

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

```
x MININGW@C:\Users\root\Desktop\BQ\S2S_DB_Assign_Sharding> java -Dlog.dir=./logs -jar cassandra-server.jar  
cassandra-server INFO [Native-Transport-Requests-7] 2025-04-20 08:07:37.692 ColumnFamilyStore.java:499 - Initializing index_keyspace.docment_index  
cassandra-server INFO [Native-Transport-Requests-7] 2025-04-20 08:07:37.778 ColumnFamilyStore.java:1052 - Enqueuing Flush of system.schema.columns_masks, Reason: INTERNALLY_FORCED, Usage: 1.096KIB (0%) on-heap,  
ob_ OB (0%) off-heap  
cassandra-server INFO [PerDiskMemtableLshWriter:0:] 2025-04-20 08:07:37.822 Flushing.java:153 - Writing Memtable-column_msk1937487320(1438 serialized bytes, 3 ops, 1.096KIB (0%) on-heap, 0B (0%) off-heap)  
eap) flushed range = [min(-9223372036854775808), max(9223372036854775807)]  
cassandra-server INFO [PerDiskMemtableLshWriter:0:] 2025-04-20 08:07:37.823 Flushing.java:179 - Completed flushing /var/lib/cassandra/data/system_schema/columns_masks-738c5cd01683268bd18534dbec278af/nb-1-big-data.db.level=0.  
Big-Data db (1258) 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 (0%) on-heap,  
OB (0%) off-heap  
cassandra-server INFO [CompactionExecutor:3] 2025-04-20 08:07:37.868 CompactionTask.java:167 - Compactng (860fcead-1dbe-11f0-9cf0-83427d4d7660) [/var/lib/cassandra/data/system_schema/columns_masks-738c5cd01683268bd18534dbec278af/nb-1-big-data.db.level=0./var/lib/cassandra/data/system_schema/columns_masks-738c5cd01683268bd18534dbec278af/nb-1-big-data.db.level=0.] Time spent writing keyspace=  
cassandra-server INFO [PerDiskMemtableLshWriter:0:] 2025-04-20 08:07:37.907 Flushing.java:153 - Writing Memtable-columns204055044(5468 serialized bytes, 3 ops, 2.712KIB (0%) on-heap, 0B (0%) off-heap).  
Flushed range = [min(-9223372036854775808), max(9223372036854775807)]  
cassandra-server INFO [PerDiskMemtableLshWriter:0:] 2025-04-20 08:07:37.908 Flushing.java:179 - Completed flushing /var/lib/cassandra/data/system_schema/columns_24101C25a2acbf787ClbOoeIaca33fb/nb-7-big-  
Data db (2528) 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 (0%) on-heap, 0B (0%) off-heap  
cassandra-server INFO [CompactionExecutor:3] 2025-04-20 08:07:38.012 CompactionTask.java:258 - Compacted (860fcead-1dbe-11f0-9cf0-83427d4d7660) 4 stable to /opt/cassandra/data/data/system_schema/column_masks-738c5cd01683268bd18534dbec278af/nb-5-big-, to level=0. 1.491KIB to 3.392KIB (~9% of original) in 14ms. Read throughput = 24.63KB/s, Write Throughput = >23.90KB/s, Row Throughput = >256/s. 5 total partitions merged to 6. Partition merge counts were [-15, 4, 1]. Time spent writing keyspace=  
cassandra-server INFO [NonPeriodicTasks:1] 2025-04-20 08:07:38.013 BiFormat.java:231 - Deleting stable: /opt/cassandra/data/data/system_schema/column_masks-738c5cd01683268bd18534dbec278af/nb-4-big-  
cassandra-server INFO [NonPeriodicTasks:1] 2025-04-20 08:07:38.019 BiFormat.java:231 - Deleting stable: /opt/cassandra/data/data/system_schema/column_masks-738c5cd01683268bd18534dbec278af/nb-1-big-  
Data db (1248) commitlog position CommitLogPosition(segmentId=1745135809599, position=200304)  
cassandra-server INFO [PerDiskMemtableLshWriter:0:] 2025-04-20 08:07:38.022 Flushing.java:153 - Writing Memtable-tables1932616997(868 serialized bytes, 1 ops, 2.878KIB (0%) on-heap, 0B (0%) off-heap).  
Flushed range = [min(-9223372036854775808), max(9223372036854775807)]  
cassandra-server INFO [PerDiskMemtableLshWriter:0:] 2025-04-20 08:07:38.024 Flushing.java:179 - Completed flushing /var/lib/cassandra/data/system_schema/tables-affdfbbdcbe1306805SeedGc302ba09/nb-7-big-D  
ata db (4248) commitlog position CommitLogPosition(segmentId=1745135809599, position=200304)  
cassandra-server INFO [CompactionExecutor:3] 2025-04-20 08:07:38.024 BiFormat.java:231 - Deleting stable: /opt/cassandra/data/data/system_schema/column_masks-738c5cd01683268bd18534dbec278af/nb-3-big-  
cassandra-server INFO [NonPeriodicTasks:1] 2025-04-20 08:07:38.029 BiFormat.java:231 - Deleting stable: /opt/cassandra/data/data/system_schema/column_masks-738c5cd01683268bd18534dbec278af/nb-2-big-  
Data db (1248) 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 (0%) on-heap, 0B (0%) off-heap  
cassandra-server INFO [PerDiskMemtableLshWriter:0:] 2025-04-20 08:07:38.208 Flushing.java:153 - Writing Memtable-keyspaces1815573366(1578 serialized bytes, 1 ops, 730B (0%) on-heap, 0B (0%) off-heap).  
