

# Assignment 2: Simple Search Engine using Hadoop MapReduce

Elena Tesmeeva; [e.tesmeeva@innopolis.university](mailto:e.tesmeeva@innopolis.university); Group: AI-01

## 1. Methodology

### A. Data collection and preparation

For prepare\_dara.py I used the script provided in the assignment description. No modifications were made.

### B. Cassandra database creation

To store indexing statistics required for search, a Cassandra database was created using the following scheme.

This was achieved using a Python script app.py, which connects to the Cassandra cluster and creates four dedicated tables in the search key `space\_engine`.

- Table term\_docs - how often a term appears in each document.

```
# Table 1: term_docs stores tf(t, d)
session.execute("""
    CREATE TABLE IF NOT EXISTS term_docs (
        term text,
        doc_id text,
        tf int,
        PRIMARY KEY (term, doc_id)
    )
""")
```

- **term**: the word/token
- **doc\_id**: the identifier of the document
- **tf**: the frequency of the term in the document

- Table term\_stats - how many documents contain the given term.

```
# Table 2: term_stats stores df(t)
session.execute("""
    CREATE TABLE IF NOT EXISTS term_stats (
        term text PRIMARY KEY,
        df int
    )
""")
```

- **term**: the word/token

- **df**: number of distinct documents that contain the term
- Table `doc_stats` - the total number of terms in each document.

```
# Table 3: doc_stats stores dl(d)
session.execute("""
    CREATE TABLE IF NOT EXISTS doc_stats (
        doc_id text PRIMARY KEY,
        dl int
    )
""")
```

- **doc\_id**: the document identifier
- **dl**: length of the document in terms (number of tokens)
- Table `corpus_stats` - holds global statistics needed for BM25 scoring:

```
# Table 4: corpus_stats stores global stats: N and dlavg
session.execute("""
    CREATE TABLE IF NOT EXISTS corpus_stats (
        id text PRIMARY KEY,
        total_docs int,
        avg_dl float
    )
""")
```

- **total\_docs (N)**: the total number of documents in the collection
- **avg\_dl**: the average document length across all documents

### C. MapReduce pipelines

Before describing the MapReduce pipelines, I would like to mention that I encountered issues with the `cassandra-driver` when running Python scripts in the Hadoop environment. To resolve this, I packaged the required Cassandra libraries using `zipimport` and included them in the job through the Hadoop streaming JAR using a zipped library file.

#### Pipeline 1: Term Frequency

In this pipeline, it was necessary to sort the output by two keys. The mapper output was sorted first by `doc_id`, and then by the token itself - word. This was achieved using the following Hadoop streaming configuration:

```
-D mapreduce.partition.keypartitioner.options=-k1,1 \
-D mapreduce.partition.keycomparator.options='-k1,1 -k2,2' \
```

- **Mapper:** It reads each line of input, expecting tab-separated values: doc\_id, title, and text. It splits the text into individual words and emits each word along with its document ID and a count of 1 in the format <<doc\_id> <word> 1>
- **Reducer:** It groups and sums up counts for each (doc\_id, term) combination to compute the term frequency. The final results are written to the term\_docs table in Cassandra with the fields term, doc\_id, and tf.

## Pipeline 2: Document Frequency

- **Mapper:** As the first stage of the pipeline, the mapper processes each line of input in the format: doc\_id, title, and text. It converts the text to lowercase, splits it into tokens, removes duplicates (using a set), and emits each unique token along with the document ID:<<term> <doc\_id>>
- **Reducer:** It processes sorted (term, doc\_id) pairs. For each unique term, it aggregates the number of unique documents it appears in. This document frequency value is stored in the term\_stats table in Cassandra with fields term and df.

## Pipeline 3: Document Statistics

- **Mapper:** It processes the same doc\_id, title, text format. It calculates the document length (dl) by counting the number of words in the text, then emits: <<doc\_id> <dl>>
- **Reducer:** It gathers all document lengths, saves each (doc\_id, dl) into the doc\_stats table, and also computes two global statistics:
  - The total number of documents (N)
  - The average document length (dlavg)

These are written to the corpus\_stats table in Cassandra.

