

BigData Assignment 2

Dmitry Dydalin, AAI-02

Methodology

The data collection and preparation phase

It remained unchanged, with only minor changes in code style.

Setting up a Cassandra database

Four tables were created:

```
stats (
    key text PRIMARY KEY,
    value double
);
```

This table stores only two items

- **N**: the total number of indexed documents
- **dl_avg**: the average document length.

```
df (
    "token" text PRIMARY KEY,
    n_docs int
);
```

This table stores the document frequency for each token.

- **token**: the word or term
- **n_docs**: the number of documents in which the token appears

```
tf (
    document_id text,
    "token" text,
    frequency int,
    PRIMARY KEY (document_id, "token")
);
```

This table holds term frequency data for each token in each document.

- **document_id**: the unique identifier of the document
- **token**: the word or term
- **frequency**: how many times the token appears in the document

```

docs (
    id text PRIMARY KEY,
    topic text,
    text text,
    len int
);

```

This table contains the raw documents being indexed.

- id: the unique identifier of the document
- topic: the topic or category of the document
- text: the full text of the document
- len: the length of the document in tokens

MapReduce jobs

- 1) The first pipeline is used for calculating the token frequency.

The mapper takes input records in the format **<doc_id | title | text>**, tokenizes the text field, and emits intermediate key-value pairs in the form **<doc_id | token | 1>**.

The reducer aggregates these values by summing the counts of each token per document and writes the resulting term frequencies to the tf table.

To sort the emitted values from the mapper using 2 keys (first by **doc_id**, then by **token**) the mapreduce was started using additional parameters as follows:

```

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

```

- 2) The second pipeline calculates document frequency.

The mapper takes input records in the format **<doc_id | title | text>**, then tokenizes the text field. Afterwards, only the unique tokens are emitted. The intermediate key-value pairs are in the form **<token | 1>**.

The reducer aggregates these values by summing the counts of each token. Thus, we obtain the number of documents containing each token. Then, the reducer writes the resulting document frequencies to the df table.

- 3) This pipeline calculates the overall statistics per documents.

The mapper takes input records in the format **<doc_id | title | text>**, then tokenizes the text field and calculates the length of the text. The mapper puts the record **<doc_id | topic | text | len>** to a docs table. For further calculation of total **N** and **dl_avg** parameters, the mapper emits a pair **<1 | len>**.

Thus, the reducer obtains the len for each text. It calculates the number of documents and the average text length. Finally, the reducer puts **N** and **dl_avg** values to the stats table.

Notes:

- Each text is processed the same way: it firstly converted to lowercase, then the regular expression '**\b\w+\b**' is applied to split text only on tokens that consists of **[a-zA-Z0-9_]** symbols.
- The only way I managed to use the external library (cassandra driver) was by using zipimopt inside the jar with the provided zipped library.

Retrieval

Key steps:

- The search query is tokenized using the same function as in the indexing stage.
- The Apache Spark's Cassandra connector is used to read each table.
- Before scoring, the script broadcasts the following variables to all workers for efficient lookup:
 - A dictionary mapping doc_id to a tuple containing document length, topic, and text
 - Document frequency
 - Total number of documents
 - The average document length
- Then, the partial bm25 score is calculated for each token. The partial bm25 score means the contribution of the token to the document's score.
- Next, the rdd reduceByKey operation is used to sum the scores per document.
- Finally, the scores are sorted and the top 10 documents are printed.

Demonstration

To run the code:

- 1) git pull <https://github.com/Nit31/big-data-assignment2.git>
- 2) Download the file [a.parquet](#) and put to an ./app directory
- 3) Run docker-compose up in the root repository directory

