Assignment №2 BigData Report

Amir Nigmatullin ”am.nigmatullin@innopolis.university”

April 15, 2025

1 Methodology

1.1 Preprocessing

First of all, I preprocess data stored in HDFS. In preprocessing I make lowercasing, removing punctuation and stopwords and tokenizing. The output is stored back in HDFS for the next stages.

1.2 MapReduce Pair 1: Document Statistics

In first Map and Reduce pair, these jobs were processed:
1) Computing unique terms (Vocabulary)
2) Computing document frequencies
3) Computing lengths and average document length

1.3 MapReduce Pair 2: Inverted Index Construction

In the second Map and Reduce pair I build inverted index and store for each term document IDs (where it appears) and the frequency of the term in each document.

This index is stored in the Cassandra table `inverted_index(term, doc_id, freq)`.

1.4 BM25 Search Engine

I deployed a search facility that did the following:

1. Preprocessed the input query the same way we handle documents.
2. For every term in the query, fetched the corresponding posting list from Cassandra.
3. Calculated the BM25 score for each candidate document.
4. Returned the top-ranked results.

I used this formula for calculating BM25. It contains from 2 parts. First:

$$\sum_{i \in q} \log\left(\frac{N - df_i + 0.5}{df_i + 0.5} + 1\right)$$

Second:

$$\frac{(k_1 + 1) \cdot tf_{i,d}}{tf_{i,d} + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl})}$$

And in the result we multiply this 2 parts and get result, our score.

1.5 Data Storage in Cassandra

- **vocabulary**: This table stores the full set of unique terms extracted from the corpus. Each row consists of a term (the word itself) and its doc_freq.

  ```
  CREATE TABLE IF NOT EXISTS vocabulary (
      term TEXT PRIMARY KEY,
      doc_freq INT
  )
  ```

- **inverted_index**: This table represents the core inverted index. It maps each term to the documents it appears in (doc_id), along with the term_freq (how often the term occurs in that document), and the positions (list of word indices where the term appears in the document).

  ```
  CREATE TABLE IF NOT EXISTS inverted_index (
      term TEXT,
      doc_id TEXT,
      term_freq INT,
      positions LIST<INT>,
      PRIMARY KEY (term, doc_id)
  )
  ```

- **doc_stats**: This table stores metadata about each document, including its doc_id, title, total number of terms (total_terms), and number of unique terms (unique_terms).

  ```
  CREATE TABLE IF NOT EXISTS doc_stats (
      doc_id TEXT PRIMARY KEY,
      title TEXT,
      total_terms INT,
      unique_terms INT
  )
  ```

1.6 About other files

- **app.py**: The main Python script handles loading the preprocessed documents into Cassandra. It reads document-term stats like term frequencies and document lengths, then stores the inverted index, document statistics, and vocabulary into their respective Cassandra tables..

- **query.py**: This script handles BM25 ranking using the indexed data stored in Cassandra. It works with free-text queries and returns the highest-ranked documents based on their BM25 scores.

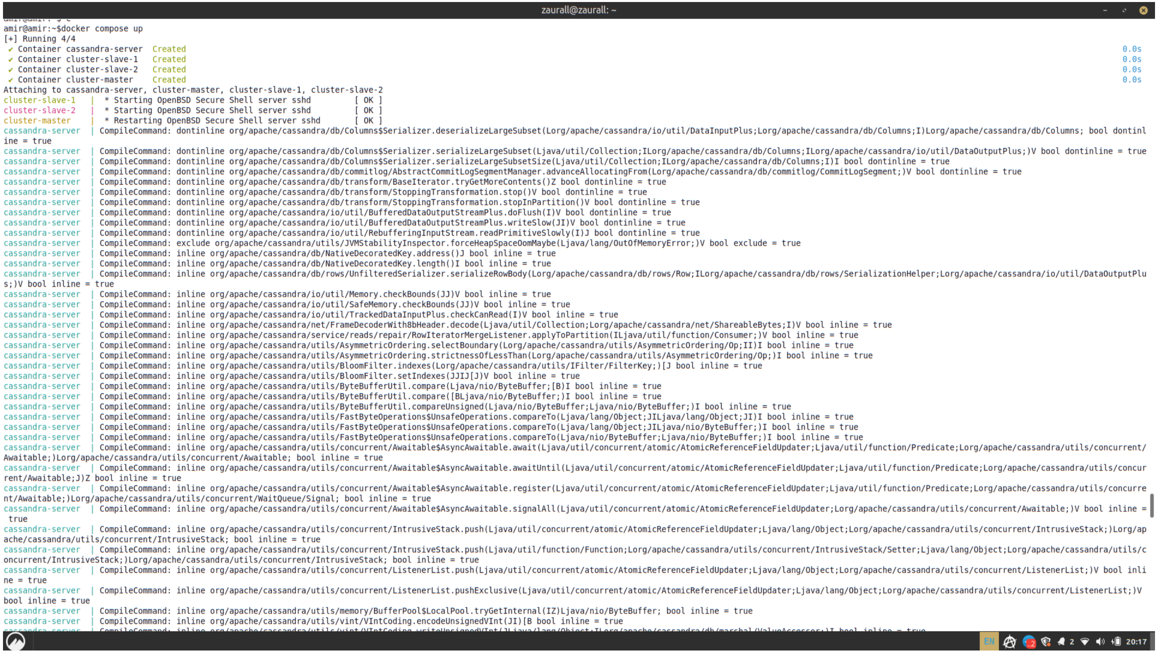
- **docker-compose.yml**: This file sets up and manages a multi-container Docker environment. It launches Hadoop for MapReduce, Spark for preprocessing, and Cassandra for storage, all in a way that's reproducible and easy to move across systems. Each service is configured with the necessary ports, volumes, and dependencies.

- **app.sh**: This shell script runs the full ingestion pipeline. It starts by launching the Docker setup, then runs preprocessing with Spark, and finally ingests the data using app.py.

- **index.sh**: This shell script handles the Hadoop-based MapReduce jobs that build the inverted index from a text corpus stored in HDFS. It takes care of compiling the Java code (if needed), runs the Map and Reduce steps, and saves the final output back into HDFS.

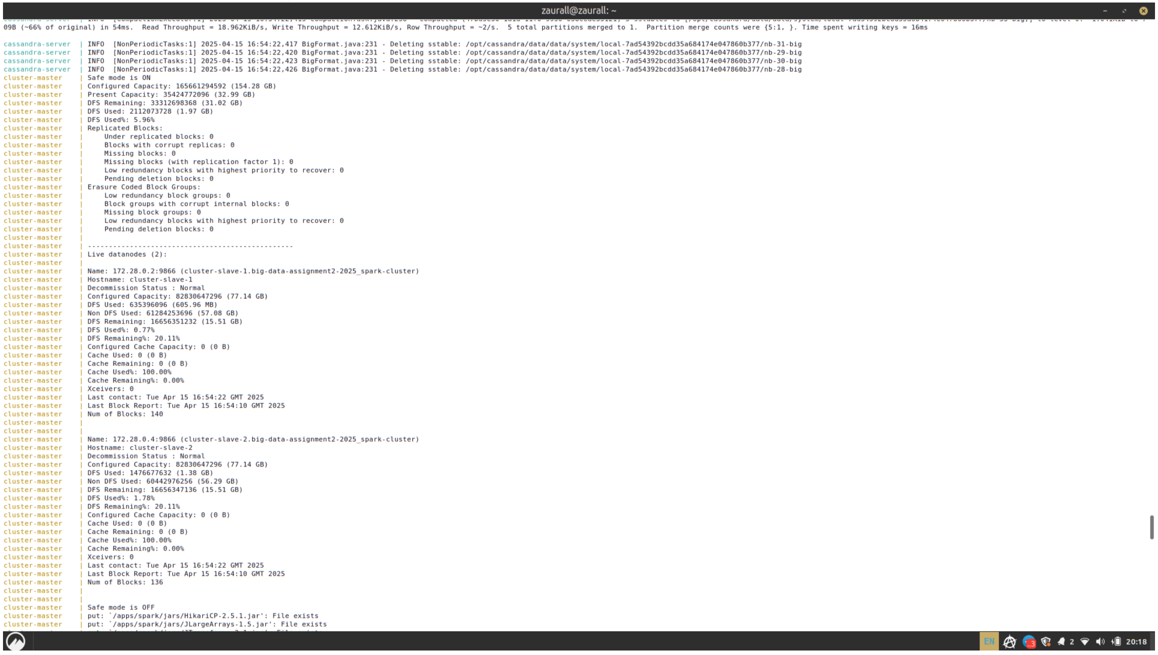
2 Demonstration

Picture 1: In this picture I run docker compose up and you can see that it starts running without errors.



