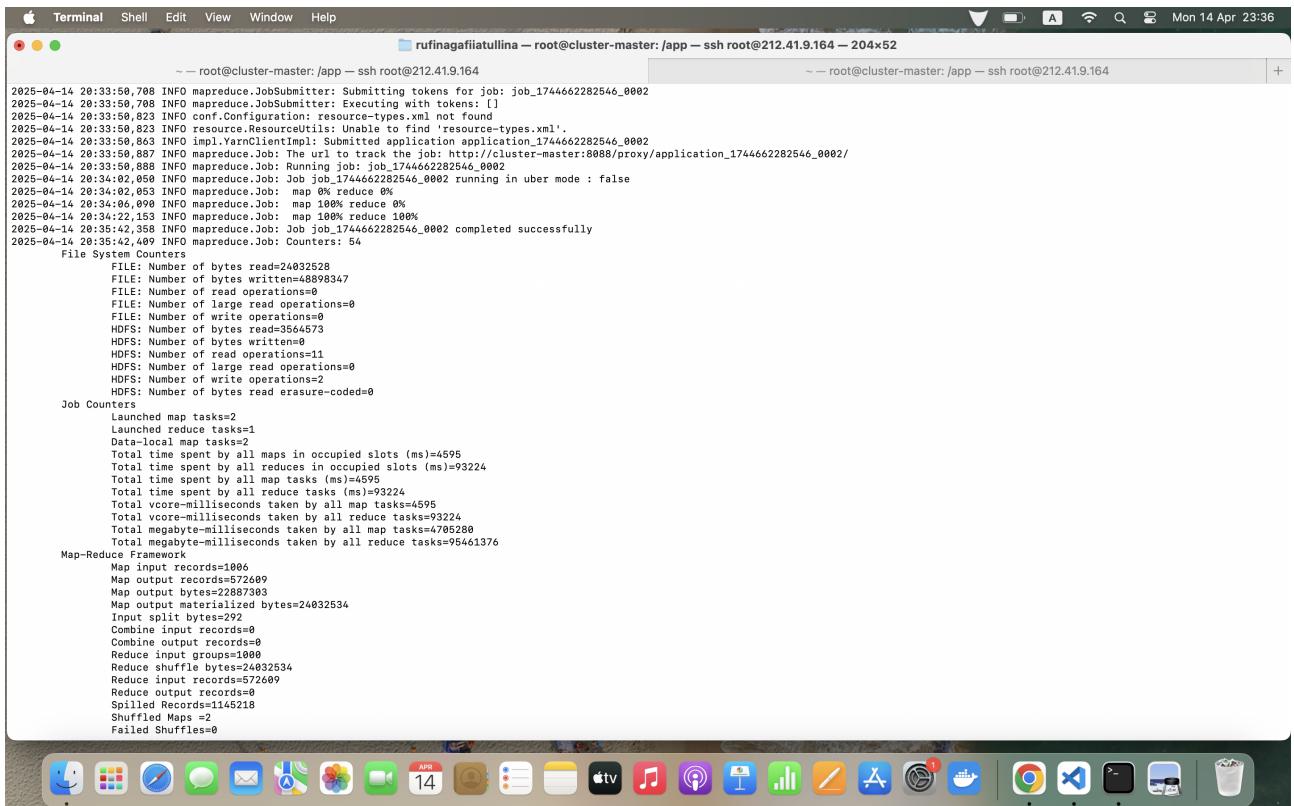
Figure 2: HDFS files

```

rufinagiafiatullina -- root@cluster-master: /app -- ssh root@212.41.9.164 -- 204x52
-- root@cluster-master: /app# bash index.sh
Total time spent by all map tasks (ms)=4545
Total time spent by all reduce tasks (ms)=43646
Total vcore-milliseconds taken by all map tasks=4545
Total vcore-milliseconds taken by all reduce tasks=43646
Total megabyte-milliseconds taken by all map tasks=465400
Total megabyte-milliseconds taken by all reduce tasks=44693504
Map-Reduce Framework
Map input records=1006
Map output records=2773
Map output bytes=2237494
Map output serialized bytes=2743052
Input split bytes=392
Combine input records=0
Combine output records=0
Reduce input groups=43379
Reduce shuffle bytes=2743052
Reduce input records=252773
Reduce output records=0
Spilled records=405546
Shuffled Maps=2
Failed Shuffles=0
Merged Map outputs=2
Physical memory (bytes) snapshot=972181504
Virtual memory (bytes) snapshot=7682843904
Total committed heap usage (bytes)=1129316352
Peak Map Physical memory (bytes)=368741800
Peak Map Virtual memory (bytes)=255818944
Peak Reduce Physical memory (bytes)=313741322
Peak Reduce Virtual memory (bytes)=3196954600
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=3564281
File Output Format Counters
Bytes Written=0
2025-04-14 20:33:46,938 INFO streaming.StreamJob: Output directory: /tmp/index/output
Running MapReduce2 job...
Deleted /tmp/index/output
Running!
packageJobJar: [/tmp/hadoop-unjar8176610180776858303/] [] /tmp/streamjob1408445843569400914.jar tmpDir=null
2025-04-14 20:33:49,572 INFO client.DefaultToHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-14 20:33:49,682 INFO client.DefaultToHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-14 20:33:49,816 INFO mapred.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744662282546_0002
2025-04-14 20:33:50,518 INFO mapred.FileInputFormat: Total input files to process : 1
2025-04-14 20:33:50,563 INFO mapreduce.JobSubmitter: number of splits:2

```

Figure 3: Indexing files



```

Terminal Shell Edit View Window Help
rufinagafiatiullina — root@cluster-master: /app — ssh root@212.41.9.164 — 204x52
~ — root@cluster-master: /app — ssh root@212.41.9.164 + Mon 14 Apr 23:36

2025-04-14 20:33:59.788 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744662282546_0002
2025-04-14 20:33:59.788 INFO mapreduce.JobSubmitter: Executing local JobRunner: []
2025-04-14 20:33:59.823 INFO conf.Configuration: resource-types.xml not found
2025-04-14 20:33:59.823 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-14 20:33:59.863 INFO impl.YarnClientImpl: Submitted application application_1744662282546_0002
2025-04-14 20:33:59.887 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744662282546_0002/
2025-04-14 20:33:59.889 INFO mapreduce.Job: Running job: job_1744662282546_0002
2025-04-14 20:34:02.055 INFO mapreduce.Job: Job job_1744662282546_0002 running in uber mode : false
2025-04-14 20:34:02.055 INFO mapreduce.Job: map 0% reduce 0%
2025-04-14 20:34:06.093 INFO mapreduce.Job: map 100% reduce 0%
2025-04-14 20:34:22.159 INFO mapreduce.Job: map 100% reduce 100%
2025-04-14 20:35:42.058 INFO mapreduce.Job: Job job_1744662282546_0002 completed successfully
2025-04-14 20:35:42.058 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=24032528
FILE: Number of bytes written=48898347
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=3564573
HDFS: Number of bytes written=0
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)=4595
Total time spent by all reduces in occupied slots (ms)=93224
Total time spent by all map tasks (ms)=4595
Total time spent by all reduce tasks (ms)=93224
Total vcore-milliseconds taken by all map tasks=4595
Total vcore-milliseconds taken by all reduce tasks=93224
Total megabyte-milliseconds taken by all map tasks=4705280
Total megabyte-milliseconds taken by all reduce tasks=95461376
Map-Reduce Framework
Map input records=1006
Map output records=572609
Map output bytes=22887303
Map output materialized bytes=24032534
Input split bytes=292
Combine input records=0
Combine output records=0
Reduce input groups=1000
Reduce shuffle bytes=24032534
Reduce input records=572609
Reduce output records=0
Spilled Records=1145218
Shuffled Maps =2
Failed Shuffles=0
Map Counters
Map input records=1006
Map output records=572609
Map output bytes=22887303
Map output materialized bytes=24032534
Input split bytes=292
Combine input records=0
Combine output records=0
Reduce input groups=1000
Reduce shuffle bytes=24032534
Reduce input records=572609
Reduce output records=0
Spilled Records=1145218
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC Local elapsed (ms)=113
CPU time spent (ms)=12498
Physical memory (bytes) snapshot=978292736
Virtual memory (bytes) snapshot=7680532480
Total committed heap usage (bytes)=1892616192
Peak Map Physical memory (bytes)=369336320
Peak Map Virtual memory (bytes)=2559463424
Peak Reduce Physical memory (bytes)=387612672
Peak Reduce Virtual memory (bytes)=3262529536
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=564281
File Output Format Counters
Bytes Written=0
2025-04-14 20:35:42.409 INFO streaming.StreamJob: Output directory: /tmp/index/output
Indexing complete. Info inserted into Cassandra.
root@cluster-master:/app# 
```

