

Report

Big Data: Assignment 2

Simple Search Engine using Hadoop MapReduce

Mariia Shmakova

April 20, 2025

I Methodology

A. *Data Collection and Preparation*

- Initially, a parquet file containing the ID, title, and text fields for each document was downloaded. Using PySpark, data was extracted from 1000 documents and saved in a specified file format.
- Each document was named using its id and title, with spaces replaced by underscores.
- The processed documents were stored in HDFS in the data folder and an intermediate RDD was created to store the document metadata in `/index/data`.

B. *Indexer task*

Documents are stored in HDFS (`/index/data`) as plain text files. Each document is processed line-by-line to extract words. Indexing implemented using Hadoop Streaming with Python-based mapper and reducer scripts.

Hadoop Streaming was chosen for its flexibility in integrating custom mapper and reducer scripts written in Python. This allows us to process large datasets efficiently while maintaining simplicity in implementation.

The mapper tokenizes the text into words and emits key-value pairs (word, document ID). The reducer aggregates these pairs to build the inverted index.

C. *Ranker task*

- **BM25 Implementation** : The ranker script (`query.py`) calculates BM25 scores for all documents based on the user query. It retrieves data from Cassandra, computes scores, and ranks the top 10 documents.
- **Broadcast Variables** : To optimize performance, we used Spark broadcast variables to share BM25 statistics across all nodes.
- **RDD API** : The implementation relies on PySpark's RDD API for distributed computation.