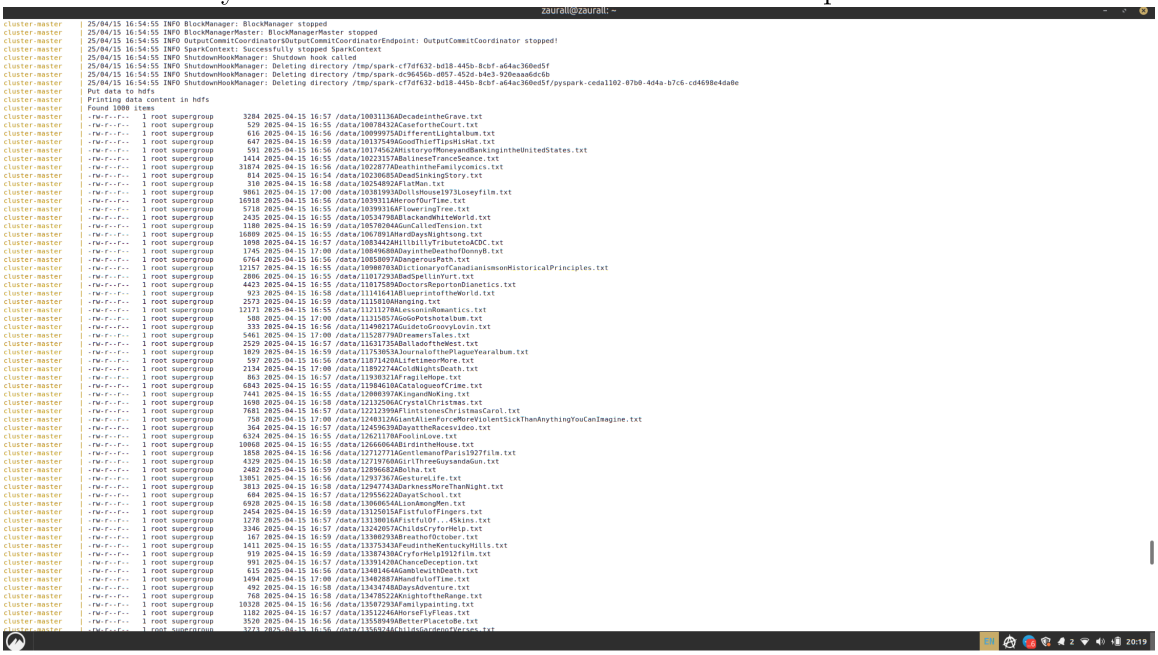
```
amir@amir:~$docker compose up
[+] Running 4/4
 ✔ Container cassandra-server  Created                                                                                                                                   0.0s
 ✔ Container cluster-slave-1   Created                                                                                                                                   0.0s
 ✔ Container cluster-slave-2   Created                                                                                                                                   0.0s
 ✔ Container cluster-master    Created                                                                                                                                   0.0s
Attaching to cassandra-server, cluster-master, cluster-slave-1, cluster-slave-2
cluster-slave-1   |  * Starting OpenBSD Secure Shell server sshd         [ OK ]
cluster-slave-2   |  * Starting OpenBSD Secure Shell server sshd         [ OK ]
cluster-master    |  * Restarting OpenBSD Secure Shell server sshd       [ OK ]
cassandra-server  | CompileCommand: dontinline org/apache/cassandra/db/Columns$Serializer.deserializeLargeSubset(Lorg/apache/cassandra/io/util/DataInputPlus;Lorg/apache/cassandra/db/Columns;I)Lorg/apache/cassandra/db/Columns; bool dontinl
ine = true
cassandra-server  | CompileCommand: dontinline org/apache/cassandra/db/Columns$Serializer.serializeLargeSubset(Ljava/util/Collection;ILorg/apache/cassandra/db/Columns;ILorg/apache/cassandra/io/util/DataOutputPlus;)V bool dontinline = true
cassandra-server  | CompileCommand: dontinline org/apache/cassandra/db/Columns$Serializer.serializeLargeSubsetSize(Ljava/util/Collection;ILorg/apache/cassandra/db/Columns;I)I bool dontinline = true
cassandra-server  | CompileCommand: dontinline org/apache/cassandra/db/commitlog/AbstractCommitLogSegmentManager.advanceAllocatingFrom(Lorg/apache/cassandra/db/commitlog/CommitLogSegment;)V bool dontinline = true
cassandra-server  | CompileCommand: dontinline org/apache/cassandra/db/transform/BaseIterator.tryGetMoreContents()Z bool dontinline = true
cassandra-server  | CompileCommand: dontinline org/apache/cassandra/db/transform/StoppingTransformation.stop()V bool dontinline = true
cassandra-server  | CompileCommand: dontinline org/apache/cassandra/db/transform/StoppingTransformation.stopInPartition()V bool dontinline = true
cassandra-server  | CompileCommand: dontinline org/apache/cassandra/io/util/BufferedDataOutputStreamPlus.doFlush(I)V bool dontinline = true
cassandra-server  | CompileCommand: dontinline org/apache/cassandra/io/util/BufferedDataOutputStreamPlus.writeSlow(JI)V bool dontinline = true
cassandra-server  | CompileCommand: dontinline org/apache/cassandra/io/util/RebufferingInputStream.readPrimitiveSlowly(I)J bool dontinline = true
cassandra-server  | CompileCommand: exclude org/apache/cassandra/utils/JVMStabilityInspector.forceHeapSpaceOomMaybe(Ljava/lang/OutOfMemoryError;)V bool exclude = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/db/NativeDecoratedKey.address()J bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/db/NativeDecoratedKey.length()I bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/db/rows/UnfilteredSerializer.serializeRowBody(Lorg/apache/cassandra/db/rows/Row;ILorg/apache/cassandra/db/rows/SerializationHelper;Lorg/apache/cassandra/io/util/DataOutputPlu
s;)V bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/io/util/Memory.checkBounds(JJ)V bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/io/util/SafeMemory.checkBounds(JJ)V bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/io/util/TrackedDataInputPlus.checkCanRead(I)V bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/net/FrameDecoderWith8bHeader.decode(Ljava/util/Collection;Lorg/apache/cassandra/net/ShareableBytes;I)V bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/service/reads/repair/RowIteratorMergeListener.applyToPartition(ILjava/util/function/Consumer;)V bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/AsymmetricOrdering.selectBoundary(Lorg/apache/cassandra/utils/AsymmetricOrdering/Op;II)I bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/AsymmetricOrdering.strictnessOfLessThan(Lorg/apache/cassandra/utils/AsymmetricOrdering/Op;)I bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/BloomFilter.indexes(Lorg/apache/cassandra/utils/IFilter/FilterKey;)[J bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/BloomFilter.setIndexes(JJIJ[J)V bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/ByteBufferUtil.compare(Ljava/nio/ByteBuffer;[B)I bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/ByteBufferUtil.compare([BLjava/nio/ByteBuffer;)I bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/ByteBufferUtil.compareUnsigned(Ljava/nio/ByteBuffer;Ljava/nio/ByteBuffer;)I bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/FastByteOperations$UnsafeOperations.compareTo(Ljava/lang/Object;JILjava/lang/Object;JI)I bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/FastByteOperations$UnsafeOperations.compareTo(Ljava/lang/Object;JILjava/nio/ByteBuffer;)I bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/FastByteOperations$UnsafeOperations.compareTo(Ljava/nio/ByteBuffer;Ljava/nio/ByteBuffer;)I bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/concurrent/Awaitable$AsyncAwaitable.await(Ljava/util/concurrent/atomic/AtomicReferenceFieldUpdater;Ljava/util/function/Predicate;Lorg/apache/cassandra/utils/concurrent/
Awaitable;)Lorg/apache/cassandra/utils/concurrent/Awaitable; bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/concurrent/Awaitable$AsyncAwaitable.awaitUntil(Ljava/util/concurrent/atomic/AtomicReferenceFieldUpdater;Ljava/util/function/Predicate;Lorg/apache/cassandra/utils/concur
rent/Awaitable;J)Z bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/concurrent/Awaitable$AsyncAwaitable.register(Ljava/util/concurrent/atomic/AtomicReferenceFieldUpdater;Ljava/util/function/Predicate;Lorg/apache/cassandra/utils/concurre
nt/Awaitable;)Lorg/apache/cassandra/utils/concurrent/WaitQueue/Signal; bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/concurrent/Awaitable$AsyncAwaitable.signalAll(Ljava/util/concurrent/atomic/AtomicReferenceFieldUpdater;Lorg/apache/cassandra/utils/concurrent/Awaitable;)V bool inline =
 true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/concurrent/IntrusiveStack.push(Ljava/util/concurrent/atomic/AtomicReferenceFieldUpdater;Ljava/lang/Object;Lorg/apache/cassandra/utils/concurrent/IntrusiveStack;)Lorg/ap
ache/cassandra/utils/concurrent/IntrusiveStack; bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/concurrent/IntrusiveStack.push(Ljava/util/function/Function;Lorg/apache/cassandra/utils/concurrent/IntrusiveStack/Setter;Ljava/lang/Object;Lorg/apache/cassandra/utils/c
oncurrent/IntrusiveStack;)Lorg/apache/cassandra/utils/concurrent/IntrusiveStack; bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/concurrent/ListenerList.push(Ljava/util/concurrent/atomic/AtomicReferenceFieldUpdater;Ljava/lang/Object;Lorg/apache/cassandra/utils/concurrent/ListenerList;)V bool inli
ne = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/concurrent/ListenerList.pushExclusive(Ljava/util/concurrent/atomic/AtomicReferenceFieldUpdater;Ljava/lang/Object;Lorg/apache/cassandra/utils/concurrent/ListenerList;)V
bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/memory/BufferPool$LocalPool.tryGetInternal(IZ)Ljava/nio/ByteBuffer; bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/vint/VIntCoding.encodeUnsignedVInt(JI)[B bool inline = true
cassandra-server  | CompileCommand: inline org/apache/cassandra/utils/vint/VIntCoding.writeUnsignedVInt(JLjava/lang/Object;ILorg/apache/cassandra/db/marshal/ValueAccessor;)I bool inline = true
```

Picture 2: In this picture you can see intermediate stage of working, where you can see cluster-slave 1 and 2, their info and etc.