## 2. Demonstration

How to run code:

1) git clone <https://github.com/elenatesm/bigdataaass2>

2) Download file a.parquet from <https://www.kaggle.com/datasets/jjinho/wikipedia-20230701?select=a.parquet> to the app folder of the repository

3) Run docker-compose up



- Putting data to hdfs

```

elenatesm@HOME-PC: /mnt/c/3year/BD/Ass2/bigdataass2
cluster-master | Putting data to hdfs
cluster-master | Found 978 items
cluster-master | -rw-r--r-- 1 root supergroup 3284 2025-04-14 09:28 /data/10031136_A_Decade_in_the_Grave.txt
cluster-master | -rw-r--r-- 1 root supergroup 529 2025-04-14 09:28 /data/10078432_A_Case_for_the_Court.txt
cluster-master | -rw-r--r-- 1 root supergroup 616 2025-04-14 09:28 /data/10099975_A_Different_Light_album.txt
cluster-master | -rw-r--r-- 1 root supergroup 647 2025-04-14 09:28 /data/10137549_A_Good_Thief_Tips_His_Hat.txt
cluster-master | -rw-r--r-- 1 root supergroup 591 2025-04-14 09:28 /data/10174562_A_History_of_Money_and_Banking_in_the_United_States.txt
cluster-master | -rw-r--r-- 1 root supergroup 1414 2025-04-14 09:28 /data/10223157_A_Balinese_Trance_Seance.txt
cluster-master | -rw-r--r-- 1 root supergroup 31874 2025-04-14 09:28 /data/10228777_A_Death_in_the_Family_comics.txt
cluster-master | -rw-r--r-- 1 root supergroup 814 2025-04-14 09:28 /data/10230685_A_Dead_Sinking_Story.txt
cluster-master | -rw-r--r-- 1 root supergroup 310 2025-04-14 09:28 /data/10254892_A_Flat_Man.txt
cluster-master | -rw-r--r-- 1 root supergroup 8861 2025-04-14 09:28 /data/10381991_A_Doll's_House_1973_Losey_film.txt
cluster-master | -rw-r--r-- 1 root supergroup 16918 2025-04-14 09:28 /data/10393111_A_Hero_of_Our_Time.txt
cluster-master | -rw-r--r-- 1 root supergroup 5718 2025-04-14 09:28 /data/10399316_A_Flowering_Tree.txt
cluster-master | -rw-r--r-- 1 root supergroup 2435 2025-04-14 09:28 /data/10534798_A_Black_and_White_World.txt
cluster-master | -rw-r--r-- 1 root supergroup 1180 2025-04-14 09:28 /data/10570204_A_Gun_Called_Tension.txt
cluster-master | -rw-r--r-- 1 root supergroup 16809 2025-04-14 09:28 /data/1067891_A_Hard_Day's_Night_song.txt
cluster-master | -rw-r--r-- 1 root supergroup 1098 2025-04-14 09:28 /data/1083442_A_Hillbilly_Tribute_to_ACDC.txt
cluster-master | -rw-r--r-- 1 root supergroup 1745 2025-04-14 09:28 /data/10849680_A_Day_in_the_Death_of_Donny_B.txt
cluster-master | -rw-r--r-- 1 root supergroup 6764 2025-04-14 09:28 /data/10858097_A_Dangerous_Path.txt
cluster-master | -rw-r--r-- 1 root supergroup 12157 2025-04-14 09:28 /data/10900703_A_Dictionary_of_Canadianisms_on_Historical_Principles.txt
cluster-master | -rw-r--r-- 1 root supergroup 2806 2025-04-14 09:28 /data/11017293_A_Bad_Spell_in_Yurt.txt
cluster-master | -rw-r--r-- 1 root supergroup 4423 2025-04-14 09:28 /data/11017589_A_Doctor's_Report_on_Dianetics.txt
cluster-master | -rw-r--r-- 1 root supergroup 923 2025-04-14 09:28 /data/11141641_A_Blueprint_of_the_World.txt
cluster-master | -rw-r--r-- 1 root supergroup 2573 2025-04-14 09:28 /data/1115810_A_Hanging.txt
cluster-master | -rw-r--r-- 1 root supergroup 12171 2025-04-14 09:28 /data/11211270_A_Lesson_in_Romantics.txt
cluster-master | -rw-r--r-- 1 root supergroup 588 2025-04-14 09:28 /data/11315857_A_Go_Go_Potshot_album.txt
cluster-master | -rw-r--r-- 1 root supergroup 333 2025-04-14 09:28 /data/11490217_A_Guide_to_Groovy_Lovin'.txt
