

BigData Assignment 2. Search Engine

Anastasia Pichugina DS-02

Repository: https://github.com/caxapb/BD_assignment2

Methodology

The search engine system is built with several technologies:

- Hadoop and Yarn: distributed data processing
- Cassandra: used database
- PySpark: efficient data processing and scores calculation

Application Workflow

The workflow is exactly the same as described in the assignment description:

- 1) .parquet file is processed to separate .txt files,
- 2) all data is moved to HDFS,
- 3) MapReduce pipeline computes required statistics, and saves it
- 4) app.py connects to the cassandra-server, creates the keyspace and tables, and inserts data from the reducer's output,
- 5) query.py reads the input query, takes data from Cassandra, calculates the BM25 score and prints the resulting queries.

Some implementation details:

- docker-compose up starts the application. The service cluster-master has the entrypoint app.sh that runs all scripts in the correct order: start-services.sh, prepare_data.sh, index.sh, and search.sh.
- start-services.sh:
 - starts all services required for Hadoop components running other scripts: HDFS, Yarn, web UI for MapReduce, and prepares Spark.
 - Sets the virtual environment and installs dependencies from the requirements.txt
- prepare_data.sh: loads a.parquet to the HDFS, prepares data using util .py files:
 - prepare_data1.py: goes through the main files and creates new separate .txt files. All files are loaded to the HDFS /data folder as <doc_id>_<doc_title>.txt.
 - prepare_data2.py creates the RDD object from those files and creates {doc_id}\t{title}\t{content} which will be saved to the /index/data/ in HDFS (part-00000 and _SUCCESS) using saveAsTextFile().

Note! index.sh doesn't accept any arguments. To change the amount of collected documents change the n parameter in the prepare_data1.py.

- MapReduce pipeline and data loading to Cassandra is conducted in the index.sh script. The script runs mapper1.py, reducer1.py, and app.py:

- `mapper1.py`: takes the input from the “/index/data” in HDFS, processes terms for each document and prints to the stdout info about terms and documents. The prints have ‘tags’: ‘TF’ and ‘DOC’. Each row starts with one of them so that the reducer could distinguish document and terms data.
- `reducer1.py`: processes `mapper1.py` output. Based on tags, it collects all required statistics: term frequencies - how many times document x meets the term y, document frequencies - how many documents contain word x, document data (doc_id - doc_length - doc_title), and global statistics so that `query.py` doesn’t compute them every time: document total count and average document length. The reducer’s output is collected in the /tmp/index_output folder (part-00000 and _SUCCESS files).
- `app.py` connects to the Cassandra-server, creates the keyspace “search_engine” and tables, takes the /tmp/index_output/part-00000 content and inserts it into tables.
- Then `search.sh` is running. This script must be run with an argument - input query. SparkSession is created and connected to the cassandra server. The input query is tokenized using regex. Cassandra tables are loaded and converted to the RDD objects:
 - `terms_rdd`: RDD with tuples of “term” and “document frequency”.
 - `term_frequencies_rdd`: RDD where each tuple is (term, (doc_id, term frequency value)). Since term frequencies are defined as “amount of term appearance id doc”, each tuple is unique,
 - `documents_rdd`: RDD with tuples of (doc_id, (length, title))
 - `stats` (global stats from Cassandra) isn’t converted to RDD but all needed values are extracted.
 - Based on obtained RDD objects, a new RDD is created: (doc_id, title, length, term frequency value, inverse document frequency value). Such tuples exist for each unique pair: (term from the query) - (document containing at least 1 word from the query).
 - The `search.sh` output is written to the console and `output.txt`

Cassandra schema and data storage:

Keyspace: “search_engine”

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

Tables:

terms:

```
CREATE TABLE IF NOT EXISTS search_engine.terms (
    term text PRIMARY KEY,
    df int
);
```

Since the BM25 formula requires document frequency for each term, this table is created.

term_frequencies:

```
CREATE TABLE IF NOT EXISTS search_engine.term_frequencies (  
    term text,  
    doc_id int,  
    tf int,  
    PRIMARY KEY (term, doc_id)  
);
```

Here the primary key is purple, because Term Frequency is defined by term and doc_id

documents:

```
CREATE TABLE IF NOT EXISTS search_engine.documents (  
    doc_id int PRIMARY KEY,  
    length int,  
    title text  
);
```

For each document we need to know its length, thus this table keeps these values.

global_stats:

```
CREATE TABLE IF NOT EXISTS search_engine.stats (  
    key text PRIMARY KEY,  
    value float  
);
```

Number of documents and average document length are constant, so we don't need to compute them every time processing input query, so this table keeps 2 of these values. The table has only 2 rows: ("key": "doc_total", "value": ...), ("key": "avg_length", "value": ...).

BM25 Search using RDD objects.

0) As it was mentioned before, Cassandra tables are loaded to RDD object and we have the following (*df* - documents frequency, *idf* - inverse document frequency, *tf* - term frequency):

N - total documents count

avg_length - average document length

terms_rdd - RDD with tuples: (term, df)

term_frequencies_rdd - RDD with tuples: (term, (doc_id, tf))

documents_rdd - RDD with tuples: (doc_id, (length, title))

- 1) Convert df values in terms_rdd to idf values. terms_rdd becomes (term, idf).
- 2) Join term_frequencies_rdd with it to get (term, ((doc_id, tf), idf)) and simplify it / transform to: (doc_id, (term, tf, idf)).
- 3) Join it with documents_rdd and transform to get: (doc_id, title, length, tf, idf). These tuples are unique for *doc_id* - *term* pairs, and this is exactly what we need. Since the formula iterates over all terms in a query and computes scores for each document for this term, this format allows map operation for each record.

- 4) Apply `bm25_scores` computing function (add a new field “score” to tuples) and reduce them by key summing scores for each document.
- 5) Sort documents by scores and extract top 10. Write them to the `output.txt` and the console.

To compute BM25 score I used the formula provided in the Assignment description:

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

The only 1 change I made: added 0.0001 to $N/\text{df}(t)$ to avoid $\log(0)$.

Demonstration.

To run this project you need to:

- Clone the git repository:

```
git clone git@github.com:caxapb/BD_assignment2.git
```

- Make sure you have a valid “.parquet” file in the `app/data/` folder. If not, create the data folder and copy the a.parquet file inside.
- Run the docker compose:

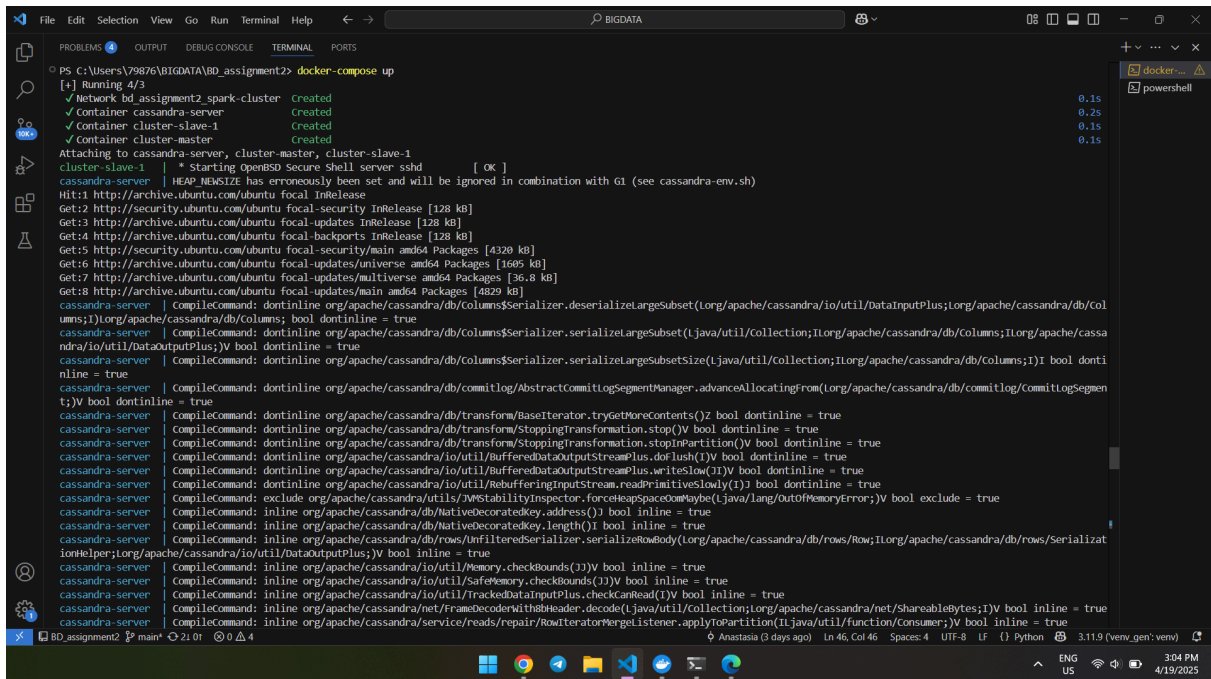
```
docker-compose up
```

This command will run `app.sh` that runs 3 containers (cluster-master, cassandra-server, and cluster-slave-1) and starts all other scripts.

