

# 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

Three tables were created:

```
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 calculates the number of tokens. After that it puts the record **<id | title | text | length>** into the docs table. Finally, the mapper 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.

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.
- The average length of documents is calculated.
- 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 per-term bm25 score is calculated for each token. This per-term bm25 score means the contribution of the token into 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

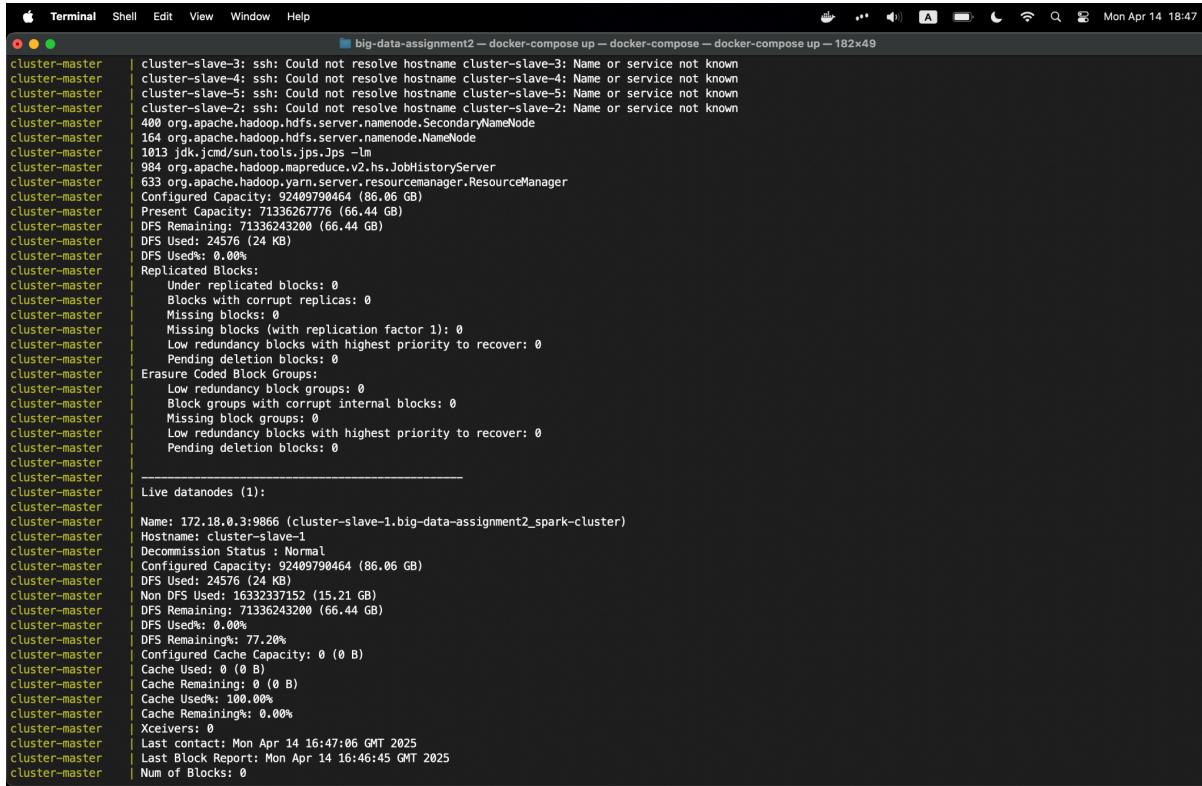
To run the code:

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

## **Starting the docker container:**

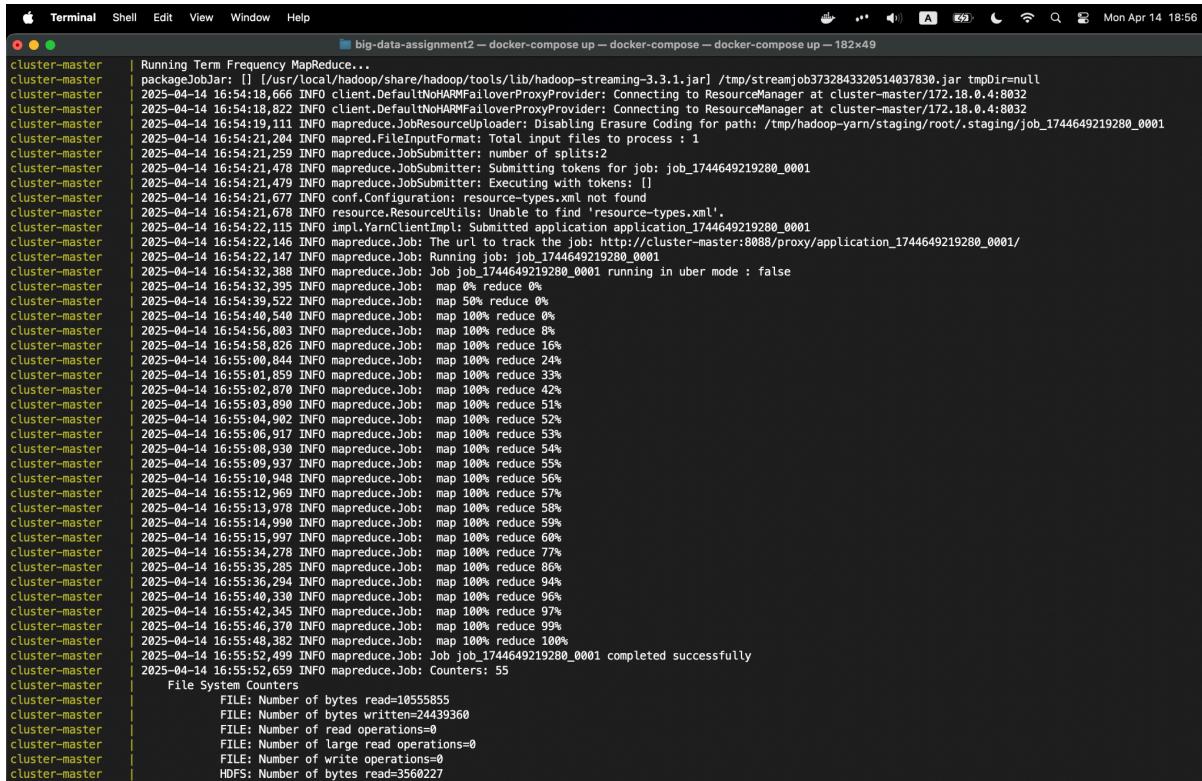
```
Terminal Shell Edit View Window Help big-data-assignment2 docker-compose up --docker-compose --docker-compose up -182x49
~/repos/big-data-assignment2 main > docker-compose up
[+] Running 6/6
✓ Network big-data-assignment2_spark-cluster
✓ Container cluster-slave-1
✓ Container cassandra-server
! Cluster-slave-1: 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$Serializer.deserializeLargeSubset(Lorg/apache/cassandra/io/util/DataInputPlus;Lorg/apache/cassandra/db/Columns;ILorg/apache/cassandra/db/Columns; boolean dontinline = true
cassandra-server | _CompileCommand: dontinline org/apache/cassandra/db/Columns$Serializer.serializeLargeSubset(Ljava/util/Collection;ILorg/apache/cassandra/db/Columns;ILorg/apache/cassandra/io/util/DataOutputPlus);V boolean dontinline = true
cassandra-server | _CompileCommand: dontinline org/apache/cassandra/db/Columns$Serializer.serializeLargeSubsetSize(Ljava/util/Collection;ILorg/apache/cassandra/db/Columns;I)V boolean dontinline = true
cassandra-server | _CompileCommand: dontinline org/apache/cassandra/db/commitlog/AbstractCommitLogSegmentManager.advanceAllocatingFrom(ILorg/apache/cassandra/db/commitlog/CommitLogSegment;V boolean dontinline = true
cassandra-server | _CompileCommand: dontinline org/apache/cassandra/db/transform/BaseIterator.tryGetMoreContents()Z boolean dontinline = true
cassandra-server | _CompileCommand: dontinline org/apache/cassandra/db/transform/StopIterationTransformation.stop()V boolean dontinline = true