Starting the docker container:

```
~/repos/big-data-assignment2 main* > docker-compose up
[+] Running 6/6
  ✓ Network big-data-assignment2_spark-cluster
  ✓ Container cluster-slave-1
  ✓ Container cassandra-server
  ! cluster-slave-1: The requested image's platform (linux/amd64) does not match the detected host platform (linux/arm64/v8) and no specific platform was requested
  ✓ Container cluster-master
  ! cluster-master: The requested image's platform (linux/amd64) does not match the detected host platform (linux/arm64/v8) and no specific platform was requested
Attaching to cassandra-server, cluster-master, cluster-slave-1
cluster-slave-1 | * Starting OpenBSD Secure Shell server sshd [ OK ]
cassandra-server |  CompileCommand: dontinline org/apache/cassandra/db/Columns$Serializer.deserializeLargeSubset(Lorg/apache/cassandra/io/util/DataInputPlus;Lorg/apache/cassandra/db/Columns;ILorg/apache/cassandra/db/Columns; boolean) boolean
cassandra-server |  CompileCommand: dontinline org/apache/cassandra/db/Columns$Serializer.serializeLargeSubset(Ljava/util/Collection;ILorg/apache/cassandra/db/Columns;ILorg/apache/cassandra/io/util/DataOutputPlus;)V void
cassandra-server |  CompileCommand: dontinline org/apache/cassandra/db/Columns$Serializer.serializeLargeSubsetSize(Ljava/util/Collection;ILorg/apache/cassandra/db/Columns;ILorg/apache/cassandra/db/Columns;)I int
cassandra-server |  CompileCommand: dontinline org/apache/cassandra/db/commitlog/AbstractCommitLogSegmentManager.advanceAllocatingFrom(Lorg/apache/cassandra/db/commitlog/CommitLogSegment;ILorg/apache/cassandra/db/commitlog/CommitLogSegment;)V void
cassandra-server |  bool dontinline = true
cassandra-server |  CompileCommand: dontinline org/apache/cassandra/db/commitlog/AbstractCommitLogSegmentManager.advanceAllocatingFrom(Lorg/apache/cassandra/db/commitlog/CommitLogSegment;ILorg/apache/cassandra/db/commitlog/CommitLogSegment;)V void
cassandra-server |  bool dontinline = true
cassandra-server |  CompileCommand: dontinline org/apache/cassandra/db/transform/BaseIterator.tryGetMoreContents()Z boolean
cassandra-server |  CompileCommand: dontinline org/apache/cassandra/db/transform/StoppingTransformation.stop()V void
cassandra-server |  bool dontinline = true
cassandra-server |  CompileCommand: dontinline org/apache/cassandra/db/transform/StoppingTransformation.stopInPartition()V void
cassandra-server |  bool dontinline = true
cassandra-server |  CompileCommand: dontinline org/apache/cassandra/io/util/BufferedDataOutputStreamPlus.doflush(I)V void
cassandra-server |  bool dontinline = true
cassandra-server |  CompileCommand: dontinline org/apache/cassandra/io/util/BufferedDataOutputStreamPlus.writeSlow(J)V void
cassandra-server |  bool dontinline = true
cassandra-server |  CompileCommand: dontinline org/apache/cassandra/io/util/RebufferingInputStream.readPrimitiveSlowly(I)V void
cassandra-server |  bool dontinline = true
cassandra-server |  CompileCommand: exclude org/apache/cassandra/utils/VMStabilityInspector.forceHeapSpaceoomMaybe(Ljava/lang/OutOfMemoryError;)V void
cassandra-server |  bool exclude = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/db/NativeDecoratedKey.address() boolean
cassandra-server |  bool inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/db/NativeDecoratedKey.length()I int
cassandra-server |  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/DataOutputPlus;)V void
cassandra-server |  bool inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/io/util/Memory.checkBounds(JJ)V void
cassandra-server |  bool inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/io/util/SafeMemory.checkBounds(JJ)V void
cassandra-server |  bool inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/io/util/TrackedDataInputPlus.checkCanRead(I)V void
cassandra-server |  inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/net/FrameDecoder.decode(Ljava/util/Collection;Lorg/apache/cassandra/net/ShareableBytes;I)V void
cassandra-server |  inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/service/reads/repair/RowIteratorMergeListener.applyToPartition(ILjava/util/function/Consumer;)V void
cassandra-server |  bool inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/AsymmetricOrdering.selectBoundary(Lorg/apache/cassandra/utils/AsymmetricOrdering/Opt;II)I int
cassandra-server |  bool inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/AsymmetricOrdering.strictnessOfLessThan(Lorg/apache/cassandra/utils/AsymmetricOrdering/Opt;I)I int
cassandra-server |  bool inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/BloomFilter.indexes(Lorg/apache/cassandra/utils/IFilter;FilterKey;)[] boolean
cassandra-server |  bool inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/BloomFilter.selIndexes(JJI[J])V void
cassandra-server |  bool inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/ByteBufferUtil.compare(Ljava/nio/ByteBuffer;[B)I int
cassandra-server |  bool inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/ByteBufferUtil.compareUnsigned(Ljava/nio/ByteBuffer;Ljava/nio/ByteBuffer;)I int
cassandra-server |  bool inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/FastByteOperations$UnsafeOperations.compareTO(Ljava/lang/Object;JLjava/lang/Object;JLjava/nio/ByteBuffer;)I int
cassandra-server |  bool inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/FastByteOperations$UnsafeOperations.compareTO(Ljava/nio/ByteBuffer;Ljava/nio/ByteBuffer;)I int
cassandra-server |  bool inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/concurrent/Awaitable.await(Ljava/util/concurrent/atomic/AtomicReferenceFieldUpdater;Ljava/util/function/Predicate;)V void
cassandra-server |  bool inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/concurrent/Awaitable.awaitUntil(Ljava/util/concurrent/atomic/AtomicReferenceFieldUpdater;Ljava/util/function/Predicate;)Z boolean
cassandra-server |  bool inline = true
```