cluster-master | -rw-r--r-- 1 root supergroup 5461 2025-04-14 09:28 /data/11528779_A_Dreamer's_Tales.txt
cluster-master | -rw-r--r-- 1 root supergroup 2529 2025-04-14 09:28 /data/11631735_A_Ballad_of_the_West.txt
cluster-master | -rw-r--r-- 1 root supergroup 1029 2025-04-14 09:28 /data/11753053_A_Journal_of_the_Plague_Year_album.txt
cluster-master | -rw-r--r-- 1 root supergroup 597 2025-04-14 09:28 /data/11871420_A_Lifetime_or_More.txt
cluster-master | -rw-r--r-- 1 root supergroup 2134 2025-04-14 09:28 /data/11892274_A_Cold_Night's_Death.txt
cluster-master | -rw-r--r-- 1 root supergroup 863 2025-04-14 09:28 /data/11930321_A_Fragile_Hope.txt
cluster-master | -rw-r--r-- 1 root supergroup 6843 2025-04-14 09:28 /data/11984610_A_Catalogue_of_Crime.txt
cluster-master | -rw-r--r-- 1 root supergroup 7441 2025-04-14 09:28 /data/12000397_A_King_and_No_King.txt
cluster-master | -rw-r--r-- 1 root supergroup 1698 2025-04-14 09:28 /data/12132506_A_Crystal_Christmas.txt
cluster-master | -rw-r--r-- 1 root supergroup 7681 2025-04-14 09:28 /data/12212389_A_Flintstones_Christmas_Carol.txt
cluster-master | -rw-r--r-- 1 root supergroup 758 2025-04-14 09:28 /data/1240312_A_Giant_Allen_Force_More_Violent_&_Sick_Than_Anything_You_Can_Imagine.txt
cluster-master | -rw-r--r-- 1 root supergroup 364 2025-04-14 09:28 /data/12459639_A_Day_at_the_Races_video.txt
cluster-master | -rw-r--r-- 1 root supergroup 6324 2025-04-14 09:28 /data/12621170_A_Fool_in_Love.txt
cluster-master | -rw-r--r-- 1 root supergroup 10068 2025-04-14 09:28 /data/12660064_A_Bird_in_the_House.txt
cluster-master | -rw-r--r-- 1 root supergroup 1858 2025-04-14 09:28 /data/12712771_A_Gentleman_of_Paris_1927_film.txt
cluster-master | -rw-r--r-- 1 root supergroup 4329 2025-04-14 09:28 /data/12719760_A_Girl_Three_Guys_and_a_Gun.txt
cluster-master | -rw-r--r-- 1 root supergroup 2482 2025-04-14 09:28 /data/12806692_A_Bolha.txt
cluster-master | -rw-r--r-- 1 root supergroup 13051 2025-04-14 09:28 /data/12937367_A_Gesture_Life.txt
cluster-master | -rw-r--r-- 1 root supergroup 3813 2025-04-14 09:28 /data/12947743_A_Darkness_More_Than_Night.txt
cluster-master | -rw-r--r-- 1 root supergroup 604 2025-04-14 09:28 /data/12955622_A_Day_at_School.txt
cluster-master | -rw-r--r-- 1 root supergroup 6928 2025-04-14 09:28 /data/13060654_A_Lion_Among_Men.txt
cluster-master | -rw-r--r-- 1 root supergroup 2454 2025-04-14 09:28 /data/13125015_A_Fistful_of_Fingers.txt
cluster-master | -rw-r--r-- 1 root supergroup 1278 2025-04-14 09:28 /data/13130815_A_Fistful_of_F..._4_Skins.txt
cluster-master | -rw-r--r-- 1 root supergroup 3346 2025-04-14 09:28 /data/13242057_A_Child's_Cry_for_Help.txt
cluster-master | -rw-r--r-- 1 root supergroup 167 2025-04-14 09:28 /data/13300293_A_Breath_of_October.txt
cluster-master | -rw-r--r-- 1 root supergroup 1411 2025-04-14 09:28 /data/13375343_A_Feud_in_the_Kentucky_Hills.txt
cluster-master | -rw-r--r-- 1 root supergroup 919 2025-04-14 09:28 /data/13387430_A_Cry_for_Help_1912_film.txt
cluster-master | -rw-r--r-- 1 root supergroup 991 2025-04-14 09:28 /data/13391420_A_Chance_Deception.txt
cluster-master | -rw-r--r-- 1 root supergroup 615 2025-04-14 09:28 /data/13401464_A_Gamble_with_Death.txt
cluster-master | -rw-r--r-- 1 root supergroup 1494 2025-04-14 09:28 /data/13402887_A_Handful_of_Time.txt