cassandra-server | _CompileCommand: dontinline org/apache/cassandra/io/util/BufferedDataOutputStreamPlus.doFlush(ILV boolean dontinline = true
cassandra-server | _CompileCommand: dontinline org/apache/cassandra/io/util/BufferedDataOutputStreamPlus.writeSlow(J)V boolean dontinline = true
cassandra-server | _CompileCommand: dontinline org/apache/cassandra/io/util/RebufferingInputStream.readPrimitiveSlowly(I)J boolean dontinline = true
cassandra-server | _CompileCommand: exclude org/apache/cassandra/utils/JMSThrowableInspector,forceHeapSpaceOnMaybe(Ljava/lang/OutOfMemoryError);V boolean exclude = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/db/NativeDecoratedKey.address()J boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/db/NativeDecoratedKey.length()I boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/db/rows/UnfilteredSerializer.serializeRowBody(Lorg/apache/cassandra/db/rows;ILorg/apache/cassandra/db/rows/SerializationHelper;Lorg/apache/cassandra/io/util/DataOutputPlus)V boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/io/util/Memory.checkBounds(JJJ)V boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/io/util/SafeMemory.checkBounds(JJJ)V boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/io/util/TrackedDataInputPlus.checkCanRead()V boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/net/FrameDecoderWithBbHeader.decode(Ljava/util/Collection;Lorg/apache/cassandra/net/ShareableBytes;I)V boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/service/reeds/repair/RowWriterMergelListener.applyToPartition(Ljava/util/function/Consumer;V boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/utils/AsymmetricOrdering.selectBoundary(Lorg/apache/cassandra/utils/AsymmetricOrdering;Op;II)V boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/utils/AsymmetricOrdering.strictnessToleranceThan(Lorg/apache/cassandra/utils/AsymmetricOrdering;Op;II)V boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/utils/BloomFilter.indexes(Lorg/apache/cassandra/utils/IFilter;FilterKey;)J boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/utils/BloomFilter.setIndexEx(LJJ;J)V boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/utils/ByteBufferUtil.compare([Ljava/nio/ByteBuffer;[B)V boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/utils/ByteBufferUtil.compare([Bljava/nio/ByteBuffer;[B)V boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/utils/ByteBufferUtil.compareUnsigned([Ljava/nio/ByteBuffer;[Ljava/nio/ByteBuffer;)I boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/utils/FastByteOperations.unsafeOperations.compareTo(Ljava/lang/Object;ILjava/lang/Object;J)J boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/utils/FastByteOperations.unsafeOperations.compareTo(Ljava/lang/Object;ILjava/nio/ByteBuffer;)I boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/utils/FastByteOperations.unsafeOperations.compareTo(Ljava/nio/ByteBuffer;Ljava/nio/ByteBuffer;)I boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/utils/concurrent/Awaitables.awaitable(Ljava/util/concurrent/atomic/AtomicReferenceFieldUpdater;Ljava/util/function/Predicate;Lorg/apache/cassandra/utils/concurrent/Awaitable;)V boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/utils/concurrent/Awaitables.awaitableAsyncAwaitable(Ljava/util/concurrent/Awaitable;boolean;)V boolean inline = true
cassandra-server | _CompileCommand: inline org/apache/cassandra/utils/concurrent/Awaitables.awaitUntil(Ljava/util/concurrent/atomic/AtomicReferenceFieldUpdater;Ljava/util/function/Predicate;Lorg/apache/cassandra/utils/concurrent/Awaitable;)Z boolean inline = true
```

Here we can see that the slave node is up:



```
cluster-master | cluster-slave-3: ssh: Could not resolve hostname cluster-slave-3: 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 | cluster-slave-5: ssh: Could not resolve hostname cluster-slave-5: Name or service not known
cluster-master | cluster-slave-2: ssh: Could not resolve hostname cluster-slave-2: Name or service not known
cluster-master | 400 org.apache.hadoop.hdfs.server.namenode.SecondaryNameNode
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: 71336267776 (66.44 GB)
cluster-master | DFS Remaining: 71336243200 (66.44 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 |         Missing block groups (with replication factor 1): 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 | Name: 172.18.0.3:9866 (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: 16332337152 (15.21 GB)
cluster-master | DFS Remaining: 71336243200 (66.44 GB)
cluster-master | DFS Used%: 0.00%
cluster-master | DFS Remaining%: 77.20%
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: Mon Apr 14 16:47:06 GMT 2025
cluster-master | Last Block Report: Mon Apr 14 16:46:45 GMT 2025
cluster-master | Num of Blocks: 0
```

The first mapreduce operation (3 pics):



```
cluster-master | Running Term Frequency MapReduce...