Here we can see that the slave node is up:

```
cluster-master | cluster-slave-2: ssh: Could not resolve hostname cluster-slave-2: Name or service not known
cluster-master | cluster-slave-3: ssh: Could not resolve hostname cluster-slave-3: Name or service not known
cluster-master | cluster-slave-5: ssh: Could not resolve hostname cluster-slave-5: Name or service not known
cluster-master | cluster-slave-4: ssh: Could not resolve hostname cluster-slave-4: Name or service not known
cluster-master | 400 org.apache.hadoop.hdfs.server.namenode.SecondaryNameNode
cluster-master | 164 org.apache.hadoop.hdfs.server.namenode.NameNode
cluster-master | 1013 jdk.jcmd/sun.tools.jps.Jps -l
cluster-master | 984 org.apache.hadoop.mapreduce.v2.hs.JobHistoryServer
cluster-master | 633 org.apache.hadoop.yarn.server.resourcemanager.ResourceManager
cluster-master | Configured Capacity: 92409790464 (86.06 GB)
cluster-master | Present Capacity: 74686726144 (69.56 GB)
cluster-master | DFS Remaining: 74686781568 (69.56 GB)
cluster-master | DFS Used: 24576 (24 KB)
cluster-master | DFS Used%: 0.00%
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 (1):
cluster-master |
cluster-master | Name: 172.18.0.2:9866 (cluster-slave-1.big-data-assignment2_spark-cluster)
cluster-master | Hostname: cluster-slave-1
cluster-master | Decommission Status : Normal
cluster-master | Configured Capacity: 92409790464 (86.06 GB)
cluster-master | DFS Used: 24576 (24 KB)
cluster-master | Non DFS Used: 12998178784 (12.09 GB)
cluster-master | DFS Remaining: 74686701568 (69.56 GB)
cluster-master | DFS Used%: 0.00%
cluster-master | DFS Remaining%: 80.82%
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 | Xcivers: 0
cluster-master | Last contact: Sat Apr 12 16:43:53 GMT 2025
cluster-master | Last Block Report: Sat Apr 12 16:43:32 GMT 2025
cluster-master | Num of Blocks: 0
```

The first mapreduce operation (2 pics):

big-data-assignment2 — docker-compose up — docker-compose — docker-compose up — 182x49

```

cluster-master | Keyspace and tables created successfully in Cassandra.
cluster-master | Running Term Frequency MapReduce...
cluster-master | packageJobJar: [/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar] /tmp/streamjob185395136790271802.jar tmpDir=null
cluster-master | 2025-04-12 16:51:01.065 INFO client.DefaultNoHDFSFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-12 16:51:01.222 INFO client.DefaultNoHDFSFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-12 16:51:03.103 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master | 2025-04-12 16:51:03.579 INFO mapreduce.JobSubmission: number of splits:2
cluster-master | 2025-04-12 16:51:03.793 INFO mapreduce.JobSubmission: Submitting tokens for job: job_1744476225945_0001
cluster-master | 2025-04-12 16:51:04.001 INFO mapreduce.JobSubmission: Executing with tokens: []
cluster-master | 2025-04-12 16:51:04.012 INFO conf.Configuration: resource-types.xml not found
cluster-master | 2025-04-12 16:51:04.013 INFO resource.ResourcesUtils: Unable to find 'resource-types.xml'.
cluster-master | 2025-04-12 16:51:04.263 INFO impl.YarnClientImpl: Submitted application application_1744476225945_0001
cluster-master | 2025-04-12 16:51:04.313 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744476225945_0001
cluster-master | 2025-04-12 16:51:04.314 INFO mapreduce.Job: Running job: job_1744476225945_0001
cluster-master | 2025-04-12 16:51:14.591 INFO mapreduce.Job: Job job_1744476225945_0001 running in uber mode : false
cluster-master | 2025-04-12 16:51:14.601 INFO mapreduce.Job: map 0% reduce 0%
cluster-master | 2025-04-12 16:51:28.750 INFO mapreduce.Job: map 100% reduce 0%
cluster-master | 2025-04-12 16:51:28.758 INFO mapreduce.Job: map 100% reduce 8%
cluster-master | 2025-04-12 16:51:39.012 INFO mapreduce.Job: map 100% reduce 52%
cluster-master | 2025-04-12 16:51:39.013 INFO mapreduce.Job: map 100% reduce 53%
cluster-master | 2025-04-12 16:51:51.141 INFO mapreduce.Job: map 100% reduce 54%
cluster-master | 2025-04-12 16:51:52.150 INFO mapreduce.Job: map 100% reduce 55%
cluster-master | 2025-04-12 16:51:54.162 INFO mapreduce.Job: map 100% reduce 56%
cluster-master | 2025-04-12 16:51:55.163 INFO mapreduce.Job: map 100% reduce 57%
cluster-master | 2025-04-12 16:51:57.191 INFO mapreduce.Job: map 100% reduce 59%
cluster-master | 2025-04-12 16:51:58.197 INFO mapreduce.Job: map 100% reduce 60%
cluster-master | 2025-04-12 16:52:16.533 INFO mapreduce.Job: map 100% reduce 77%
cluster-master | 2025-04-12 16:52:17.545 INFO mapreduce.Job: map 100% reduce 86%
cluster-master | 2025-04-12 16:52:18.552 INFO mapreduce.Job: map 100% reduce 94%
cluster-master | 2025-04-12 16:52:22.575 INFO mapreduce.Job: map 100% reduce 96%
cluster-master | 2025-04-12 16:52:24.588 INFO mapreduce.Job: map 100% reduce 97%
cluster-master | 2025-04-12 16:52:28.610 INFO mapreduce.Job: map 100% reduce 99%
cluster-master | 2025-04-12 16:52:30.623 INFO mapreduce.Job: map 100% reduce 100%
cluster-master | 2025-04-12 16:52:33.723 INFO mapreduce.Job: Job job_1744476225945_0001 completed successfully
cluster-master | 2025-04-12 16:52:33.831 INFO mapreduce.Job: Counters: 55
cluster-master |     File System Counters
cluster-master |         FILE: Number of bytes read=10555855
cluster-master |         FILE: Number of bytes written=24439348
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=4346244
cluster-master |         HDFS: Number of read operations=56
cluster-master |         HDFS: Number of large read operations=20
cluster-master |         HDFS: Number of write operations=20
cluster-master |         HDFS: Number of bytes read erasure-coded=0
cluster-master | 
cluster-master |     Job Counters
cluster-master |         Killed reduce tasks=1
cluster-master |         Launched map tasks=2
cluster-master |         Launched reduce tasks=11
cluster-master |         Data-local map tasks=2
cluster-master |         Total time spent by all maps in occupied slots (ms)=8488
cluster-master |         Total time spent by all reduces in occupied slots (ms)=354945
cluster-master |         Total time spent by all map tasks (ms)=8488
cluster-master |         Total time spent by all reduce tasks (ms)=354945
cluster-master |         Total vcore-milliseconds taken by all map tasks=8488
cluster-master |         Total vcore-milliseconds taken by all reduce tasks=8488
cluster-master |         Total megabyte-milliseconds taken by all map tasks=8691712
cluster-master |         Total megabyte-milliseconds taken by all reduce tasks=363463680
cluster-master | 
cluster-master |     Map-Reduce Framework
cluster-master |         Map input records=1003
cluster-master |         Map output records=578378
cluster-master |         Map output bytes=9399039
cluster-master |         Map output materialized bytes=10555915
cluster-master |         Input split bytes=292
cluster-master |         Combine input records=0
cluster-master |         Combine output records=0
cluster-master |         Reduce input groups=250180
cluster-master |         Reduce shuffle bytes=10555915
cluster-master |         Reduce input records=578378
cluster-master |         Reduce output records=250180
cluster-master |         Spilled Records=1156756
cluster-master |         Shuffled Maps =20
cluster-master |         Failed Shuffles=0
cluster-master |         Merged Map outputs=20
cluster-master |         GC time elapsed (ms)=1259
cluster-master |         CPU time spent (ms)=108380
cluster-master |         Physical memory (bytes) snapshot=3874922496
cluster-master |         Virtual memory (bytes) snapshot=40994349056
cluster-master |         Total committed heap usage (bytes)=2933391360
cluster-master |         Peak Map Physical memory (bytes)=328134656
cluster-master |         Peak Map Virtual memory (bytes)=3412107264
cluster-master |         Peak Reduce Physical memory (bytes)=370585600