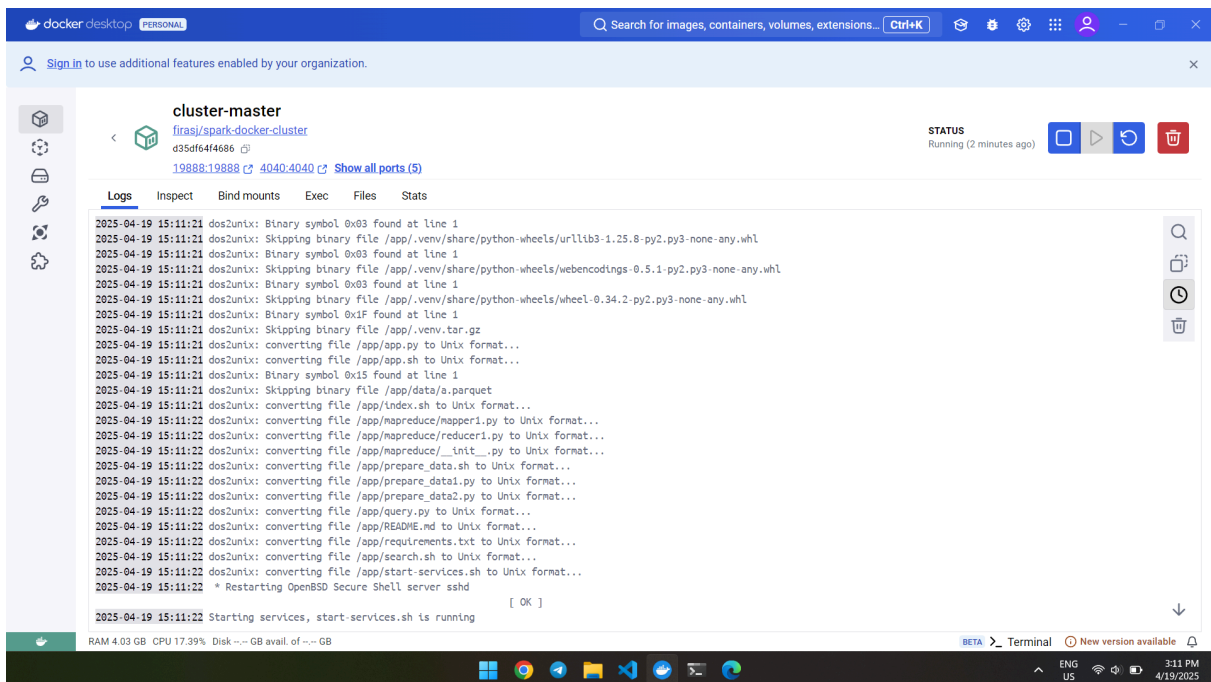
The screenshot shows the Docker Desktop application window. The left sidebar contains navigation options: Containers, Images, Volumes, Builds, Docker Scout, and Extensions. The main panel displays the 'Containers' section with a search bar and a toggle for 'Only show running containers'. Below this, a table lists four running containers:

	Name	Container ID	Image	Port(s)	CPU (%)	Memory usage...	Memory (%)	Disk read/w	Actions
<input type="checkbox"/>	bd_assignment2 -		-	-	117.31%	4.48GB / 16.93GE	120.58%	0B / 0B	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<input type="checkbox"/>	cluster-slave- 608706c914d2		firasj/spark -		0.6%	1.01GB / 7.46GB	13.59%	0B / 0B	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<input type="checkbox"/>	cluster-maste 3ca2403a2940		firasj/spark 19888:19888		65.02%	1.81GB / 7.46GB	24.26%	0B / 0B	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<input type="checkbox"/>	cassandra-se e2097406e610		cassandra	7000:7000	51.69%	1.65GB / 2GB	82.73%	0B / 0B	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

At the bottom of the window, a status bar shows 'Engine running', system resources (RAM 5.41 GB, CPU 1.63%, Disk --- GB avail. of --- GB), and a terminal window with the text 'BETA Terminal' and 'New version available'. The system clock indicates 2:58 PM on 4/19/2023.



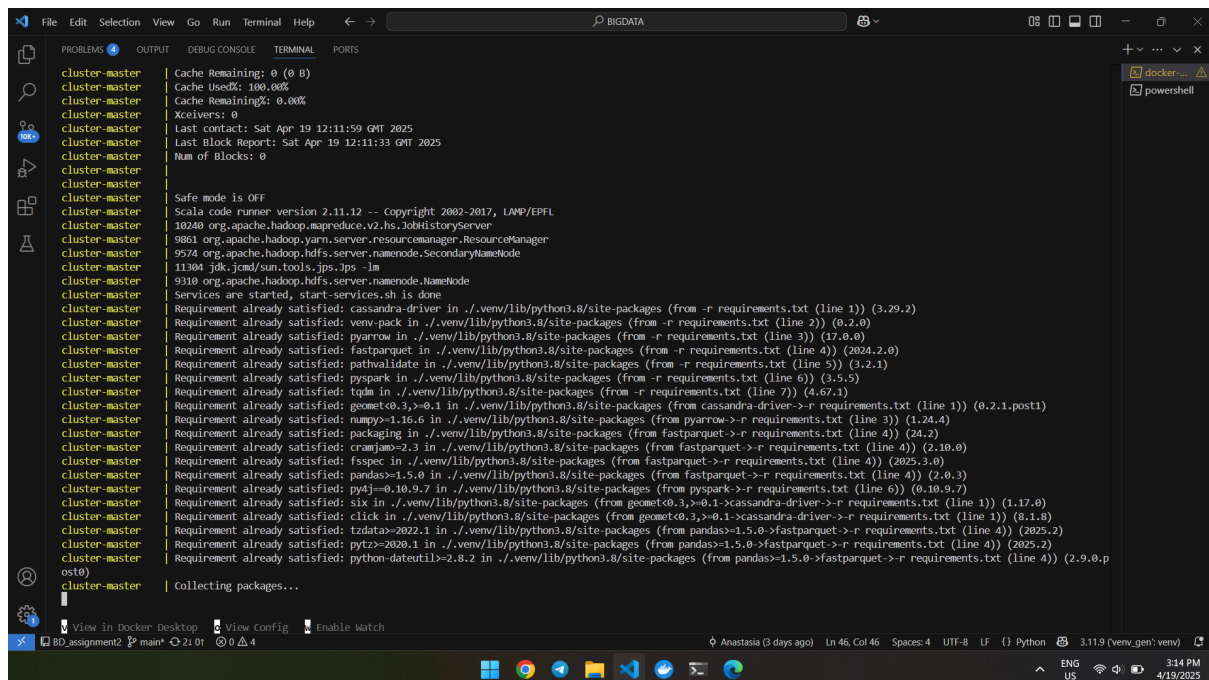
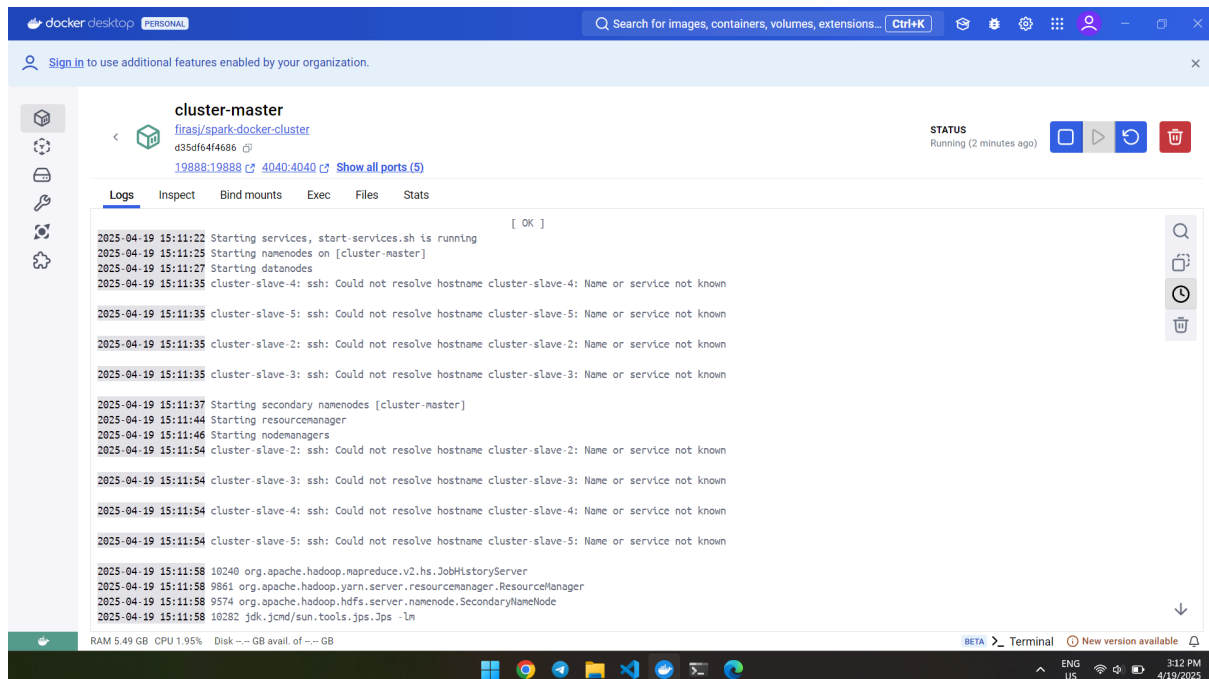
While loading you can see logs about converting files:



This is done to avoid errors in Windows-written files which have ‘\r\n’ at the ends of lines and took about 6 minutes for me. Since I already have the .venv folder, all files from .venv are also checked. However when cloning the repository, .venv doesn’t exist and will be created later. So, the conversion time will be reduced after cloning.

app.sh runs start-services.sh:

- starts Hadoop services
- sets the virtual environment and installs all packages from requirements.txt. This will take about 10 minutes.



Then `app.sh` runs `prepare_data.sh` (here you can see logs like “load parquet file”, “prepare_data1.py is running”, etc.).

```
File Edit View Go Run Terminal Help
cluster-master | prepare_data.sh has started...
cluster-master | load parquet file
cluster-master | prepare_data.py is running
cluster-master | 25/04/19 12:16:52 INFO SparkContext: Running Spark version 3.5.4
cluster-master | 25/04/19 12:16:52 INFO SparkContext: OS info Linux, 5.15.167.4-microsoft-standard-WSL2, amd64
cluster-master | 25/04/19 12:16:52 INFO SparkContext: Java version 1.8.0_442
cluster-master | 25/04/19 12:16:52 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
cluster-master | 25/04/19 12:16:53 INFO ResourceUtils:
cluster-master | 25/04/19 12:16:53 INFO ResourceUtils: No custom resources configured for spark.driver.
cluster-master | 25/04/19 12:16:53 INFO ResourceUtils:
cluster-master | 25/04/19 12:16:53 INFO SparkContext: Submitted application: Data preparation
cluster-master | 25/04/19 12:16:53 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name:
memory, amount: 4096, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
cluster-master | 25/04/19 12:16:53 INFO ResourceProfile: Limiting resource is cpu
cluster-master | 25/04/19 12:16:53 INFO ResourceProfileManager: Added ResourceProfile id: 0
cluster-master | 25/04/19 12:16:53 INFO SecurityManager: Changing view acls to: root
cluster-master | 25/04/19 12:16:53 INFO SecurityManager: Changing modify acls to: root
cluster-master | 25/04/19 12:16:53 INFO SecurityManager: Changing view acls groups to:
cluster-master | 25/04/19 12:16:53 INFO SecurityManager: Changing modify acls groups to:
cluster-master | 25/04/19 12:16:53 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
cluster-master | 25/04/19 12:16:53 INFO Utils: Successfully started service 'sparkDriver' on port 45525.
cluster-master | 25/04/19 12:16:53 INFO SparkEnv: Registering MapOutputTracker
cluster-master | 25/04/19 12:16:54 INFO SparkEnv: Registering BlockManagerMaster
cluster-master | 25/04/19 12:16:54 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
cluster-master | 25/04/19 12:16:54 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
cluster-master | 25/04/19 12:16:54 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
cluster-master | 25/04/19 12:16:54 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-fcace890-ea84-49b4-893b-849634fa59af
cluster-master | 25/04/19 12:16:54 INFO MemoryStore: MemoryStore started with capacity 366.3 MiB
cluster-master | 25/04/19 12:16:54 INFO SparkEnv: Registering OutputCommitCoordinator
cluster-master | 25/04/19 12:16:54 INFO JettyUtils: Start Jetty 9.0.0.v20140404 for SparkUI
cluster-master | 25/04/19 12:16:54 INFO Utils: Successfully started service 'SparkUI' on port 4040.
cluster-master | 25/04/19 12:16:54 INFO Executor: Starting executor ID driver on host cluster-master
cluster-master | 25/04/19 12:16:54 INFO Executor: OS info Linux, 5.15.167.4-microsoft-standard-WSL2, amd64
cluster-master | 25/04/19 12:16:54 INFO Executor: Java version 1.8.0_442
cluster-master | 25/04/19 12:16:54 INFO Executor: Starting executor with user classpath (userClasspathFirst = false): ''
cluster-master | 25/04/19 12:16:54 INFO Executor: Created or updated repl class loader org.apache.spark.util.MutableClassLoaderHolder@6e8cb0 for default.
cluster-master | 25/04/19 12:16:54 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on host cluster-master, 38833.
cluster-master | 25/04/19 12:16:54 INFO NettyBlockTransferService: Server created on cluster-master:38833
cluster-master | 25/04/19 12:16:54 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
cluster-master | 25/04/19 12:16:55 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, cluster-master, 38833, None)
```