```

- Pipeline1



```
elenatesm@HOME-PC: /mnt/c/3year/BD/Ass2/bigdataass2
[INFO] Starting Pipeline 1: Term Frequency
cluster-master | packageJobJar: [ [/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1-jar] /tmp/streamjob2163867850578855021.jar tmpDir=null
cluster-master | 2025-04-14 09:34:36,979 INFO client.DefaultHARMPaloverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-14 09:34:37,076 INFO client.DefaultHARMPaloverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-14 09:34:37,224 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744622728245_0001
cluster-master | 2025-04-14 09:34:48,415 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master | 2025-04-14 09:34:48,474 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master | 2025-04-14 09:34:48,567 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744622728245_0001
cluster-master | 2025-04-14 09:34:48,567 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master | 2025-04-14 09:34:48,691 INFO conf.Configuration: resource-types.xml not found
cluster-master | 2025-04-14 09:34:48,691 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master | 2025-04-14 09:34:48,852 INFO impl.YarnClientImpl: Submitted application application_1744622728245_0001
cluster-master | 2025-04-14 09:34:48,893 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744622728245_0001/
cluster-master | 2025-04-14 09:34:48,894 INFO mapreduce.Job: Running job: job_1744622728245_0001
cluster-master | 2025-04-14 09:34:54,975 INFO mapreduce.Job: Job job_1744622728245_0001 running in uber mode : false
cluster-master | 2025-04-14 09:34:54,976 INFO mapreduce.Job: map 0% reduce 0%
cluster-master | 2025-04-14 09:34:58,921 INFO mapreduce.Job: map 50% reduce 0%
cluster-master | 2025-04-14 09:34:59,934 INFO mapreduce.Job: map 100% reduce 0%
cassandra-server | INFO [Native-Transport-Requests-16] 2025-04-14 09:35:02,100 QueryProcessor.java:654 - Fully upgraded to at least 5.0.3
cluster-master | 2025-04-14 09:35:15,010 INFO mapreduce.Job: map 100% reduce 16%
cluster-master | 2025-04-14 09:35:16,015 INFO mapreduce.Job: map 100% reduce 31%
cluster-master | 2025-04-14 09:35:17,021 INFO mapreduce.Job: map 100% reduce 47%
cluster-master | 2025-04-14 09:35:19,030 INFO mapreduce.Job: map 100% reduce 62%
cluster-master | 2025-04-14 09:35:20,035 INFO mapreduce.Job: map 100% reduce 78%
cluster-master | 2025-04-14 09:35:21,042 INFO mapreduce.Job: map 100% reduce 76%
cluster-master | 2025-04-14 09:35:22,048 INFO mapreduce.Job: map 100% reduce 80%
cluster-master | 2025-04-14 09:35:23,054 INFO mapreduce.Job: map 100% reduce 81%
cluster-master | 2025-04-14 09:35:25,063 INFO mapreduce.Job: map 100% reduce 82%
cluster-master | 2025-04-14 09:35:26,069 INFO mapreduce.Job: map 100% reduce 83%
cluster-master | 2025-04-14 09:35:27,074 INFO mapreduce.Job: map 100% reduce 85%
cluster-master | 2025-04-14 09:35:28,080 INFO mapreduce.Job: map 100% reduce 86%
cluster-master | 2025-04-14 09:35:28,986 INFO mapreduce.Job: map 100% reduce 87%
cluster-master | 2025-04-14 09:35:30,999 INFO mapreduce.Job: map 100% reduce 88%
cluster-master | 2025-04-14 09:35:32,003 INFO mapreduce.Job: map 100% reduce 89%
cluster-master | 2025-04-14 09:35:33,009 INFO mapreduce.Job: map 100% reduce 90%
cluster-master | 2025-04-14 09:35:34,014 INFO mapreduce.Job: map 100% reduce 91%
cluster-master | 2025-04-14 09:35:35,020 INFO mapreduce.Job: map 100% reduce 92%
cluster-master | 2025-04-14 09:35:37,029 INFO mapreduce.Job: map 100% reduce 93%
cluster-master | 2025-04-14 09:35:38,035 INFO mapreduce.Job: map 100% reduce 94%
cluster-master | 2025-04-14 09:35:39,040 INFO mapreduce.Job: map 100% reduce 95%