```

The second mapreduce operation (2 pics):

```

Terminal Shell Edit View Window Help
big-data-assignment2 - docker-compose up - docker-compose - docker-compose up - 182x49
cluster-master | Running Document Frequency MapReduce...
cluster-master | packageJobJar: [] [/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar] /tmp/streamjob24052673417192064.jar tmpDir=null
cluster-master | 2025-04-12 16:52:36,582 INFO client.DefaultNoHDFSFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-12 16:52:36,745 INFO client.DefaultNoHDFSFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-12 16:52:36,988 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744476225945_0002
cluster-master | 2025-04-12 16:52:39,039 INFO mapreduce.JobSubmitter: Total input files to process : 1
cluster-master | 2025-04-12 16:52:39,067 INFO mapreduce.reduce.tasks: 1
cluster-master | 2025-04-12 16:52:39,252 INFO mapreduce.JobSubmitter: INFO Configuration deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
cluster-master | 2025-04-12 16:52:39,252 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744476225945_0002
cluster-master | 2025-04-12 16:52:39,252 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master | 2025-04-12 16:52:39,432 INFO conf.Configuration: resource-types.xml not found
cluster-master | 2025-04-12 16:52:39,433 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master | 2025-04-12 16:52:39,528 INFO impl.YarnClientImpl: Submitted application application_1744476225945_0002
cluster-master | 2025-04-12 16:52:39,558 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744476225945_0002
cluster-master | 2025-04-12 16:52:47,782 INFO mapreduce.Job: Job: job_1744476225945_0002 running in uber mode : false
cluster-master | 2025-04-12 16:52:47,786 INFO mapreduce.Job: map 0% reduce 0%
cluster-master | 2025-04-12 16:52:53,883 INFO mapreduce.Job: map 100% reduce 0%
cluster-master | 2025-04-12 16:53:04,967 GCInspector.java:285 - G1 Young Generation GC in 422ms. G1 Eden Space: 1912602624 -> 0; G1 Old Gen: 157900800
Cassandra-server | INFO [Notification Thread] 2025-04-12 16:53:04,967 GCIInspector.java:285 - G1 Young Generation GC in 422ms. G1 Eden Space: 1912602624 -> 0; G1 Old Gen: 157900800
-> 189863936; G1 Survivor Space: 34467896 -> 45003296
cluster-master | 2025-04-12 16:53:12,152 INFO mapreduce.Job: map 100% reduce 25%
cluster-master | 2025-04-12 16:53:13,194 INFO mapreduce.Job: map 100% reduce 50%
cluster-master | 2025-04-12 16:53:14,284 INFO mapreduce.Job: map 100% reduce 75%
cluster-master | 2025-04-12 16:53:15,219 INFO mapreduce.Job: map 100% reduce 100%
cluster-master | 2025-04-12 16:53:16,269 INFO mapreduce.Job: Job job_1744476225945_0002 completed successfully
cluster-master | 2025-04-12 16:53:16,383 INFO mapreduce.Job: Counters: 55
cluster-master |     File System Counters
cluster-master |         FILE: Number of bytes read=2703565
cluster-master |         FILE: Number of bytes written=7067806
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=417343
cluster-master |         HDFS: Number of read operations=26
cluster-master |         HDFS: Number of large read operations=0
cluster-master |         HDFS: Number of write operations=8
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=4
cluster-master |         Data-local map tasks=2
cluster-master |         Total time spent by all maps in occupied slots (ms)=7812
cluster-master |         Total time spent by all reduces in occupied slots (ms)=63835
cluster-master |         Total time spent by all map tasks (ms)=7812
cluster-master |         Total time spent by all reduce tasks (ms)=63835
cluster-master |         Total vcore-milliseconds taken by all map tasks=7812
cluster-master |         Total vcore-milliseconds taken by all reduce tasks=63835
cluster-master |         Total vcore-milliseconds taken by all map tasks=799488
cluster-master |         Total megabyte-milliseconds taken by all map tasks=65367040