As a result hdfs will collect all needed data ready for MapReduce pipeline in /data folder:

```
root@cluster-master:/app % docker exec -it cluster-master bash
root@cluster-master:/app# hdfs dfs -ls /data
Found 346 items
-rw-r--r-- 1 root supergroup 3284 2025-04-19 12:17 /data/10031136_A_Decade_in_the_Grave.txt
-rw-r--r-- 1 root supergroup 529 2025-04-19 12:17 /data/10078432_A_Case_for_the_Court.txt
-rw-r--r-- 1 root supergroup 616 2025-04-19 12:17 /data/10099975_A_Different_Light_(album).txt
-rw-r--r-- 1 root supergroup 647 2025-04-19 12:17 /data/10137549_A_Good_Thief_Tips_His_Hat.txt
-rw-r--r-- 1 root supergroup 591 2025-04-19 12:17 /data/10170662_A_History_of_Money_and_Banking_in_the_United_States.txt
-rw-r--r-- 1 root supergroup 1414 2025-04-19 12:17 /data/10233157_A_Balinese_Trance_Scene.txt
-rw-r--r-- 1 root supergroup 31870 2025-04-19 12:17 /data/10228777_A_Death_in_the_Family_(comics).txt
-rw-r--r-- 1 root supergroup 814 2025-04-19 12:17 /data/10230685_A_Dead_Sinking_Story.txt
-rw-r--r-- 1 root supergroup 310 2025-04-19 12:17 /data/10254892_A_Flat_Man.txt
-rw-r--r-- 1 root supergroup 9861 2025-04-19 12:17 /data/10381993_A_Doll's_House_(1973_Losey_film).txt
-rw-r--r-- 1 root supergroup 16918 2025-04-19 12:17 /data/10393111_A_Hero_of_Our_Time.txt
-rw-r--r-- 1 root supergroup 5718 2025-04-19 12:17 /data/10399316_A_Flowering_Tree.txt
-rw-r--r-- 1 root supergroup 2435 2025-04-19 12:17 /data/10534798_A_Black_and_White_World.txt
-rw-r--r-- 1 root supergroup 1180 2025-04-19 12:17 /data/10570204_A_Gun_Called_Tension.txt
-rw-r--r-- 1 root supergroup 16809 2025-04-19 12:17 /data/1067891_A_Hard_Day's_Night_(song).txt
-rw-r--r-- 1 root supergroup 1098 2025-04-19 12:17 /data/1083442_A_Hillbilly_Tribute_to_ADC.txt
-rw-r--r-- 1 root supergroup 1745 2025-04-19 12:17 /data/10849680_A_Day_in_the_Death_of_Donny_B.txt
-rw-r--r-- 1 root supergroup 6764 2025-04-19 12:17 /data/10858097_A_Dangerous_Path.txt
-rw-r--r-- 1 root supergroup 12157 2025-04-19 12:17 /data/10990703_A_Dictionary_of_Canadianisms_on_Historical_Principles.txt
-rw-r--r-- 1 root supergroup 2886 2025-04-19 12:17 /data/11017293_A_Bad_Spell_in_Yurt.txt
-rw-r--r-- 1 root supergroup 4423 2025-04-19 12:17 /data/11075899_A_Doctor's_Report_on_Dianetics.txt
-rw-r--r-- 1 root supergroup 923 2025-04-19 12:17 /data/11141641_A_Blueprint_of_the_World.txt
-rw-r--r-- 1 root supergroup 2573 2025-04-19 12:17 /data/1115810_A_Hanging.txt
-rw-r--r-- 1 root supergroup 12171 2025-04-19 12:17 /data/11211270_A_Lesson_in_Romantics.txt
-rw-r--r-- 1 root supergroup 588 2025-04-19 12:17 /data/11315857_A_Go_Go_(Potshot_album).txt
-rw-r--r-- 1 root supergroup 24342 2025-04-19 12:17 /data/1136819_A_Game_at_Chess.txt
-rw-r--r-- 1 root supergroup 333 2025-04-19 12:17 /data/11490217_A_Guide_to_Groovy_Lovin'.txt
-rw-r--r-- 1 root supergroup 5461 2025-04-19 12:17 /data/11528779_A_Dreamer's_Tales.txt
-rw-r--r-- 1 root supergroup 2529 2025-04-19 12:17 /data/11631735_A_Ballad_of_the_West.txt
-rw-r--r-- 1 root supergroup 1029 2025-04-19 12:17 /data/11753853_A_Journal_of_the_Plague_Year_(album).txt
-rw-r--r-- 1 root supergroup 597 2025-04-19 12:17 /data/11871420_A_Lifetime_or_More.txt
-rw-r--r-- 1 root supergroup 2134 2025-04-19 12:17 /data/11892274_A_Cold_Night's_Death.txt
-rw-r--r-- 1 root supergroup 863 2025-04-19 12:17 /data/11930321_A_Fragile_Hope.txt
```

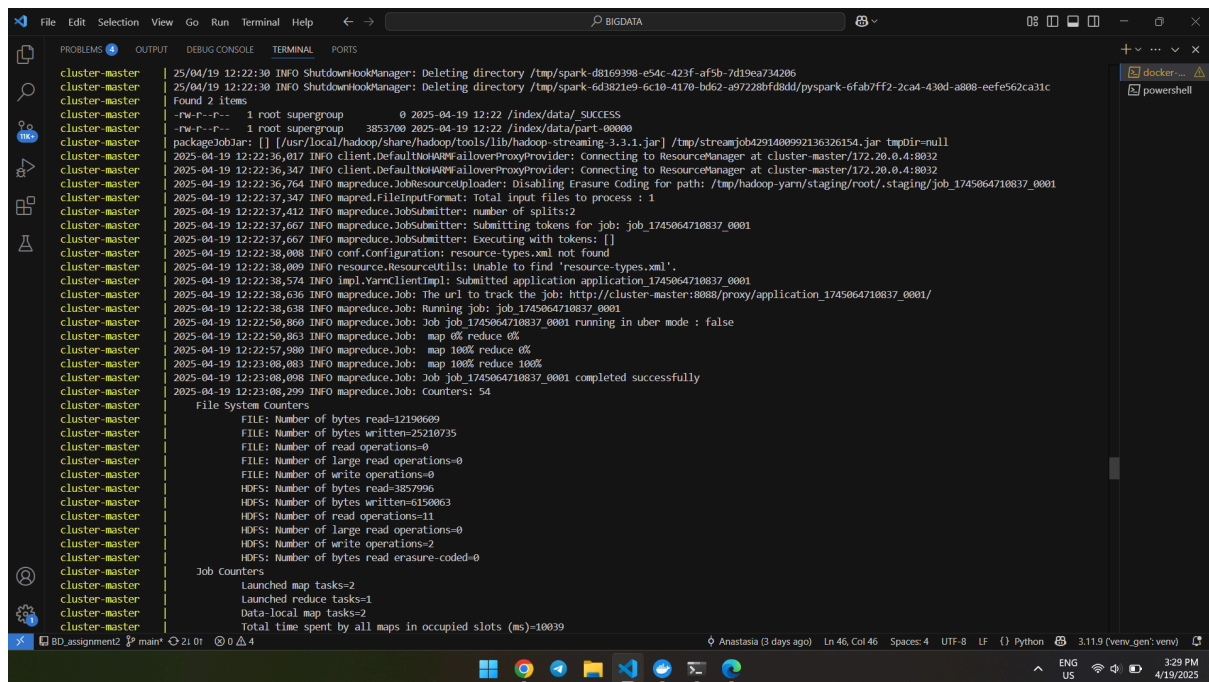
And RDD object after prepare_data2.py is saved to /index/data:

```
root@cluster-master: /app  X  Windows PowerShell  X  root@cluster-master: /app  X  +  v  X
Windows PowerShell
(C) Корпорация Майкрософт (Microsoft Corporation). Все права защищены.

Установите последнюю версию PowerShell для новых функций и улучшения! https://aka.ms/PSWindows

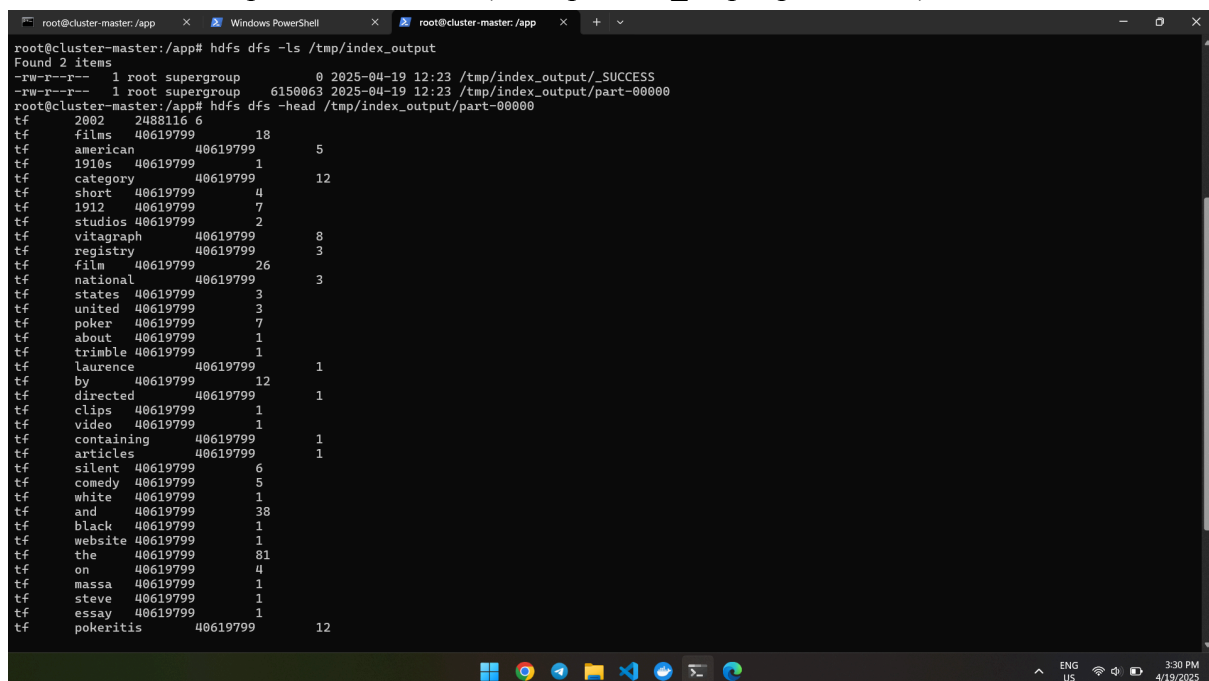
PS C:\Users\79876> docker exec -it cluster-master bash
root@cluster-master:/app# hdfs dfs -ls /index/data
Found 2 items
-rw-r--r--  1 root supergroup            0 2025-04-19 12:22 /index/data/_SUCCESS
-rw-r--r--  1 root supergroup 3853700 2025-04-19 12:22 /index/data/part-00000
root@cluster-master:/app# hdfs dfs -head /index/data/part-00000
10031136      A Decade in the Grave  A Decade In The Grave is a box set (4 CDs + 1 DVD) by death metal band Six Feet Under. It was released in 2005 on Me
tal Blade Records. Ten years after the formation of Six Feet Under, Metal Blade celebrated the band's longevity with A Decade in the Grave, a five-disc box
set (four audio CDs plus a DVD). The first two discs is SFU's best-of, disc three contains rare demos as well as live performances, while the fourth disc co
ntains demos and rehearsal material by Leviathan (the band Barnes sang for before Cannibal Corpse or SFU). The DVD offers a blend of videos and live perform
ances. The box set also includes 4 new songs, "Dead and Buried (Living Life in the Grave)", "From Flesh Bone", "A Knife Fight to the Death", and "Burned at
the Stake", the latter 3 being previously unreleased, and the former being recorded for the album. ==Track listing== ===CD 1 (Best Of Vol. 1)=== #"Feasting
On The Blood Of The Insane" #"Revenge Of The Zombies" #"Impulse To Disembowel" #"Bonesaw" #"Dead root@cluster-master:/app#
```