II Demonstration

A. Data Preparation section

```
MINGW64/C:/Users/rover/Desktop/BD_ASSIGN/Smokov...
cassandra-server INFO [Native-Transport-Requests-7] 2025-04-20 08:07:37.692 ColumnFamilyStore.java:499 - Initializing index_keyspace.document_index
ap, OB (OK) off-head INFO [Native-Transport-Requests-7] 2025-04-20 08:07:37.778 ColumnFamilyStore.java:1052 - Enqueuing flush of system.schema.columns, Reason: INTERNALLY_FORCED, Usage: 1.096Kib (OK) on-hea
cassandra-server INFO [PeriodicMentableLshWriter-0:4] 2025-04-20 08:07:37.822 Flushing-Java:153 - Writing Memtable-column_masks1397487320(1438 serialized bytes, 3 ops, 1.096Kib OK) on-head, OB (OK) off-h
eap, } Flushed range = [min(-922337203685475808), max(922337203685475807)]
cassandra-server INFO [PeriodicMentableLshWriter-0:4] 2025-04-20 08:07:37.823 Flushing-Java:179 - Completed Flushing /var/lib/cassandra/data/system_schema/column_masks-738c5ed01683268bd9d183dbdc278af/nb-4
-big-data.db (21248) for commitlog position CommitLogPosition(segmentId=1745135809599, position=200304)
cassandra-server INFO [Native-Transport-Requests-7] 2025-04-20 08:07:37.866 ColumnFamilyStore.java:1052 - Enqueuing flush of system.schema.columns, Reason: INTERNALLY_FORCED, Usage: 2.712Kib (OK) on-head, g
(OK) off-head INFO [PeriodicMentableLshWriter-0:3] 2025-04-20 08:07:37.868 CompactionTask.java:167 - Compacting (860fce0-1dbe-11f0-9cfo-8342740d7660) [/var/lib/cassandra/data/system_schema/column_masks-738c5ed01683268bd9d183dbdc278af/nb-4-big-data.db; level=0, /var/lib/cassandra/data/system_schema/column_masks-738c5ed01683268bd9d183dbdc278af/nb-3-big-data.db; level=0, /var/lib/cassandra/data/system_schema/column_masks-738c5ed01683268bd9d183dbdc278af/nb-2-big-data.db; level=0, /var/lib/cassandra/data/system_schema/column_masks-738c5ed01683268bd9d183dbdc278af/nb-1-big-data.db; level=0, ]
cassandra-server INFO [PeriodicMentableLshWriter-0:3] 2025-04-20 08:07:37.907 Flushing-Java:153 - Writing Memtable-columns204055044(5468 serialized bytes, 3 ops, 2.712Kib OK) on-head, OB (OK) off-head,
cassandra-server INFO [PeriodicMentableLshWriter-0:3] 2025-04-20 08:07:37.908 Flushing-Java:179 - Completed Flushing /var/lib/cassandra/data/system_schema/columns-2410125a2ae3af787c1b40eac3a33f/nb-7-big-
Data.db (2526) for commitlog position CommitLogPosition(segmentId=1745135809599, position=200304)
cassandra-server INFO [Native-Transport-Requests-7] 2025-04-20 08:07:37.976 ColumnFamilyStore.java:1052 - Enqueuing Flush of system.schema.tables, Reason: INTERNALLY_FORCED, Usage: 2.878Kib (OK) on-head, OB
(OK) off-head INFO [CompactionSector-3] 2025-04-20 08:07:38.012 CompactionTask.java:258 - Compacted (860fce0-1dbe-11f0-9cfo-8342740d7660) 4 sstables to [/opt/cassandra/data/data/system.schema.columns-m
cassandra-server INFO [PeriodicMentableLshWriter-0:3] 2025-04-20 08:07:38.013 BigFormat-Java:231 - Deleting stable: /opt/cassandra/data/data/system.schema/column_masks-738c5ed01683268bd9d183dbdc278af/nb-4-big-
Data.db (21248) for commitlog position CommitLogPosition(segmentId=1745135809599, position=200304)
cassandra-server INFO [NonPeriodicTasks:] 2025-04-20 08:07:38.013 BigFormat-Java:231 - Deleting stable: /opt/cassandra/data/data/system.schema/column_masks-738c5ed01683268bd9d183dbdc278af/nb-4-big-
Data.db (21248) for commitlog position CommitLogPosition(segmentId=1745135809599, position=200304)
cassandra-server INFO [PeriodicMentableLshWriter-0:4] 2025-04-20 08:07:38.02 Flushing-Java:153 - Writing Memtable-tables139261697(7868 serialized bytes, 1 op, 2.878Kib OK) on-head, OB (OK) off-head,
Flushed range = [min(-922337203685475808), max(922337203685475807)]
cassandra-server INFO [PeriodicMentableLshWriter-0:4] 2025-04-20 08:07:38.024 Flushing-Java:179 - Completed Flushing /var/lib/cassandra/data/system_schema/tables-addffdf9dbcc3068605eedcd302ba09/nb-7-big-D
cassandra-server INFO [NonPeriodicTasks:] 2025-04-20 08:07:38.024 BigFormat-Java:231 - Deleting stable: /opt/cassandra/data/data/system.schema/column_masks-738c5ed01683268bd9d183dbdc278af/nb-3-big-
cassandra-server INFO [NonPeriodicTasks:] 2025-04-20 08:07:38.029 BigFormat-Java:231 - Deleting stable: /opt/cassandra/data/data/system.schema/column_masks-738c5ed01683268bd9d183dbdc278af/nb-2-big-
Data.db (21248) for commitlog position CommitLogPosition(segmentId=1745135809599, position=200304)