```

```

Terminal Shell Edit View Window Help
big-data-assignment2 - docker-compose up - docker-compose - docker-compose up - 182x49
cluster-master | Killed reduce tasks=1
cluster-master | Launched map tasks=2
cluster-master | Launched reduce tasks=4
cluster-master | Data-local map tasks=2
cluster-master | Total time spent by all maps in occupied slots (ms)=7812
cluster-master | Total time spent by all reduces in occupied slots (ms)=63835
cluster-master | Total time spent by all map tasks (ms)=7812
cluster-master | Total time spent by all reduce tasks (ms)=63835
cluster-master | Total vcore-milliseconds taken by all map tasks=7812
cluster-master | Total vcore-milliseconds taken by all reduce tasks=63835
cluster-master | Total megabyte-milliseconds taken by all map tasks=799488
cluster-master | Total megabyte-milliseconds taken by all reduce tasks=65367040
Map-Reduce Framework
cluster-master |     Map input records=1003
cluster-master |     Map output records=250180
cluster-master |     Map output bytes=2203181
cluster-master |     Map output materialized bytes=2703589
cluster-master |     Input split bytes=292
cluster-master |     Combine input records=0
cluster-master |     Combine output records=0
cluster-master |     Reduce input groups=40803
cluster-master |     Reduce shuffle bytes=2793589
cluster-master |     Reduce input records=250180
cluster-master |     Reduce output records=40803
cluster-master |     Spilled Records=500360
cluster-master |     Shuffled Maps =8
cluster-master |     Failed Shuffles=0
cluster-master |     Merged Map outputs=8
cluster-master |     GC time elapsed (ms)=20980
cluster-master |     CPU time spent (ms)=20980
cluster-master |     Physical memory (bytes) snapshot=1924489216
cluster-master |     Virtual memory (bytes) snapshot=20487897088
cluster-master |     Total committed heap usage (bytes)=1510998016
cluster-master |     Peak Map Physical memory (bytes)=325218304
cluster-master |     Peak Map Virtual memory (bytes)=3411214336
cluster-master |     Peak Reduce Physical memory (bytes)=363315200
cluster-master |     Peak Reduce Virtual memory (bytes)=4109934592
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=417343
cluster-master | 2025-04-12 16:53:16,383 INFO streaming.StreamJob: Output directory: /tmp/index/output_d1