Next, the index.sh script is running:



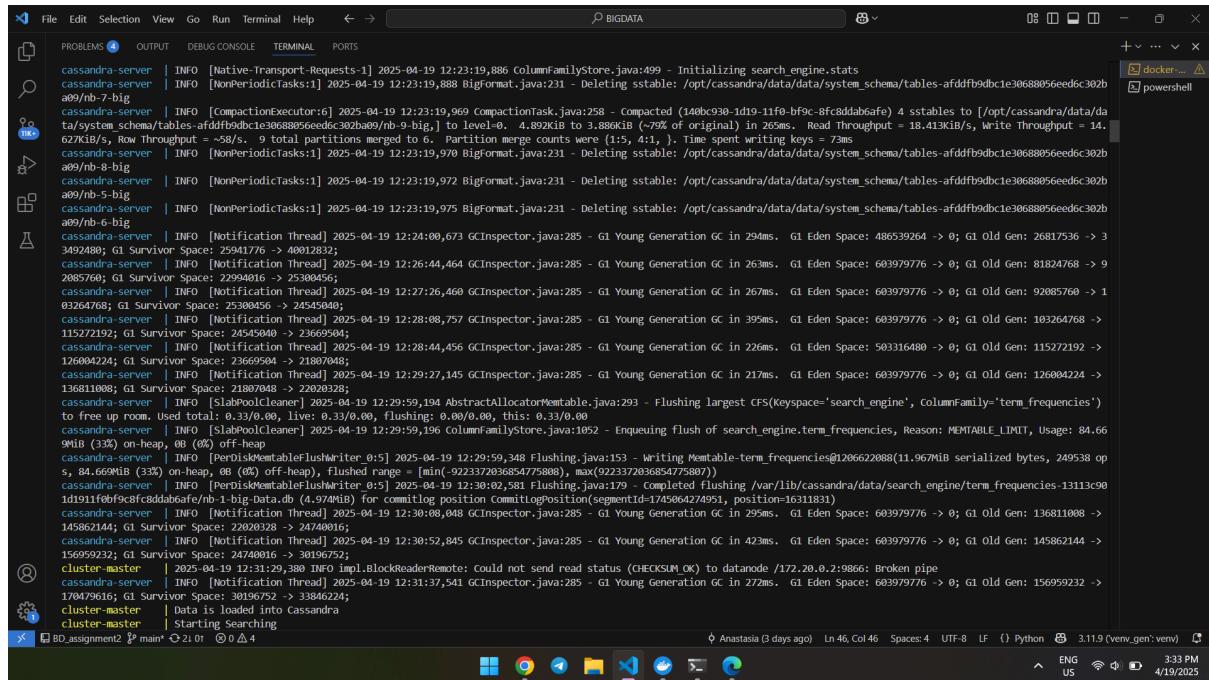
```
cluster-master 25/04/19 12:22:30 INFO ShutdownHookManager: Deleting directory /tmp/spark-d8169398-e54c-423f-af5b-7d19ea734206
cluster-master 25/04/19 12:22:30 INFO ShutdownHookManager: Deleting directory /tmp/spark-6d3821e9-6c10-4170-bd52-a97228bfd8dd/pyspark-6fab7ff2-2ca4-430d-a808-eefe562ca31c
cluster-master Found 2 items
cluster-master -rw-r--r-- 1 root supergroup 0 2025-04-19 12:22 /index/data/_SUCCESS
cluster-master -rw-r--r-- 1 root supergroup 3853700 2025-04-19 12:22 /index/data/part-00000
cluster-master packageJobJar: [ [ /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar] /tmp/streamjob4291480992136326154.jar tmpDir=null
cluster-master 2025-04-19 12:22:36,817 INFO client.DefaultHadoopFileSystemProvider: Connecting to ResourceManager at cluster-master/172.20.0.4:8032
cluster-master 2025-04-19 12:22:36,347 INFO client.DefaultHadoopFileSystemProvider: Connecting to ResourceManager at cluster-master/172.20.0.4:8032
cluster-master 2025-04-19 12:22:36,764 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1745064710837_0001
cluster-master 2025-04-19 12:22:37,347 INFO mapreduce.FileInputFormat: Total input files to process : 1
cluster-master 2025-04-19 12:22:37,412 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master 2025-04-19 12:22:37,667 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1745064710837_0001
cluster-master 2025-04-19 12:22:37,667 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master 2025-04-19 12:22:38,088 INFO conf.Configuration: resource-types.xml not found
cluster-master 2025-04-19 12:22:38,089 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master 2025-04-19 12:22:38,574 INFO impl.YarnClientImpl: Submitted application application_1745064710837_0001
cluster-master 2025-04-19 12:22:38,636 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1745064710837_0001/
cluster-master 2025-04-19 12:22:38,638 INFO mapreduce.Job: Running job: job_1745064710837_0001
cluster-master 2025-04-19 12:22:50,860 INFO mapreduce.Job: Job job_1745064710837_0001 running in uber mode : false
cluster-master 2025-04-19 12:22:50,863 INFO mapreduce.Job: map 0% reduce 0%
cluster-master 2025-04-19 12:22:57,980 INFO mapreduce.Job: map 100% reduce 0%
cluster-master 2025-04-19 12:23:08,083 INFO mapreduce.Job: map 100% reduce 100%
cluster-master 2025-04-19 12:23:08,098 INFO mapreduce.Job: Job job_1745064710837_0001 completed successfully
cluster-master 2025-04-19 12:23:08,299 INFO mapreduce.Job: Counters: 54
cluster-master File System Counters
cluster-master FILE: Number of bytes read=12190609
cluster-master FILE: Number of bytes written=25210735
cluster-master FILE: Number of read operations=0
cluster-master FILE: Number of large read operations=0
cluster-master FILE: Number of write operations=0
cluster-master HDFS: Number of bytes read=3857996
cluster-master HDFS: Number of bytes written=6150063
cluster-master HDFS: Number of read operations=11
cluster-master HDFS: Number of large read operations=0
cluster-master HDFS: Number of write operations=2
cluster-master HDFS: Number of bytes read erasure-coded=0
cluster-master Job Counters
cluster-master Launched map tasks=2
cluster-master Launched reduce tasks=1
cluster-master Data-local map tasks=2
cluster-master Total time spent by all maps in occupied slots (ms)=10039
```