Figure 4: Indexing files



```

Terminal Shell Edit View Window Help
rufinagafiatiullina — root@cluster-master: /app — ssh root@212.41.9.164 — 204x52
~ — root@cluster-master: /app — ssh root@212.41.9.164 + Mon 14 Apr 23:36

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)=4595
Total time spent by all reduces in occupied slots (ms)=93224
Total time spent by all map tasks (ms)=4595
Total time spent by all reduce tasks (ms)=93224
Total vcore-milliseconds taken by all map tasks=4595
Total vcore-milliseconds taken by all reduce tasks=93224
Total megabyte-milliseconds taken by all map tasks=4705280
Total megabyte-milliseconds taken by all reduce tasks=95461376
Map-Reduce Framework
Map input records=1006
Map output records=572609
Map output bytes=22887303
Map output materialized bytes=24032534
Input split bytes=292
Combine input records=0
Combine output records=0
Reduce input groups=1000
Reduce shuffle bytes=24032534
Reduce input records=572609
Reduce output records=0
Spilled Records=1145218
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC Local elapsed (ms)=113
CPU time spent (ms)=12498
Physical memory (bytes) snapshot=978292736
Virtual memory (bytes) snapshot=7680532480
Total committed heap usage (bytes)=1892616192
Peak Map Physical memory (bytes)=369336320
Peak Map Virtual memory (bytes)=2559463424
Peak Reduce Physical memory (bytes)=387612672
Peak Reduce Virtual memory (bytes)=3262529536
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=564281
File Output Format Counters
Bytes Written=0
2025-04-14 20:35:42.409 INFO streaming.StreamJob: Output directory: /tmp/index/output
Indexing complete. Info inserted into Cassandra.
root@cluster-master:/app# 
```

Figure 5: Indexing files

After that I check Cassandra again. Fig.6 shows that now I have `index_keyspace` with four tables. Output from tables of some examples you can see in picture. In total I have 1000 indexed documents.

```

Terminal Shell Edit View Window Help
rufinagafiatullina — root@cluster-master:/app — ssh root@212.41.9.164 — 204x52
~ — root@cluster-master:/app — ssh root@212.41.9.164
(.venv) root@cluster-master:/app# cqlsh cassandra-server 9042
WARNING: cqlsh was built against 5.0.0, but this server is 5.0.4. All features may not work!
Connected to Test Cluster at cassandra-server:9042
[cqlsh 6.2.0 | Cassandra 5.0.4 | QOL spec 3.4.7 | Native protocol v5]
Use HELP for help.
cqlsh> desc keyspaces;
system      system_distributed  system_traces  system_virtual_schema
system_auth  system_schema      system_views
cqlsh> desc keyspaces;
index_keyspace  system_auth      system_schema  system_views
system        system_distributed system_traces  system_virtual_schema
cqlsh> use index_keyspace;
cqlsh:index_keyspace> desc tables;
corpus_info  doc_frequency_of_term  doc_info  term_frequency_in_doc
cqlsh:index_keyspace> select * from corpus_info where corpus_name='whole_corpus';
corpus_name | doc_n | total_doc_length
whole_corpus | 1000 | 572689
(1 rows)
cqlsh:index_keyspace> select * from doc_frequency_of_term where corpus_name='whole_corpus' and term='satisfy';
corpus_name | term | doc_frequency
whole_corpus | satisfy | 4
(1 rows)
cqlsh:index_keyspace> select * from term_frequency_in_doc where corpus_name='whole_corpus' and term='satisfy';
term | corpus_name | doc_id | doc_title | term.frequency
satisfy | whole_corpus | 4136530 | A Bay of Blood | 1
satisfy | whole_corpus | 33228023 | A Guided Tour of Madness | 1
satisfy | whole_corpus | 43986815 | A Game of Chance | 1
satisfy | whole_corpus | 47388207 | A Lego Brickumentary | 1
(4 rows)
cqlsh:index_keyspace> select * from doc_info where doc_id=4136530 and doc_title='A Bay of Blood';
doc_id | doc_title | doc_length
4136530 | A Bay of Blood | 2726
(1 rows)
cqlsh:index_keyspace>

```

Figure 6: Checking Cassandra after indexing

2.2.2 Indexing Local File

I want to index my small sample.txt file. The content of this file is following:

```

1 Hello everyone! Nice to meet you!
2 It is my sample text file.

```

Since I want to track that my script update information in Cassnadle properly, then I output information about word 'hello' from Cassandra tables. Result you can see in fig.7. These results I will use to compare with results after indexing sample.txt file. I execute index.sh with /data/sample.txt argument. The logs and outputs you can see in fig.8, fig.9. After that I check in HDFS /local_files directory where index.sh stores local files that should be indexed. The results you can see in fig.10. The file not only appeared on the HDFS, but also received a random 6-digit ID (doc_id), the original file name (doc_title), and its contents (doc_text).

```

Terminal Shell Edit View Window Help
rufinagafiatullina — root@cluster-master: /app — ssh root@212.41.9.164 — 204x52
~ — root@cluster-master: /app — ssh root@212.41.9.164
cqlsh:index_keyspace> select * from doc_frequency_of_term where corpus_name='whole_corpus' and term='hello';
corpus_name | term | doc_frequency
whole_corpus | hello | 2
(1 rows)
cqlsh:index_keyspace> select * from term_frequency_in_doc where corpus_name='whole_corpus' and term='hello';
term | corpus_name | doc_id | doc_title | term_frequency
hello | whole_corpus | 55178618 | A Boogie wit da Hoodie discography | 1
hello | whole_corpus | 64251145 | A Couple of Cuckoos | 6
(2 rows)
cqlsh:index_keyspace>

```

The screenshot shows a macOS desktop with a Terminal window open. The window title is "rufinagafiatullina — root@cluster-master: /app — ssh root@212.41.9.164". The terminal content displays two CQLSH queries. The first query selects the document frequency of the term 'hello' in the 'whole_corpus' keyspace, resulting in one row with a corpus name of 'whole_corpus', a term of 'hello', and a doc frequency of 2. The second query selects the term frequency in the documents containing the term 'hello' in the same keyspace, resulting in two rows with doc IDs 55178618 and 64251145, and their respective term frequencies of 1 and 6. Below the terminal window is a dock with various application icons.