```

The third mapreduce operation (2 pics):

```
cluster-master | Running Stats Aggregation MapReduce...
cluster-master | packageJobJar: [] [/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar] /tmp/streamjob5555417172455689390.jar tmpDir=null
2025-04-12 16:53:19,242 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-12 16:53:19,381 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-12 16:53:19,620 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/job_1744476225945_0003
2025-04-12 16:53:21,234 INFO mapred.FileInputFormat: Total input files to process : 1
2025-04-12 16:53:21,707 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-12 16:53:21,733 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2025-04-12 16:53:21,899 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744476225945_0003
2025-04-12 16:53:21,900 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-12 16:53:22,106 INFO conf.Configuration: resource-types.xml not found
2025-04-12 16:53:22,187 INFO Resource.ResourcesUtil: Unable to find 'resource-types.xml'
2025-04-12 16:53:22,186 INFO Impl.YarnClientImpl: Submitted application application_1744476225945_0003
2025-04-12 16:53:22,224 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744476225945_0003/
2025-04-12 16:53:22,225 INFO mapreduce.Job: Running job: job_1744476225945_0003
2025-04-12 16:53:29,519 INFO mapreduce.Job: Job job_1744476225945_0003 running in uber mode : false
2025-04-12 16:53:29,521 INFO mapreduce.Job: map 0% reduce 0%
2025-04-12 16:53:36,671 INFO mapreduce.Job: map 100% reduce 0%
2025-04-12 16:53:42,883 INFO mapreduce.Job: map 100% reduce 10%
2025-04-12 16:53:44,955 INFO mapreduce.Job: map 100% reduce 20%
2025-04-12 16:53:45,989 INFO mapreduce.Job: map 100% reduce 30%
2025-04-12 16:53:49,107 INFO mapreduce.Job: map 100% reduce 50%
2025-04-12 16:53:50,128 INFO mapreduce.Job: map 100% reduce 60%
2025-04-12 16:53:51,136 INFO mapreduce.Job: map 100% reduce 80%
2025-04-12 16:53:53,162 INFO mapreduce.Job: map 100% reduce 90%
2025-04-12 16:53:54,188 INFO mapreduce.Job: map 100% reduce 100%
2025-04-12 16:53:54,254 INFO mapreduce.Job: Job job_1744476225945_0003 completed successfully
2025-04-12 16:53:54,354 INFO mapreduce.Job: Counters: 55
cluster-master | File System Counters
cluster-master |   FILE: Number of bytes read=8066
cluster-master |   FILE: Number of bytes written=3337590
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=31
cluster-master |   HDFS: Number of read operations=56
cluster-master |   HDFS: Number of large read operations=0
cluster-master |   HDFS: Number of write operations=20
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=10
cluster-master |   Data-local map tasks=2
cluster-master |   Total time spent by all maps in occupied slots (ms)=8893
cluster-master |   Total time spent by all reduces in occupied slots (ms)=53592
cluster-master |   Total time spent by all map tasks (ms)=8893
cluster-master |   Total time spent by all reduce tasks (ms)=53592
```

```
Terminal Shell Edit View Window Help big-data-assignment2 — docker-compose up — docker-compose — docker-compose up — 182x49
cluster-master | Killed reduce tasks=1
cluster-master | Launched map tasks=2
cluster-master | Launched reduce tasks=10
cluster-master | Data-local map tasks=2
cluster-master | Total time spent by all maps in occupied slots (ms)=8893
cluster-master | Total time spent by all reduces in occupied slots (ms)=53592
cluster-master | Total time spent by all map tasks (ms)=8893
cluster-master | Total time spent by all reduce tasks (ms)=53592
cluster-master | Total vcore-milliseconds taken by all map tasks=8893
cluster-master | Total vcore-milliseconds taken by all reduce tasks=53592
cluster-master | Total megabyte-milliseconds taken by all map tasks=9106432
cluster-master | Total megabyte-milliseconds taken by all reduce tasks=54878208
cluster-master | Map-Reduce Framework
cluster-master |   Map input records=1003
cluster-master |   Map output records=997
cluster-master |   Map output bytes=6012
cluster-master |   Map output materialized bytes=8126
cluster-master |   Input split bytes=292
cluster-master |   Combine input records=0
cluster-master |   Combine output records=0
cluster-master |   Reduce input groups=1
cluster-master |   Reduce shuffle bytes=8126
cluster-master |   Reduce input records=997
cluster-master |   Reduce output records=2
cluster-master |   Spilled Records=1994
cluster-master |   Shuffled Maps =20
cluster-master |   Failed Shuffles=0
cluster-master |   Merged Map outputs=20
cluster-master |   GC time elapsed (ms)=1219
cluster-master |   CPU time spent (ms)=8480
cluster-master |   Physical memory (bytes) snapshot=3762417664
cluster-master |   Virtual memory (bytes) snapshot=40961970176
cluster-master |   Total committed heap usage (bytes)=2999451648
cluster-master |   Peak Map Physical memory (bytes)=321040384
cluster-master |   Peak Map Virtual memory (bytes)=3410620416
cluster-master |   Peak Reduce Physical memory (bytes)=316637184
cluster-master |   Peak Reduce Virtual memory (bytes)=3415748608
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=31
2025-04-12 16:53:54,354 INFO streaming.StreamJob: Output directory: /tmp/index/output_stats
```

