Assignment 2: Simple Search Engine using Hadoop MapReduce

Elena Tesmeeva; 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.

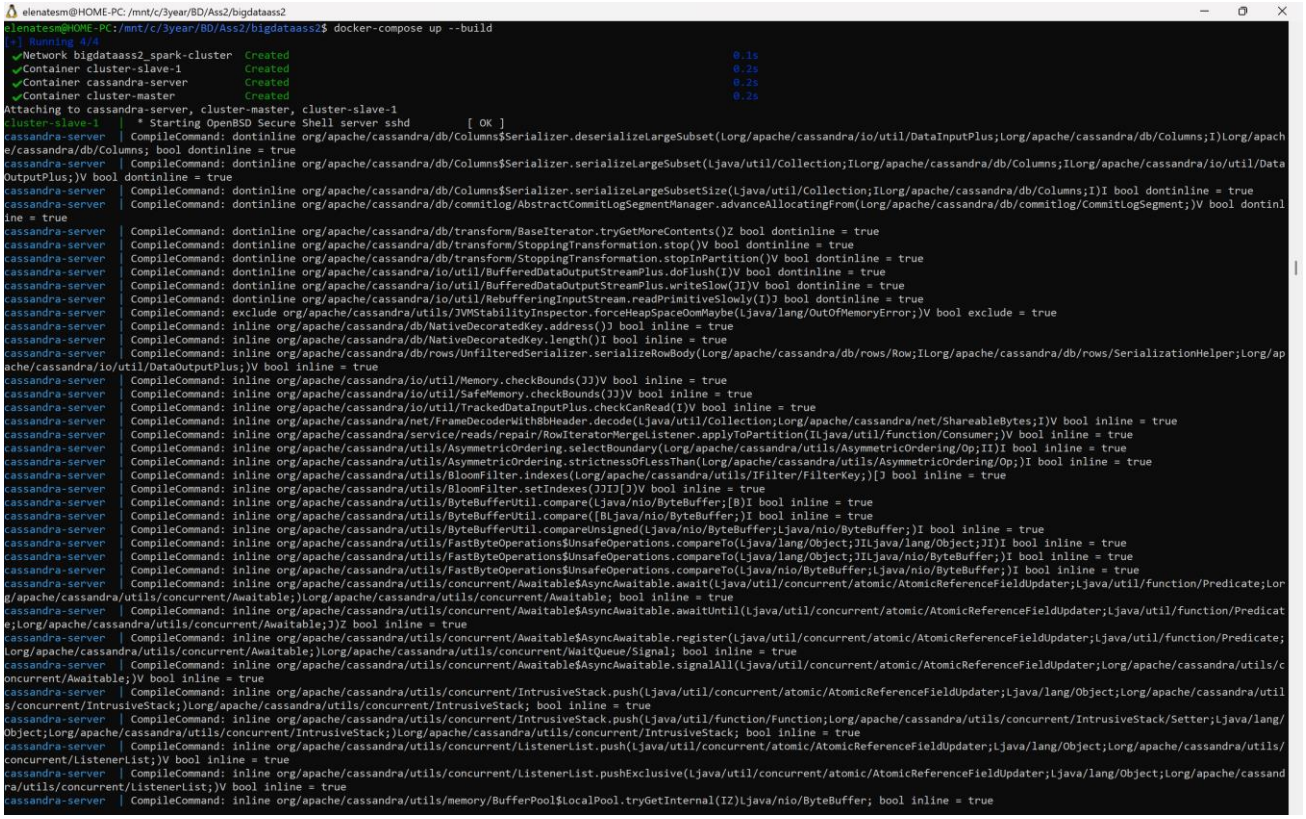
2. Demonstration

How to run code:
1) git clone https://github.com/elenatesm/bigdataass2

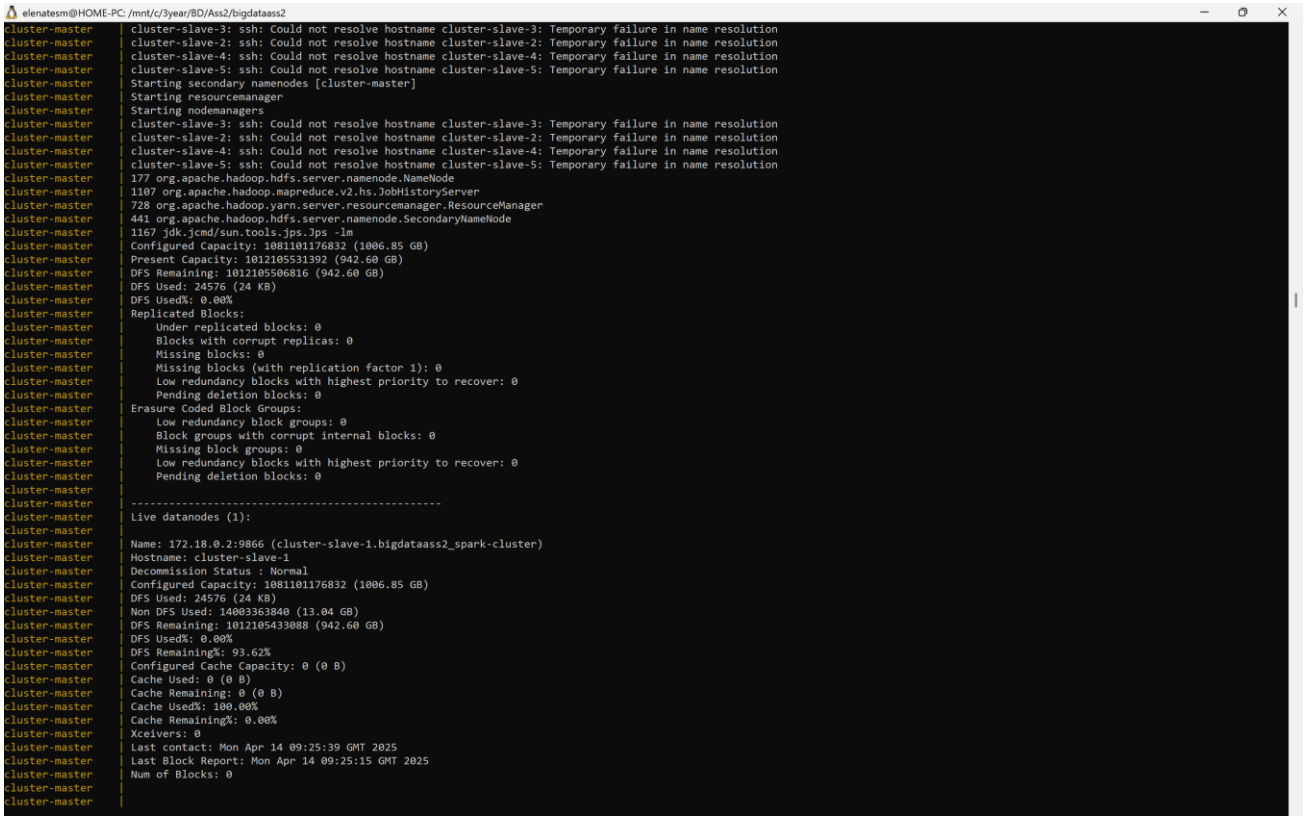
2) Download file a.parquet from https://www.kaggle.com/datasets/jjinho/wikipedia-20230701?select=a.parquet to the app folder of the repository