cluster-master | 2025-04-14 09:35:40,044 INFO mapreduce.Job: map 100% reduce 97%
cluster-master | 2025-04-14 09:35:41,049 INFO mapreduce.Job: map 100% reduce 98%
cluster-master | 2025-04-14 09:35:43,054 INFO mapreduce.Job: map 100% reduce 99%
cluster-master | 2025-04-14 09:35:44,062 INFO mapreduce.Job: map 100% reduce 100%
cluster-master | 2025-04-14 09:35:46,076 INFO mapreduce.Job: Job job_1744622728245_0001 completed successfully
cluster-master | 2025-04-14 09:35:46,148 INFO mapreduce.Job: Counters: 54
cluster-master | File System Counters
cluster-master | FILE: Number of bytes read=11166286
cluster-master | FILE: Number of bytes written=24276566
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=3560227
cluster-master | HDFS: Number of bytes written=0
cluster-master | HDFS: Number of read operations=31
cluster-master | HDFS: Number of large read operations=0
cluster-master | HDFS: Number of write operations=10
cluster-master | HDFS: Number of bytes read erasure-coded=0
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=3560227
cluster-master | HDFS: Number of bytes written=0
cluster-master | HDFS: Number of read operations=31
cluster-master | HDFS: Number of large read operations=0
cluster-master | HDFS: Number of write operations=10
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=5
cluster-master | Data-local map tasks=2
cluster-master | Total time spent by all maps in occupied slots (ms)=5128
cluster-master | Total time spent by all reduces in occupied slots (ms)=203374
cluster-master | Total time spent by all map tasks (ms)=5128
cluster-master | Total time spent by all reduce tasks (ms)=203374
cluster-master | Total vcore-milliseconds taken by all map tasks=5128
cluster-master | Total vcore-milliseconds taken by all reduce tasks=203374
cluster-master | Total megabyte-milliseconds taken by all map tasks=5251072
cluster-master | Total megabyte-milliseconds taken by all reduce tasks=208254976
cluster-master | Map-Reduce Framework
cluster-master | Map input records=1003
cluster-master | Map output records=568330
cluster-master | Map output bytes=10029595
cluster-master | Map output materialized bytes=11166316
cluster-master | Input split bytes=292
cluster-master | Combine input records=0
cluster-master | Combine output records=0
cluster-master | Reduce input groups=293705
cluster-master | Reduce shuffle bytes=11166316
cluster-master | Reduce input records=568330
cluster-master | Reduce output records=0
cluster-master | Spilled Records=1136660
cluster-master | Shuffled Maps =10
cluster-master | Failed Shuffles=0
cluster-master | Merged Map outputs=10
cluster-master | GC time elapsed (ms)=306
cluster-master | CPU time spent (ms)=131670
cluster-master | Physical memory (bytes) snapshot=1948946432
cluster-master | Virtual memory (bytes) snapshot=18161057792
cluster-master | Total committed heap usage (bytes)=1923612672
cluster-master | Peak Map Physical memory (bytes)=473530368
cluster-master | Peak Map Virtual memory (bytes)=2593521664
cluster-master | Peak Reduce Physical memory (bytes)=288378880
cluster-master | Peak Reduce Virtual memory (bytes)=2867228672
cluster-master | Shuffle Errors
cluster-master | BAD_ID=0
cluster-master | CONNECTION=0
cluster-master | IO_ERROR=0
cluster-master | WRONG_LENGTH=0
cluster-master | WRONG_MAP=0
cluster-master | WRONG_REDUCE=0
cluster-master | File Input Format Counters
cluster-master | Bytes Read=3559935
cluster-master | File Output Format Counters
cluster-master | Bytes Written=0
cluster-master | 2025-04-14 09:35:46,148 INFO streaming.StreamJob: Output directory: /tmp/index/xf
```