cluster-master | packageJobJar: [] /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar] /tmp/streamjob3732843320514037830.jar tmpDir=null
cluster-master | 2025-04-14 16:54:18,666 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-14 16:54:18,822 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-14 16:54:19,111 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744649219280_0001
cluster-master | 2025-04-14 16:54:21,284 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master | 2025-04-14 16:54:21,258 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master | 2025-04-14 16:54:21,478 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744649219280_0001
cluster-master | 2025-04-14 16:54:21,479 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master | 2025-04-14 16:54:21,677 INFO conf.Configuration: resource-types.xml not found
cluster-master | 2025-04-14 16:54:21,678 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master | 2025-04-14 16:54:22,115 INFO impl.YarnClientImpl: Submitted application application_1744649219280_0001
cluster-master | 2025-04-14 16:54:22,146 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744649219280_0001/
cluster-master | 2025-04-14 16:54:22,147 INFO mapreduce.Job: Running job: job_1744649219280_0001
cluster-master | 2025-04-14 16:54:32,388 INFO mapreduce.Job: Job job_1744649219280_0001 running in uber mode : false
cluster-master | 2025-04-14 16:54:32,395 INFO mapreduce.Job: map 0% reduce 0%
cluster-master | 2025-04-14 16:54:39,522 INFO mapreduce.Job: map 50% reduce 0%
cluster-master | 2025-04-14 16:54:40,540 INFO mapreduce.Job: map 100% reduce 0%
cluster-master | 2025-04-14 16:54:56,883 INFO mapreduce.Job: map 100% reduce 8%
cluster-master | 2025-04-14 16:54:58,822 INFO mapreduce.Job: map 100% reduce 16%
cluster-master | 2025-04-14 16:55:00,844 INFO mapreduce.Job: map 100% reduce 24%
cluster-master | 2025-04-14 16:55:01,859 INFO mapreduce.Job: map 100% reduce 33%
cluster-master | 2025-04-14 16:55:02,878 INFO mapreduce.Job: map 100% reduce 42%
cluster-master | 2025-04-14 16:55:03,894 INFO mapreduce.Job: map 100% reduce 51%
cluster-master | 2025-04-14 16:55:04,982 INFO mapreduce.Job: map 100% reduce 52%
cluster-master | 2025-04-14 16:55:06,917 INFO mapreduce.Job: map 100% reduce 53%
cluster-master | 2025-04-14 16:55:08,934 INFO mapreduce.Job: map 100% reduce 54%
cluster-master | 2025-04-14 16:55:09,937 INFO mapreduce.Job: map 100% reduce 55%
cluster-master | 2025-04-14 16:55:10,948 INFO mapreduce.Job: map 100% reduce 56%
cluster-master | 2025-04-14 16:55:12,968 INFO mapreduce.Job: map 100% reduce 57%
cluster-master | 2025-04-14 16:55:13,978 INFO mapreduce.Job: map 100% reduce 58%
cluster-master | 2025-04-14 16:55:14,990 INFO mapreduce.Job: map 100% reduce 59%
cluster-master | 2025-04-14 16:55:15,997 INFO mapreduce.Job: map 100% reduce 60%
cluster-master | 2025-04-14 16:55:34,278 INFO mapreduce.Job: map 100% reduce 77%
cluster-master | 2025-04-14 16:55:35,285 INFO mapreduce.Job: map 100% reduce 86%
cluster-master | 2025-04-14 16:55:36,294 INFO mapreduce.Job: map 100% reduce 94%
cluster-master | 2025-04-14 16:55:40,330 INFO mapreduce.Job: map 100% reduce 96%
cluster-master | 2025-04-14 16:55:42,345 INFO mapreduce.Job: map 100% reduce 97%
cluster-master | 2025-04-14 16:55:46,370 INFO mapreduce.Job: map 100% reduce 99%
cluster-master | 2025-04-14 16:55:48,382 INFO mapreduce.Job: map 100% reduce 100%
cluster-master | 2025-04-14 16:55:52,499 INFO mapreduce.Job: Job job_1744649219280_0001 completed successfully
cluster-master | 2025-04-14 16:55:52,659 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=24439360
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
```

```
Terminal Shell Edit View Window Help big-data-assignment2 -- docker-compose up -- docker-compose up -- 182x49
cluster-master | File System Counters
cluster-master |     FILE: Number of bytes read=10555855
cluster-master |     FILE: Number of bytes written=24439360
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=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=11
cluster-master |     Data-local map tasks=2