3) Run docker-compose up

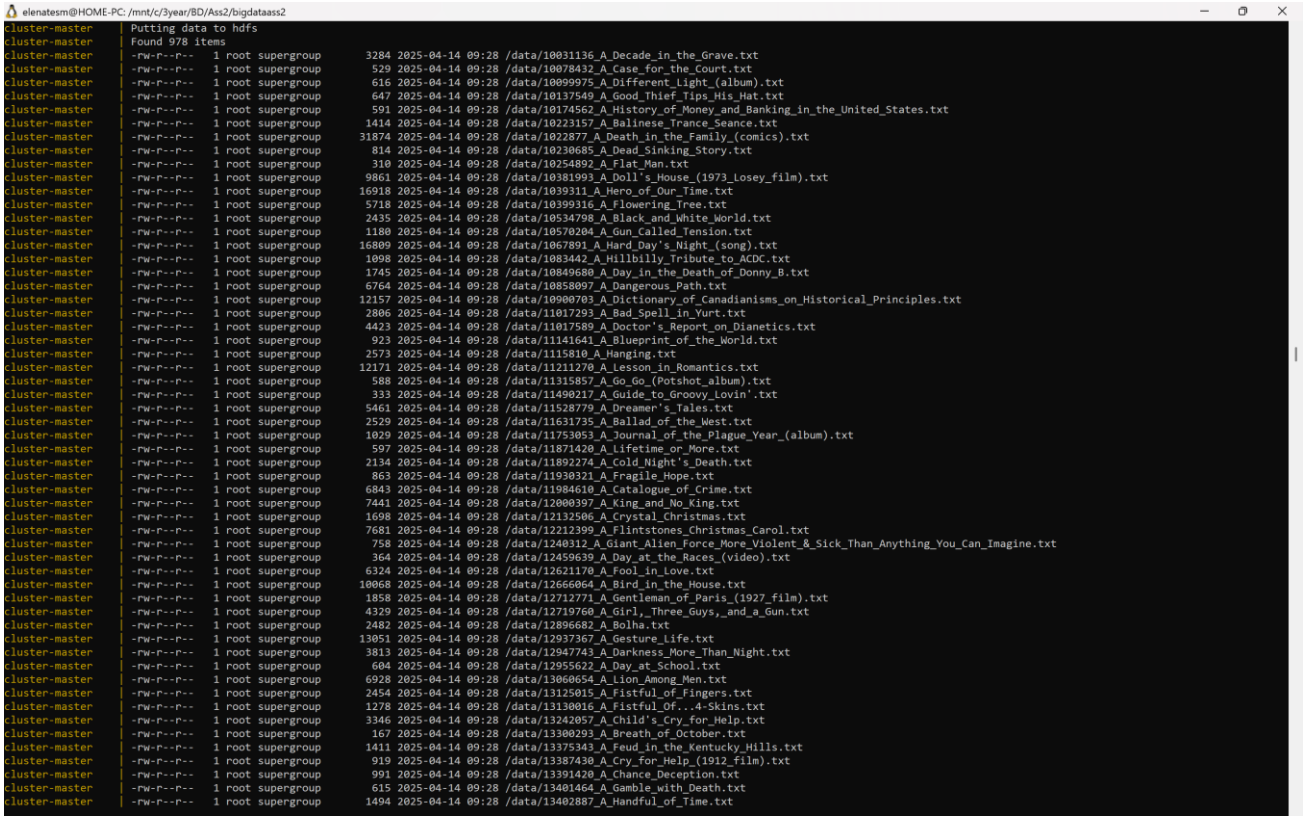
- Creating containers in Docker



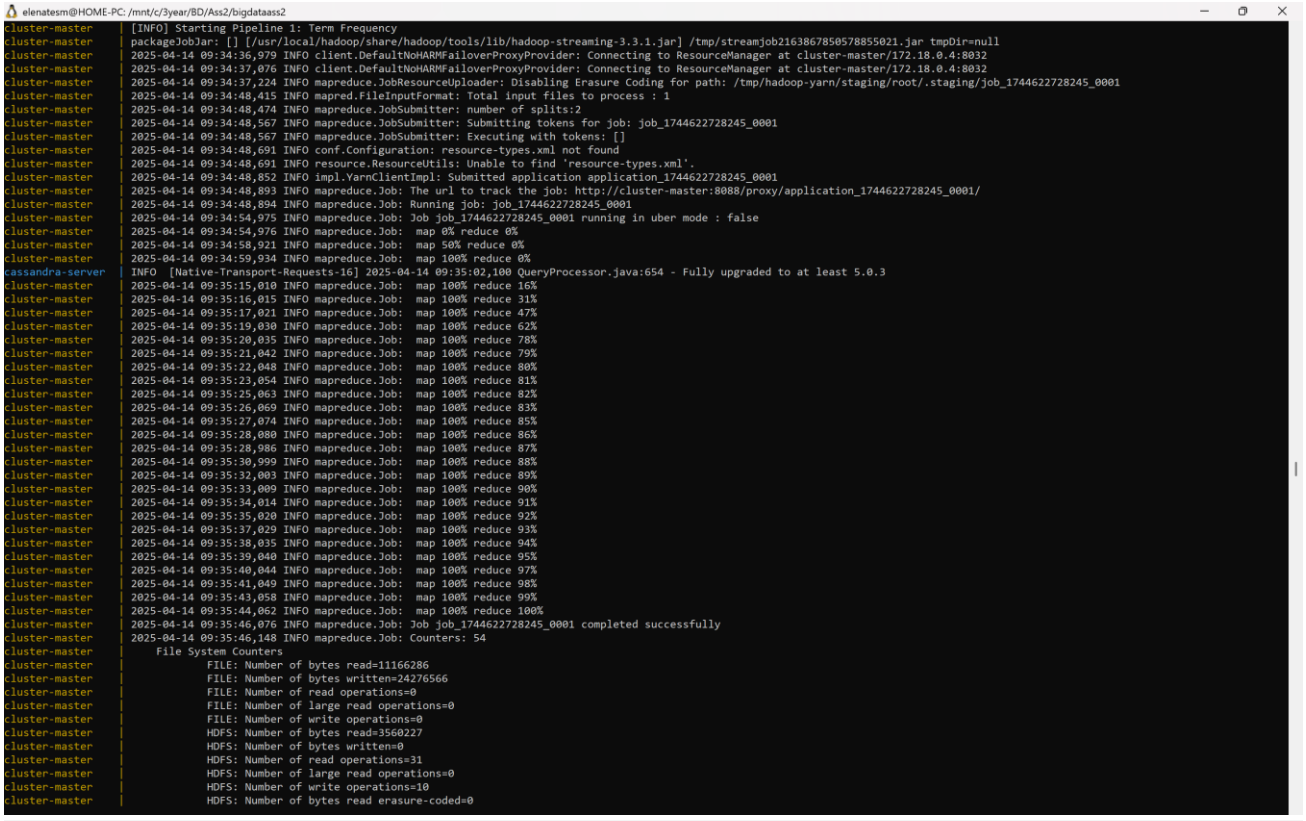
- Live Datanodes

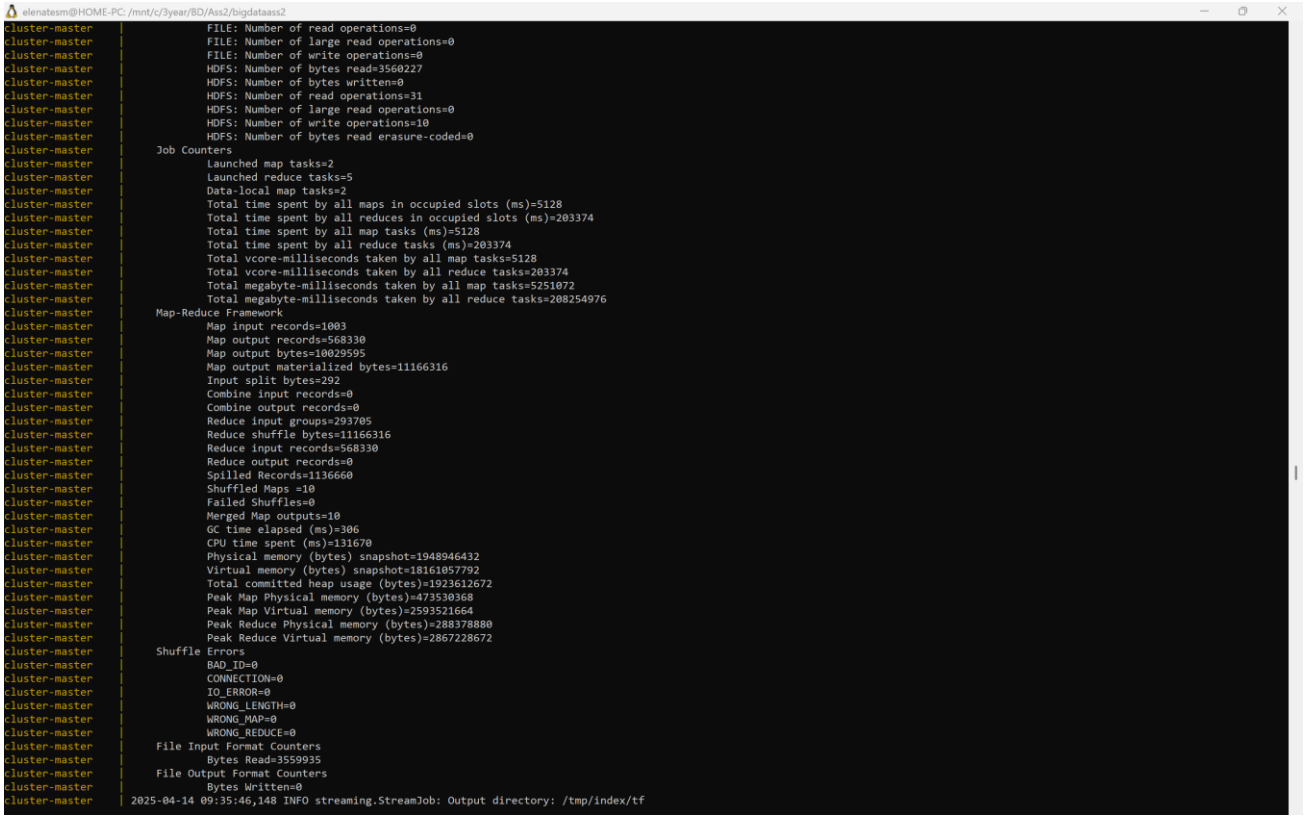


- Putting data to hdfs