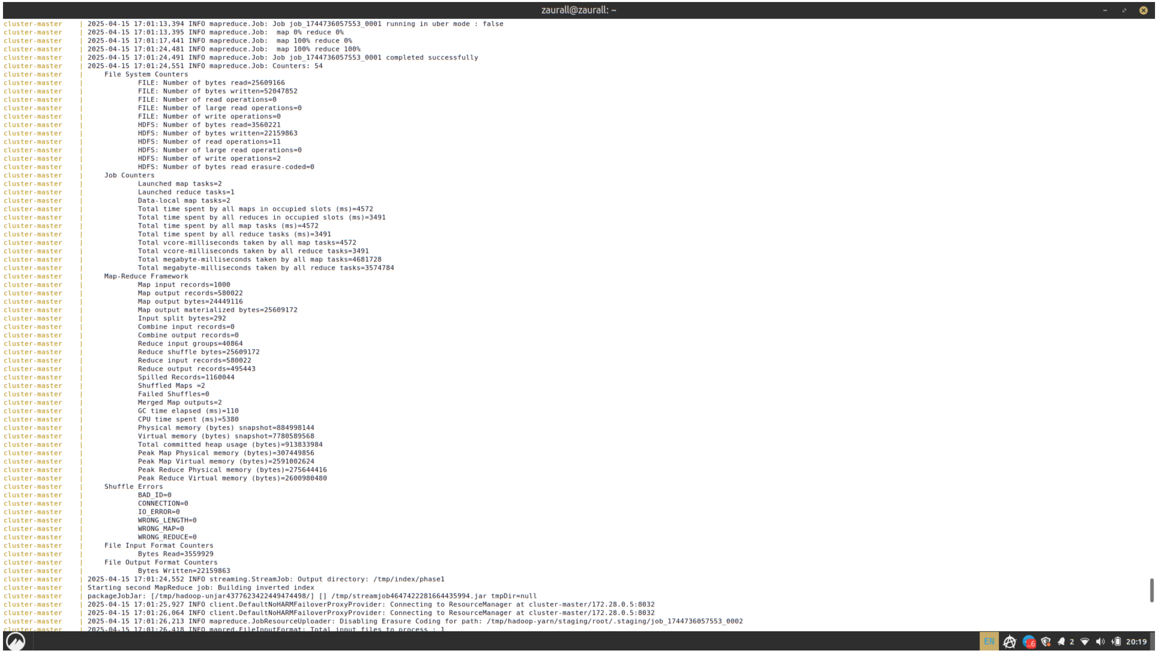
```
09B (~66% of original) in 54ms.  Read Throughput = 18.962KiB/s, Write Throughput = 12.612KiB/s, Row Throughput = ~2/s.  5 total partitions merged to 1.  Partition merge counts were {5:1, }. Time spent writing keys = 16ms

cassandra-server  | INFO  [NonPeriodicTasks:1] 2025-04-15 16:54:22,417 BigFormat.java:231 - Deleting sstable: /opt/cassandra/data/data/system/local-7ad54392bcdd35a684174e047860b377/nb-31-big
cassandra-server  | INFO  [NonPeriodicTasks:1] 2025-04-15 16:54:22,420 BigFormat.java:231 - Deleting sstable: /opt/cassandra/data/data/system/local-7ad54392bcdd35a684174e047860b377/nb-29-big
cassandra-server  | INFO  [NonPeriodicTasks:1] 2025-04-15 16:54:22,423 BigFormat.java:231 - Deleting sstable: /opt/cassandra/data/data/system/local-7ad54392bcdd35a684174e047860b377/nb-30-big
cassandra-server  | INFO  [NonPeriodicTasks:1] 2025-04-15 16:54:22,426 BigFormat.java:231 - Deleting sstable: /opt/cassandra/data/data/system/local-7ad54392bcdd35a684174e047860b377/nb-28-big
cluster-master    | Safe mode is ON
cluster-master    | Configured Capacity: 165661294592 (154.28 GB)
cluster-master    | Present Capacity: 35424772096 (32.99 GB)
cluster-master    | DFS Remaining: 33312698368 (31.02 GB)
cluster-master    | DFS Used: 2112073728 (1.97 GB)
cluster-master    | DFS Used%: 5.96%
cluster-master    | Replicated Blocks:
cluster-master    |     Under replicated blocks: 0
cluster-master    |     Blocks with corrupt replicas: 0
cluster-master    |     Missing blocks: 0
cluster-master    |     Missing blocks (with replication factor 1): 0
cluster-master    |     Low redundancy blocks with highest priority to recover: 0
cluster-master    |     Pending deletion blocks: 0
cluster-master    | Erasure Coded Block Groups:
cluster-master    |     Low redundancy block groups: 0
cluster-master    |     Block groups with corrupt internal blocks: 0
cluster-master    |     Missing block groups: 0
cluster-master    |     Low redundancy blocks with highest priority to recover: 0
cluster-master    |     Pending deletion blocks: 0
cluster-master    |
cluster-master    | -------------------------------------------------
cluster-master    | Live datanodes (2):
cluster-master    |
cluster-master    | Name: 172.28.0.2:9866 (cluster-slave-1.big-data-assignment2-2025_spark-cluster)
cluster-master    | Hostname: cluster-slave-1
cluster-master    | Decommission Status : Normal
cluster-master    | Configured Capacity: 82830647296 (77.14 GB)
cluster-master    | DFS Used: 635396096 (605.96 MB)
cluster-master    | Non DFS Used: 61284253696 (57.08 GB)
cluster-master    | DFS Remaining: 16656351232 (15.51 GB)
cluster-master    | DFS Used%: 0.77%
cluster-master    | DFS Remaining%: 20.11%
cluster-master    | Configured Cache Capacity: 0 (0 B)
cluster-master    | Cache Used: 0 (0 B)
cluster-master    | Cache Remaining: 0 (0 B)
cluster-master    | Cache Used%: 100.00%
cluster-master    | Cache Remaining%: 0.00%
cluster-master    | Xceivers: 0
cluster-master    | Last contact: Tue Apr 15 16:54:22 GMT 2025
cluster-master    | Last Block Report: Tue Apr 15 16:54:10 GMT 2025
cluster-master    | Num of Blocks: 140
cluster-master    |
cluster-master    |
cluster-master    | Name: 172.28.0.4:9866 (cluster-slave-2.big-data-assignment2-2025_spark-cluster)
cluster-master    | Hostname: cluster-slave-2
cluster-master    | Decommission Status : Normal
cluster-master    | Configured Capacity: 82830647296 (77.14 GB)
cluster-master    | DFS Used: 1476677632 (1.38 GB)
cluster-master    | Non DFS Used: 60442976256 (56.29 GB)
cluster-master    | DFS Remaining: 16656347136 (15.51 GB)
cluster-master    | DFS Used%: 1.78%
cluster-master    | DFS Remaining%: 20.11%
cluster-master    | Configured Cache Capacity: 0 (0 B)
cluster-master    | Cache Used: 0 (0 B)
cluster-master    | Cache Remaining: 0 (0 B)
cluster-master    | Cache Used%: 100.00%
cluster-master    | Cache Remaining%: 0.00%
cluster-master    | Xceivers: 0
cluster-master    | Last contact: Tue Apr 15 16:54:22 GMT 2025
cluster-master    | Last Block Report: Tue Apr 15 16:54:10 GMT 2025
cluster-master    | Num of Blocks: 136
cluster-master    |
cluster-master    |
cluster-master    | Safe mode is OFF
cluster-master    | put: `/apps/spark/jars/HikariCP-2.5.1.jar': File exists
cluster-master    | put: `/apps/spark/jars/JLargeArrays-1.5.jar': File exists
```

Picture 3: Here you can see that I took 1000 documents and print the content of hdfs.