cassandra-server INFO [Native-Transport-Requests-7] 2025-04-20 08:07:38.165 ColumnFamilyStore.java:1052 - Enqueuing flush of system.schema.keyspaces, Reason: INTERNALLY_FORCED, Usage: 730B (OK) on-head, OB
(OK) off-head INFO [PeriodicMentableLshWriter-0:3] 2025-04-20 08:07:38.208 Flushing-Java:153 - Writing Memtable-keyspaces8185157366(1578 serialized bytes, 1 ops, 730B OK) on-head, OB (OK) off-head, f
Flushed range = [min(-922337203685475808), max(922337203685475807)]
cassandra-server INFO [PeriodicMentableLshWriter-0:3] 2025-04-20 08:07:38.209 Flushing-Java:179 - Completed Flushing /var/lib/cassandra/data/system_schema/keyspaces-abc5682de6a31cb553b3dbcfdf0fb/nb-8-bi
g-data.db (1248) for commitlog position CommitLogPosition(segmentId=1745135809599, position=201293)
INFO [CompactionSector-3] 2025-04-20 08:07:38.257 CompactionTask.java:167 - Compacting (8646b90-1dbe-11f0-9cfo-8342740d7660) 4 sstables to [/opt/cassandra/data/system_schema/keyspaces-abc5682de6a31cb553b3dbcfdf0fb/nb-8-big-data.db; level=0, /var/lib/cassandra/data/system_schema/keyspaces-abc5682de6a31cb553b3dbcfdf0fb/nb-7-big-data.db; level=0, /var/lib/cassandra/data/system_schema/keyspaces-abc5682de6a31cb553b3dbcfdf0fb/nb-6-big-data.db; level=0, /var/lib/cassandra/data/system_schema/keyspaces-abc5682de6a31cb553b3dbcfdf0fb/nb-5-big-data.db; level=0, /var/lib/cassandra/data/system_schema/keyspaces-abc5682de6a31cb553b3dbcfdf0fb/nb-4-big-data.db; level=0, /var/lib/cassandra/data/system_schema/keyspaces-abc5682de6a31cb553b3dbcfdf0fb/nb-3-big-data.db; level=0, /var/lib/cassandra/data/system_schema/keyspaces-abc5682de6a31cb553b3dbcfdf0fb/nb-2-big-data.db; level=0, /var/lib/cassandra/data/system_schema/keyspaces-abc5682de6a31cb553b3dbcfdf0fb/nb-1-big-data.db; level=0, ]
cassandra-server INFO [Native-Transport-Requests-7] 2025-04-20 08:07:38.278 Keyspace-Java:379 - Creating replication strategy index_keyspace.bm2_stats params KeyspaceParams(durable_writers=true, replicationReplica
tor=SimpleStrategy(factor=1))
cassandra-server INFO [Native-Transport-Requests-7] 2025-04-20 08:07:38.286 ColumnFamilyStore.java:499 - Initializing index_keyspace.bm2_stats
cluster-master INFO [Keyspace] Keyspace and tables created successfully!
cluster-master INFO [CassandraSetup] Cassandra setup completed!
cassandra-server INFO [CompactionSector-3] 2025-04-20 08:07:38.471 CompactionTask.java:258 - Compacted (8646b90-1dbe-11f0-9cfo-8342740d7660) 4 sstables to [/opt/cassandra/data/data/system.schema/keyspace
s-abc5682de6a31cb553b3dbcfdf0fb/nb-9-big-data.db; level=0, 6868 to 2908 (-42% of original) in 210ms. Read throughput = 3.183Kib/s, Write Throughput = 1.346Kib/s, Row Throughput = -12/s, 9 total partitions mer
ged to 6. Partition merge ratios were {15, 41, 41}. Time spent writing keys = 77ms
cassandra-server INFO [NonPeriodicTasks:] 2025-04-20 08:07:38.472 BigFormat-Java:231 - Deleting stable: /opt/cassandra/data/data/system.schema/keyspaces-abc5682de6a31cb553b3dbcfdf0fb/nb-8-big-
Data.db (1248) for commitlog position CommitLogPosition(segmentId=1745135809599, position=201293)
cassandra-server INFO [NonPeriodicTasks:] 2025-04-20 08:07:38.477 BigFormat-Java:231 - Deleting stable: /opt/cassandra/data/data/system.schema/keyspaces-abc5682de6a31cb553b3dbcfdf0fb/nb-7-big-
Data.db (1248) for commitlog position CommitLogPosition(segmentId=1745135809599, position=201293)
cassandra-server INFO [NonPeriodicTasks:] 2025-04-20 08:07:38.482 BigFormat-Java:231 - Deleting stable: /opt/cassandra/data/data/system.schema/keyspaces-abc5682de6a31cb553b3dbcfdf0fb/nb-6-big-
Data.db (1248) for commitlog position CommitLogPosition(segmentId=1745135809599, position=201293)
cassandra-server INFO [NonPeriodicTasks:] 2025-04-20 08:07:38.493 BigFormat-Java:231 - Deleting stable: /opt/cassandra/data/data/system.schema/keyspaces-abc5682de6a31cb553b3dbcfdf0fb/nb-5-big-
Data.db (1248) for commitlog position CommitLogPosition(segmentId=1745135809599, position=201293)
Preparing data...
Activating virtual environment...
Checking source parquet file...
Setting up HDFS directories...
Copying parquet file to HDFS...
Verifying HDFS file...
Starting Spark data preparation job...
Successfully created 1000 documents in /app/data
Data preparation completed successfully!
Copying local input to HDFS...
```