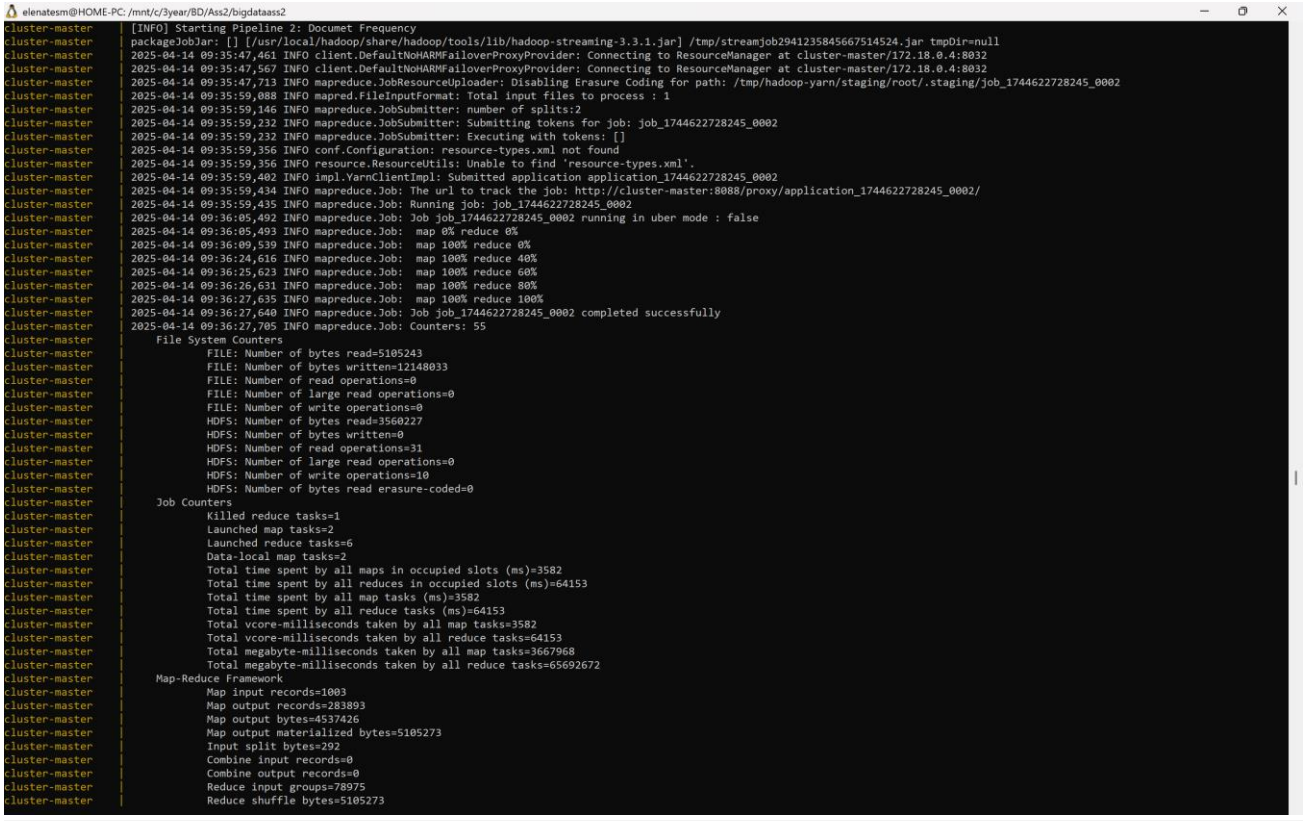


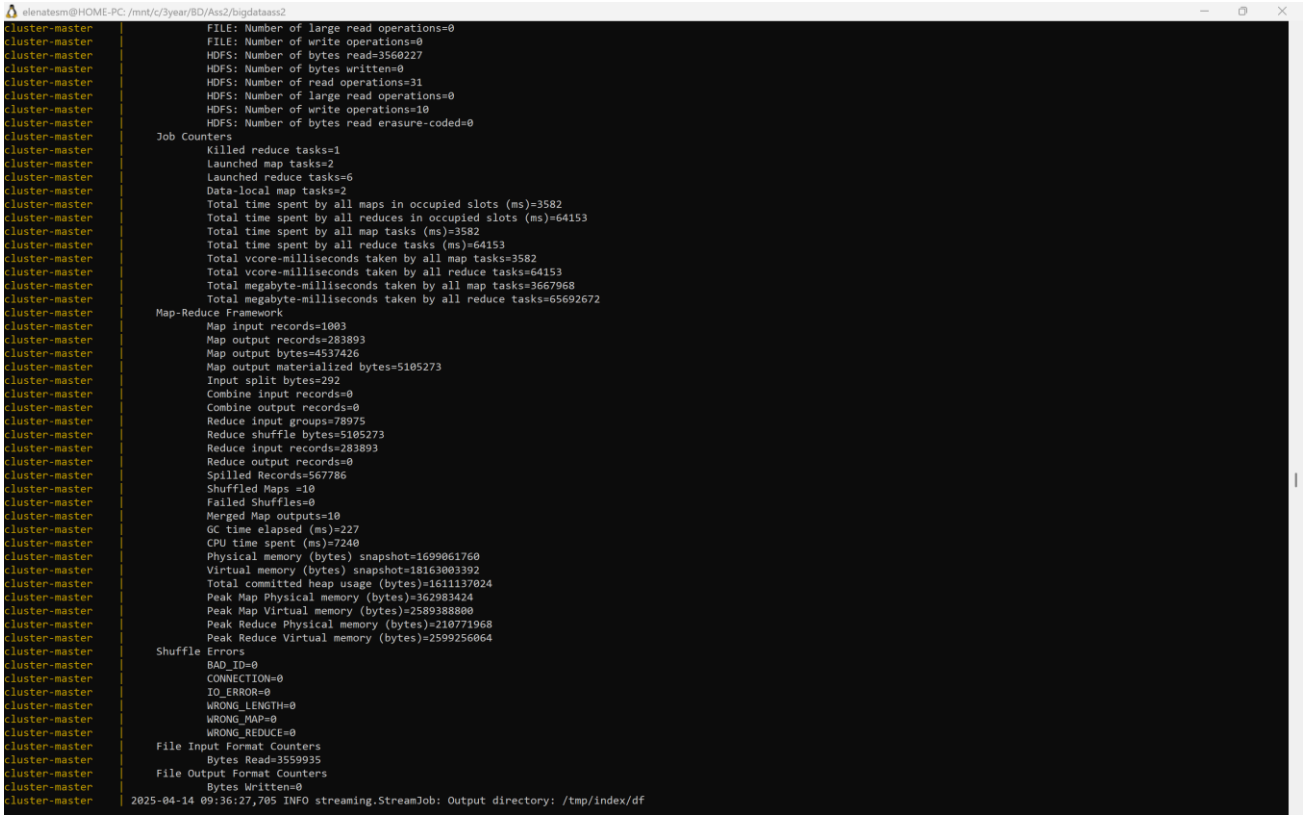
- Pipeline1


elenatesm@HOME-PC: /mnt/c/3year/BD/Ass2/bigdataass2
cluster-master | [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.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-14 09:34:37,076 INFO client.DefaultNoHARMFailoverProxyProvider: 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 79%
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 09:35:43,058 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