Figure 7: Term 'hello'

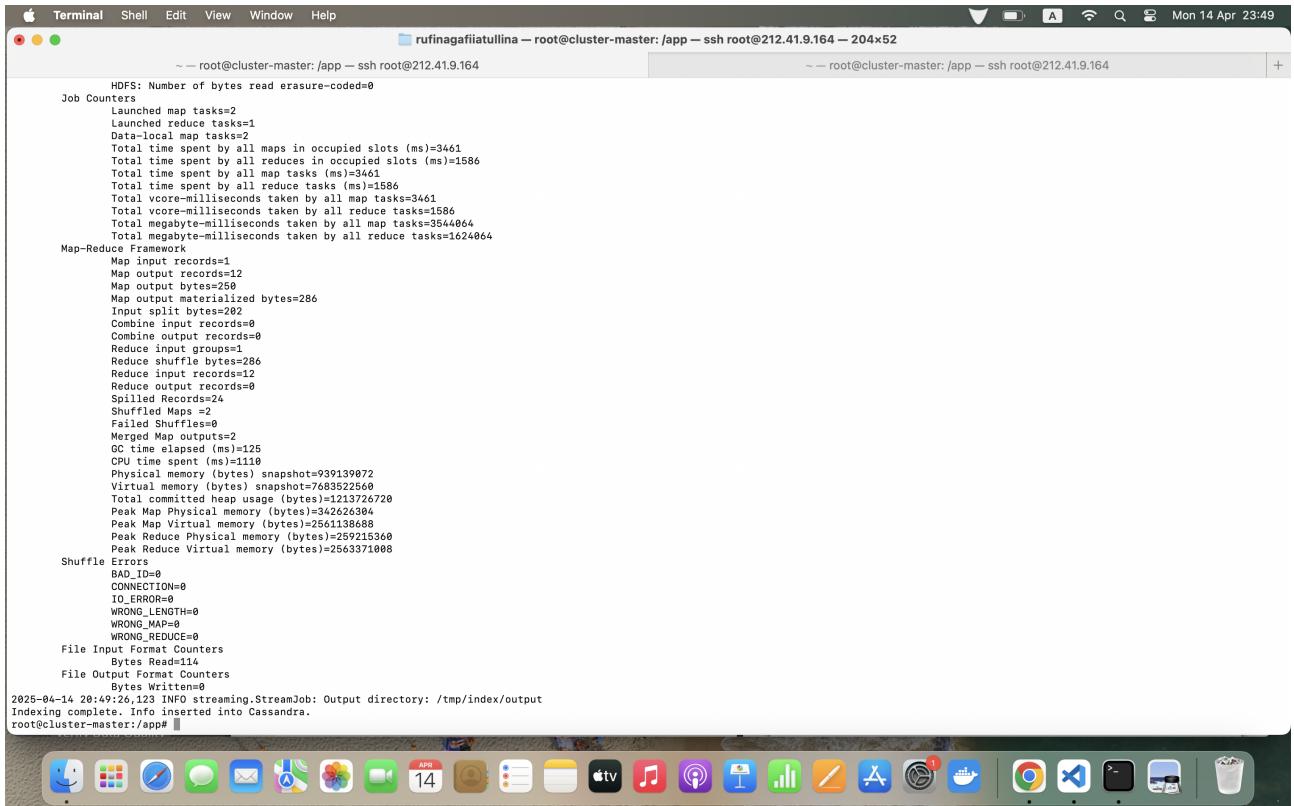
```

Terminal Shell Edit View Window Help
rufinagafiatullina — root@cluster-master: /app — ssh root@212.41.9.164 — 204x52
~ — root@cluster-master: /app — ssh root@212.41.9.164
root@cluster-master:/app# cat data/sample.txt
Hello everyone! Nice to meet you!
It is my sample text file.
root@cluster-master:/app# bash index.sh data/sample.txt
Put file from data/sample.txt to hdfs
rm: '/local_files': No such file or directory
Found 0 items
-rw-r--r-- 1 root supergroup 76 2025-04-14 20:48 /local_files/sample.txt
Setting up Cassandra schema...
WARNING: cqlsh was built against 5.0.0, but this server is 5.0.4. All features may not work!
Running MapReduce job...
Deleted /tmp/index/output
Running!
packageJobJar: [/tmp/hadoop-unjar5157896812931920797/] [] /tmp/streamjob5278404338327802928.jar tmpDir=null
2025-04-14 20:48:44,442 INFO client.DefaultTNHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8083
2025-04-14 20:48:44,558 INFO client.DefaultTNHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8083
2025-04-14 20:48:44,596 INFO mapreduce.JobResourceUploader: Duplicating Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744662282546_0003
2025-04-14 20:48:45,378 INFO mapred.FileInputFormat: Total input files: 1
2025-04-14 20:48:45,408 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-14 20:48:45,483 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744662282546_0003
2025-04-14 20:48:45,483 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-14 20:48:45,588 INFO conf.Configuration: resource-types.xml not found
2025-04-14 20:48:45,588 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-14 20:48:45,625 INFO impl.YarnClientImpl: Submitted application application_1744662282546_0003
2025-04-14 20:48:45,645 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744662282546_0003/
2025-04-14 20:48:45,658 INFO mapreduce.Job: Job ID: job_1744662282546_0003
2025-04-14 20:48:45,728 INFO mapreduce.Job: Job: map 0% reduce 0%
2025-04-14 20:48:45,739 INFO mapreduce.Job: map 100% reduce 100%
2025-04-14 20:48:45,884 INFO mapreduce.Job: map 100% reduce 100%
2025-04-14 20:49:02,911 INFO mapreduce.Job: map 100% reduce 100%
2025-04-14 20:49:02,928 INFO mapreduce.Job: Job job_1744662282546_0003 completed successfully
2025-04-14 20:49:03,024 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=112
FILE: Number of bytes written=833851
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=314
HDFS: Number of bytes written=0
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)=3529
Total time spent by all reduces in occupied slots (ms)=1614
Total time spent by all map tasks (ms)=3529
Total time spent by all reduce tasks (ms)=1614

```

The screenshot shows a macOS desktop with a Terminal window open. The window title is "rufinagafiatullina — root@cluster-master: /app — ssh root@212.41.9.164". The terminal content shows the execution of a Hadoop MapReduce job. It starts by creating a local file 'sample.txt' with the text "Hello everyone! Nice to meet you! It is my sample text file.". Then it runs a script 'index.sh' to copy this file to HDFS. The log then details the submission of the job to the ResourceManager, the creation of the job ID, and the tracking URL. It also provides detailed file system and job counters, including metrics for map and reduce tasks, and the total time spent in occupied slots for both stages. Below the terminal window is a dock with various application icons.