B. Indexer tasks section

Running indexer:

```

MINGW64/c/Users/user/Desktop/BD/S25_BD_assign2_Shmakova
cassandra-server INFO [NonPeriodicTasks:1] 2025-04-20 08:33:08.383 Bigformat.java:231 - Deleting sstable: /opt/cassandra/data/data/system_schema/keyspaces-abac5682dea631c5b353b3d6cfff0fb6/nb-7-big
cassandra-server INFO [NonPeriodicTasks:1] 2025-04-20 08:33:08.387 Bigformat.java:231 - Deleting sstable: /opt/cassandra/data/data/system_schema/keyspaces-abac5682dea631c5b353b3d6cfff0fb6/nb-6-big
cassandra-server INFO [NonPeriodicTasks:1] 2025-04-20 08:33:08.389 Bigformat.java:231 - Deleting sstable: /opt/cassandra/data/data/system_schema/keyspaces-abac5682dea631c5b353b3d6cfff0fb6/nb-5-big
Cluster-master Preparing data...
Cluster-master Activating virtual environment...
Cluster-master Checking source parquet file...
Cluster-master Error: Source parquet file not found at /data/a.parquet
Cluster-master Copying local input to HDFS...
Cluster-master Indexing files from: /index/tmp/input
Cluster-master Collecting packages...
Cluster-master Packing environment at '/app/.venv' to '/app/.venv.tar.gz'
Cluster-master [#####] 100% Completed | 2min 0.9s
Cluster-master packagejob.jar: /tmp/hadoop-unjar7331683581074837399/ [ ] /tmp/streamjob895161070147320.jar tmpDir=null
Cluster-master 2025-04-20 08:35:33.469 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.3:8032
Cluster-master 2025-04-20 08:35:33.744 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.3:8032
Cluster-master 2025-04-20 08:36:06.543 INFO mapred.FileInputFormat: Total input files to process : 20
Cluster-master 2025-04-20 08:36:06.609 INFO mapreduce.JobSubmitter: number of splits:20
Cluster-master 2025-04-20 08:36:07.221 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1745137721999_0001
Cluster-master 2025-04-20 08:36:07.321 INFO mapreduce.JobSubmitter: Executing with tokens: [ ]
Cluster-master 2025-04-20 08:36:07.590 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'
Cluster-master 2025-04-20 08:36:08.161 INFO impl.YarnClientImpl: Submitted application application_1745137721999_0001
Cluster-master 2025-04-20 08:36:08.214 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1745137721999_0001/
Cluster-master 2025-04-20 08:36:08.216 INFO mapreduce.Job: Running job: job_1745137721999_0001
Cluster-master 2025-04-20 08:36:30.415 INFO mapreduce.Job: Job job_1745137721999_0001 running in uber mode : false
Cluster-master 2025-04-20 08:36:30.418 INFO mapreduce.Job: map 0% reduce 0%
Cluster-master 2025-04-20 08:36:42.760 INFO mapreduce.Job: map 15% reduce 0%
Cluster-master 2025-04-20 08:36:52.004 INFO mapreduce.Job: map 25% reduce 0%
Cluster-master 2025-04-20 08:36:53.011 INFO mapreduce.Job: map 30% reduce 0%
Cluster-master 2025-04-20 08:37:03.232 INFO mapreduce.Job: map 45% reduce 0%
Cluster-master 2025-04-20 08:37:13.514 INFO mapreduce.Job: map 60% reduce 0%
Cluster-master 2025-04-20 08:37:23.811 INFO mapreduce.Job: map 70% reduce 0%
Cluster-master 2025-04-20 08:37:29.846 INFO mapreduce.Job: map 80% reduce 0%
Cluster-master 2025-04-20 08:37:34.932 INFO mapreduce.Job: map 80% reduce 7%
Cluster-master 2025-04-20 08:37:35.939 INFO mapreduce.Job: map 90% reduce 7%
Cluster-master 2025-04-20 08:37:40.009 INFO mapreduce.Job: map 90% reduce 8%
Cluster-master 2025-04-20 08:37:41.016 INFO mapreduce.Job: map 100% reduce 8%