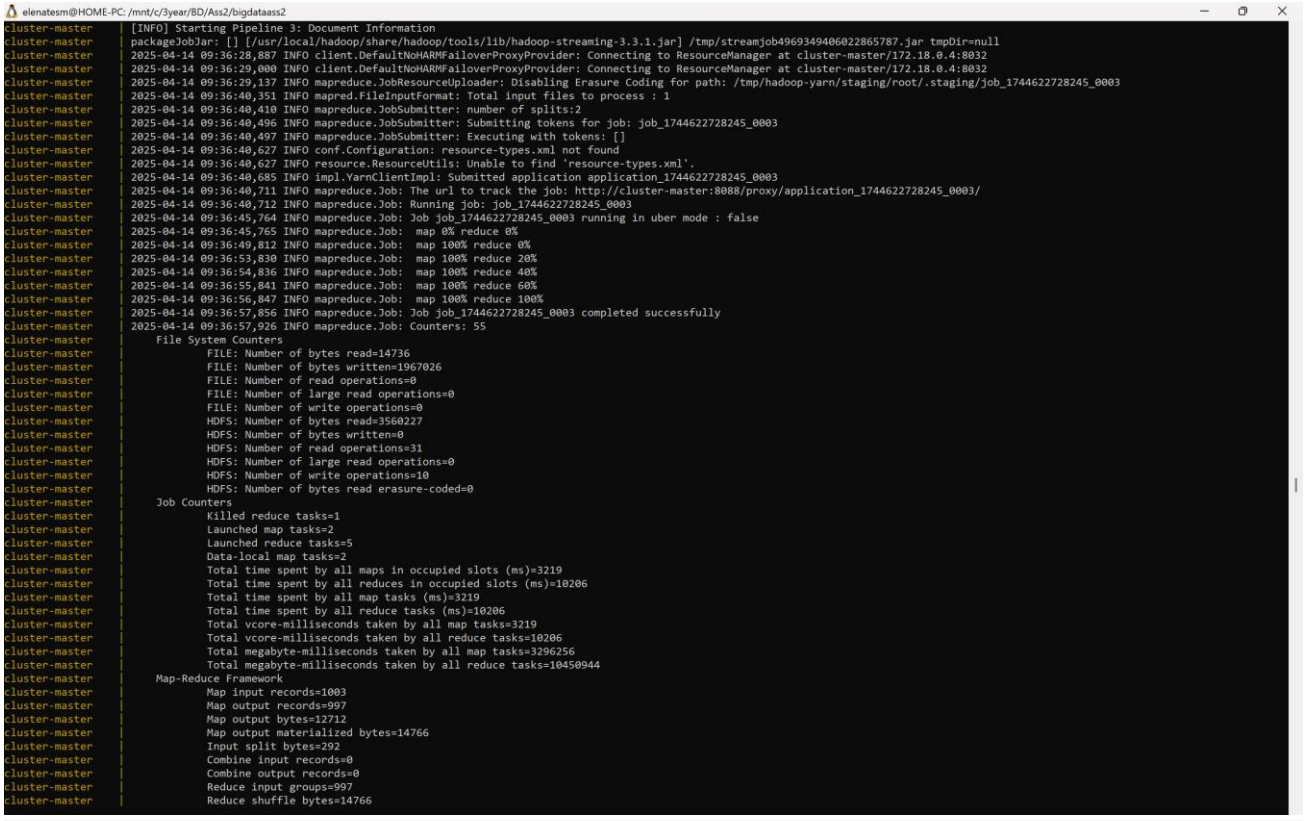

elenatesm@HOME-PC: /mnt/c/3year/BD/Ass2/bigdataass2
cluster-master | FILE: Number of read operations=0
cluster-master | FILE: Number of large read operations=0
cluster-master | FILE: Number of write operations=0
cluster-master | HDFS: Number of bytes read=3560227
cluster-master | HDFS: Number of bytes 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/tf