Figure 8: Indexing local file



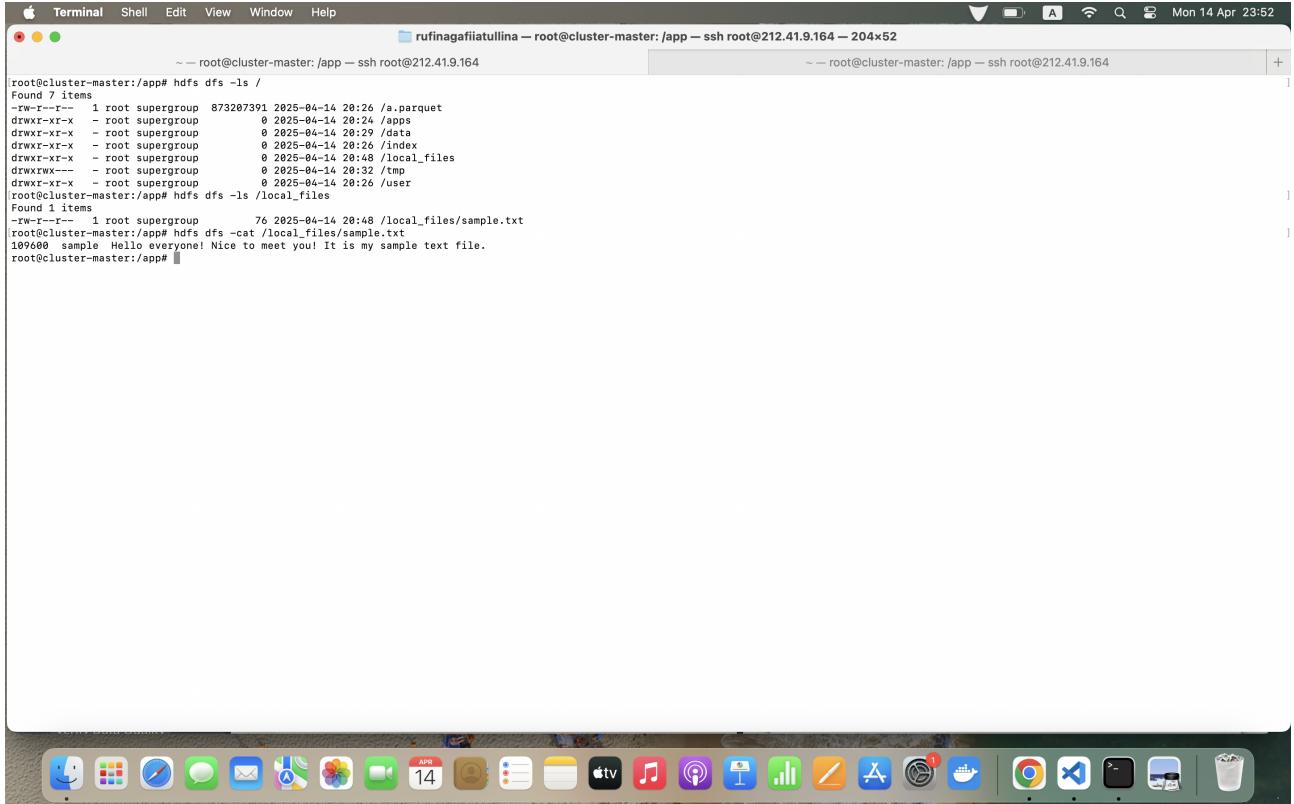
```

Terminal Shell Edit View Window Help
rufinagafiatullina — root@cluster-master: /app — ssh root@212.41.9.164 — 204x52
~ — root@cluster-master: /app — ssh root@212.41.9.164 + Mon 14 Apr 23:49

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)=3461
  Total time spent by all reduces in occupied slots (ms)=1586
  Total time spent by all map tasks (ms)=3461
  Total time spent by all reduce tasks (ms)=1586
  Total core-milliseconds taken by all map tasks=3461
  Total core-milliseconds taken by all reduce tasks=1586
  Total megabyte-milliseconds taken by all map tasks=354.064
  Total megabyte-milliseconds taken by all reduce tasks=1624.064
Map-Reduce Framework
  Map input records=1
  Map output records=12
  Map output bytes=250
  Map output materialized bytes=286
  Input splits=1, records=202
  Combine input records=0
  Combine output records=8
  Reduce input groups=1
  Reduce input records=286
  Reduce shuffle bytes=286
  Reduce input records=12
  Reduce output records=0
  Spilled Records=24
  Shuffled Maps = 1
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=1125
  CPU time spent (ms)=1110
  Physical memory (bytes) snapshot=939139072
  Virtual memory (bytes) snapshot=7683522560
  Total committed heap usage (bytes)=1213726720
  Peak Map Physical memory (bytes)=342626304
  Peak Map Virtual memory (bytes)=2561138688
  Peak Reduce Physical memory (bytes)=259215360
  Peak Reduce Virtual memory (bytes)=2563371008
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=14
  File Output Format Counters
    Bytes Written=0
2025-04-14 20:49:26,123 INFO streaming.StreamingJob: Output directory: /tmp/index/output
Indexing complete. Inf inserted into Cassandra.
root@cluster-master:/app# 

```

Figure 9: Indexing local file



```

Terminal Shell Edit View Window Help
rufinagafiatullina — root@cluster-master: /app — ssh root@212.41.9.164 — 204x52
~ — root@cluster-master: /app — ssh root@212.41.9.164 + Mon 14 Apr 23:52

root@cluster-master:/app# hdfs dfs -ls /
Found 7 items
-rw-r--r-- 1 root supergroup 873207391 2025-04-14 20:26 a.parquet
drwxr-xr-x - root supergroup 0 2025-04-14 20:24 /apps
drwxr-xr-x - root supergroup 0 2025-04-14 20:29 /data
drwxr-xr-x - root supergroup 0 2025-04-14 20:26 /index
drwxr-xr-x - root supergroup 0 2025-04-14 20:48 /local_files
drwxr-xr-x - root supergroup 0 2025-04-14 20:32 /tmp
drwxr-xr-x - root supergroup 0 2025-04-14 20:26 /user
root@cluster-master:/app# hdfs dfs -ls /local_files
Found 1 items
-rw-r--r-- 1 root supergroup 76 2025-04-14 20:48 /local_files/sample.txt
root@cluster-master:/app# hdfs dfs -cat /local_files/sample.txt
109600 sample Hello everyone! Nice to meet you! It is my sample text file.
root@cluster-master:/app# 

```

Figure 10: Check HDFS

After that I again check Cassandra tables for term 'hello'. As the result, values changed. Moreover, using doc_id from HDFS result, I find sample.txt in doc_info table. It means that index.sh index local file properly.



The screenshot shows a Mac OS X desktop with a Terminal window open. The window title is 'rufinagafiatullina — root@cluster-master: /app — ssh root@212.41.9.164 — 204x52'. The terminal session is running cqlsh: index_keystore>. It displays several queries and their results:

```

cqlsh:index_keystore> select * from doc_info where doc_id=109600 and doc_title='sample';
doc_id | doc_title | doc_length
109600 | sample | 12

(1 rows)

cqlsh:index_keystore> select * from corpus_info where corpus_name='whole_corpus';
corpus_name | doc_n | total_doc_length
whole_corpus | 1001 | 572621

(1 rows)

cqlsh:index_keystore> select * from term_frequency_in_doc where corpus_name='whole_corpus' and term='hello';
term | corpus_name | doc_id | doc_title | term_frequency
hello | whole_corpus | 109600 | sample | 1
hello | whole_corpus | 55178618 | A Boogie wit da Hoodie discography | 1
hello | whole_corpus | 64251145 | A Couple of Cuckoos | 6

(3 rows)

cqlsh:index_keystore> select * from doc_frequency_of_term where corpus_name='whole_corpus' and term='hello';
corpus_name | term | doc_frequency
whole_corpus | hello | 3

(1 rows)

cqlsh:index_keystore>

```

The desktop dock at the bottom contains icons for various Mac OS X applications like Mail, Safari, and Finder.