And here is the output of the reducer (in /tmp/undex_output/part-00000)



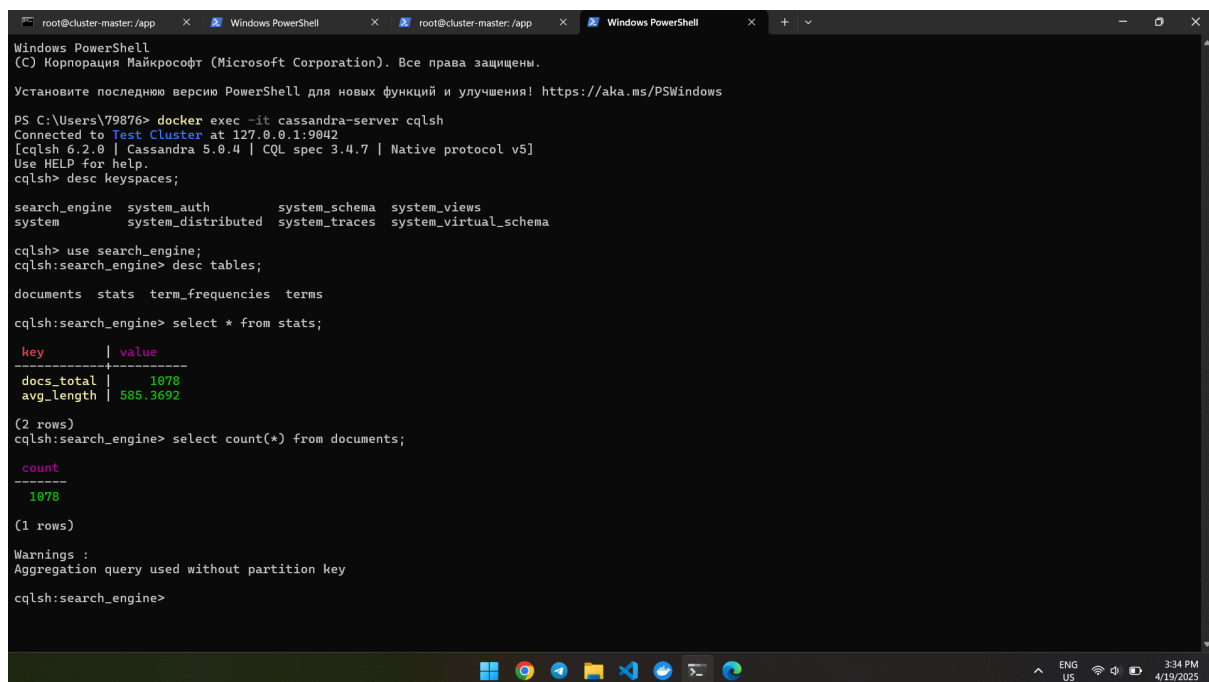
```
root@cluster-master:/app# hdfs dfs -ls /tmp/index_output
Found 2 items
-rw-r--r-- 1 root supergroup 0 2025-04-19 12:23 /tmp/index_output/_SUCCESS
-rw-r--r-- 1 root supergroup 6150063 2025-04-19 12:23 /tmp/index_output/part-00000
root@cluster-master:/app# hdfs dfs -head /tmp/index_output/part-00000
tf 2002 2488116 6
tf films 40619799 18
tf american 40619799 5
tf 1910s 40619799 1
tf category 40619799 12
tf short 40619799 4
tf 1912 40619799 7
tf studios 40619799 2
tf vitagraph 40619799 8
tf registry 40619799 3
tf film 40619799 26
tf national 40619799 3
tf states 40619799 3
tf united 40619799 3
tf poker 40619799 7
tf about 40619799 1
tf trimble 40619799 1
tf laurence 40619799 1
tf by 40619799 12
tf directed 40619799 1
tf clips 40619799 1
tf video 40619799 1
tf containing 40619799 1
tf articles 40619799 1
tf silent 40619799 6
tf comedy 40619799 5
tf white 40619799 1
tf and 40619799 38
tf black 40619799 1
tf website 40619799 1
tf the 40619799 81
tf on 40619799 4
tf massa 40619799 1
tf steve 40619799 1
tf essay 40619799 1
tf pokeritis 40619799 12
```

After the MapReduce pipeline app.py is running. It creates the keyspace, the tables, and inserts data (in the end we see “Data is loaded into Cassandra”):



```
File Edit Selection View Go Run Terminal Help
BIGDATA
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
cassandra-server INFO [Native-Transport-Requests-1] 2025-04-19 12:23:19,886 ColumnFamilyStore.java:499 - Initializing search_engine.stats
cassandra-server INFO [NonPeriodicTasks:1] 2025-04-19 12:23:19,888 BigFormat.java:231 - Deleting sstable: /opt/cassandra/data/data/system_schema/tables-afddfb9dbc1e30688056eeddc302b
a09/nb-7-big
cassandra-server INFO [CompactionExecutor:6] 2025-04-19 12:23:19,969 CompactionTask.java:258 - Compacted (140b:c30-1d19-11f0-bf9c-8fc8ddab6afe) 4 sstables to [/opt/cassandra/data/data/system_schema/tables-afddfb9dbc1e30688056eeddc302ba09/nb-9-big.] to level=0. 4.092KiB to 3.896KiB (~79% of original) in 265ms. Read Throughput = 18.413KiB/s, Write Throughput = 14.627KiB/s, Row Throughput = ~58/s. 9 total partitions merged to 6. Partition merge counts were {1:5, 4:1, }. Time spent writing keys = 73ms
cassandra-server INFO [NonPeriodicTasks:1] 2025-04-19 12:23:19,970 BigFormat.java:231 - Deleting sstable: /opt/cassandra/data/data/system_schema/tables-afddfb9dbc1e30688056eeddc302b
a09/nb-8-big
cassandra-server INFO [NonPeriodicTasks:1] 2025-04-19 12:23:19,972 BigFormat.java:231 - Deleting sstable: /opt/cassandra/data/data/system_schema/tables-afddfb9dbc1e30688056eeddc302b
a09/nb-5-big
cassandra-server INFO [NonPeriodicTasks:1] 2025-04-19 12:23:19,975 BigFormat.java:231 - Deleting sstable: /opt/cassandra/data/data/system_schema/tables-afddfb9dbc1e30688056eeddc302b
a09/nb-6-big
cassandra-server INFO [Notification Thread] 2025-04-19 12:24:00,673 GCInspector.java:285 - G1 Young Generation GC in 294ms. G1 Eden Space: 486539264 -> 0; G1 Old Gen: 26817536 -> 3
3492480; G1 Survivor Space: 25941776 -> 40012832;
cassandra-server INFO [Notification Thread] 2025-04-19 12:26:44,464 GCInspector.java:285 - G1 Young Generation GC in 263ms. G1 Eden Space: 603979776 -> 0; G1 Old Gen: 81824768 -> 9
2085760; G1 Survivor Space: 22994016 -> 25300456;
cassandra-server INFO [Notification Thread] 2025-04-19 12:27:26,460 GCInspector.java:285 - G1 Young Generation GC in 267ms. G1 Eden Space: 603979776 -> 0; G1 Old Gen: 92085760 -> 1
03264768; G1 Survivor Space: 25300456 -> 24545040;
cassandra-server INFO [Notification Thread] 2025-04-19 12:28:00,757 GCInspector.java:285 - G1 Young Generation GC in 395ms. G1 Eden Space: 603979776 -> 0; G1 Old Gen: 103264768 ->
115272192; G1 Survivor Space: 24545040 -> 23669904;
cassandra-server INFO [Notification Thread] 2025-04-19 12:28:44,456 GCInspector.java:285 - G1 Young Generation GC in 226ms. G1 Eden Space: 503316480 -> 0; G1 Old Gen: 115272192 ->
126004224; G1 Survivor Space: 23669904 -> 21807048;
cassandra-server INFO [Notification Thread] 2025-04-19 12:29:17,145 GCInspector.java:285 - G1 Young Generation GC in 217ms. G1 Eden Space: 603979776 -> 0; G1 Old Gen: 126004224 ->
136811008; G1 Survivor Space: 21807048 -> 22020328;
cassandra-server INFO [SlabPoolCleaner] 2025-04-19 12:29:59,194 AbstractAllocatorMemtable.java:293 - Flushing largest CFS(keyspace='search_engine', columnFamily='term_frequencies')
to free up room. Used total: 0.33/0.00, Live: 0.33/0.00, flushing: 0.00/0.00, this: 0.33/0.00
cassandra-server INFO [SlabPoolCleaner] 2025-04-19 12:29:59,196 ColumnFamilyStore.java:1052 - Enqueuing flush of search_engine.term_frequencies, Reason: MEMTABLE_LIMIT, Usage: 84.66
9MiB (33%) on-heap, 00 (0%) off-heap
cassandra-server INFO [PerDiskMemtableFlushWriter 0:5] 2025-04-19 12:29:59,348 Flushing.java:153 - Writing Memtable-term_frequencies@1206622088(11,967MiB serialized bytes, 249538 op
s, 84.669MiB (33%) on-heap, 00 (0%) off-heap), flushed range = [min(-9223372036854775808), max(9223372036854775807)]
cassandra-server INFO [PerDiskMemtableFlushWriter 0:5] 2025-04-19 12:30:02,581 Flushing.java:179 - Completed flushing /var/lib/cassandra/data/search_engine/term_frequencies-13113c90
1d1911f0b9c8fc8ddab6afe/nb-1-big-data.db (4.974MiB) for commitlog position commitlogposition(segmentId=1745064274951, position=16311831)
cassandra-server INFO [Notification Thread] 2025-04-19 12:30:08,048 GCInspector.java:285 - G1 Young Generation GC in 295ms. G1 Eden Space: 603979776 -> 0; G1 Old Gen: 136811008 ->
145862144; G1 Survivor Space: 22020328 -> 24740016;
cassandra-server INFO [Notification Thread] 2025-04-19 12:30:52,845 GCInspector.java:285 - G1 Young Generation GC in 423ms. G1 Eden Space: 603979776 -> 0; G1 Old Gen: 145862144 ->
156959232; G1 Survivor Space: 24740016 -> 30196752;
cluster-master INFO 2025-04-19 12:31:29,380 INFO impl.BlockReaderRemote: Could not send read status (CHECKSUM OK) to datanode /172.20.0.2:9866: Broken pipe
cassandra-server INFO [Notification Thread] 2025-04-19 12:31:37,541 GCInspector.java:285 - G1 Young Generation GC in 272ms. G1 Eden Space: 603979776 -> 0; G1 Old Gen: 156959232 ->
170479616; G1 Survivor Space: 30196752 -> 33846224;
cluster-master INFO Data is loaded into Cassandra
cluster-master INFO Starting Searching
```