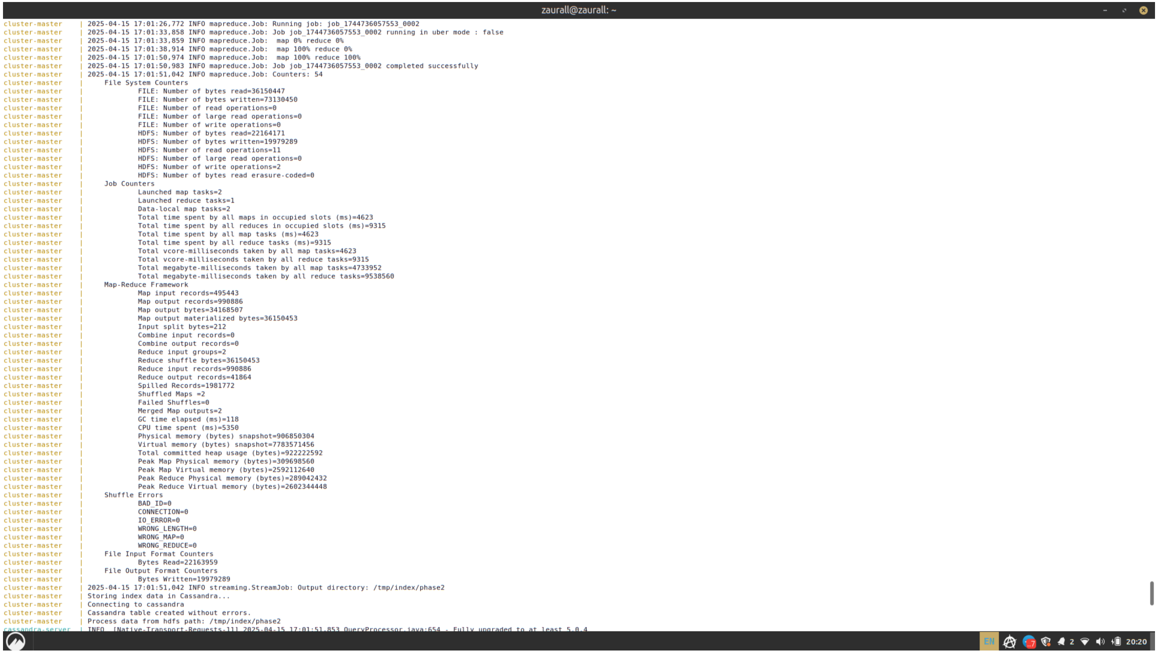
```
cluster-master    | 25/04/15 16:54:55 INFO BlockManager: BlockManager stopped
cluster-master    | 25/04/15 16:54:55 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master    | 25/04/15 16:54:55 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master    | 25/04/15 16:54:55 INFO SparkContext: Successfully stopped SparkContext
cluster-master    | 25/04/15 16:54:55 INFO ShutdownHookManager: Shutdown hook called
cluster-master    | 25/04/15 16:54:55 INFO ShutdownHookManager: Deleting directory /tmp/spark-cf7df632-bd18-445b-8cbf-a64ac360ed5f
cluster-master    | 25/04/15 16:54:55 INFO ShutdownHookManager: Deleting directory /tmp/spark-dc96456b-d057-452d-b4e3-920eaaa6dc6b
cluster-master    | 25/04/15 16:54:55 INFO ShutdownHookManager: Deleting directory /tmp/spark-cf7df632-bd18-445b-8cbf-a64ac360ed5f/pyspark-ceda1102-07b0-4d4a-b7c6-cd4698e4da0e
cluster-master    | Put data to hdfs
cluster-master    | Printing data content in hdfs
cluster-master    | Found 1000 items
cluster-master    | -rw-r--r--   1 root supergroup       3284 2025-04-15 16:57 /data/10031136ADecadeintheGrave.txt
cluster-master    | -rw-r--r--   1 root supergroup        529 2025-04-15 16:55 /data/10078432ACasefortheCourt.txt
cluster-master    | -rw-r--r--   1 root supergroup        616 2025-04-15 16:56 /data/10099975ADifferentLightalbum.txt
cluster-master    | -rw-r--r--   1 root supergroup        647 2025-04-15 16:59 /data/10137549AGoodThiefTipsHisHat.txt
cluster-master    | -rw-r--r--   1 root supergroup        591 2025-04-15 16:56 /data/10174562AHistoryofMoneyandBankingintheUnitedStates.txt
cluster-master    | -rw-r--r--   1 root supergroup       1414 2025-04-15 16:55 /data/10223157ABalineseTranceSeance.txt
cluster-master    | -rw-r--r--   1 root supergroup      31874 2025-04-15 16:56 /data/1022877ADeathintheFamilycomics.txt
cluster-master    | -rw-r--r--   1 root supergroup        814 2025-04-15 16:54 /data/10230685ADeadSinkingStory.txt
cluster-master    | -rw-r--r--   1 root supergroup        310 2025-04-15 16:58 /data/10254892AFlatMan.txt
cluster-master    | -rw-r--r--   1 root supergroup       9861 2025-04-15 17:00 /data/10381993ADollsHouse1973Loseyfilm.txt
cluster-master    | -rw-r--r--   1 root supergroup      16918 2025-04-15 16:56 /data/1039311AHeroofOurTime.txt
cluster-master    | -rw-r--r--   1 root supergroup       5718 2025-04-15 16:55 /data/10399316AFloweringTree.txt
cluster-master    | -rw-r--r--   1 root supergroup       2435 2025-04-15 16:55 /data/10534798ABlackandWhiteWorld.txt
cluster-master    | -rw-r--r--   1 root supergroup       1180 2025-04-15 16:59 /data/10570204AGunCalledTension.txt
cluster-master    | -rw-r--r--   1 root supergroup      16809 2025-04-15 16:55 /data/1067891AHardDaysNightsong.txt
cluster-master    | -rw-r--r--   1 root supergroup       1098 2025-04-15 16:57 /data/1083442AHillbillyTributetoACDC.txt
cluster-master    | -rw-r--r--   1 root supergroup       1745 2025-04-15 17:00 /data/10849680ADayintheDeathofDonnyB.txt
cluster-master    | -rw-r--r--   1 root supergroup       6764 2025-04-15 16:56 /data/10858097ADangerousPath.txt
cluster-master    | -rw-r--r--   1 root supergroup      12157 2025-04-15 16:55 /data/10900703ADictionaryofCanadianismsonHistoricalPrinciples.txt
cluster-master    | -rw-r--r--   1 root supergroup       2806 2025-04-15 16:55 /data/11017293ABadSpellinYurt.txt
cluster-master    | -rw-r--r--   1 root supergroup       4423 2025-04-15 16:55 /data/11017589ADoctorsReportonDianetics.txt
cluster-master    | -rw-r--r--   1 root supergroup        923 2025-04-15 16:58 /data/11141641ABlueprintoftheWorld.txt
cluster-master    | -rw-r--r--   1 root supergroup       2573 2025-04-15 16:59 /data/1115810AHanging.txt
cluster-master    | -rw-r--r--   1 root supergroup      12171 2025-04-15 16:55 /data/11211270ALessoninRomantics.txt
cluster-master    | -rw-r--r--   1 root supergroup        588 2025-04-15 17:00 /data/11315857AGoGoPotshotalbum.txt
cluster-master    | -rw-r--r--   1 root supergroup        333 2025-04-15 16:56 /data/11490217AGuidetoGroovyLovin.txt
cluster-master    | -rw-r--r--   1 root supergroup       5461 2025-04-15 17:00 /data/11528779ADreamersTales.txt
cluster-master    | -rw-r--r--   1 root supergroup       2529 2025-04-15 16:57 /data/11631735ABalladoftheWest.txt
cluster-master    | -rw-r--r--   1 root supergroup       1029 2025-04-15 16:59 /data/11753053AJournalofthePlagueYearalbum.txt
cluster-master    | -rw-r--r--   1 root supergroup        597 2025-04-15 16:56 /data/11871420ALifetimeorMore.txt
cluster-master    | -rw-r--r--   1 root supergroup       2134 2025-04-15 17:00 /data/11892274AColdNightsDeath.txt
cluster-master    | -rw-r--r--   1 root supergroup        863 2025-04-15 16:57 /data/11930321AFragileHope.txt
cluster-master    | -rw-r--r--   1 root supergroup       6843 2025-04-15 16:55 /data/11984610ACatalogueofCrime.txt
cluster-master    | -rw-r--r--   1 root supergroup       7441 2025-04-15 16:55 /data/12000397AKingandNoKing.txt
cluster-master    | -rw-r--r--   1 root supergroup       1698 2025-04-15 16:58 /data/12132506ACrystalChristmas.txt
cluster-master    | -rw-r--r--   1 root supergroup       7681 2025-04-15 16:57 /data/12212399AFlintstonesChristmasCarol.txt
cluster-master    | -rw-r--r--   1 root supergroup        758 2025-04-15 17:00 /data/1240312AGiantAlienForceMoreViolentSickThanAnythingYouCanImagine.txt
cluster-master    | -rw-r--r--   1 root supergroup        364 2025-04-15 16:57 /data/12459639ADayattheRacesvideo.txt
cluster-master    | -rw-r--r--   1 root supergroup       6324 2025-04-15 16:55 /data/12621170AFoolinLove.txt
cluster-master    | -rw-r--r--   1 root supergroup      10068 2025-04-15 16:55 /data/12666064ABirdintheHouse.txt
cluster-master    | -rw-r--r--   1 root supergroup       1858 2025-04-15 16:56 /data/12712771AGentlemanofParis1927film.txt
cluster-master    | -rw-r--r--   1 root supergroup       4329 2025-04-15 16:58 /data/12719760AGirlThreeGuysandaGun.txt
cluster-master    | -rw-r--r--   1 root supergroup       2482 2025-04-15 16:59 /data/12896682ABolha.txt
cluster-master    | -rw-r--r--   1 root supergroup      13051 2025-04-15 16:56 /data/12937367AGestureLife.txt
cluster-master    | -rw-r--r--   1 root supergroup       3813 2025-04-15 16:58 /data/12947743ADarknessMoreThanNight.txt
cluster-master    | -rw-r--r--   1 root supergroup        604 2025-04-15 16:57 /data/12955622ADayatSchool.txt
cluster-master    | -rw-r--r--   1 root supergroup       6928 2025-04-15 16:58 /data/13060654ALionAmongMen.txt
cluster-master    | -rw-r--r--   1 root supergroup       2454 2025-04-15 16:59 /data/13125015AFistfulofFingers.txt
cluster-master    | -rw-r--r--   1 root supergroup       1278 2025-04-15 16:57 /data/13130016AFistfulOf...4Skins.txt
cluster-master    | -rw-r--r--   1 root supergroup       3346 2025-04-15 16:57 /data/13242057AChildsCryforHelp.txt
cluster-master    | -rw-r--r--   1 root supergroup        167 2025-04-15 16:59 /data/13300293ABreathofOctober.txt
cluster-master    | -rw-r--r--   1 root supergroup       1411 2025-04-15 16:55 /data/13375343AFeudintheKentuckyHills.txt
cluster-master    | -rw-r--r--   1 root supergroup        919 2025-04-15 16:59 /data/13387430ACryforHelp1912film.txt
cluster-master    | -rw-r--r--   1 root supergroup        991 2025-04-15 16:57 /data/13391420AChanceDeception.txt
cluster-master    | -rw-r--r--   1 root supergroup        615 2025-04-15 16:56 /data/13401464AGamblewithDeath.txt
cluster-master    | -rw-r--r--   1 root supergroup       1494 2025-04-15 17:00 /data/13402887AHandfulofTime.txt
cluster-master    | -rw-r--r--   1 root supergroup        492 2025-04-15 16:58 /data/13434748ADaysAdventure.txt
cluster-master    | -rw-r--r--   1 root supergroup        768 2025-04-15 16:58 /data/13478522AKnightoftheRange.txt
cluster-master    | -rw-r--r--   1 root supergroup      10328 2025-04-15 16:56 /data/13507293AFamilypainting.txt
cluster-master    | -rw-r--r--   1 root supergroup       1182 2025-04-15 16:57 /data/13512246AHorseFlyFleas.txt
cluster-master    | -rw-r--r--   1 root supergroup       3520 2025-04-15 16:56 /data/13558949ABetterPlacetoBe.txt
cluster-master    | -rw-r--r--   1 root supergroup       3273 2025-04-15 16:56 /data/1356924AChildsGardenofVerses.txt
```

Picture 4: Here you can see my mapping and reducing jobs, they are successful and result of their job is printed.



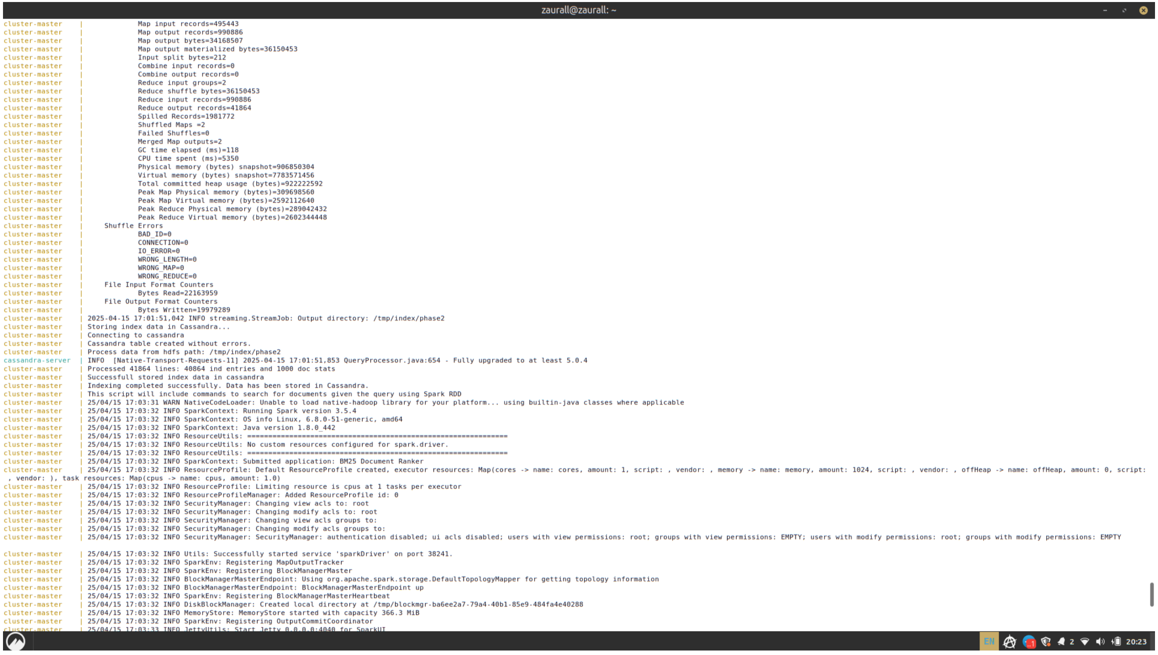
```
cluster-master    | 2025-04-15 17:01:13,394 INFO mapreduce.Job: Job job_1744736057553_0001 running in uber mode : false
cluster-master    | 2025-04-15 17:01:13,395 INFO mapreduce.Job:  map 0% reduce 0%
cluster-master    | 2025-04-15 17:01:17,441 INFO mapreduce.Job:  map 100% reduce 0%
cluster-master    | 2025-04-15 17:01:24,481 INFO mapreduce.Job:  map 100% reduce 100%
cluster-master    | 2025-04-15 17:01:24,491 INFO mapreduce.Job: Job job_1744736057553_0001 completed successfully
cluster-master    | 2025-04-15 17:01:24,551 INFO mapreduce.Job: Counters: 54
cluster-master    |     File System Counters
cluster-master    |             FILE: Number of bytes read=25609166
cluster-master    |             FILE: Number of bytes written=52047852
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=3560221
cluster-master    |             HDFS: Number of bytes written=22159863
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)=4572
cluster-master    |             Total time spent by all reduces in occupied slots (ms)=3491
cluster-master    |             Total time spent by all map tasks (ms)=4572
cluster-master    |             Total time spent by all reduce tasks (ms)=3491
cluster-master    |             Total vcore-milliseconds taken by all map tasks=4572
cluster-master    |             Total vcore-milliseconds taken by all reduce tasks=3491
cluster-master    |             Total megabyte-milliseconds taken by all map tasks=4681728
cluster-master    |             Total megabyte-milliseconds taken by all reduce tasks=3574784
cluster-master    |     Map-Reduce Framework
cluster-master    |             Map input records=1000
cluster-master    |             Map output records=580022
cluster-master    |             Map output bytes=24449116
cluster-master    |             Map output materialized bytes=25609172
cluster-master    |             Input split bytes=292
cluster-master    |             Combine input records=0
cluster-master    |             Combine output records=0
cluster-master    |             Reduce input groups=40864
cluster-master    |             Reduce shuffle bytes=25609172
cluster-master    |             Reduce input records=580022
cluster-master    |             Reduce output records=495443
cluster-master    |             Spilled Records=1160044
cluster-master    |             Shuffled Maps =2
cluster-master    |             Failed Shuffles=0
cluster-master    |             Merged Map outputs=2
cluster-master    |             GC time elapsed (ms)=110
cluster-master    |             CPU time spent (ms)=5380
cluster-master    |             Physical memory (bytes) snapshot=884998144
cluster-master    |             Virtual memory (bytes) snapshot=7780589568
cluster-master    |             Total committed heap usage (bytes)=913833984
cluster-master    |             Peak Map Physical memory (bytes)=307449856
cluster-master    |             Peak Map Virtual memory (bytes)=2591002624
cluster-master    |             Peak Reduce Physical memory (bytes)=275644416
cluster-master    |             Peak Reduce Virtual memory (bytes)=2600980480
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=3559929
cluster-master    |     File Output Format Counters
cluster-master    |             Bytes Written=22159863
cluster-master    | 2025-04-15 17:01:24,552 INFO streaming.StreamJob: Output directory: /tmp/index/phase1
cluster-master    | Starting second MapReduce job: Building inverted index
cluster-master    | packageJobJar: [/tmp/hadoop-unjar4377623422449474498/] [] /tmp/streamjob4647422281664435994.jar tmpDir=null
cluster-master    | 2025-04-15 17:01:25,927 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.28.0.5:8032
cluster-master    | 2025-04-15 17:01:26,064 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.28.0.5:8032
cluster-master    | 2025-04-15 17:01:26,213 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744736057553_0002
cluster-master    | 2025-04-15 17:01:26,418 INFO mapred.FileInputFormat: Total input files to process : 1
```