- Pipeline 2

```
elenatesm@HOME-PC: /mnt/c/3year/BD/Ass2/bigdataass2
[INFO] Starting Pipeline 2: Documet Frequency
packageJobJar: [ [/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1-jar] /tmp/streamjob2941235845667514524.jar tmpDir=null
2025-04-14 09:35:47,461 INFO client.DefaultHARMPFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-14 09:35:47,567 INFO client.DefaultHARMPFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-14 09:35:47,713 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744622728245_0002
2025-04-14 09:35:59,088 INFO mapred.FileInputFormat: Total input files to process : 1
2025-04-14 09:35:59,146 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-14 09:35:59,232 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744622728245_0002
2025-04-14 09:35:59,232 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-14 09:35:59,356 INFO conf.Configuration: resource-types.xml not found
2025-04-14 09:35:59,356 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-14 09:35:59,402 INFO impl.VarnClientImpl: Submitted application application_1744622728245_0002
2025-04-14 09:35:59,434 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744622728245_0002/
2025-04-14 09:35:59,435 INFO mapreduce.Job: Running job: job_1744622728245_0002
2025-04-14 09:36:05,492 INFO mapreduce.Job: Job job_1744622728245_0002 running in uber mode : false
2025-04-14 09:36:05,493 INFO mapreduce.Job: map 0% reduce 0%
2025-04-14 09:36:09,530 INFO mapreduce.Job: map 100% reduce 0%
2025-04-14 09:36:24,616 INFO mapreduce.Job: map 100% reduce 40%
2025-04-14 09:36:25,623 INFO mapreduce.Job: map 100% reduce 60%
2025-04-14 09:36:26,631 INFO mapreduce.Job: map 100% reduce 80%
2025-04-14 09:36:27,635 INFO mapreduce.Job: map 100% reduce 100%
2025-04-14 09:36:27,640 INFO mapreduce.Job: Job job_1744622728245_0002 completed successfully
2025-04-14 09:36:27,705 INFO mapreduce.Job: Counters: 55
File System Counters
FILE: Number of bytes read=5105243
FILE: Number of bytes written=12148033
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=3560227
HDFS: Number of bytes written=0
HDFS: Number of read operations=31
HDFS: Number of large read operations=0
HDFS: Number of write operations=10
HDFS: Number of bytes read erasure-coded=0
Job Counters
Killed reduce tasks=1
Launched map tasks=2
Launched reduce tasks=6
Data-local map tasks=2
Total time spent by all maps in occupied slots (ms)=3582
Total time spent by all reduces in occupied slots (ms)=64153
Total time spent by all map tasks (ms)=3582
Total time spent by all reduce tasks (ms)=64153
Total vcore-milliseconds taken by all map tasks=3582
Total vcore-milliseconds taken by all reduce tasks=64153
Total megabyte-milliseconds taken by all map tasks=3667968
Total megabyte-milliseconds taken by all reduce tasks=65692672
Map-Reduce Framework
Map input records=1003
Map output records=283893
Map output bytes=4537426
Map output materialized bytes=5105273
Input split bytes=292
Combine input records=0
Combine output records=0
Reduce input groups=78975
Reduce shuffle bytes=5105273
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=3560227
HDFS: Number of bytes written=0
HDFS: Number of read operations=31
HDFS: Number of large read operations=0
HDFS: Number of write operations=10
HDFS: Number of bytes read erasure-coded=0
Job Counters
Killed reduce tasks=1
Launched map tasks=2
Launched reduce tasks=6
Data-local map tasks=2
Total time spent by all maps in occupied slots (ms)=3582
Total time spent by all reduces in occupied slots (ms)=64153
Total time spent by all map tasks (ms)=3582
Total time spent by all reduce tasks (ms)=64153
Total vcore-milliseconds taken by all map tasks=3582
Total vcore-milliseconds taken by all reduce tasks=64153
Total megabyte-milliseconds taken by all map tasks=3667968
Total megabyte-milliseconds taken by all reduce tasks=65692672
Map-Reduce Framework
Map input records=1003
Map output records=283893
Map output bytes=4537426
Map output materialized bytes=5105273
Input split bytes=292
Combine input records=0
Combine output records=0
Reduce input groups=78975
Reduce shuffle bytes=5105273
Reduce input records=283893
Reduce output records=0
Spilled Records=567786
Shuffled Maps =10
Failed Shuffles=0
Merged Map outputs=10
GC time elapsed (ms)=227
CPU time spent (ms)=7240
Physical memory (bytes) snapshot=1699061760
Virtual memory (bytes) snapshot=18163003392
Total committed heap usage (bytes)=1611137024
Peak Map Physical memory (bytes)=360983424
Peak Map Virtual memory (bytes)=2589388800
Peak Reduce Physical memory (bytes)=210771968
Peak Reduce Virtual memory (bytes)=2599256064
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=3559935
File Output Format Counters
Bytes Written=0
2025-04-14 09:36:27,705 INFO streaming.StreamJob: Output directory: /tmp/index/df
```