I have created a simple python script (app/test_cassandra.py) to print several examples from the created tables:

```

cluster-master | Indexing done. Testing the index...
cluster-master | First 5 rows in ff table:
cluster-master |     Row(document_id='10230685', token='18', frequency=1)
cluster-master |     Row(document_id='10230685', token='2003', frequency=2)
cluster-master |     Row(document_id='10230685', token='a', frequency=1)
cluster-master |     Row(document_id='10230685', token='after', frequency=1)
cluster-master |     Row(document_id='10230685', token='album', frequency=5)
cluster-master |
cluster-master | First 5 rows in df table:
cluster-master |     Row(tokens='dobsion', n_docs=1)
cluster-master |     Row(tokens='sain', n_docs=1)
cluster-master |     Row(tokens='bessus', n_docs=1)
cluster-master |     Row(tokens='ix', n_docs=4)
cluster-master |     Row(tokens='await', n_docs=2)
cluster-master |
cluster-master | Example from docs table:
cluster-master |     10230685
cluster-master |     115
cluster-master |     A Dead Sinking Story is an album by the Japanese band Envy. The album was released on August 18, 2003 on Level Plane Records. It is their only album recorded with three guitarists: Daichi Takasugi joined before the creation of the album and left after the related tour. ==Track listing== ==Personnel== *Dairoku Seki - Drums *Tetsuya Fukagawa - Sequencer, Vocals *Nobukata Kawai - Guitar *Masahiro Tobita - Guitar *Manabu Nakagawa - Bass Guitar *Daichi Takasugi - Guitar *Takashi Kitaguchi - Engineering *Tatsuya Kase - Mastering ==Reception== Reception was almost uniformly positive, with Allmusic, Stylus Magazine, and Punknews.org all praising the album's hybrid of emo, screamo, and post-rock. ==References== * Category:2003 albums Category:Level Plane Records albums Category:Envy (band) albums
cluster-master |     A Dead Sinking Story
cluster-master |
cluster-master | Stats from stats table:
cluster-master |     Row(key='dl_avg', value=580.1183550651956)
cluster-master |     Row(key='N', value=997.0)
cluster-master | Running bm25 search for query: James Dearden films
cluster-master | :: loading settings :: url = jarfile:/usr/local/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
cluster-master | Ivy Default Cache set to: /root/.ivy2/cache
cluster-master | The jars for the packages stored in: /root/.ivy2/jars
com.datastax.spark#spark-cassandra-connector_2.12 added as a dependency
cluster-master | :: resolving dependencies :: org.apache.spark$spark-submit-parent-9cc651c2-0623-4b70-9bf7-c927ee973c7d;1.0
conf: [default]
found com.datastax.spark#spark-cassandra-connector_2.12;3.1.0 in central
found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.1.0 in central
found com.datastax.oss#java-driver-core-shaded;4.12.0 in central
found com.datastax.oss#native-protocol;1.5.0 in central
found com.datastax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 in central
found com.typesafe#config;1.4.1 in central
found org.slf4j#slf4j-api;1.7.26 in central
found io.dropwizard.metrics#metrics-core;4.1.18 in central
found org.hdrhistogram#HdrHistogram;2.1.12 in central
found org.reactivestreams#reactive-streams;1.0.3 in central
found com.github.stephenhc.jcip#jcip-annotations;1.0-1 in central
found com.github.spotbugs#spotbugs-annotations;3.1.12 in central
found com.google.code.findbugs#jsr305;3.0.2 in central
found com.datastax.oss#java-driver-mapper-runtime;4.12.0 in central

```

Finally, we got to the retrieval! Below you can see results for 3 queries

- 1) The first result is a perfect hit. The documents 2-4 are films directed by some "James" and the documents 5-10 were chosen only for the "film" hit

```

cluster-master | Top 10 Documents for query: 'James Dearden films'
cluster-master |
1. Document ID: 1667210
Topic: A Kiss Before Dying (1991 film)
Score: 13.6715
Text: "A Kiss Before Dying is a 1991 American romantic thriller film directed by James Dearden, and based ...
cluster-master |
2. Document ID: 63002694
Topic: A Close Call for Ellery Queen
Score: 5.7367
Text: A Close Call for Ellery Queen is a 1942 American mystery film directed by James P. Hogan and written...
cluster-master |
3. Document ID: 63016837
Topic: A Desperate Chance for Ellery Queen
Score: 5.7087
Text: A Desperate Chance for Ellery Queen is a 1942 American mystery film directed by James P. Hogan and w...
cluster-master |
4. Document ID: 57704768
Topic: A Gentleman of Quality
Score: 5.6547
Text: A Gentleman of Quality is a 1919 American silent drama film directed by James Young and starring Ear...
cluster-master |
5. Document ID: 19749985
Topic: A Glimpse of Hell (film)
Score: 5.5722
Text: "A Glimpse of Hell is a 2001 American-Canadian made-for-television drama film directed by Mikael Sal...
cluster-master |
6. Document ID: 32997171
Topic: A Girl of the Bush
Score: 5.5438
Text: A Girl of the Bush is a 1921 Australian silent film directed by Franklyn Barrett. It is one of the f...
cluster-master |
7. Document ID: 65891184
Topic: A Cry from Within
Score: 5.4639
Text: A Cry From Within is a 2015 American horror film directed by Deborah Twiss and Zach Miller and starr...
cluster-master |
8. Document ID: 30011413
Topic: A Harlot's Progress (film)
Score: 5.4331
Text: A Harlot's Progress is a 2006 British television film directed by Justin Hardy and starring Zoe Tapp...
cluster-master |
9. Document ID: 46908278
Topic: A Gamble for Love
Score: 5.4220
Text: A Gamble for Love is a 1917 British silent sports film directed by Frank Wilson and starring Gerald ...
cluster-master |
10. Document ID: 64378948
Topic: A Gift from Bob