And the Cassandra tables check:



```
root@cluster-master:/app x Windows PowerShell x root@cluster-master:/app x Windows PowerShell x
Windows PowerShell
(C) Корпорация Майкрософт (Microsoft Corporation). Все права защищены.

Установите последнюю версию PowerShell для новых функций и улучшения! https://aka.ms/PSWindows

PS C:\Users\79876> docker exec -it cassandra-server cqlsh
Connected to test Cluster at 127.0.0.1:9042
[cqlsh 6.2.0 | Cassandra 5.0.4 | CQL spec 3.4.7 | Native protocol v5]
Use HELP for help.
cqlsh> desc keyspaces;

search_engine system_auth system_schema system_views
system system_distributed system_traces system_virtual_schema

cqlsh> use search_engine;
cqlsh:search_engine> desc tables;

documents stats term_frequencies terms

cqlsh:search_engine> select * from stats;

key | value
-----|-----
docs_total | 1078
avg_length | 585.3692

(2 rows)
cqlsh:search_engine> select count(*) from documents;

count
-----
1078

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:search_engine>
```

We don't see 1100 documents in total (as written in data_preparation1.py: n=1100) because some documents have a problem with their titles: characters are not ASCII. As we are required to use at least 1000 documents I decided to ignore documents with not relevant titles and load 1100 (1078 in real life) documents.

Finally, the search.sh is running. According to the Assignment requirements, I ran this script 3 times for different queries: “my query” to see the ability to find anything, “asian family page memior kirkus krueger bao phi”, and “Tim Chamberlain, a doctoral candidate at BirkBeck, University of London wrote in his review for the London School”. The last 2 are from particular Wikipedia pages to see if the search engine found them (yes).

```

cluster-master 25/04/19 12:36:01 INFO YarnScheduler: Adding task set 13.0 with 1 tasks resource profile 0
cluster-master 25/04/19 12:36:01 INFO TaskSetManager: Starting task 0.0 in stage 13.0 (TID 126) (cluster-slave-1, executor 1, partition 0, NODE_LOCAL, 8828 bytes)
cluster-master 25/04/19 12:36:01 INFO BlockManagerInfo: Added broadcast 7 piece0 in memory on cluster-slave-1:42380 (size: 6.5 KiB, free: 366.3 MiB)
cluster-master 25/04/19 12:36:01 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 3 to 172.20.0.2:48678
cluster-master 25/04/19 12:36:01 INFO TaskSetManager: Finished task 0.0 in stage 13.0 (TID 126) in 83 ms on cluster-slave-1 (executor 1) (1/1)
cluster-master 25/04/19 12:36:01 INFO YarnScheduler: Removed TaskSet 13.0, whose tasks have all completed, from pool
cluster-master 25/04/19 12:36:01 INFO DAGScheduler: ResultStage 13 (runJob at PythonRDD.scala:181) finished in 0.106 s
cluster-master 25/04/19 12:36:01 INFO DAGScheduler: Job 3 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master 25/04/19 12:36:01 INFO YarnScheduler: Killing all running tasks in stage 13: Stage finished
cluster-master 25/04/19 12:36:01 INFO DAGScheduler: Job 3 finished: runJob at PythonRDD.scala:181, took 1.500110 s
cluster-master Document ID: 72327259, title: "A Historical Atlas of Tibet", score: 45.39
cluster-master Document ID: 41252900, title: "A Far Cry", score: 17.23
cluster-master Document ID: 18237033, title: "A Beautiful Child", score: 15.29
cluster-master Document ID: 73670, title: "A Christmas Carol", score: 13.84
cluster-master Document ID: 69979031, title: "A History of Science, Technology, and Philosophy in the 16th and 17thCenturies", score: 13.22
cluster-master Document ID: 56880098, title: "A J Balliol Salmon", score: 13.07
cluster-master Document ID: 48385917, title: "A J M Nasir Uddin", score: 12.30
cluster-master Document ID: 24772931, title: "A Culture of Conspiracy", score: 12.26
cluster-master Document ID: 60207932, title: "A History of Modern Yoga", score: 11.98
cluster-master Document ID: 4186594, title: "A History of the Early Part of the Reign of James II", score: 11.51
cluster-master 25/04/19 12:36:02 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master 25/04/19 12:36:02 INFO SparkUI: Stopped spark web UI at http://cluster-master:4040
cluster-master 25/04/19 12:36:02 INFO YarnClientSchedulerBackend: Interrupting monitor thread
cluster-master 25/04/19 12:36:02 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster-master 25/04/19 12:36:02 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
cluster-master 25/04/19 12:36:02 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
cluster-master 25/04/19 12:36:02 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master 25/04/19 12:36:02 INFO MemoryStore: MemoryStore cleared
cluster-master 25/04/19 12:36:02 INFO BlockManager: BlockManager stopped
cluster-master 25/04/19 12:36:02 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master 25/04/19 12:36:02 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master 25/04/19 12:36:02 INFO SparkContext: Successfully stopped SparkContext
cluster-master 25/04/19 12:36:03 INFO ShutdownHookManager: Shutdown hook called
cluster-master 25/04/19 12:36:03 INFO ShutdownHookManager: Deleting directory /tmp/spark-a531f22f-dc4b-4c06-b5af-fee35479a8db
cluster-master 25/04/19 12:36:03 INFO ShutdownHookManager: Deleting directory /tmp/spark-0e0d7cda-95e3-4c70-a5ed-f4d8ca16dfba/pyspark-88183905-9d41-4719-89f0-fa050c1eab24
cluster-master 25/04/19 12:36:03 INFO ShutdownHookManager: Deleting directory /tmp/spark-0e0d7cda-95e3-4c70-a5ed-f4d8ca16dfba
cluster-master 25/04/19 12:36:03 INFO CassandraConnector: Disconnected from Cassandra cluster.
cluster-master 25/04/19 12:36:03 INFO SerialShutdownHooks: Successfully executed shutdown hook: Clearing session cache for C* connector