Picture 5: Here you can see my second mapping and reducing jobs, they are successful and result of their job is also printed.



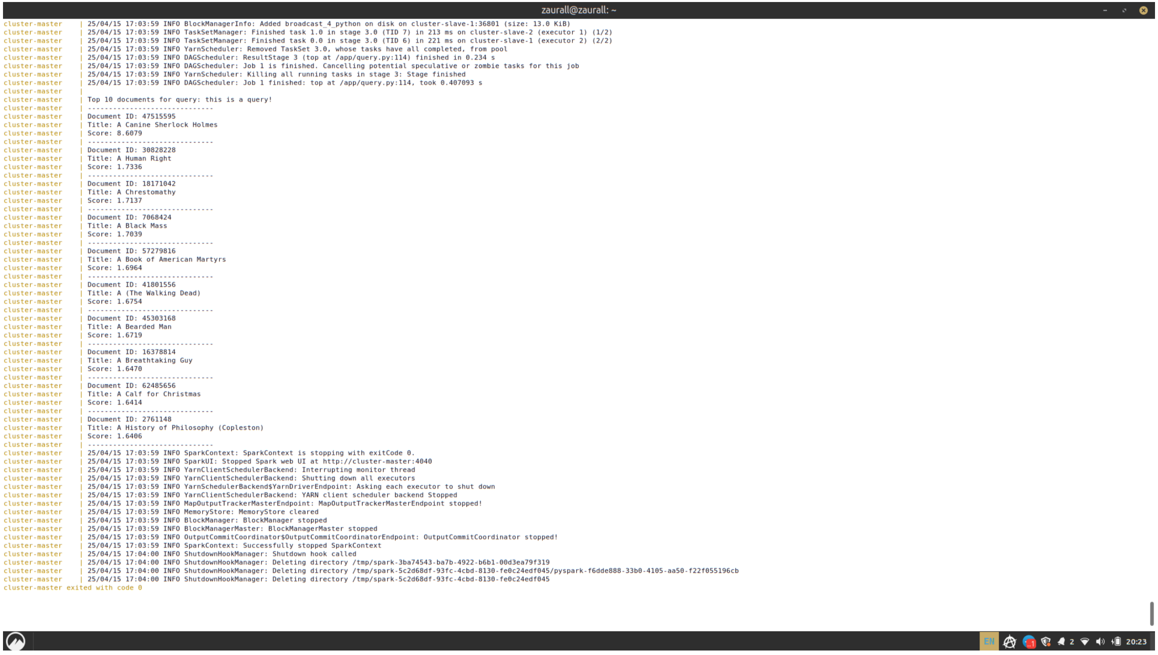
```
cluster-master    | 2025-04-15 17:01:26,772 INFO mapreduce.Job: Running job: job_1744736057553_0002
cluster-master    | 2025-04-15 17:01:33,858 INFO mapreduce.Job: Job job_1744736057553_0002 running in uber mode : false
cluster-master    | 2025-04-15 17:01:33,859 INFO mapreduce.Job:  map 0% reduce 0%
cluster-master    | 2025-04-15 17:01:38,914 INFO mapreduce.Job:  map 100% reduce 0%
cluster-master    | 2025-04-15 17:01:50,974 INFO mapreduce.Job:  map 100% reduce 100%
cluster-master    | 2025-04-15 17:01:50,983 INFO mapreduce.Job: Job job_1744736057553_0002 completed successfully
cluster-master    | 2025-04-15 17:01:51,042 INFO mapreduce.Job: Counters: 54
cluster-master    |     File System Counters
cluster-master    |             FILE: Number of bytes read=36150447
cluster-master    |             FILE: Number of bytes written=73130450
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=22164171
cluster-master    |             HDFS: Number of bytes written=19979289
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)=4623
cluster-master    |             Total time spent by all reduces in occupied slots (ms)=9315
cluster-master    |             Total time spent by all map tasks (ms)=4623
cluster-master    |             Total time spent by all reduce tasks (ms)=9315
cluster-master    |             Total vcore-milliseconds taken by all map tasks=4623
cluster-master    |             Total vcore-milliseconds taken by all reduce tasks=9315
cluster-master    |             Total megabyte-milliseconds taken by all map tasks=4733952
cluster-master    |             Total megabyte-milliseconds taken by all reduce tasks=9538560
cluster-master    |     Map-Reduce Framework
cluster-master    |             Map input records=495443
cluster-master    |             Map output records=990886
cluster-master    |             Map output bytes=34168507
cluster-master    |             Map output materialized bytes=36150453
cluster-master    |             Input split bytes=212
cluster-master    |             Combine input records=0
cluster-master    |             Combine output records=0
cluster-master    |             Reduce input groups=2
cluster-master    |             Reduce shuffle bytes=36150453
cluster-master    |             Reduce input records=990886
cluster-master    |             Reduce output records=41864
cluster-master    |             Spilled Records=1981772
cluster-master    |             Shuffled Maps =2
cluster-master    |             Failed Shuffles=0
cluster-master    |             Merged Map outputs=2
cluster-master    |             GC time elapsed (ms)=118
cluster-master    |             CPU time spent (ms)=5350
cluster-master    |             Physical memory (bytes) snapshot=906850304
cluster-master    |             Virtual memory (bytes) snapshot=7783571456
cluster-master    |             Total committed heap usage (bytes)=922222592
cluster-master    |             Peak Map Physical memory (bytes)=309698560
cluster-master    |             Peak Map Virtual memory (bytes)=2592112640
cluster-master    |             Peak Reduce Physical memory (bytes)=289042432
cluster-master    |             Peak Reduce Virtual memory (bytes)=2602344448
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=22163959
cluster-master    |     File Output Format Counters
cluster-master    |             Bytes Written=19979289
cluster-master    | 2025-04-15 17:01:51,042 INFO streaming.StreamJob: Output directory: /tmp/index/phase2
cluster-master    | Storing index data in Cassandra...
cluster-master    | Connecting to cassandra
cluster-master    | Cassandra table created without errors.
cluster-master    | Process data from hdfs path: /tmp/index/phase2
cassandra-server  | INFO  [Native-Transport-Requests-1]] 2025-04-15 17:01:51,853 QueryProcessor.java:654 - Fully upgraded to at least 5.0.4
```

Picture 6: Here you can see that I connect to Cassandra server and indexing completed successfully and data stored in Cassandra.