Cluster-master 2025-04-20 08:37:43.100 INFO mapreduce.Job: map 100% reduce 23%
Cluster-master 2025-04-20 08:37:53.234 INFO mapreduce.Job: map 100% reduce 100%
Cluster-master 2025-04-20 08:37:54.250 INFO mapreduce.Job: Job job_1745137721999_0001 completed successfully
Cluster-master 2025-04-20 08:37:54.363 INFO mapreduce.Job: Counters: 54
Cluster-master File System Counters
Cluster-master FILE: Number of bytes read=24
Cluster-master FILE: Number of bytes written=6676614
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=104743
Cluster-master HDFS: Number of bytes written=0
Cluster-master HDFS: Number of read operations=80
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 Launched map tasks=20
Cluster-master Launched reduce tasks=4
Cluster-master Data-local map tasks=20
Cluster-master Total time spent by all maps in occupied slots (ms)=324396
Cluster-master Total time spent by all reduces in occupied slots (ms)=119332
Cluster-master Total time spent by all map tasks (ms)=162198
Cluster-master FILE: Number of bytes written=6676614
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=104743
Cluster-master HDFS: Number of bytes written=0
Cluster-master HDFS: Number of read operations=80
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 Launched map tasks=20
Cluster-master Launched reduce tasks=4
Cluster-master Data-local map tasks=20
Cluster-master Total time spent by all maps in occupied slots (ms)=324396
Cluster-master Total time spent by all reduces in occupied slots (ms)=119332
Cluster-master Total time spent by all map tasks (ms)=162198
Cluster-master Total time spent by all reduce tasks (ms)=59666
Cluster-master Total vcore-milliseconds taken by all map tasks=162198
Cluster-master Total vcore-milliseconds taken by all reduce tasks=59666
Cluster-master Total megabyte-milliseconds taken by all map tasks=332181504
Cluster-master Total megabyte-milliseconds taken by all reduce tasks=122195968
Cluster-master Map-Reduce Framework
Cluster-master Map input records=20
Cluster-master Map output records=0
Cluster-master Map output bytes=0
Cluster-master Map output materialized bytes=480
Cluster-master Input split bytes=2323
Cluster-master Combine input records=0
Cluster-master Combine output records=0
Cluster-master Reduce input groups=0
Cluster-master Reduce shuffle bytes=480
Cluster-master Reduce input records=0
Cluster-master Reduce output records=0
Cluster-master Spilled Records=0
Cluster-master Shuffled Maps=80
Cluster-master Failed Shuffles=0
Cluster-master Merged Map outputs=80
Cluster-master GC time elapsed (ms)=5007
Cluster-master CPU time spent (ms)=16960
Cluster-master Physical memory (bytes) snapshot=6360211456
Cluster-master Virtual memory (bytes) snapshot=8214243280
Cluster-master Total committed heap usage (bytes)=5068816384
Cluster-master Peak Map Physical memory (bytes)=295190528
Cluster-master Peak Map Virtual memory (bytes)=3423092786
Cluster-master Peak Reduce Physical memory (bytes)=192126976
Cluster-master Peak Reduce Virtual memory (bytes)=3427926016
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=102220
Cluster-master File Output Format Counters
Cluster-master Bytes Written=0
Cluster-master 2025-04-20 08:37:54.363 INFO streaming.StreamJob: Output directory: /index/output
Cluster-master Deleted /index/tmp/input
Cluster-master Indexing completed successfully