Flushed range = [min(-9223372036854775808), max(9223372036854775807)]  
cassandra-server INFO [PerDiskMemtableLshWriter:0:] 2025-04-20 08:07:38.209 Flushing.java:179 - Completed flushing /var/lib/cassandra/data/system_schema/keyspaces-abac5682deaf31cb5353bdcff0df/bn-8-big-  
Data db (1248) commitlog position CommitLogPosition(segmentId=1745135809599, position=201293)  
cassandra-server INFO [CompactionExecutor:3] 2025-04-20 08:07:38.257 CompactionTask.java:167 - Compactng (86ab29f0-1dbe-11f0-9cf0-83427d4d7660) [/var/lib/cassandra/data/system_schema/keyspaces-abac5682deaf31cb5353bdcff0df/nb-8-big-  
Data db (1248) commitlog position CommitLogPosition(segmentId=1745135809599, position=201293)]  
cassandra-server INFO [CompactionExecutor:3] 2025-04-20 08:07:38.257 CompactionTask.java:167 - Compactng (86ab29f0-1dbe-11f0-9cf0-83427d4d7660) 4 stable to /opt/cassandra/data/data/system_schema/keyspaces-abac5682deaf31cb5353bdcff0df/nb-8-big-  
Data db (1248) commitlog position CommitLogPosition(segmentId=1745135809599, position=201293)  
cassandra-server INFO [Native-Transport-Requests-7] 2025-04-20 08:07:38.278 Keyspace.java:379 - Creating replication strategy IndexKeySpaceParams {keyspaceParam {durable_writes=true, replication={replicationFactor=1}}}  
params {class org.apache.cassandra.locator.SimpleStrategy, replication_factor=1}  
cassandra-server INFO [Native-Transport-Requests-7] 2025-04-20 08:07:38.286 ColumnFamilyStore.java:499 - Initializing index_keyspace.bm25_stats  
Cluster-master INFO Keyspaces and tables created successfully!  
Cluster-master INFO Cassandra setup completed!  
Cluster-master INFO [CompactionExecutor:3] 2025-04-20 08:07:38.471 CompactionTask.java:258 - Compacted (86ab29f0-1dbe-11f0-9cf0-83427d4d7660) 4 stable to /opt/cassandra/data/data/system_schema/keyspaces-abac5682deaf31cb5353bdcff0df/nb-9-big-Data db (1248) commitlog position CommitLogPosition(segmentId=1745135809599, position=201293). Read throughput = s.183KB/s, Write throughput = 1.346KB/s, Row Throughput = ~12/s. 9 total partitions merged to 6. Partition merge counts were [-15, 4, 1]. Time spent writing keyspace = 77ms  
Cluster-master INFO [NonPeriodicTasks:1] 2025-04-20 08:07:38.472 BiFormat.java:231 - Deleting stable: /opt/cassandra/data/data/system_schema/keyspaces-abac5682deaf31cb5353bdcff0df/nb-8-big-Data db (1248) commitlog position CommitLogPosition(segmentId=1745135809599, position=201293)  
Cluster-master INFO [NonPeriodicTasks:1] 2025-04-20 08:07:38.477 BiFormat.java:231 - Deleting stable: /opt/cassandra/data/data/system_schema/keyspaces-abac5682deaf31cb5353bdcff0df/nb-7-big-Data db (1248) commitlog position CommitLogPosition(segmentId=1745135809599, position=201293)  
Cluster-master INFO [NonPeriodicTasks:1] 2025-04-20 08:07:38.482 BiFormat.java:231 - Deleting stable: /opt/cassandra/data/data/system_schema/keyspaces-abac5682deaf31cb5353bdcff0df/nb-6-big-Data db (1248) commitlog position CommitLogPosition(segmentId=1745135809599, position=201293)  
Cluster-master INFO [NonPeriodicTasks:1] 2025-04-20 08:07:38.493 BiFormat.java:231 - Deleting stable: /opt/cassandra/data/data/system_schema/keyspaces-abac5682deaf31cb5353bdcff0df/nb-5-big-Data db (1248) commitlog position CommitLogPosition(segmentId=1745135809599, position=201293)  
Cluster-master INFO Preparing data...  
Cluster-master INFO Activating virtual environment...  
Cluster-master INFO Checking source parquet file...  
Cluster-master INFO Setting up HDFS directories...  
Cluster-master INFO Copying parquet file to HDFS...  
Cluster-master INFO Verifying HDFS file...  
Cluster-master INFO Starting Spark data preparation job...  