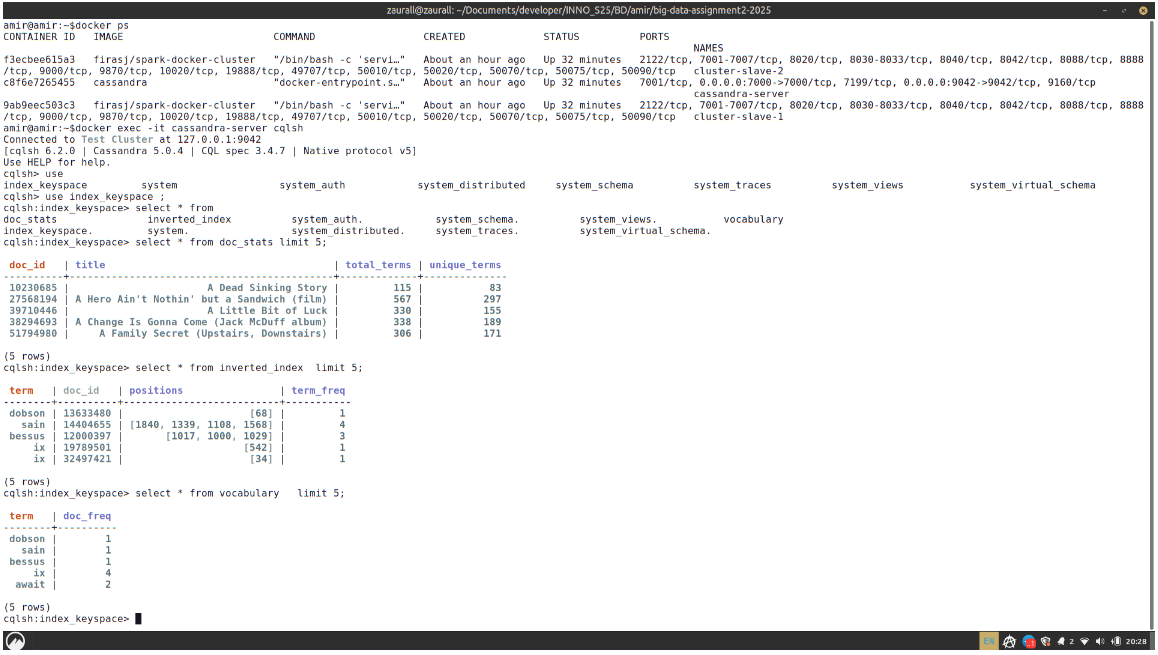
```
cluster-master    |             Map input records=495443
cluster-master    |             Map output records=990886
cluster-master    |             Map output bytes=34168507
cluster-master    |             Map output materialized bytes=36150453
cluster-master    |             Input split bytes=212
cluster-master    |             Combine input records=0
cluster-master    |             Combine output records=0
cluster-master    |             Reduce input groups=2
cluster-master    |             Reduce shuffle bytes=36150453
cluster-master    |             Reduce input records=990886
cluster-master    |             Reduce output records=41864
cluster-master    |             Spilled Records=1981772
cluster-master    |             Shuffled Maps =2
cluster-master    |             Failed Shuffles=0
cluster-master    |             Merged Map outputs=2
cluster-master    |             GC time elapsed (ms)=118
cluster-master    |             CPU time spent (ms)=5350
cluster-master    |             Physical memory (bytes) snapshot=906850304
cluster-master    |             Virtual memory (bytes) snapshot=7783571456
cluster-master    |             Total committed heap usage (bytes)=922222592
cluster-master    |             Peak Map Physical memory (bytes)=309698560
cluster-master    |             Peak Map Virtual memory (bytes)=2592112640
cluster-master    |             Peak Reduce Physical memory (bytes)=289042432
cluster-master    |             Peak Reduce Virtual memory (bytes)=2602344448
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=22163959
cluster-master    |     File Output Format Counters
cluster-master    |             Bytes Written=19979289
cluster-master    | 2025-04-15 17:01:51,042 INFO streaming.StreamJob: Output directory: /tmp/index/phase2
cluster-master    | Storing index data in Cassandra...
cluster-master    | Connecting to cassandra
cluster-master    | Cassandra table created without errors.
cluster-master    | Process data from hdfs path: /tmp/index/phase2
cassandra-server  | INFO  [Native-Transport-Requests-11] 2025-04-15 17:01:51,853 QueryProcessor.java:654 - Fully upgraded to at least 5.0.4
cluster-master    | Processed 41864 lines: 40864 ind entries and 1000 doc stats
cluster-master    | Successfull stored index data in cassandra
cluster-master    | Indexing completed successfully. Data has been stored in Cassandra.
cluster-master    | This script will include commands to search for documents given the query using Spark RDD
cluster-master    | 25/04/15 17:03:31 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
cluster-master    | 25/04/15 17:03:32 INFO SparkContext: Running Spark version 3.5.4
cluster-master    | 25/04/15 17:03:32 INFO SparkContext: OS info Linux, 6.8.0-51-generic, amd64
cluster-master    | 25/04/15 17:03:32 INFO SparkContext: Java version 1.8.0_442
cluster-master    | 25/04/15 17:03:32 INFO ResourceUtils: ==============================================================
cluster-master    | 25/04/15 17:03:32 INFO ResourceUtils: No custom resources configured for spark.driver.
cluster-master    | 25/04/15 17:03:32 INFO ResourceUtils: ==============================================================
cluster-master    | 25/04/15 17:03:32 INFO SparkContext: Submitted application: BM25 Document Ranker
cluster-master    | 25/04/15 17:03:32 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script:
 , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
cluster-master    | 25/04/15 17:03:32 INFO ResourceProfile: Limiting resource is cpus at 1 tasks per executor
cluster-master    | 25/04/15 17:03:32 INFO ResourceProfileManager: Added ResourceProfile id: 0
cluster-master    | 25/04/15 17:03:32 INFO SecurityManager: Changing view acls to: root
cluster-master    | 25/04/15 17:03:32 INFO SecurityManager: Changing modify acls to: root
cluster-master    | 25/04/15 17:03:32 INFO SecurityManager: Changing view acls groups to:
cluster-master    | 25/04/15 17:03:32 INFO SecurityManager: Changing modify acls groups to:
cluster-master    | 25/04/15 17:03:32 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    | 25/04/15 17:03:32 INFO Utils: Successfully started service 'sparkDriver' on port 38241.
cluster-master    | 25/04/15 17:03:32 INFO SparkEnv: Registering MapOutputTracker
cluster-master    | 25/04/15 17:03:32 INFO SparkEnv: Registering BlockManagerMaster
cluster-master    | 25/04/15 17:03:32 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
cluster-master    | 25/04/15 17:03:32 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
cluster-master    | 25/04/15 17:03:32 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
cluster-master    | 25/04/15 17:03:32 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-ba6ee2a7-79a4-40b1-85e9-484fa4e40288
cluster-master    | 25/04/15 17:03:32 INFO MemoryStore: MemoryStore started with capacity 366.3 MiB
cluster-master    | 25/04/15 17:03:32 INFO SparkEnv: Registering OutputCommitCoordinator
cluster-master    | 25/04/15 17:03:33 INFO JettyUtils: Start Jetty 0.0.0.0:4040 for SparkUI
```

Picture 7: Here you can see the result of the first query.



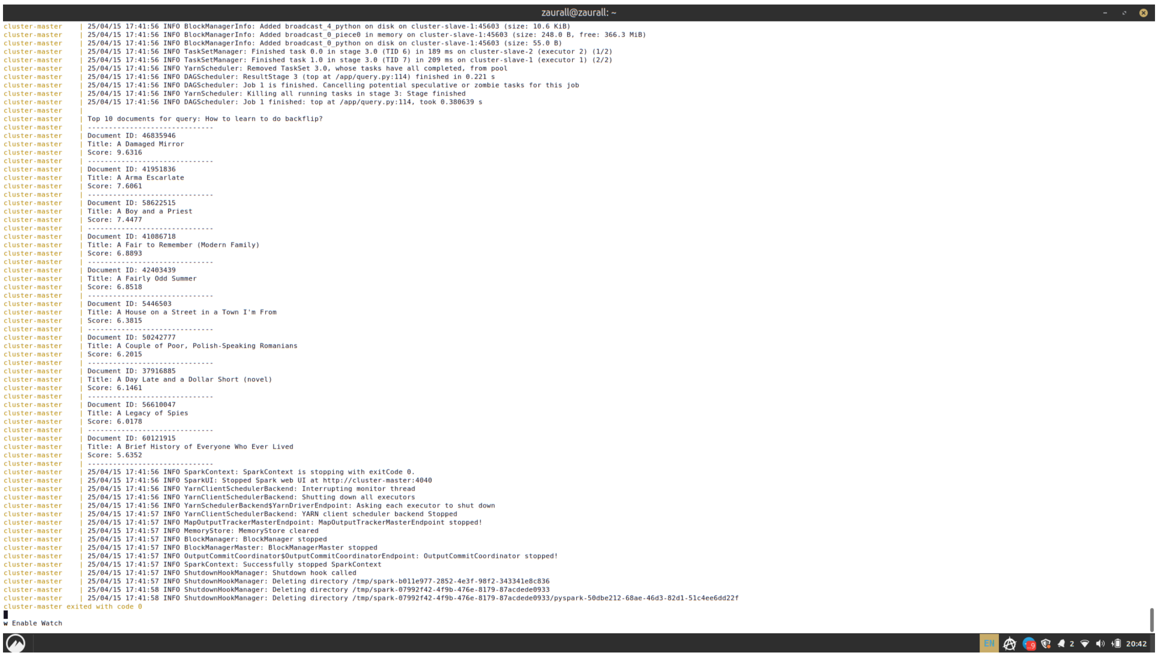
```
cluster-master    | 25/04/15 17:03:59 INFO BlockManagerInfo: Added broadcast_4_python on disk on cluster-slave-1:36801 (size: 13.0 KiB)
cluster-master    | 25/04/15 17:03:59 INFO TaskSetManager: Finished task 1.0 in stage 3.0 (TID 7) in 213 ms on cluster-slave-2 (executor 1) (1/2)
cluster-master    | 25/04/15 17:03:59 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 6) in 221 ms on cluster-slave-1 (executor 2) (2/2)
cluster-master    | 25/04/15 17:03:59 INFO YarnScheduler: Removed TaskSet 3.0, whose tasks have all completed, from pool
cluster-master    | 25/04/15 17:03:59 INFO DAGScheduler: ResultStage 3 (top at /app/query.py:114) finished in 0.234 s
cluster-master    | 25/04/15 17:03:59 INFO DAGScheduler: Job 1 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master    | 25/04/15 17:03:59 INFO YarnScheduler: Killing all running tasks in stage 3: Stage finished
cluster-master    | 25/04/15 17:03:59 INFO DAGScheduler: Job 1 finished: top at /app/query.py:114, took 0.407093 s
cluster-master    |
cluster-master    | Top 10 documents for query: this is a query!
cluster-master    | ------------------------------
cluster-master    | Document ID: 47515595
cluster-master    | Title: A Canine Sherlock Holmes
cluster-master    | Score: 8.6079
cluster-master    | ------------------------------
cluster-master    | Document ID: 30828228
cluster-master    | Title: A Human Right
cluster-master    | Score: 1.7336
cluster-master    | ------------------------------
cluster-master    | Document ID: 18171042
cluster-master    | Title: A Chrestomathy
cluster-master    | Score: 1.7137
cluster-master    | ------------------------------
cluster-master    | Document ID: 7068424
cluster-master    | Title: A Black Mass
cluster-master    | Score: 1.7039
cluster-master    | ------------------------------
cluster-master    | Document ID: 57279816
cluster-master    | Title: A Book of American Martyrs
cluster-master    | Score: 1.6964
cluster-master    | ------------------------------
cluster-master    | Document ID: 41801556
cluster-master    | Title: A (The Walking Dead)
cluster-master    | Score: 1.6754
cluster-master    | ------------------------------
cluster-master    | Document ID: 45303168
cluster-master    | Title: A Bearded Man
cluster-master    | Score: 1.6719
cluster-master    | ------------------------------
cluster-master    | Document ID: 16378814
cluster-master    | Title: A Breathtaking Guy
cluster-master    | Score: 1.6470
cluster-master    | ------------------------------
cluster-master    | Document ID: 62485656
cluster-master    | Title: A Calf for Christmas
cluster-master    | Score: 1.6414
cluster-master    | ------------------------------
cluster-master    | Document ID: 2761148
cluster-master    | Title: A History of Philosophy (Copleston)
cluster-master    | Score: 1.6406
cluster-master    | ------------------------------
cluster-master    | 25/04/15 17:03:59 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master    | 25/04/15 17:03:59 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master    | 25/04/15 17:03:59 INFO YarnClientSchedulerBackend: Interrupting monitor thread
cluster-master    | 25/04/15 17:03:59 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster-master    | 25/04/15 17:03:59 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
cluster-master    | 25/04/15 17:03:59 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
cluster-master    | 25/04/15 17:03:59 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master    | 25/04/15 17:03:59 INFO MemoryStore: MemoryStore cleared
cluster-master    | 25/04/15 17:03:59 INFO BlockManager: BlockManager stopped
cluster-master    | 25/04/15 17:03:59 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master    | 25/04/15 17:03:59 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master    | 25/04/15 17:03:59 INFO SparkContext: Successfully stopped SparkContext
cluster-master    | 25/04/15 17:04:00 INFO ShutdownHookManager: Shutdown hook called
cluster-master    | 25/04/15 17:04:00 INFO ShutdownHookManager: Deleting directory /tmp/spark-3ba74543-ba7b-4922-b6b1-00d3ea79f319
cluster-master    | 25/04/15 17:04:00 INFO ShutdownHookManager: Deleting directory /tmp/spark-5c2d68df-93fc-4cbd-8130-fe0c24edf045/pyspark-f6dde888-33b0-4105-aa50-f22f055196cb
cluster-master    | 25/04/15 17:04:00 INFO ShutdownHookManager: Deleting directory /tmp/spark-5c2d68df-93fc-4cbd-8130-fe0c24edf045
cluster-master exited with code 0
```