cluster-master |     Total time spent by all maps in occupied slots (ms)=11183
cluster-master |     Total time spent by all reduces in occupied slots (ms)=351324
cluster-master |     Total time spent by all map tasks (ms)=11183
cluster-master |     Total time spent by all reduce tasks (ms)=351324
cluster-master |     Total vcore-milliseconds taken by all map tasks=11183
cluster-master |     Total vcore-milliseconds taken by all reduce tasks=351324
cluster-master |     Total megabyte-milliseconds taken by all map tasks=11451392
cluster-master |     Total megabyte-milliseconds taken by all reduce tasks=359755776
cluster-master | Map-Reduce Framework
cluster-master |     Map input records=1003
cluster-master |     Map output records=578378
cluster-master |     Map output bytes=9390039
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)=107150
cluster-master |     CPU time spent (ms)=107150
cluster-master |     Physical memory (bytes) snapshot=3793678336
cluster-master |     Virtual memory (bytes) snapshot=40988930048
cluster-master |     Total committed heap usage (bytes)=3007840256
cluster-master |     Peak Map Physical memory (bytes)=322379776
cluster-master |     Peak Map Virtual memory (bytes)=3411992576
cluster-master |     Peak Reduce Physical memory (bytes)=357208064
cluster-master |     Peak Reduce Virtual memory (bytes)=4109815808
```

```
Terminal Shell Edit View Window Help big-data-assignment2 -- docker-compose up -- docker-compose up -- 182x49
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)=11183
cluster-master | Total time spent by all reduces in occupied slots (ms)=351324
cluster-master | Total time spent by all map tasks (ms)=11183
cluster-master | Total time spent by all reduce tasks (ms)=351324
cluster-master | Total vcore-milliseconds taken by all map tasks=11183
cluster-master | Total vcore-milliseconds taken by all reduce tasks=351324
cluster-master | Total megabyte-milliseconds taken by all map tasks=11451392
cluster-master | Total megabyte-milliseconds taken by all reduce tasks=359755776
cluster-master | Map-Reduce Framework
cluster-master |     Map input records=1003
cluster-master |     Map output records=578378
cluster-master |     Map output bytes=9390039
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)=1266
cluster-master |     CPU time spent (ms)=107150
cluster-master |     Physical memory (bytes) snapshot=3793678336
cluster-master |     Virtual memory (bytes) snapshot=40988930048
cluster-master |     Total committed heap usage (bytes)=3007840256
cluster-master |     Peak Map Physical memory (bytes)=322379776
cluster-master |     Peak Map Virtual memory (bytes)=3411992576
cluster-master |     Peak Reduce Physical memory (bytes)=357208064
cluster-master |     Peak Reduce Virtual memory (bytes)=4109815808
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=4346244
2025-04-14 16:55:52,659 INFO streaming.StreamJob: Output directory: /tmp/index/output_tf
```

## 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/streamjob5074372695233011396.jar tmpDir=null
cluster-master | 2025-04-14 16:55:55.542 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-14 16:55:55.677 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-14 16:55:55.914 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744649219280_0002
cluster-master | 2025-04-14 16:55:57.002 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master | 2025-04-14 16:55:57.035 INFO mapreduce.JobSubmissionHandler: number of splits=2
cluster-master | 2025-04-14 16:55:57.063 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
cluster-master | 2025-04-14 16:55:57.228 INFO mapreduce.JobSubmissionHandler: Submitting tokens for job: job_1744649219280_0002
cluster-master | 2025-04-14 16:55:57.229 INFO mapreduce.JobSubmissionHandler: Executing with tokens: []
cluster-master | 2025-04-14 16:55:57.369 INFO conf.Configuration: resource-types.xml not found
cluster-master | 2025-04-14 16:55:57.378 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master | 2025-04-14 16:55:57.453 INFO impl.YarnClientImpl: Submitted application application_1744649219280_0002