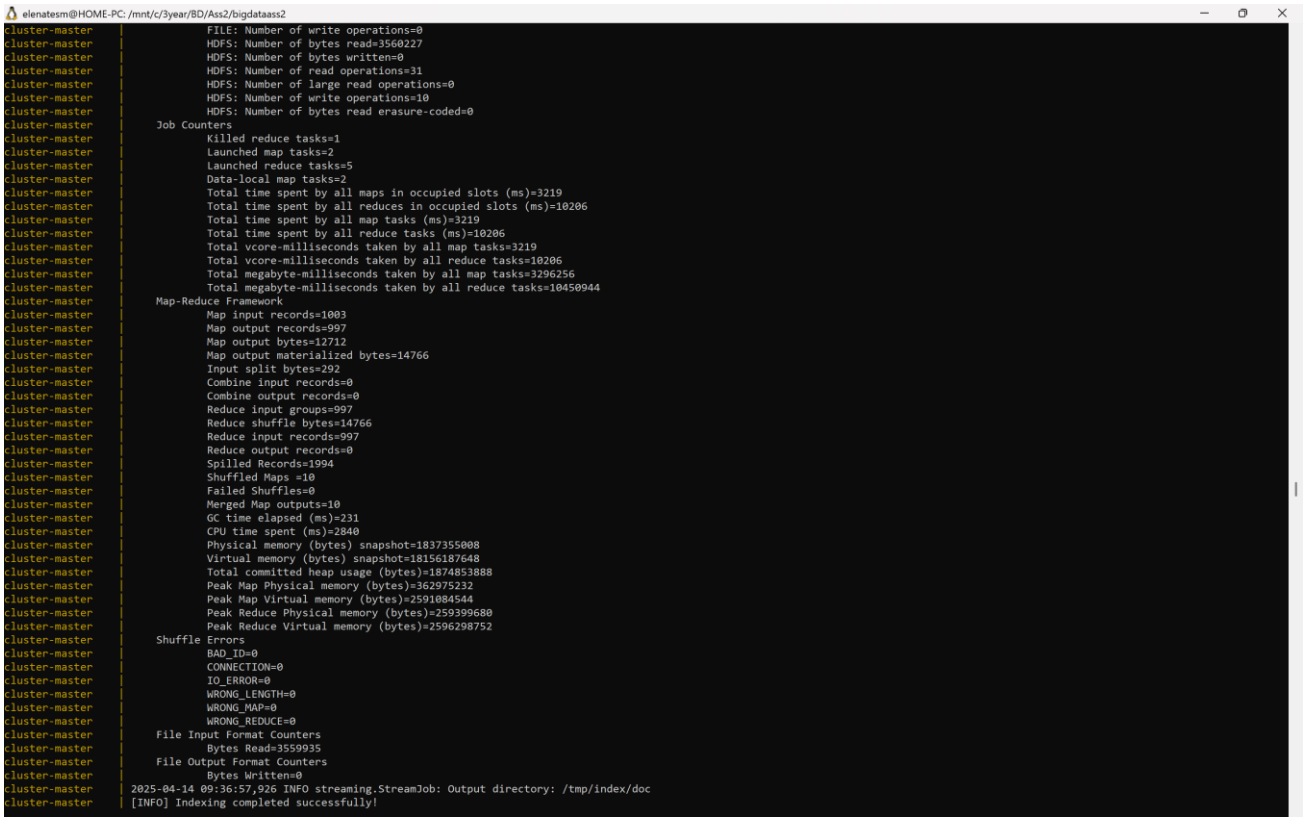
- Pipeline 2


elenatesm@HOME-PC: /mnt/c/3year/BD/Ass2/bigdataass2
cluster-master | [INFO] Starting Pipeline 2: Documet Frequency
cluster-master | packageJobJar: [] [/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar] /tmp/streamjob2941235845667514524.jar tmpDir=null
cluster-master | 2025-04-14 09:35:47,461 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-14 09:35:47,567 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-14 09:35:47,713 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744622728245_0002
cluster-master | 2025-04-14 09:35:59,088 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master | 2025-04-14 09:35:59,146 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master | 2025-04-14 09:35:59,232 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744622728245_0002
cluster-master | 2025-04-14 09:35:59,232 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master | 2025-04-14 09:35:59,356 INFO conf.Configuration: resource-types.xml not found
cluster-master | 2025-04-14 09:35:59,356 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master | 2025-04-14 09:35:59,402 INFO impl.YarnClientImpl: Submitted application application_1744622728245_0002
cluster-master | 2025-04-14 09:35:59,434 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744622728245_0002/
cluster-master | 2025-04-14 09:35:59,435 INFO mapreduce.Job: Running job: job_1744622728245_0002
cluster-master | 2025-04-14 09:36:05,492 INFO mapreduce.Job: Job job_1744622728245_0002 running in uber mode : false
cluster-master | 2025-04-14 09:36:05,493 INFO mapreduce.Job: map 0% reduce 0%
cluster-master | 2025-04-14 09:36:09,539 INFO mapreduce.Job: map 100% reduce 0%
cluster-master | 2025-04-14 09:36:24,616 INFO mapreduce.Job: map 100% reduce 40%
cluster-master | 2025-04-14 09:36:25,623 INFO mapreduce.Job: map 100% reduce 60%
cluster-master | 2025-04-14 09:36:26,631 INFO mapreduce.Job: map 100% reduce 80%
cluster-master | 2025-04-14 09:36:27,635 INFO mapreduce.Job: map 100% reduce 100%
cluster-master | 2025-04-14 09:36:27,640 INFO mapreduce.Job: Job job_1744622728245_0002 completed successfully
cluster-master | 2025-04-14 09:36:27,705 INFO mapreduce.Job: Counters: 55
cluster-master | File System Counters
cluster-master | FILE: Number of bytes read=5105243
cluster-master | FILE: Number of bytes written=12148033
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 | Killed reduce tasks=1
cluster-master | Launched map tasks=2
cluster-master | Launched reduce tasks=6
cluster-master | Data-local map tasks=2
cluster-master | Total time spent by all maps in occupied slots (ms)=3582
cluster-master | Total time spent by all reduces in occupied slots (ms)=64153
cluster-master | Total time spent by all map tasks (ms)=3582
cluster-master | Total time spent by all reduce tasks (ms)=64153
cluster-master | Total vcore-milliseconds taken by all map tasks=3582
cluster-master | Total vcore-milliseconds taken by all reduce tasks=64153
cluster-master | Total megabyte-milliseconds taken by all map tasks=3667968
cluster-master | Total megabyte-milliseconds taken by all reduce tasks=65692672
cluster-master | Map-Reduce Framework
cluster-master | Map input records=1003
cluster-master | Map output records=283893
cluster-master | Map output bytes=4537426
cluster-master | Map output materialized bytes=5105273
cluster-master | Input split bytes=292
cluster-master | Combine input records=0
cluster-master | Combine output records=0
cluster-master | Reduce input groups=78975
cluster-master | Reduce shuffle bytes=5105273