Picture 8: Here you can see my tables in Cassandra. I have 3 tables, as I described in the methodology. And you can see how they look like.


zaurall@zaurall: ~/Documents/developer/INNO_S25/BD/amir/big-data-assignment2-2025
EN
20:28

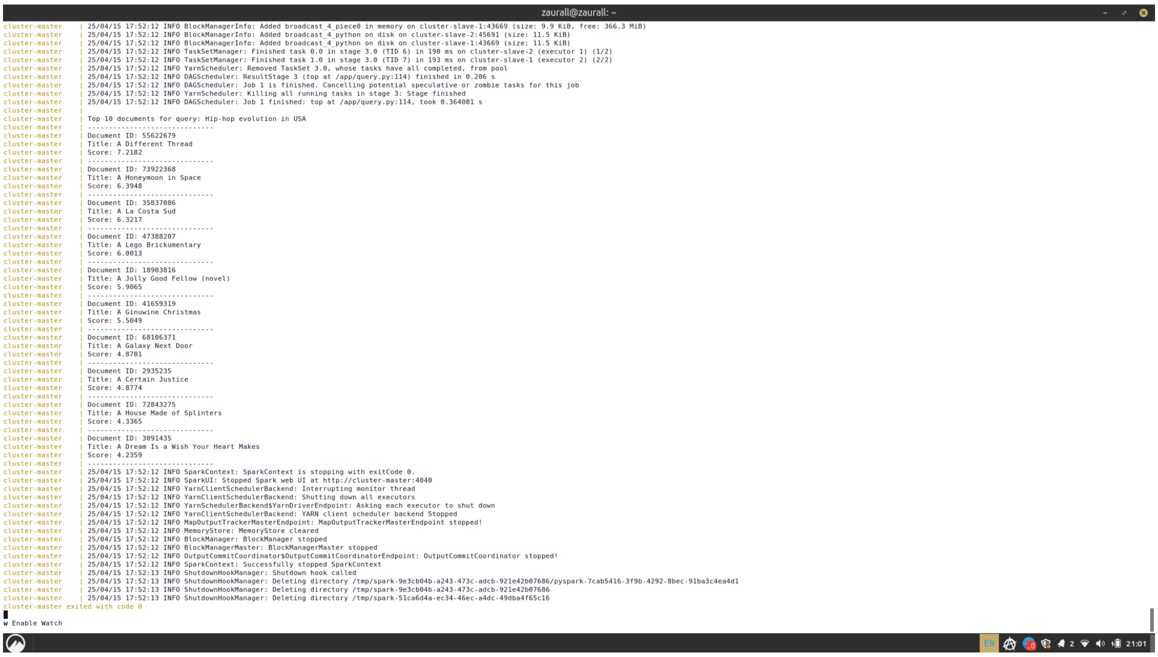
```
amir@amir:~$docker ps
CONTAINER ID   IMAGE                          COMMAND                  CREATED             STATUS          PORTS
                                                                                                                            NAMES
f3ecbee615a3   firasj/spark-docker-cluster   "/bin/bash -c 'servi…"   About an hour ago   Up 32 minutes   2122/tcp, 7001-7007/tcp, 8020/tcp, 8030-8033/tcp, 8040/tcp, 8042/tcp, 8088/tcp, 8888
/tcp, 9000/tcp, 9870/tcp, 10020/tcp, 19888/tcp, 49707/tcp, 50010/tcp, 50020/tcp, 50070/tcp, 50075/tcp, 50090/tcp   cluster-slave-2
c8f6e7265455   cassandra                      "docker-entrypoint.s…"   About an hour ago   Up 32 minutes   7001/tcp, 0.0.0.0:7000->7000/tcp, 7199/tcp, 0.0.0.0:9042->9042/tcp, 9160/tcp
                                                                                                                            cassandra-server
9ab9eec503c3   firasj/spark-docker-cluster   "/bin/bash -c 'servi…"   About an hour ago   Up 32 minutes   2122/tcp, 7001-7007/tcp, 8020/tcp, 8030-8033/tcp, 8040/tcp, 8042/tcp, 8088/tcp, 8888
/tcp, 9000/tcp, 9870/tcp, 10020/tcp, 19888/tcp, 49707/tcp, 50010/tcp, 50020/tcp, 50070/tcp, 50075/tcp, 50090/tcp   cluster-slave-1
amir@amir:~$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> use
index_keyspace        system                 system_auth            system_distributed     system_schema          system_traces          system_views           system_virtual_schema
cqlsh> use index_keyspace ;
cqlsh:index_keyspace> select * from
doc_stats               inverted_index          system_auth.            system_schema.          system_views.           vocabulary
index_keyspace.         system.                 system_distributed.     system_traces.          system_virtual_schema.
cqlsh:index_keyspace> select * from doc_stats limit 5;

 doc_id   | title                                      | total_terms | unique_terms
----------+--------------------------------------------+-------------+--------------
 10230685 |                        A Dead Sinking Story |         115 |           83
 27568194 | A Hero Ain't Nothin' but a Sandwich (film) |         567 |          297
 39710446 |                        A Little Bit of Luck |         330 |          155
 38294693 | A Change Is Gonna Come (Jack McDuff album) |         338 |          189
 51794980 |     A Family Secret (Upstairs, Downstairs) |         306 |          171

(5 rows)
cqlsh:index_keyspace> select * from inverted_index  limit 5;

 term   | doc_id   | positions                | term_freq
--------+----------+--------------------------+-----------
 dobson | 13633480 |                     [68] |         1
   sain | 14404655 | [1840, 1339, 1108, 1568] |         4
 bessus | 12000397 |       [1017, 1000, 1029] |         3
     ix | 19789501 |                    [542] |         1
     ix | 32497421 |                     [34] |         1

(5 rows)
cqlsh:index_keyspace> select * from vocabulary   limit 5;

 term   | doc_freq
--------+----------
 dobson |        1
   sain |        1
 bessus |        1
     ix |        4
  await |        2

(5 rows)
cqlsh:index_keyspace> █
```

Picture 9: Here you can see the result of the second (custom) query, that I wrote by myself.