```

C. Ranker tasks section

```
MINGW64/C:/Users/user/Desktop/BD/S25_BD_assign2_Shmakova
cluster-master | 25/04/20 08:38:40 INFO Executor: Fetching spark://cluster-master:39385/jars/com.github.spotbugs.spotbugs-annotations-3.1.12.jar with timestamp 1745138318435
cluster-master | 25/04/20 08:38:40 INFO Utils: Fetching spark://cluster-master:39385/jars/com.github.spotbugs.spotbugs-annotations-3.1.12.jar to /tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/fetchFileTemp8102227884798840462.tmp
cluster-master | 25/04/20 08:38:40 INFO Utils: /tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/fetchFileTemp8102227884798840462.tmp has been previously copied to /tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/com.github.spotbugs.spotbugs-annotations-3.1.12.jar
cluster-master | 25/04/20 08:38:40 INFO Executor: Adding file:/tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/com.github.spotbugs.spotbugs-annotations-3.1.12.jar to class loader default
cluster-master | 25/04/20 08:38:40 INFO Executor: Fetching spark://cluster-master:39385/jars/com.thoughtworks.paramater-paramater-2.8.jar with timestamp 1745138318435
cluster-master | 25/04/20 08:38:40 INFO Utils: Fetching spark://cluster-master:39385/jars/com.thoughtworks.paramater-paramater-2.8.jar to /tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/fetchFileTemp312932582072543839.tmp
cluster-master | 25/04/20 08:38:40 INFO Utils: /tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/fetchFileTemp312932582072543839.tmp has been previously copied to /tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/com.thoughtworks.paramater-paramater-2.8.jar
cluster-master | 25/04/20 08:38:40 INFO Executor: Adding file:/tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/com.thoughtworks.paramater-paramater-2.8.jar to class loader default
cluster-master | 25/04/20 08:38:40 INFO Executor: Fetching spark://cluster-master:39385/jars/com.github.stephenc.jcip.jcip-annotations-1.0-1.jar with timestamp 1745138318435
cluster-master | 25/04/20 08:38:40 INFO Utils: Fetching spark://cluster-master:39385/jars/com.github.stephenc.jcip.jcip-annotations-1.0-1.jar to /tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/fetchFileTemp6961536798234623957.tmp
cluster-master | 25/04/20 08:38:40 INFO Utils: /tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/fetchFileTemp6961536798234623957.tmp has been previously copied to /tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/com.github.stephenc.jcip.jcip-annotations-1.0-1.jar
cluster-master | 25/04/20 08:38:40 INFO Executor: Adding file:/tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/com.github.stephenc.jcip.jcip-annotations-1.0-1.jar to class loader default
cluster-master | 25/04/20 08:38:40 INFO Executor: Fetching spark://cluster-master:39385/jars/io.dropwizard.metrics_metrics-core-4.1.18.jar with timestamp 1745138318435
cluster-master | 25/04/20 08:38:40 INFO Utils: Fetching spark://cluster-master:39385/jars/io.dropwizard.metrics_metrics-core-4.1.18.jar to /tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/fetchFileTemp7061278391386563980.tmp
cluster-master | 25/04/20 08:38:40 INFO Utils: /tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/fetchFileTemp7061278391386563980.tmp has been previously copied to /tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/io.dropwizard.metrics_metrics-core-4.1.18.jar
cluster-master | 25/04/20 08:38:40 INFO Executor: Adding file:/tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/userFiles-c5fb82d7-31a7-4729-80f0-e58db744a454/io.dropwizard.metrics_metrics-core-4.1.18.jar to class loader default
cluster-master | 25/04/20 08:38:40 INFO NettyBlockTransferService: Server created on cluster-master:45923.
cluster-master | 25/04/20 08:38:40 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
cluster-master | 25/04/20 08:38:40 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, cluster-master, 45923, None)
cluster-master | 25/04/20 08:38:40 INFO BlockManagerMasterEndpoint: Registering block manager cluster-master:45923 with 366.3 MiB RAM, BlockManagerId(driver, cluster-master, 45923, None)
cluster-master | 25/04/20 08:38:40 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, cluster-master, 45923, None)
cluster-master | 25/04/20 08:38:40 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, cluster-master, 45923, None)
cassandra-server | WARN [Native-Transport-Requests-1] 2025-04-20 08:38:41,963 SelectStatement.java:557 - Aggregation query used without partition key on table index_keyspace.document_index, aggregation type: AGGREGATE EVERYTHING
cassandra-server | WARN [Native-Transport-Requests-2] 2025-04-20 08:38:42,014 SelectStatement.java:557 - Aggregation query used without partition key on table index_keyspace.bm25_stats, aggregation type: AGGREGATE EVERYTHING
cluster-master | Traceback (most recent call last):
cluster-master |   File "/app/query.py", line 95, in <module>
cluster-master |     main()
cluster-master |   File "/app/query.py", line 41, in main
cluster-master |     query_terms = re.findall(r'ws', query.lower())
cluster-master | NameError: name 're' is not defined
cluster-master | 25/04/20 08:38:42 INFO SparkContext: Invoking stop() from shutdown hook
cluster-master | 25/04/20 08:38:42 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master | 25/04/20 08:38:42 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master | 25/04/20 08:38:42 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master | 25/04/20 08:38:42 INFO MemoryStore: MemoryStore cleared
cluster-master | 25/04/20 08:38:42 INFO BlockManager: BlockManager stopped
cluster-master | 25/04/20 08:38:42 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master | 25/04/20 08:38:42 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master | 25/04/20 08:38:42 INFO SparkContext: Successfully stopped SparkContext
cluster-master | 25/04/20 08:38:42 INFO ShutdownHookManager: Shutdown hook called
cluster-master | 25/04/20 08:38:42 INFO ShutdownHookManager: Deleting directory /tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6/pyspark-9aab6d4-fdf1-4ce1-9375-9373dc290cc3
cluster-master | 25/04/20 08:38:42 INFO ShutdownHookManager: Deleting directory /tmp/spark-2ca5ad29-a406-456e-9ae4-c7e0051f44f4
cluster-master | 25/04/20 08:38:42 INFO ShutdownHookManager: Deleting directory /tmp/spark-22fc2b25-1da2-4a8a-bbaf-6a3a46ff97c6
cluster-master | Successfully search!
```