- Pipeline 3

```
elenatesm@HOME-PC: /mnt/c/3year/BD/Ass2/bigdataass2
[INFO] Starting Pipeline 3: Document Information
packageJobJar: [ [/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1-jar] /tmp/streamjob4969349406022865787.jar tmpDir=null
2025-04-14 09:36:28,887 INFO client.DefaultHARMPaloverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-14 09:36:29,000 INFO client.DefaultHARMPaloverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-14 09:36:29,137 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744622728245_0003
2025-04-14 09:36:40,351 INFO mapred.FileInputFormat: Total input files to process : 1
2025-04-14 09:36:40,410 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-14 09:36:40,496 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744622728245_0003
2025-04-14 09:36:40,497 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-14 09:36:40,627 INFO conf.Configuration: resource-types.xml not found
2025-04-14 09:36:40,627 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-14 09:36:40,685 INFO impl.VarnClientImpl: Submitted application application_1744622728245_0003
2025-04-14 09:36:40,711 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744622728245_0003/
2025-04-14 09:36:40,712 INFO mapreduce.Job: Running job: job_1744622728245_0003
2025-04-14 09:36:45,764 INFO mapreduce.Job: Job job_1744622728245_0003 running in uber mode : false
2025-04-14 09:36:45,765 INFO mapreduce.Job: map 0% reduce 0%
2025-04-14 09:36:49,812 INFO mapreduce.Job: map 100% reduce 0%
2025-04-14 09:36:53,830 INFO mapreduce.Job: map 100% reduce 20%
2025-04-14 09:36:54,836 INFO mapreduce.Job: map 100% reduce 40%
2025-04-14 09:36:55,841 INFO mapreduce.Job: map 100% reduce 60%
2025-04-14 09:36:56,847 INFO mapreduce.Job: map 100% reduce 100%
2025-04-14 09:36:57,856 INFO mapreduce.Job: Job job_1744622728245_0003 completed successfully
2025-04-14 09:36:57,926 INFO mapreduce.Job: Counters: 55
File System Counters
  FILE: Number of bytes read=14736
  FILE: Number of bytes written=1967026
  FILE: Number of read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=3560227
  HDFS: Number of bytes written=0
  HDFS: Number of read operations=31
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=10
  HDFS: Number of bytes read erasure-coded=0
Job Counters
  Killed reduce tasks=1
  Launched map tasks=2
  Launched reduce tasks=5
  Data-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=3219
  Total time spent by all reduces in occupied slots (ms)=10206
  Total time spent by all map tasks (ms)=3219
  Total time spent by all reduce tasks (ms)=10206
  Total vcore-milliseconds taken by all map tasks=3219
  Total vcore-milliseconds taken by all reduce tasks=10206
  Total megabyte-milliseconds taken by all map tasks=3296256
  Total megabyte-milliseconds taken by all reduce tasks=10450944
Map-Reduce Framework
  Map input records=1003
  Map output records=997
  Map output bytes=12712
  Map output materialized bytes=14766
  Input split bytes=292