Figure 11: Checking sample.txt in Cassandra

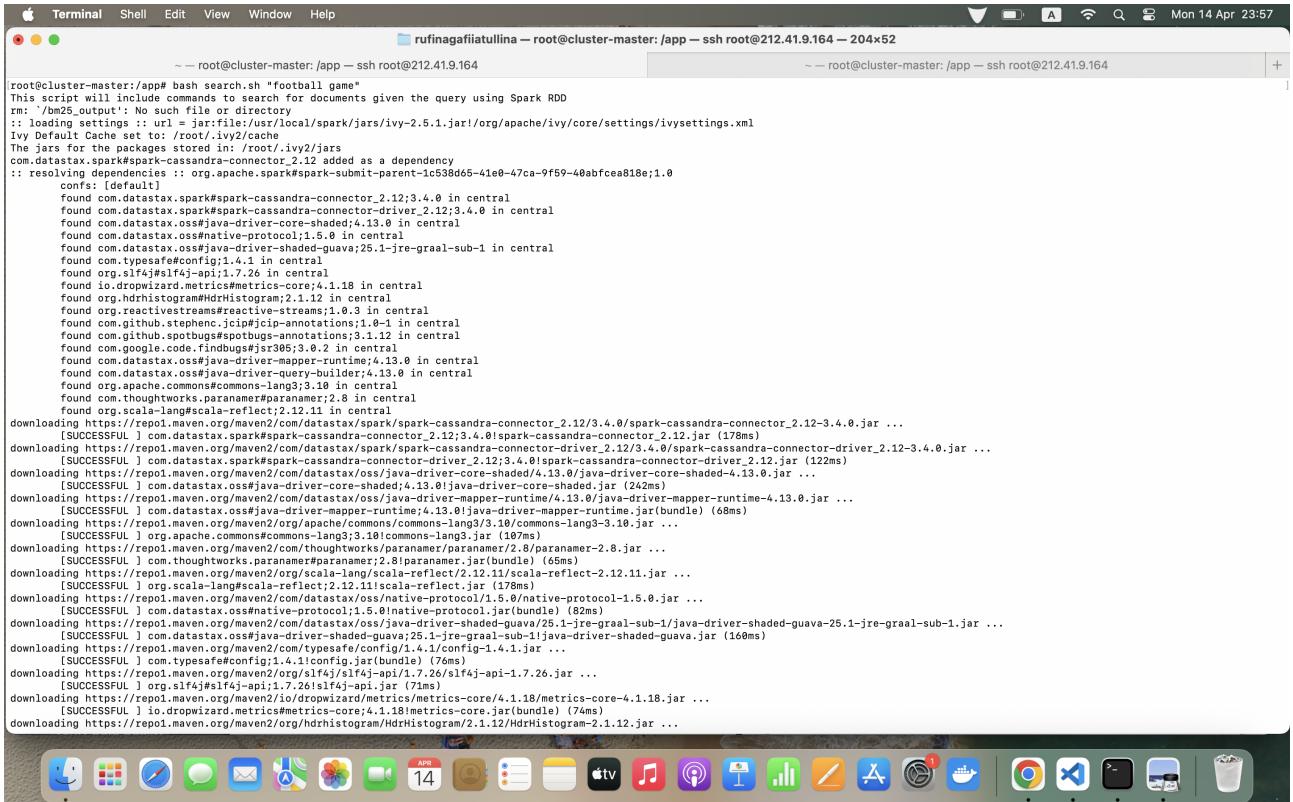
2.2.3 Searching

I search documents for query "football game" that shown in fig.12. The results present in fig.13. Here we can notice that our search engine gives us relevant documents (e.g. "A Football Life", "A Fan's Notes (film)").

I also search documents for non-existing term in query that shown in fig.14. The result is empty file that shown in fig.15 (nothing is printed to output).

I search documents for query "hello" that shown in fig.16. In the results you can see our local file that we indexed in previous examples that shown in fig.17. This further proves that the file has been indexed.

I search documents for query "create innovation" that shown in fig.18. The results present in fig.19. Here we can notice that our search engine gives us relevant documents.

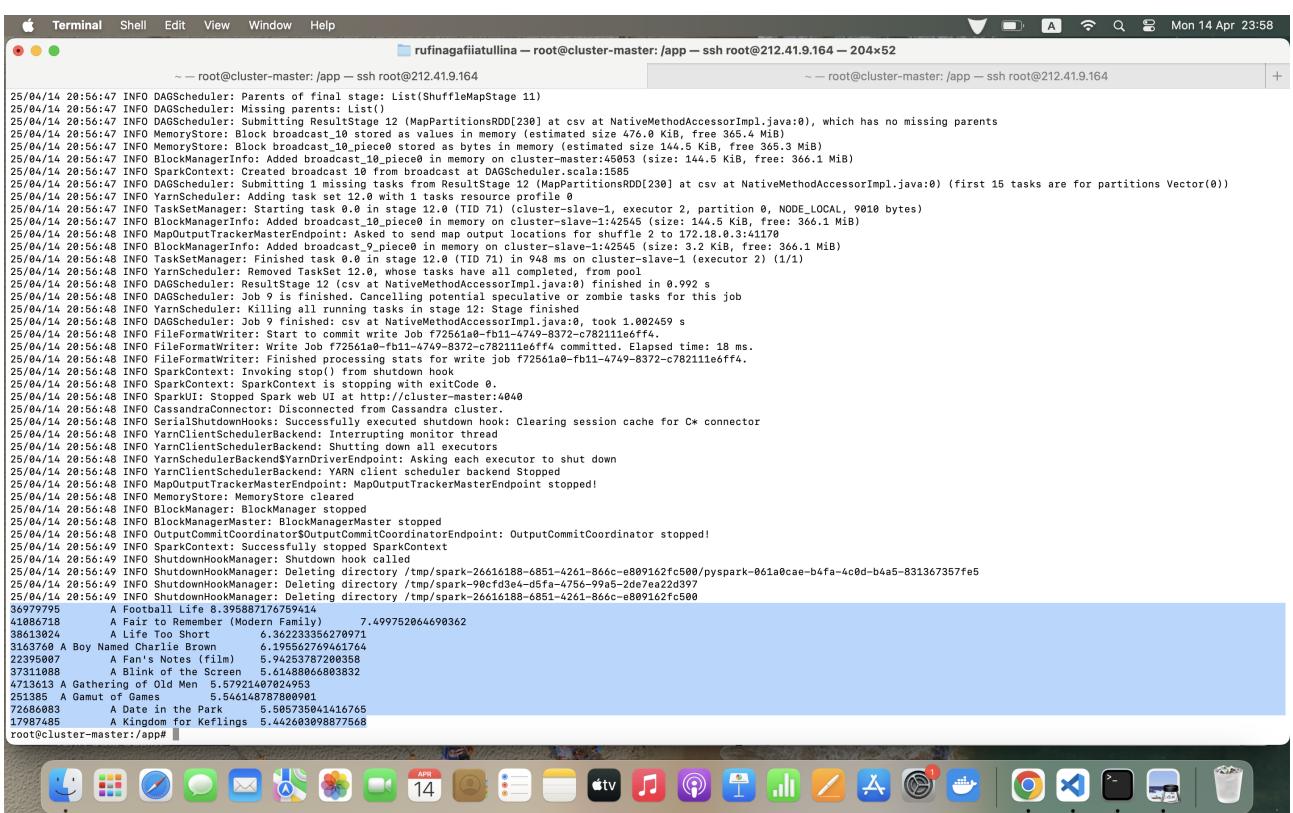


```
rufinagafiatullina - root@cluster-master:/app - ssh root@212.41.9.164 - 204x52
~ -- root@cluster-master:/app - ssh root@212.41.9.164
```

```
root@cluster-master:/app# bash search.sh "football game"
This script will index commands and search for documents given the query using Spark RDD
rm -f ./bm25_output*: No such file or directory
:: loading settings :: url = jarfile:/usr/local/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
com.datashax.spark#spark-cassandra-connector_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-lcs38d65-41e0-47ca-9f59-40abfce818e;1.0
  confs: [default]
    found com.datashax.spark#spark-cassandra-connector_2.12;3.4.0 in central
    found com.datashax.spark#spark-cassandra-connector-driver_2.12;3.4.0 in central
    found com.datashax.spark#spark-cassandra-connector_2.12;4.13.0 in central
    found com.datashax.oss#native-protocol;1.5.0 in central
    found com.datashax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 in central
    found com.typesafe#config;1.4.1 in central
    found org.slf4j#slf4j-api;1.7.26 in central
    found io.dropwizard.metrics#metrics-core;4.1.18 in central
    found org.hdrhistogram#HdrHistogram;2.1.12 in central
    found org.reactivestreams#reactive-streams;1.0.3 in central
    found com.github.sbt#jmh-plugin;1.2.0 in central
    found com.cubrid#cassandra-connector;2.12 in central
    found com.google.code.findbugs#jsr305;3.0.2 in central
    found com.datashax.oss#java-driver-mapper-runtime;4.13.0 in central
    found com.datashax.oss#java-driver-query-builder;4.13.0 in central
    found org.apache.commons#commons-lang3;3.1.0 in central
    found com.thoughtworks.parameter#parameter;2.8 in central
    found org.scala-lang#scala-reflect;2.12.11 in central
downloading https://repo1.maven.org/maven2/com/datashax/spark/spark-cassandra-connector_2.12/3.4.0/spark-cassandra-connector_2.12-3.4.0.jar ...
[SUCCESSFUL ] com.datashax.spark#spark-cassandra-connector_2.12;3.4.0!spark-cassandra-connector_2.12.jar (178ms)
downloading https://repo1.maven.org/maven2/com/datashax/spark/spark-cassandra-connector_2.12/3.4.0/spark-cassandra-connector-driver_2.12-3.4.0.jar ...
[SUCCESSFUL ] com.datashax.spark#spark-cassandra-connector-driver_2.12;3.4.0!spark-cassandra-connector-driver_2.12.jar (122ms)
downloading https://repo1.maven.org/maven2/com/datashax/oss#java-driver-core-shaded;4.13.0!java-driver-core-shaded.jar ...
[SUCCESSFUL ] com.datashax.oss#java-driver-core-shaded;4.13.0!java-driver-core-shaded.jar (242ms)
downloading https://repo1.maven.org/maven2/com/datashax/oss/java-driver-mapper-runtime;4.13.0!java-driver-mapper-runtime.jar ...
[SUCCESSFUL ] com.datashax.oss#java-driver-mapper-runtime;4.13.0!java-driver-mapper-runtime.jar(bundle) (68ms)
downloading https://repo1.maven.org/maven2/org/apache/commons#commons-lang3;3.10!commons-lang3.jar (107ms)
[SUCCESSFUL ] org.apache.commons#commons-lang3;3.10!commons-lang3.jar ...
downloading https://repo1.maven.org/maven2/com/thoughtworks/parameter#parameter;2.8!parameter-2.8.jar ...
[SUCCESSFUL ] com.thoughtworks.parameter#parameter;2.8!parameter-2.8.jar (65ms)
downloading https://repo1.maven.org/maven2/org/scalaj/scalaj-reflect/2.12.11/scalaj-reflect-2.12.11.jar ...
[SUCCESSFUL ] org.scala-lang#scala-reflect;2.12.11!scala-reflect.jar (378ms)
downloading https://repo1.maven.org/maven2/com/datashax/oss/native-protocol/1.5.0/native-protocol-1.5.0.jar ...
[SUCCESSFUL ] com.datashax.oss#native-protocol;1.5.0!native-protocol.jar(bundle) (82ms)
downloading https://repo1.maven.org/maven2/com/datashax/oss/java-driver-shaded-guava;25.1-jre-graal-sub-1!java-driver-shaded-guava.jar ...
[SUCCESSFUL ] com.datashax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1!java-driver-shaded-guava.jar (160ms)
downloading https://repo1.maven.org/maven2/com/typesafe/config/1.4.1/config-1.4.1.jar ...
[SUCCESSFUL ] com.typesafe#config;1.4.1!config.jar(bundle) (76ms)
downloading https://repo1.maven.org/maven2/com/datashax/oss#java-driver-mapper-runtime;4.13.0!java-driver-mapper-runtime.jar ...
[SUCCESSFUL ] com.datashax.oss#java-driver-mapper-runtime;4.13.0!java-driver-mapper-runtime.jar ...
downloading https://repo1.maven.org/maven2/io/dropwizard/metrics#metrics-core;4.1.18!metrics-core-4.1.18.jar ...
[SUCCESSFUL ] io.dropwizard.metrics#metrics-core;4.1.18!metrics-core.jar(bundle) (74ms)
downloading https://repo1.maven.org/maven2/org/hdrhistogram/HdrHistogram/2.1.12/HdrHistogram-2.1.12.jar ...