Cluster-master INFO Successfully created 1000 documents in /app/data/  
Cluster-master INFO Data preparation completed successfully!  
Cluster-master INFO Copying local input to HDFS...
```

B. Indexer tasks section

Running indexer:

```
INFO [NonPeriodicTasks-1] 2025-04-20 08:33:08,383 BigFormat.java:231 - Deleting stable:/opt/cassandra/data/data/system_schema/keyspaces-abac5682dea631c5b353b3dcffdf0b6/nb-7-big
INFO [NonPeriodicTasks-1] 2025-04-20 08:33:08,387 BigFormat.java:231 - Deleting stable:/opt/cassandra/data/data/system_schema/keyspaces-abac5682dea631c5b353b3dcffdf0b6/nb-6-big
INFO [NonPeriodicTasks-1] 2025-04-20 08:33:08,389 BigFormat.java:231 - Deleting stable:/opt/cassandra/data/data/system_schema/keyspaces-abac5682dea631c5b353b3dcffdf0b6/nb-5-big
Preparing data...
Activating virtual environment...
Checking source parquet file...
Error: Source parquet file not found at /data/a.parquet
Copying local input to HDFS...
Indexing files from: /index/tmp/input
Collecting packages...
Packing environment at '/app/.venv' to '/app/.venv.tar.gz'
##### | 100% completed | 2min 0.9s
##### | [mp/stream/job895161070147320.jar tmpdnull]
cluster-master [mp/stream/job895161070147320.jar tmpdnull]
2025-04-20 08:35:33,469 INFO client.DefaultNOHARMAIoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.3:8032
2025-04-20 08:35:33,744 INFO client.DefaultNOHARMAIoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.3:8032
2025-04-20 08:36:06,543 INFO mapreduce.JobResourceTracker: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1745137721999_0001
2025-04-20 08:36:06,543 INFO mapped.FileInputFormat: Total Input Files to process = 20
2025-04-20 08:36:06,609 INFO mapreduce.JobSubmitter: number of splits=20
2025-04-20 08:36:07,321 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1745137721999_0001
2025-04-20 08:36:07,321 INFO mapreduce.JobSubmitter: Executing with tokens=[]
2025-04-20 08:36:07,590 INFO conf.Configuration: resource-types.xml not found
2025-04-20 08:36:07,590 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-20 08:36:08,161 INFO ToolRunnerImpl: Submitted application application_1745137721999_0001
2025-04-20 08:36:08,214 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1745137721999_0001/
2025-04-20 08:36:08,216 INFO mapreduce.Job: Running job: job_1745137721999_0001
2025-04-20 08:36:30,415 INFO mapreduce.Job: Job: job_1745137721999_0001 running in uber mode : false
2025-04-20 08:36:30,418 INFO mapreduce.Job: map 0% reduce 0%
2025-04-20 08:36:42,760 INFO mapreduce.Job: map 15% reduce 0%
2025-04-20 08:36:52,004 INFO mapreduce.Job: map 25% reduce 0%
2025-04-20 08:36:53,011 INFO mapreduce.Job: map 30% reduce 0%
2025-04-20 08:37:03,232 INFO mapreduce.Job: map 45% reduce 0%
2025-04-20 08:37:13,514 INFO mapreduce.Job: map 60% reduce 0%
2025-04-20 08:37:23,811 INFO mapreduce.Job: map 70% reduce 0%
2025-04-20 08:37:29,846 INFO mapreduce.Job: map 80% reduce 0%
2025-04-20 08:37:34,932 INFO mapreduce.Job: map 80% reduce 7%
2025-04-20 08:37:35,930 INFO mapreduce.Job: map 80% reduce 7%
2025-04-20 08:37:40,009 INFO mapreduce.Job: map 90% reduce 8%
2025-04-20 08:37:41,016 INFO mapreduce.Job: map 100% reduce 8%
2025-04-20 08:37:43,100 INFO mapreduce.Job: map 100% reduce 25%
2025-04-20 08:37:53,234 INFO mapreduce.Job: map 100% reduce 100%
2025-04-20 08:37:54,250 INFO mapreduce.Job: Job: job_1745137721999_0001 completed successfully
2025-04-20 08:37:54,363 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=24
FILE: Number of bytes written=6676614
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=104743
HDFS: Number of bytes written=0
HDFS: Number of read operations=80
HDFS: Number of large read operations=0
HDFS: Number of write operations=8
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=20
Launched reduce tasks=4
Data-local map tasks=20
Total time spent by all maps in occupied slots (ms)=324396
Total time spent by all reduces in occupied slots (ms)=119332
Total time spent by all map tasks (ms)=162198
=====
MINGW64/C:/Users/user/Desktop/BD/S2S_BD_assign2_Shmakova
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=32181504
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=2523
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=82142433280
cluster-master Total committed heap usage (bytes)=308865384
cluster-master Peak Map Physical memory (bytes)=295190528
cluster-master Peak Map Virtual memory (bytes)=3423092736
cluster-master Peak Reduce Physical memory (bytes)=152126976
cluster-master Peak Reduce Virtual memory (bytes)=3423266016
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!
```