```

2) Again, the first result is the most relevant for the query. Other documents were chosen for film/date/genre hits.

```
cluster-master | Top 10 Documents for query: 'I want to find the American comedy created in 1916'
cluster-master |
cluster-master | 1. Document ID: 12955622
cluster-master |   Topic: A Day at School
cluster-master |   Score: 13.5713
cluster-master |   Text: A Day at School is a 1916 American silent comedy film featuring Oliver Hardy. ==Cast== * Oliver Hard...
cluster-master |
cluster-master | 2. Document ID: 2544021
cluster-master |   Topic: A Jitney Elopement
cluster-master |   Score: 11.5660
cluster-master |   Text: "thumb|Full film A Jitney Elopement was Charlie Chaplin's fifth film for Essanay Films. It starr...
cluster-master |
cluster-master | 3. Document ID: 71344775
cluster-master |   Topic: A Death in Bed No. 12
cluster-master |   Score: 10.7017
cluster-master |   Text: "A Death in Bed No.12 is the first collection of stories written by Ghassan Kanafani. It was publish...
cluster-master |
cluster-master | 4. Document ID: 44853014
cluster-master |   Topic: A Gutter Magdalene
cluster-master |   Score: 9.9184
cluster-master |   Text: A Gutter Magdalene is a lostThe Library of Congress American Silent Feature Film Survival Catalog:....
cluster-master |
cluster-master | 5. Document ID: 40474725
cluster-master |   Topic: A Corner in Cotton
cluster-master |   Score: 9.5547
cluster-master |   Text: A Corner in Cotton is a five-reel silent film melodrama produced in 1916 by Quality Pictures and dis...
cluster-master |
cluster-master | 6. Document ID: 54317905
cluster-master |   Topic: A Horse Walks into a Bar
cluster-master |   Score: 9.5458
cluster-master |   Text: "A Horse Walks into a Bar () is a novel by Israeli author David Grossman. First published in Hebrew ...
cluster-master |
cluster-master | 7. Document ID: 1356924
cluster-master |   Topic: A Child's Garden of Verses
cluster-master |   Score: 9.3771
cluster-master |   Text: "thumb|Title Page of a 1916 US edition A Child's Garden of Verses is an 1885 volume of 64 poems for ...
cluster-master |
cluster-master | 8. Document ID: 37284994
cluster-master |   Topic: A Girl's Folly
cluster-master |   Score: 9.3466
cluster-master |   Text: thumb|The film A Girl's Folly A Girl's Folly is a 1917 American silent comedy film directed by Mauri...
cluster-master |
cluster-master | 9. Document ID: 13478522
cluster-master |   Topic: A Knight of the Range
cluster-master |   Score: 9.2222
cluster-master |   Text: A Knight of the Range is a 1916 American Western film, featuring Harry Carey.A Knight of the Range a...
cluster-master |
cluster-master | 10. Document ID: 50813860
cluster-master |   Topic: A History of Garage and Frat Bands in Memphis 1968-1975, Volume 1
```

3) Nothing to add. Video games are found perfectly.

```
cluster-master | Top 10 Documents for query: 'Video game'
cluster-master | -----
cluster-master | 1. Document ID: 4887308
cluster-master |   Topic: A Fork in the Tale
cluster-master |   Score: 9.9184
cluster-master |   Text: "A Fork in the Tale is a full motion video (FMV) comedic adventure game developed by Any River Enter...
cluster-master | -----
cluster-master | 2. Document ID: 28566279
cluster-master |   Topic: A Game of Concentration
cluster-master |   Score: 9.9105
cluster-master |   Text: "A Game of Concentration (also known as Concentration and Hunt & Score) is a video game developed by...
cluster-master | -----
cluster-master | 3. Document ID: 72686083
cluster-master |   Topic: A Date in the Park
cluster-master |   Score: 9.7936
cluster-master |   Text: "A Date in the Park is a 2014 video game by British independent developer Cloak and Dagger Games. De...
cluster-master | -----
cluster-master | 4. Document ID: 58465868
cluster-master |   Topic: A Blind Legend
cluster-master |   Score: 9.6464
cluster-master |   Text: "A Blind Legend is an action-adventure Audio game. The game was Supported financially by Centre natio...
cluster-master | -----
cluster-master | 5. Document ID: 58455505
cluster-master |   Topic: A Bastard's Tale
cluster-master |   Score: 9.4497
cluster-master |   Text: "A Bastard's Tale is an action 2D video game, published by Swedish studio No Pest Productions for Pla...
cluster-master | -----
cluster-master | 6. Document ID: 17987485
cluster-master |   Topic: A Kingdom for Keflings
cluster-master |   Score: 9.4324
cluster-master |   Text: "A Kingdom for Keflings is a video game developed by NinjaBee for the Xbox Live Arcade which was rel...
cluster-master | -----
cluster-master | 7. Document ID: 35456304
cluster-master |   Topic: A Good Librarian Like a Good Shepherd
cluster-master |   Score: 7.9424
cluster-master |   Text: "A Good Librarian Like a Good Shepherd, known in Japan as , is a Japanese adult visual novel develop...
cluster-master | -----
cluster-master | 8. Document ID: 251385
cluster-master |   Topic: A Gamut of Games
cluster-master |   Score: 6.0321
cluster-master |   Text: "A Gamut of Games is an innovative book of games written by Sid Sackson and first published in 1969....
cluster-master | -----
cluster-master | 9. Document ID: 800911
cluster-master |   Topic: A Cold Wind Blows (game)
cluster-master |   Score: 5.7198
cluster-master |   Text: "A Cold Wind Blows or The Big Wind Blows is a noncompetitive substitute for the game of musical chal...
cluster-master | -----
cluster-master | 10. Document ID: 60809985
cluster-master |   Topic: A Little Game
```

Conclusion

The BM25 algorithm performs well, delivering the expected results. However, in my experience, MapReduce feels limited in terms of functionality - it is constrained on how data can be processed.

Additionally, the execution speed is currently slower than running the equivalent logic in pure Python without Spark. This is primarily due to the relatively small size of the dataset. For smaller workloads, the overhead of initializing and coordinating distributed tasks outweighs the benefits. I understand that with the number of documents growing, the distributed approach becomes more apparent, and Spark must outperform single-node Python execution significantly.