cluster-master | 2025-04-14 16:55:57.488 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744649219280_0002/
cluster-master | 2025-04-14 16:55:57.489 INFO mapreduce.Job: Running job: job_1744649219280_0002
cluster-master | 2025-04-14 16:56:05.707 INFO mapreduce.Job: Job job_1744649219280_0002 running in uber mode : false
cluster-master | 2025-04-14 16:56:05.711 INFO mapreduce.Job: map 0% reduce 0%
cluster-master | 2025-04-14 16:56:12.899 INFO mapreduce.Job: map 100% reduce 0%
cluster-master | 2025-04-14 16:56:12.899 INFO mapreduce.Job: map 100% reduce 0%
cassandra-server | INFO [Notification Thread] 2025-04-14 16:56:24,234 GCInspector.java:285 - G1 Young Generation GC in 438ms. G1 Eden Space: 1895825408 -> 0; G1 Old Gen: 158603776
-> 192588288; G1 Survivor Space: 39847016 -> 45479968;
cluster-master | 2025-04-14 16:56:30.172 INFO mapreduce.Job: map 100% reduce 25%
cluster-master | 2025-04-14 16:56:31.182 INFO mapreduce.Job: map 100% reduce 50%
cluster-master | 2025-04-14 16:56:33.284 INFO mapreduce.Job: map 100% reduce 75%
cluster-master | 2025-04-14 16:56:34.214 INFO mapreduce.Job: map 100% reduce 100%
cluster-master | 2025-04-14 16:56:34.254 INFO mapreduce.Job: Job job_1744649219280_0002 completed successfully
cluster-master | 2025-04-14 16:56:34.354 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)=8120
cluster-master |         Total time spent by all reduces in occupied slots (ms)=64322
cluster-master |         Total time spent by all map tasks (ms)=8120
cluster-master |         Total time spent by all reduce tasks (ms)=64322
cluster-master |         Total vcore-milliseconds taken by all map tasks=8120
cluster-master |         Total vcore-milliseconds taken by all reduce tasks=64322
cluster-master |         Total megabyte-milliseconds taken by all map tasks=8314880
cluster-master |         Total megabyte-milliseconds taken by all reduce tasks=65865728
cluster-master |     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=2703589
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)=551
cluster-master |         CPU time spent (ms)=2020
cluster-master |         Physical memory (bytes) snapshot=1887485952
cluster-master |         Virtual memory (bytes) snapshot=20487966720
cluster-master |         Total committed heap usage (bytes)=1518338048
cluster-master |         Peak Map Physical memory (bytes)=324669920
cluster-master |         Peak Map Virtual memory (bytes)=3413942272
cluster-master |         Peak Reduce Physical memory (bytes)=359900224
cluster-master |         Peak Reduce Virtual memory (bytes)=4108648448
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
2025-04-14 16:56:34.354 INFO streaming.StreamJob: Output directory: /tmp/index/output_df
```

```
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)=8120
cluster-master | Total time spent by all reduces in occupied slots (ms)=64322
cluster-master | Total time spent by all map tasks (ms)=8120
cluster-master | Total time spent by all reduce tasks (ms)=64322
cluster-master | Total vcore-milliseconds taken by all map tasks=8120
cluster-master | Total vcore-milliseconds taken by all reduce tasks=64322
cluster-master | Total megabyte-milliseconds taken by all map tasks=8314880
cluster-master | Total megabyte-milliseconds taken by all reduce tasks=65865728
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=2703589
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)=551
cluster-master | CPU time spent (ms)=2020
cluster-master | Physical memory (bytes) snapshot=1887485952
cluster-master | Virtual memory (bytes) snapshot=20487966720
cluster-master | Total committed heap usage (bytes)=1518338048
cluster-master | Peak Map Physical memory (bytes)=324669920
cluster-master | Peak Map Virtual memory (bytes)=3413942272
cluster-master | Peak Reduce Physical memory (bytes)=359900224
cluster-master | Peak Reduce Virtual memory (bytes)=4108648448
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
2025-04-14 16:56:34.354 INFO streaming.StreamJob: Output directory: /tmp/index/output_df
```

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 indexer...