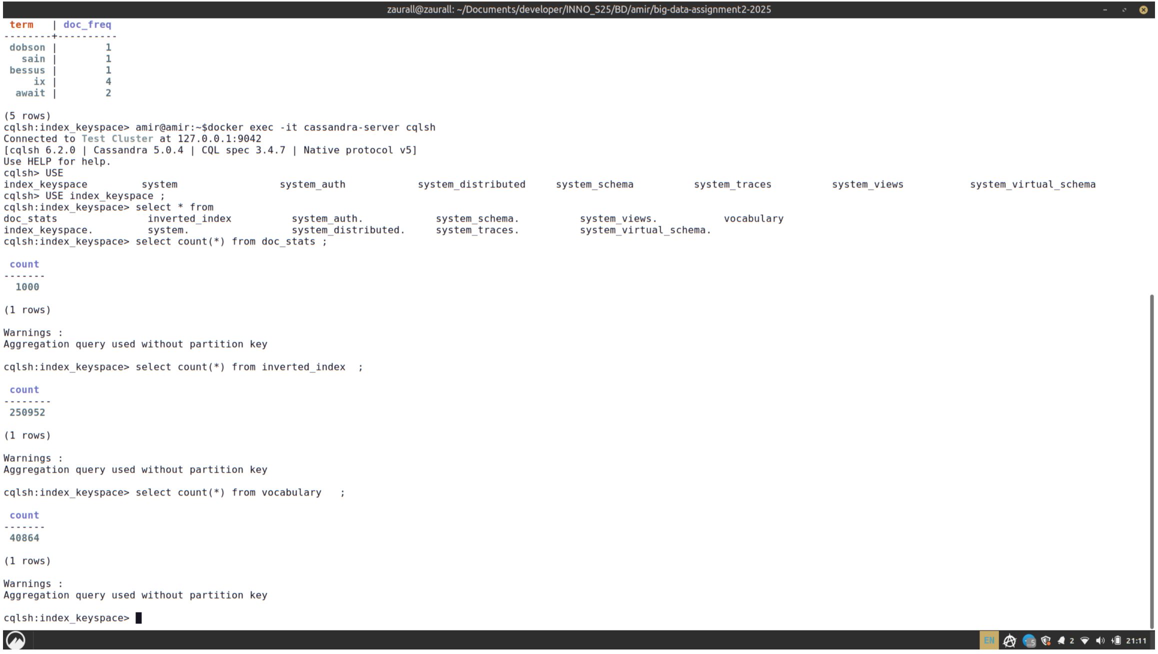
```
cluster-master    | 25/04/15 17:41:56 INFO BlockManagerInfo: Added broadcast_4_python on disk on cluster-slave-1:45603 (size: 10.6 KiB)
cluster-master    | 25/04/15 17:41:56 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on cluster-slave-1:45603 (size: 248.0 B, free: 366.3 MiB)
cluster-master    | 25/04/15 17:41:56 INFO BlockManagerInfo: Added broadcast_0_python on disk on cluster-slave-1:45603 (size: 55.0 B)
cluster-master    | 25/04/15 17:41:56 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 6) in 189 ms on cluster-slave-2 (executor 2) (1/2)
cluster-master    | 25/04/15 17:41:56 INFO TaskSetManager: Finished task 1.0 in stage 3.0 (TID 7) in 209 ms on cluster-slave-1 (executor 1) (2/2)
cluster-master    | 25/04/15 17:41:56 INFO YarnScheduler: Removed TaskSet 3.0, whose tasks have all completed, from pool
cluster-master    | 25/04/15 17:41:56 INFO DAGScheduler: ResultStage 3 (top at /app/query.py:114) finished in 0.221 s
cluster-master    | 25/04/15 17:41:56 INFO DAGScheduler: Job 1 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master    | 25/04/15 17:41:56 INFO YarnScheduler: Killing all running tasks in stage 3: Stage finished
cluster-master    | 25/04/15 17:41:56 INFO DAGScheduler: Job 1 finished: top at /app/query.py:114, took 0.380639 s
cluster-master    |
cluster-master    | Top 10 documents for query: How to learn to do backflip?
cluster-master    | ------------------------------
cluster-master    | Document ID: 46835946
cluster-master    | Title: A Damaged Mirror
cluster-master    | Score: 9.6316
cluster-master    | ------------------------------
cluster-master    | Document ID: 41951836
cluster-master    | Title: A Arma Escarlate
cluster-master    | Score: 7.6061
cluster-master    | ------------------------------
cluster-master    | Document ID: 58622515
cluster-master    | Title: A Boy and a Priest
cluster-master    | Score: 7.4477
cluster-master    | ------------------------------
cluster-master    | Document ID: 41086718
cluster-master    | Title: A Fair to Remember (Modern Family)
cluster-master    | Score: 6.8893
cluster-master    | ------------------------------
cluster-master    | Document ID: 42403439
cluster-master    | Title: A Fairly Odd Summer
cluster-master    | Score: 6.8518
cluster-master    | ------------------------------
cluster-master    | Document ID: 5446503
cluster-master    | Title: A House on a Street in a Town I'm From
cluster-master    | Score: 6.3815
cluster-master    | ------------------------------
cluster-master    | Document ID: 50242777
cluster-master    | Title: A Couple of Poor, Polish-Speaking Romanians
cluster-master    | Score: 6.2015
cluster-master    | ------------------------------
cluster-master    | Document ID: 37916885
cluster-master    | Title: A Day Late and a Dollar Short (novel)
cluster-master    | Score: 6.1461
cluster-master    | ------------------------------
cluster-master    | Document ID: 56610047
cluster-master    | Title: A Legacy of Spies
cluster-master    | Score: 6.0178
cluster-master    | ------------------------------
cluster-master    | Document ID: 60121915
cluster-master    | Title: A Brief History of Everyone Who Ever Lived
cluster-master    | Score: 5.6352
cluster-master    | ------------------------------
cluster-master    | 25/04/15 17:41:56 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master    | 25/04/15 17:41:56 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master    | 25/04/15 17:41:56 INFO YarnClientSchedulerBackend: Interrupting monitor thread
cluster-master    | 25/04/15 17:41:56 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster-master    | 25/04/15 17:41:56 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
cluster-master    | 25/04/15 17:41:57 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
cluster-master    | 25/04/15 17:41:57 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master    | 25/04/15 17:41:57 INFO MemoryStore: MemoryStore cleared
cluster-master    | 25/04/15 17:41:57 INFO BlockManager: BlockManager stopped
cluster-master    | 25/04/15 17:41:57 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master    | 25/04/15 17:41:57 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master    | 25/04/15 17:41:57 INFO SparkContext: Successfully stopped SparkContext
cluster-master    | 25/04/15 17:41:57 INFO ShutdownHookManager: Shutdown hook called
cluster-master    | 25/04/15 17:41:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-b011e977-2852-4e3f-98f2-343341e8c836
cluster-master    | 25/04/15 17:41:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-07992f42-4f9b-476e-8179-87acdede0933
cluster-master    | 25/04/15 17:41:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-07992f42-4f9b-476e-8179-87acdede0933/pyspark-50dbe212-68ae-46d3-82d1-51c4ee6dd22f
cluster-master exited with code 0
█
w Enable Watch
```

Picture 10: Here you can see the result of the third (custom) query, that I wrote by myself.



```
cluster-master     | 25/04/15 17:52:12 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory on cluster-slave-1:43669 (size: 9.9 KiB, free: 366.3 MiB)
cluster-master     | 25/04/15 17:52:12 INFO BlockManagerInfo: Added broadcast_4_python on disk on cluster-slave-2:45691 (size: 11.5 KiB)
cluster-master     | 25/04/15 17:52:12 INFO BlockManagerInfo: Added broadcast_4_python on disk on cluster-slave-1:43669 (size: 11.5 KiB)
cluster-master     | 25/04/15 17:52:12 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 6) in 190 ms on cluster-slave-2 (executor 1) (1/2)
cluster-master     | 25/04/15 17:52:12 INFO TaskSetManager: Finished task 1.0 in stage 3.0 (TID 7) in 193 ms on cluster-slave-1 (executor 2) (2/2)
cluster-master     | 25/04/15 17:52:12 INFO YarnScheduler: Removed TaskSet 3.0, whose tasks have all completed, from pool
cluster-master     | 25/04/15 17:52:12 INFO DAGScheduler: ResultStage 3 (top at /app/query.py:114) finished in 0.206 s
cluster-master     | 25/04/15 17:52:12 INFO DAGScheduler: Job 1 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master     | 25/04/15 17:52:12 INFO YarnScheduler: Killing all running tasks in stage 3: Stage finished
cluster-master     | 25/04/15 17:52:12 INFO DAGScheduler: Job 1 finished: top at /app/query.py:114, took 0.364081 s
cluster-master     |
cluster-master     | Top 10 documents for query: Hip-hop evolution in USA
cluster-master     | ------------------------------
cluster-master     | Document ID: 55622679
cluster-master     | Title: A Different Thread
cluster-master     | Score: 7.2182
cluster-master     | ------------------------------
cluster-master     | Document ID: 73922368
cluster-master     | Title: A Honeymoon in Space
cluster-master     | Score: 6.3948
cluster-master     | ------------------------------
cluster-master     | Document ID: 35837086
cluster-master     | Title: A La Costa Sud
cluster-master     | Score: 6.3217
cluster-master     | ------------------------------
cluster-master     | Document ID: 47388207
cluster-master     | Title: A Lego Brickumentary
cluster-master     | Score: 6.0013
cluster-master     | ------------------------------
cluster-master     | Document ID: 18903816
cluster-master     | Title: A Jolly Good Fellow (novel)
cluster-master     | Score: 5.9065
cluster-master     | ------------------------------
cluster-master     | Document ID: 41659319
cluster-master     | Title: A Ginuwine Christmas
cluster-master     | Score: 5.5049
cluster-master     | ------------------------------
cluster-master     | Document ID: 68106371
cluster-master     | Title: A Galaxy Next Door
cluster-master     | Score: 4.8781
cluster-master     | ------------------------------
cluster-master     | Document ID: 2935235
cluster-master     | Title: A Certain Justice
cluster-master     | Score: 4.8774
cluster-master     | ------------------------------
cluster-master     | Document ID: 72843275
cluster-master     | Title: A House Made of Splinters
cluster-master     | Score: 4.3365
cluster-master     | ------------------------------
cluster-master     | Document ID: 3091435
cluster-master     | Title: A Dream Is a Wish Your Heart Makes
cluster-master     | Score: 4.2359
cluster-master     | ------------------------------
cluster-master     | 25/04/15 17:52:12 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master     | 25/04/15 17:52:12 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master     | 25/04/15 17:52:12 INFO YarnClientSchedulerBackend: Interrupting monitor thread
cluster-master     | 25/04/15 17:52:12 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster-master     | 25/04/15 17:52:12 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
cluster-master     | 25/04/15 17:52:12 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
cluster-master     | 25/04/15 17:52:12 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master     | 25/04/15 17:52:12 INFO MemoryStore: MemoryStore cleared
cluster-master     | 25/04/15 17:52:12 INFO BlockManager: BlockManager stopped
cluster-master     | 25/04/15 17:52:12 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master     | 25/04/15 17:52:12 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master     | 25/04/15 17:52:12 INFO SparkContext: Successfully stopped SparkContext
cluster-master     | 25/04/15 17:52:13 INFO ShutdownHookManager: Shutdown hook called
cluster-master     | 25/04/15 17:52:13 INFO ShutdownHookManager: Deleting directory /tmp/spark-9e3cb04b-a243-473c-adcb-921e42b07686/pyspark-7cab5416-3f9b-4292-8bec-91ba3c4ea4d1
cluster-master     | 25/04/15 17:52:13 INFO ShutdownHookManager: Deleting directory /tmp/spark-9e3cb04b-a243-473c-adcb-921e42b07686
cluster-master     | 25/04/15 17:52:13 INFO ShutdownHookManager: Deleting directory /tmp/spark-51ca6d4a-ec34-46ec-a4dc-49dba4f65c16
cluster-master exited with code 0

w Enable Watch
```

Picture 11: I forgot to check it before, but here you can see the size of the my tables in Cassandra.


zaurall@zaurall: ~/Documents/developer/INNO_S25/BD/amir/big-data-assignment2-2025
EN
21:11

```
 term   | doc_freq
--------+----------
 dobson |        1
   sain |        1
 bessus |        1
     ix |        4
  await |        2

(5 rows)
cqlsh:index_keyspace> amir@amir:~$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> USE
index_keyspace       system                 system_auth            system_distributed     system_schema          system_traces          system_views           system_virtual_schema
cqlsh> USE index_keyspace ;
cqlsh:index_keyspace> select * from
doc_stats              inverted_index         system_auth.           system_schema.         system_views.          vocabulary
index_keyspace.        system.                system_distributed.    system_traces.         system_virtual_schema.
cqlsh:index_keyspace> select count(*) from doc_stats ;

 count
-------
  1000

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:index_keyspace> select count(*) from inverted_index  ;

 count
--------
 250952

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:index_keyspace> select count(*) from vocabulary   ;

 count
-------
 40864

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:index_keyspace> 
```