```

Figure 12: search.sh "football game"



```
rufinagafiatullina - root@cluster-master:/app - ssh root@212.41.9.164 - 204x52
~ -- root@cluster-master:/app - ssh root@212.41.9.164
```

```
25/04/14 20:56:47 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 11)
25/04/14 20:56:47 INFO DAGScheduler: Missing parents: List()
25/04/14 20:56:47 INFO DAGScheduler: Submitting ResultStage 12 (MapPartitionsRDD[230] at csv at NativeMethodAccessorImpl.java:0), which has no missing parents
25/04/14 20:56:47 INFO MemoryStore: Block broadcast_10 stored as values in memory (estimated size 476.0 KiB, free 365.4 MiB)
25/04/14 20:56:47 INFO BlockManagerInfo: Block broadcast_10_piece0 stored as bytes in memory (estimated size 144.5 KiB, free 365.3 MiB)
25/04/14 20:56:47 INFO BlockManagerInfo: Added broadcast_10_piece0 in memory on cluster-master:45953 (size: 144.5 KiB, free: 366.1 MiB)
25/04/14 20:56:47 INFO DAGScheduler: Submitting 1 pending tasks from ResultStage 12 (MapPartitionsRDD[230] at csv at NativeMethodAccessorImpl.java:0) (first 15 tasks are for partitions Vector(0))
25/04/14 20:56:47 INFO YarnScheduler: Adding task set 12.0 with 1 tasks resource profile 0
25/04/14 20:56:47 INFO TaskSetManager: Starting task 0.0 in stage 12.0 (TID 71) (cluster-slave-1, executor 2, partition 0, NODE_LOCAL, 9010 bytes)
25/04/14 20:56:48 INFO BlockManagerInfo: Added broadcast_10_piece0 in memory on cluster-slave-1:42546 (size: 3.2 KiB, free: 366.1 MiB)
25/04/14 20:56:48 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 2 to 172.18.0.3:14170
25/04/14 20:56:48 INFO ClockManagerInfo: Added broadcast_9_piece0 in memory on cluster-slave-1:42546 (size: 3.2 KiB, free: 366.1 MiB)
25/04/14 20:56:48 INFO TaskSetManager: Finished task 0.0 in stage 12.0 (TID 71) in 948 ms on cluster-slave-1 (executor 2) (1/1)
25/04/14 20:56:48 INFO YarnScheduler: Removed TaskSet 12.0, whose tasks have all completed, from pool
25/04/14 20:56:48 INFO BlockManagerInfo: Added broadcast_10_piece0 in memory on cluster-slave-1:42546 (size: 144.5 KiB, free: 366.1 MiB)
25/04/14 20:56:48 INFO TaskSetManager: Removed TaskSet 12.0, whose tasks have all completed, from pool
25/04/14 20:56:48 INFO DAGScheduler: 10.0 as finished. Cancelling remaining tasks for this job
25/04/14 20:56:48 INFO DAGScheduler: Killing all running tasks in stage 12: Stage finished
25/04/14 20:56:48 INFO DAGScheduler: Job 9 finished: csv at NativeMethodAccessorImpl.java:0, took 1.002459 s
25/04/14 20:56:48 INFO FileFormatWriter: Start to commit write Job f72561a0-fb11-4749-8372-c782111e6ff4.
25/04/14 20:56:48 INFO FileFormatWriter: Write Job f72561a0-fb11-4749-8372-c782111e6ff4 committed. Elapsed time: 18 ms.
25/04/14 20:56:48 INFO FileFormatWriter: Finished processing stats for write Job f72561a0-fb11-4749-8372-c782111e6ff4.
25/04/14 20:56:48 INFO SparkContext: Invoking stop() from shutdown hook
25/04/14 20:56:48 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/14 20:56:48 INFO SparkContext: Stopped Spark Web UI at http://cluster-master:4040
25/04/14 20:56:48 INFO SparkSession: Disconnected from master cluster
25/04/14 20:56:48 INFO SerialShutdownHook: Successfully executed shutdown hook! Clearing session cache for C* connector
25/04/14 20:56:48 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/14 20:56:48 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/14 20:56:48 INFO YarnSchedulerBackendYarnDriverEndpoint: Asking each executor to shut down
25/04/14 20:56:48 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
25/04/14 20:56:48 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/14 20:56:48 INFO MemoryStore: MemoryStore cleared
25/04/14 20:56:48 INFO BlockManager: BlockManager stopped
25/04/14 20:56:48 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/14 20:56:49 INFO SparkContext: Successfully stopped SparkContext
25/04/14 20:56:49 INFO ShutdownHookManager: Deleting directory /tmp/spark-26616188-6851-4261-866c-e809162fc500/pyspark-061a0cae-b4fa-4c0d-b4a5-831367357fe5
25/04/14 20:56:49 INFO ShutdownHookManager: Deleting directory /tmp/spark-26616188-6851-4261-866c-e809162fc500
3697795 A Football Life 8.39588717675944
4188616 A Fair to Remember (Modern Times) 7.499752064690362
8663624 A Little Too Shy 6.36223358270971
3163768 A Boy Named Charlie Brown 6.1955427694461764
22395087 A Faron's Notes (film) 5.94253787289358
37311988 A Blink of the Screen 5.6148806883832
4713613 A Gathering of Old Men 5.7921407824953
251385 A Gamut of Games 5.54614878780091
72686083 A Date in the Park 5.5085735041416765
17987485 A Kingdom for Keflings 5.44260398877565
root@cluster-master:/app#
```