cluster-master | First 5 rows in tf 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='dobson', 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 Tobe - 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 |     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
cluster-master |     com.datastax.spark#spark-cassandra-connector_2.12 added as a dependency
cluster-master |     :: resolving dependencies :: org.apache.spark#spark-submit-parent-a623c008-9277-42c7-b25d-bb7bd6180fc7;1.0
cluster-master |     confs: [default]
cluster-master |     found com.datastax.spark#spark-cassandra-connector_2.12;3.1.0 in central
cluster-master |     found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.1.0 in central
cluster-master |     found com.datastax.oss#java-driver-core-shaded;4.12.0 in central
cluster-master |     found com.datastax.oss#native-protocol;15.0 in central
cluster-master |     found com.datastax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 in central
cluster-master |     found com.typesafe#config;1.4.1 in central
cluster-master |     found org.slf4j#slf4j-api;1.7.26 in central
cluster-master |     found io.dropwizard.metrics#metrics-core;4.1.18 in central
cluster-master |     found org.hdrhistogram#HdrHistogram;2.1.12 in central
cluster-master |     found org.reactivestreams#reactive-streams;1.0.3 in central
cluster-master |     found com.github.stephenc.jcip#jcip-annotations;1.0-1 in central
cluster-master |     found com.github.spotbugs#spotbugs-annotations;3.1.12 in central
cluster-master |     found com.google.code.findbugs#jsr305;3.0.2 in central
cluster-master |     found com.datastax.oss#java-driver-mapper-runtime;4.12.0 in central
cluster-master |     found com.datastax.oss#java-driver-query-builder;4.12.0 in central
cluster-master |     found org.apache.commons#commons-lang3;3.10 in central
cluster-master |     found com.thoughtworks.parameter#parameter;2.8 in central
cluster-master |     found org.scala-lang#scala-reflect;2.12.11 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
cluster-master |     Topic: A Kiss Before Dying (1991 film)
cluster-master |     Score: 13.6715
cluster-master |     Text: "A Kiss Before Dying is a 1991 American romantic thriller film directed by James Dearden, and based ...
cluster-master |
cluster-master | 2. Document ID: 63002694
cluster-master |     Topic: A Close Call for Ellery Queen
cluster-master |     Score: 5.7367
cluster-master |     Text: A Close Call for Ellery Queen is a 1942 American mystery film directed by James P. Hogan and written...
cluster-master |
cluster-master | 3. Document ID: 63016837
cluster-master |     Topic: A Desperate Chance for Ellery Queen
cluster-master |     Score: 5.7087
cluster-master |     Text: A Desperate Chance for Ellery Queen is a 1942 American mystery film directed by James P. Hogan and w...
cluster-master |
cluster-master | 4. Document ID: 57704768
cluster-master |     Topic: A Gentleman of Quality
cluster-master |     Score: 5.6547
cluster-master |     Text: A Gentleman of Quality is a 1919 American silent drama film directed by James Young and starring Ear...
cluster-master |
cluster-master | 5. Document ID: 19749985
cluster-master |     Topic: A Glimpse of Hell (film)
cluster-master |     Score: 5.5722
cluster-master |     Text: "A Glimpse of Hell is a 2001 American-Canadian made-for-television drama film directed by Mikael Sal...
cluster-master |
cluster-master | 6. Document ID: 32997171
cluster-master |     Topic: A Girl of the Bush
cluster-master |     Score: 5.5438
cluster-master |     Text: A Girl of the Bush is a 1921 Australian silent film directed by Franklyn Barrett. It is one of the f...
cluster-master |
cluster-master | 7. Document ID: 65891184
cluster-master |     Topic: A Cry from Within
cluster-master |     Score: 5.4639
cluster-master |     Text: A Cry From Within is a 2015 American horror film directed by Deborah Twiss and Zach Miller and starr...
cluster-master |
cluster-master | 8. Document ID: 30011413
cluster-master |     Topic: A Harlot's Progress (film)
cluster-master |     Score: 5.4331
cluster-master |     Text: A Harlot's Progress is a 2006 British television film directed by Justin Hardy and starring Zoe Tapp...
cluster-master |
cluster-master | 9. Document ID: 46908278
cluster-master |     Topic: A Gamble for Love
cluster-master |     Score: 5.4220
cluster-master |     Text: A Gamble for Love is a 1917 British silent sports film directed by Frank Wilson and starring Gerald ...
cluster-master |
cluster-master | 10. Document ID: 64378948
cluster-master |     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: 9.9185
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 chai...
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.