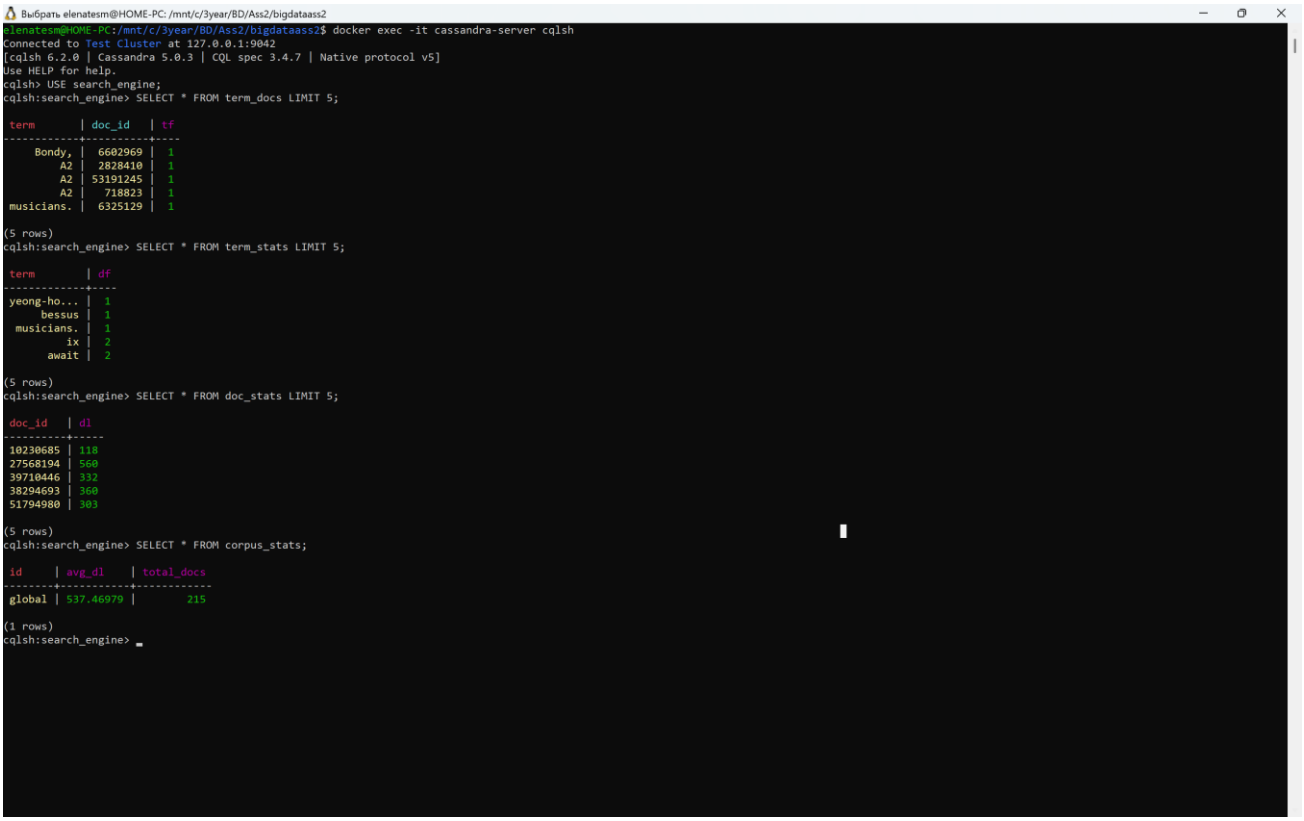

elenatesm@HOME-PC: /mnt/c/3year/BD/Ass2/bigdataass2
cluster-master | FILE: Number of large read operations=0
cluster-master | FILE: Number of write operations=0
cluster-master | HDFS: Number of bytes read=3560227
cluster-master | HDFS: Number of bytes written=0
cluster-master | HDFS: Number of read operations=31
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 | Killed reduce tasks=1
cluster-master | Launched map tasks=2
cluster-master | Launched reduce tasks=6
cluster-master | Data-local map tasks=2
cluster-master | Total time spent by all maps in occupied slots (ms)=3582
cluster-master | Total time spent by all reduces in occupied slots (ms)=64153
cluster-master | Total time spent by all map tasks (ms)=3582
cluster-master | Total time spent by all reduce tasks (ms)=64153
cluster-master | Total vcore-milliseconds taken by all map tasks=3582
cluster-master | Total vcore-milliseconds taken by all reduce tasks=64153
cluster-master | Total megabyte-milliseconds taken by all map tasks=3667968
cluster-master | Total megabyte-milliseconds taken by all reduce tasks=65692672
cluster-master | Map-Reduce Framework
cluster-master | Map input records=1003
cluster-master | Map output records=283893
cluster-master | Map output bytes=4537426
cluster-master | Map output materialized bytes=5105273
cluster-master | Input split bytes=292
cluster-master | Combine input records=0
cluster-master | Combine output records=0
cluster-master | Reduce input groups=78975
cluster-master | Reduce shuffle bytes=5105273
cluster-master | Reduce input records=283893
cluster-master | Reduce output records=0
cluster-master | Spilled Records=567786
cluster-master | Shuffled Maps =10
cluster-master | Failed Shuffles=0
cluster-master | Merged Map outputs=10
cluster-master | GC time elapsed (ms)=227
cluster-master | CPU time spent (ms)=7240
cluster-master | Physical memory (bytes) snapshot=1699061760
cluster-master | Virtual memory (bytes) snapshot=18163003392
cluster-master | Total committed heap usage (bytes)=1611137024
cluster-master | Peak Map Physical memory (bytes)=362983424
cluster-master | Peak Map Virtual memory (bytes)=2589388800
cluster-master | Peak Reduce Physical memory (bytes)=210771968
cluster-master | Peak Reduce Virtual memory (bytes)=2599256064
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 09:36:27,705 INFO streaming.StreamJob: Output directory: /tmp/index/df

- Pipeline 3





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



Conclusion

In this assignment, a basic search engine was successfully developed utilizing the Hadoop MapReduce framework and Cassandra database. The system accurately indexes documents by computing term frequency, document frequency, and document statistics, with results stored in dedicated Cassandra tables.

Despite encountering integration challenges, particularly with the Cassandra driver in the Hadoop environment, these issues were resolved through the packaging and deployment of the necessary libraries. The three MapReduce pipelines—Term Frequency, Document Frequency, and Document Statistics—were executed successfully, enabling the creation of a comprehensive index facilitating efficient document retrieval.