  Combine input records=0
  Combine output records=0
  Reduce input groups=997
  Reduce shuffle bytes=14766
  FILE: Number of write operations=0
  HDFS: Number of bytes read=3560227
  HDFS: Number of bytes written=0
  HDFS: Number of read operations=31
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=10
  HDFS: Number of bytes read erasure-coded=0
Job Counters
  Killed reduce tasks=1
  Launched map tasks=2
  Launched reduce tasks=5
  Data-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=3219
  Total time spent by all reduces in occupied slots (ms)=10206
  Total time spent by all map tasks (ms)=3219
  Total time spent by all reduce tasks (ms)=10206
  Total vcore-milliseconds taken by all map tasks=3219
  Total vcore-milliseconds taken by all reduce tasks=10206
  Total megabyte-milliseconds taken by all map tasks=3296256
  Total megabyte-milliseconds taken by all reduce tasks=10450944
Map-Reduce Framework
  Map input records=1003
  Map output records=997
  Map output bytes=12712
  Map output materialized bytes=14766
  Input split bytes=292
  Combine input records=0
  Combine output records=0
  Reduce input groups=997
  Reduce shuffle bytes=14766
  Reduce input records=997
  Reduce output records=0
  Spilled Records=1994
  Shuffled Maps =10
  Failed Shuffles=0
  Merged Map outputs=10
  GC time elapsed (ms)=231
  CPU time spent (ms)=2840
  Physical memory (bytes) snapshot=1837355008
  Virtual memory (bytes) snapshot=18156187648
  Total committed heap usage (bytes)=1874853888
  Peak Map Physical memory (bytes)=362975232
  Peak Map Virtual memory (bytes)=2591084544
  Peak Reduce Physical memory (bytes)=259399680
  Peak Reduce Virtual memory (bytes)=2596298752
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=3559935
File Output Format Counters
  Bytes Written=0
2025-04-14 09:36:57,926 INFO streaming.StreamJob: Output directory: /tmp/index/doc
[INFO] Indexing completed successfully!
```

- To verify that the data was correctly written to Cassandra, I used the following CQL commands.



```

Выбратъ elenatesm@HOME-PC: /mnt/c/3year/BD/Ass2/bigdataass2$ docker exec -it cassandra-server cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.2.0 | Cassandra 5.0.3 | CQL spec 3.4.7 | Native protocol v5]
Use HELP for help.
cqlsh> USE search_engine;
cqlsh:search_engine> SELECT * FROM term_docs LIMIT 5;

term      | doc_id | tf
-----
Bondy,    | 6602969 | 1
A2        | 2820410 | 1
A2        | 53191245 | 1
A2        | 718823 | 1
musicians. | 6325129 | 1

(5 rows)
cqlsh:search_engine> SELECT * FROM term_stats LIMIT 5;

term      | df
-----
yeong-ho... | 1
bessus     | 1
musicians. | 1
ix         | 2
await      | 2

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

doc_id | d1
-----
10230685 | 118
27568194 | 560
39710446 | 332
38294693 | 360
51794980 | 303

(5 rows)
cqlsh:search_engine> SELECT * FROM corpus_stats;

id      | avg_d1 | total_docs
-----
global | 537.46979 | 215

(1 rows)
cqlsh:search_engine>

```