```

The lists of found files are appended to the output.txt:

```

1 Your query after the processing: {'my', 'query'}
2 The list of found papers:
3 Document ID: 47515595, title: "A Canine Sherlock Holmes", score: 8.94.
4 Document ID: 1356924, title: "A Child's Garden of Verses", score: 4.25.
5 Document ID: 48508834, title: "A Fist Within Four Walls", score: 4.12.
6 Document ID: 55178618, title: "A Boogie wit da Hoodie discography", score: 4.03.
7 Document ID: 47595311, title: "A Copy of My Mind", score: 3.98.
8 Document ID: 39333458, title: "A Bucketful of Soul", score: 3.87.
9 Document ID: 9919932, title: "A Family Affair (musical)", score: 3.71.
10 Document ID: 33456030, title: "A Journey Through Time", score: 3.65.
11 Document ID: 1030311, title: "A Hero of Our Time", score: 3.62.
12 Document ID: 47948097, title: "A Far Cry from You", score: 3.61.
13
14
15 Your query after the processing: {'family', 'page', 'krueger', 'kirkus', 'phi', 'bao', 'memior', 'asian'}
16 The list of found papers:
17 Document ID: 58850847, title: "A Different Pond", score: 40.86.
18 Document ID: 929153, title: "A Bao A Qu (album)", score: 9.49.
19 Document ID: 30932963, title: "A History of Venice", score: 8.89.
20 Document ID: 11315857, title: "A Go Go (Potshot album)", score: 8.16.
21 Document ID: 31387700, title: "A History of Marriage", score: 8.10.
22 Document ID: 62701269, title: "A Feast of Snakes", score: 7.44.
23 Document ID: 5870694, title: "A Gathering of Days", score: 6.62.
24 Document ID: 62485656, title: "A Calf for Christmas", score: 6.44.
25 Document ID: 35739004, title: "A Dog's Purpose", score: 6.29.
26 Document ID: 45387446, title: "A Chinese Life", score: 6.12.
27
28
29 Your query after the processing: {'doctoral', 'tim', 'at', 'school', 'chamberlain', 'candidate', 'university', 'his', 'review', 'a', 'wrote', 'for', 'in', 'the', 'of', 'l'}
30 The list of found papers:
31 Document ID: 72327259, title: "A Historical Atlas of Tibet", score: 45.39.
32 Document ID: 41252900, title: "A Far Cry", score: 17.23.
33 Document ID: 18237033, title: "A Beautiful Child", score: 15.29.
34 Document ID: 73670, title: "A Christmas Carol", score: 13.84.
35 Document ID: 69979031, title: "A History of Science, Technology, and Philosophy in the 16th and 17thCenturies", score: 13.22.
36 Document ID: 56880098, title: "A J Balliol Salmon", score: 13.07.
37 Document ID: 48385917, title: "A J M Nasir Uddin", score: 12.30.

```

To run your own query you can run the following in a new terminal:

```
docker exec -it cluster-master bash
./search.sh "your own query"
```

To avoid cluster-master stops I added “tail -f /dev/null” at the end of the app.sh. It allows connection to the container from other terminals even after the app.sh is done.

Some web UIs:

hadoop

Cluster

About Nodes Node Labels Applications NEW SAVING SUBMITTED ACCEPTED RUNNING FINISHED FAILED KILLED Scheduler Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources
4	0	0	4	0	<memory 0 B, vCores 0>	<memory 8 GB, vCores 8>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes
1	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory 1024, vCores 1>	<memory 8192, vCores 4>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB
application_1745064710837_0004	root	TFIDF search	SPARK		default	0	Sat Apr 19 15:35:01 +0300 2025	Sat Apr 19 15:35:02 +0300 2025	Sat Apr 19 15:36:02 +0300 2025	FINISHED	SUCCEEDED	N/A	N/A	N/A
application_1745064710837_0003	root	TFIDF search	SPARK		default	0	Sat Apr 19 15:33:45 +0300 2025	Sat Apr 19 15:33:46 +0300 2025	Sat Apr 19 15:34:45 +0300 2025	FINISHED	SUCCEEDED	N/A	N/A	N/A
application_1745064710837_0002	root	TFIDF search	SPARK		default	0	Sat Apr 19 15:32:12 +0300 2025	Sat Apr 19 15:32:13 +0300 2025	Sat Apr 19 15:33:26 +0300 2025	FINISHED	SUCCEEDED	N/A	N/A	N/A
application_1745064710837_0001	root	streamjob4291400992136326154.jar	MAPREDUCE		default	0	Sat Apr 19 15:22:38 +0300 2025	Sat Apr 19 15:22:40 +0300 2025	Sat Apr 19 15:23:06 +0300 2025	FINISHED	SUCCEEDED	N/A	N/A	N/A

Showing 1 to 4 of 4 entries

JobHistory

Retired Jobs

Show 20 entries

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed	Elapsed Time
2025-04-19 12:22:38 GMT	2025-04-19 12:22:49 GMT	2025-04-19 12:23:06 GMT	job_1745064710837_0001	streamjob4291400992136326154.jar	root	default	SUCCEEDED	2	2	1	1	00hrs, 00mins, 17sec

Showing 1 to 1 of 1 entries

hadoop

JobHistory

Retired Jobs

Show 20 entries

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed	Elapsed Time
2025-04-19 12:22:38 GMT	2025-04-19 12:22:49 GMT	2025-04-19 12:23:06 GMT	job_1745064710837_0001	streamjob4291400992136326154.jar	root	default	SUCCEEDED	2	2	1	1	00hrs, 00mins, 17sec

Showing 1 to 1 of 1 entries

The screenshot shows the Hadoop NameNode Overview page for the cluster 'cluster-master:9000' (active). The page is accessed via a web browser at localhost:9870/dfshealth.html#tab-overview. The browser's address bar and tabs are visible at the top. The page has a green header with navigation links: Hadoop, Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. The main content area is divided into two sections: Overview and Summary.

Overview 'cluster-master:9000' (✓active)

Started:	Sat Apr 19 15:11:29 +0300 2025
Version:	3.3.1, ra365c37a397ad4188041dd80621bdeefc468852
Compiled:	Tue Jun 15 08:12:00 +0300 2021 by ubuntu from (HEAD detached at release-3.3.1-RC3)
Cluster ID:	CID-63e395b-a2a1-46ea-b7cb-699c4d3062a
Block Pool ID:	BP-1836921958-127.0.0.1-1741262854324

Summary

Security is off.
 Safemode is off.
 1361 files and directories, 1341 blocks (1341 replicated blocks, 0 erasure coded block groups) = 2702 total filesystem object(s).
 Heap Memory used 173.12 MB of 345 MB Heap Memory. Max Heap Memory is 1.66 GB.
 Non Heap Memory used 67.48 MB of 69 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	1006.85 GB
Configured Remote Capacity:	0 B
DFS Used:	1.16 GB (0.12%)
Non DFS Used:	6.96 GB
DFS Remaining:	947.52 GB (94.11%)
Block Pool Used:	1.16 GB (0.12%)
DataNodes usages% (Min/Median/Max/stdDev):	0.12% / 0.12% / 0.12% / 0.00%

1 DataNodes

The bottom of the screenshot shows the Windows taskbar with various application icons and the system clock indicating 3:48 PM on 4/19/2025.

In total the application run took about 30 minutes (with searching for 3 queries but without packages installation). The most time consuming: Dos2unix conversion (6 minutes for me), packages installation (about 10 minutes), data loading to HDFS (5 minutes), data loading to Cassandra (8 minutes). Query search takes about 1.5 minutes.