Figure 13: Results for "football game"

```

rufinagafiatullina - root@cluster-master: /app - ssh root@212.41.9.164 - 204x52
~ -- root@cluster-master: /app - ssh root@212.41.9.164
root@cluster-master:/app# bash search.sh "wefwrgwg"
This script will include commands to search for documents given the query using Spark RDD
Deleted /var/log/output
:: loading settings :: url = jar:file:/usr/local/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
com.datastax.spark#spark-cassandra-connector_2.12 added as a dependency
:: resolving dependencies :: [default]
  confs: [default]
    found com.datastax.spark#spark-cassandra-connector_2.12;3.4.0 in central
    found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.4.0 in central
    found com.github.sbt#minitest-interface_2.12;0.13.0 in central
    found com.datastax.oss#native-protocol_1.0.0 in central
    found com.datastax.oss#java-driver-shaded-quava;25.1-jre-graal-sub-1 in central
    found com.typesafe#config;1.4.1 in central
    found org.slf4j#slf4j-api;1.7.26 in central
    found io.dropwizard.metrics#metrics-core;4.1.18 in central
    found org.hdrhistogram#HdrHistogram;2.1.12 in central
    found org.reactivestreams#reactive-streams;1.0.3 in central
    found com.github.sbt#jcliprjciepAnnotations;1.0.0 in central
    found com.github.sbt#junit-interface_2.12;0.13.0 in central
    found com.google.code.findbugs#jsr305;3.0.2 in central
    found com.datastax.oss#java-driver-mapper-runtime;4.13.0 in central
    found com.datastax.oss#java-driver-query-builder;4.13.0 in central
    found com.org.apache.commons#lang3;3.1.10 in central
    found com.thoughtworks.parameter#parameter;2.8 in central
    found org.scala-lang#scala-reflect;2.12.11 in central
:: resolution report :: resolved 335ms :: artifacts id 12ms
:: modules in use ::

|   conf      | number | searchDownloaded|evicted| numberDownloaded |
|           |       |               |       |               |
| default    |   18   |       0        | 0     |       0        |
|           |       |               |       |               |
+-----+-----+-----+-----+-----+

```

Figure 14: search.sh "wefwrgwg"

```

rufinagafiatullina - root@cluster-master: /app - ssh root@212.41.9.164
~ -- root@cluster-master: /app - ssh root@212.41.9.164
25/04/14 20:59:33 INFO YarnScheduler: Killing all running tasks in stage 0: Stage finished
25/04/14 20:59:33 INFO DAGScheduler: Job 0 finished; head at /app/Server.py:38, took 3.087006 s
25/04/14 20:59:33 INFO FileOutputCommitter: File Output Committer Algorithm version is 1
25/04/14 20:59:33 INFO FileOutputCommitter: skip cleanup _tempory_files under output directory:false, ignore cleanup failures: false
25/04/14 20:59:33 INFO FileOutputCommitter: Using output committer class org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
25/04/14 20:59:33 INFO CodeGenerator: Code generated in 8.920265 ms
25/04/14 20:59:33 INFO DAGScheduler: Submitting job 1:csv at NativeMethodAccessorImpl.java:0 with 2 output partitions
25/04/14 20:59:33 INFO DAGScheduler: Final stage: ResultStage 1 (csv at NativeMethodAccessorImpl.java:0)
25/04/14 20:59:33 INFO DAGScheduler: Parents of final stage: List()
25/04/14 20:59:33 INFO DAGScheduler: Missing parents: List()
25/04/14 20:59:33 INFO DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD10) at csv at NativeMethodAccessorImpl.java:0, which has no missing parents
25/04/14 20:59:33 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 233.2 KB, free: 366.8 MiB)
25/04/14 20:59:33 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 85.8 KiB, free: 366.8 MiB)
25/04/14 20:59:33 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on cluster-master:42145 (size: 85.8 KiB, free: 366.2 MiB)
25/04/14 20:59:33 INFO DAGScheduler: Submitting job 2:csv at NativeMethodAccessorImpl.java:0 with 1 output partition
25/04/14 20:59:33 INFO DAGScheduler: Starting task 0 in stage 1 (NativeMethodAccessorImpl.java:0) for partition 0 (partitions Vector(0, 1))
25/04/14 20:59:33 INFO YarnScheduler: Adding task set 1.0 with 2 tasks, resource profile 0
25/04/14 20:59:33 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1) (cluster-slave-1, executor 1, partition 0, PROCESS_LOCAL, 8990 bytes)
25/04/14 20:59:33 INFO TaskSetManager: Starting task 1.0 in stage 1.0 (TID 2) (cluster-slave-1, executor 2, partition 1, PROCESS_LOCAL, 8990 bytes)
25/04/14 20:59:33 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on cluster-slave-1:34655 (size: 85.8 KiB, free: 366.2 MiB)
25/04/14 20:59:33 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on cluster-slave-1:33529 (size: 85.8 KiB, free: 366.2 MiB)
25/04/14 20:59:35 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 1422 ms on cluster-slave-1 (executor 1) (1/2)
25/04/14 20:59:35 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (TID 2) in 1953 ms on cluster-slave-1 (executor 2) (2/2)
25/04/14 20:59:35 INFO YarnScheduler: RequestStage 0 finished. All tasks have 0 pending tasks in front pool
25/04/14 20:59:35 INFO DAGScheduler: Stage 0 finished. Cancelling potential speculative or zombie tasks for this job
25/04/14 20:59:35 INFO YarnScheduler: Job 1 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/14 20:59:35 INFO DAGScheduler: Killing all running tasks in stage 1: Stage finished
25/04/14 20:59:35 INFO DAGScheduler: Job 1 finished: csv at NativeMethodAccessorImpl.java:0, took 2.006979 s
25/04/14 20:59:35 INFO FileFormatWriter: Start to commit write Job ed1204fe-99dd-4098-b831-adecd335a89e.
25/04/14 20:59:35 INFO FileFormatWriter: Write Job ed1204fe-99dd-4098-b831-adecd335a89e committed. Elapsed time: 18 ms.
25/04/14 20:59:35 INFO FileFormatWriter: Finished processing stats for write job ed1204fe-99dd-4098-b831-adecd335a89e.
25/04/14 20:59:35 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/14 20:59:35 INFO SparkContext: Stopping SparkContext at http://cluster-slave-1:4040
25/04/14 20:59:35 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/14 20:59:35 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/14 20:59:35 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
25/04/14 20:59:35 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/14 20:59:35 INFO MemoryStore: MemoryStore cleared
25/04/14 20:59:35 INFO BlockManager: BlockManager stopped
25/04/14 20:59:35 INFO BlockManager: BlockManager stopped
25/04/14 20:59:35 INFO OutputCommitCoordinatorEndpoint: OutputCommitCoordinatorEndpoint stopped!
25/04/14 20:59:35 INFO SparkContext: Successfully stopped SparkContext
25/04/14 20:59:36 INFO ShutdownHookManager: Shutdown hook called
25/04/14 20:59:36 INFO ShutdownHookManager: Deleting directory /tmp/spark-ff39a018-c862-4d9b-9303-d47c9808d706
25/04/14 20:59:36 INFO ShutdownHookManager: Deleting directory /tmp/spark-0b9e37c0-4eb4-ac25-84f9-090184fb49ae/pyspark-b7e39e1c-06e7-4931-be9c-4d6ba6f8c988
25/04/14 20:59:36 INFO CassandraConnector: Disconnected from Cassandra cluster.
25/04/14 20:59:36 INFO SerialShutdownHooks: Successfully executed shutdown hook: Clearing session cache for C* connector
root@cluster-master:/app#

```

Figure 15: Results for "wefwrgwg"

	modules	artifacts
conf	number	searched downloaded evicted number downloaded
default	18	0 0 0 18 0

Figure 16: search.sh "hello"

Figure 17: Results for "hello"

```

root@cluster-master:/app# bash search.sh "create innovation"
This script will include commands to search for documents given the query using Spark RDD
Deleted /bin/sh.output
:: loading settings :: url = jar:file:/usr/local/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
com.datastax.spark#spark-cassandra-connector_2.12 added as a dependency
:: resolving dependencies :: [default]
  confs: [default]
    found com.datastax.spark#spark-cassandra-connector_2.12;3.4.0 in central
    found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.4.0 in central
    found com.github.sbt#minitest-interface_2.12;0.13.0 in central
    found com.datastax.ossfive-protocol_1.0.0 in central
    found com.datastax.ossfivejava-driver-shaded-quava;25.1-jre-graal-sub-1 in central
    found com.typesafeconfig;1.4.1 in central
    found org.slf4j#slf4j-api;1.7.26 in central
    found io.dropwizard.metrics#metrics-core;4.1.18 in central
    found org.hdrhistogram#HdrHistogram;2.1.12 in central
    found org.reactivestreams#reactive-streams;1.0.3 in central
    found com.github.jclippings#jcl-annotations;1.0.0 in central
    found com.github.sbt#minitest-interface_2.12;0.13.0 in central
    found com.google.code.findbugs#jsr305;3.0.2 in central
    found com.datastax.ossfivejava-driver-mapper-runtime;4.13.0 in central
    found com.datastax.ossfivejava-driver-query-builder;4.13.0 in central
    found org.apache.commons#commons-lang3;3.1.0 in central
    found com.thoughtworks.parameter#parameter;2.8 in central
    found org.scala-lang#scala-reflect;2.12.11 in central
:: resolution report :: resolve 3377 :: artifacts d10mns
:: modules in use ::
com.datastax.oss#java-driver-core-shaded;4.13.0 from central in [default]
com.datastax.oss#java-driver-mapper-runtime;4.13.0 from central in [default]
com.datastax.oss#java-driver-query-builder;4.13.0 from central in [default]
com.datastax.oss#java-driver-shaded-quava;25.1-jre-graal-sub-1 from central in [default]
com.datastax.oss#native-protocol;1.5.0 from central in [default]
com.datastax.spark#spark-cassandra-connector-driver_2.12;3.4.0 from central in [default]
com.github.spotbugs#spotbugs-annotations;3.1.12 from central in [default]
com.github.jclippings#jcl-annotations;1.0.1 from central in [default]
com.google.code.findbugs#jsr305;3.0.0 from central in [default]
com.thoughtworks.parameter#parameter;2.8 from central in [default]
com.typesafeconfig;1.4.1 from central in [default]
io.dropwizard.metrics#metrics-core;4.1.18 from central in [default]
org.apache.commons#commons-lang3;3.1.0 from central in [default]
org.hdrhistogram#HdrHistogram;2.1.12 from central in [default]
org.reactivestreams#reactive-streams;1.0.3 from central in [default]
org.scala-lang#scala-reflect;2.12.11 from central in [default]
org.slf4j#slf4j-api;1.7.26 from central in [default]
+-----+-----+-----+-----+
|       | modules |       | artifacts |
| conf | number| searchDownloaded| evicted | numberDownloaded |
+-----+-----+-----+-----+
| default | 18 | 0 | 0 | 0 | 18 | 0 |

```

Figure 18: search.sh "create innovation"

```

root@cluster-master:/app# bash search.sh "create innovation"
This script will include commands to search for documents given the query using Spark RDD
Deleted /bin/sh.output
:: loading settings :: url = jar:file:/usr/local/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
com.datastax.spark#spark-cassandra-connector_2.12 added as a dependency
:: resolving dependencies :: [default]
  confs: [default]
    found com.datastax.spark#spark-cassandra-connector_2.12;3.4.0 in central
    found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.4.0 in central
    found com.github.sbt#minitest-interface_2.12;0.13.0 in central
    found com.datastax.ossfive-protocol_1.0.0 in central
    found com.datastax.ossfivejava-driver-shaded-quava;25.1-jre-graal-sub-1 in central
    found com.typesafeconfig;1.4.1 in central
    found org.slf4j#slf4j-api;1.7.26 in central
    found io.dropwizard.metrics#metrics-core;4.1.18 in central
    found org.hdrhistogram#HdrHistogram;2.1.12 in central
    found org.reactivestreams#reactive-streams;1.0.3 in central
    found com.github.jclippings#jcl-annotations;1.0.0 in central
    found com.google.code.findbugs#jsr305;3.0.2 in central
    found com.datastax.ossfivejava-driver-mapper-runtime;4.13.0 in central
    found com.datastax.ossfivejava-driver-query-builder;4.13.0 in central
    found org.apache.commons#commons-lang3;3.1.0 in central
    found com.thoughtworks.parameter#parameter;2.8 in central
    found org.scala-lang#scala-reflect;2.12.11 in central
    found org.slf4j#slf4j-api;1.7.26 from central in [default]
:: modules in use ::
com.datastax.oss#java-driver-core-shaded;4.13.0 from central in [default]
com.datastax.oss#java-driver-mapper-runtime;4.13.0 from central in [default]
com.datastax.oss#java-driver-query-builder;4.13.0 from central in [default]
com.datastax.oss#java-driver-shaded-quava;25.1-jre-graal-sub-1 from central in [default]
com.datastax.oss#native-protocol;1.5.0 from central in [default]
com.datastax.spark#spark-cassandra-connector-driver_2.12;3.4.0 from central in [default]
com.github.spotbugs#spotbugs-annotations;3.1.12 from central in [default]
com.github.jclippings#jcl-annotations;1.0.1 from central in [default]
com.google.code.findbugs#jsr305;3.0.0 from central in [default]
com.thoughtworks.parameter#parameter;2.8 from central in [default]
com.typesafeconfig;1.4.1 from central in [default]
io.dropwizard.metrics#metrics-core;4.1.18 from central in [default]
org.apache.commons#commons-lang3;3.1.0 from central in [default]
org.hdrhistogram#HdrHistogram;2.1.12 from central in [default]
org.reactivestreams#reactive-streams;1.0.3 from central in [default]
org.scala-lang#scala-reflect;2.12.11 from central in [default]
org.slf4j#slf4j-api;1.7.26 from central in [default]
+-----+-----+-----+-----+
|       | modules |       | artifacts |
| conf | number| searchDownloaded| evicted | numberDownloaded |
+-----+-----+-----+-----+
| default | 18 | 0 | 0 | 0 | 18 | 0 |

```

Figure 19: Results for "create innovation"