

## Big Data Assignment 2 report

The indexing process is split into:

### **Mapper1.py**

Parses each input document (doc\_id, title, text).

Combines title and text into a single string.

Tokenizes the content into lowercase words using regex.

Emits key-value pairs: word<TAB>doc\_id<TAB>1.

### **Reducer1.py**

Receives sorted mapper output as input.

Aggregates term frequencies for each (term, doc\_id) pair.

Calculates document lengths by summing frequencies.

Connects to Cassandra and stores: inverted\_index(term, doc\_id, freq), doc\_stats(doc\_id, length)

The reducer writes two types of information into Cassandra: inverted\_index stores the frequency of each term in each document. doc\_stats stores the length (total number of words) of each document.

### **Index.sh**

This bash script automates the indexing pipeline:

Accepts a file path to documents as input (default: /index/data).

Executes a Hadoop streaming job with mapper1.py.

Collects the output from HDFS and feeds it into reducer1.py via standard input.

reducer1.py writes the final index data into Cassandra.

## Prepare data

```
root@cluster-master: /app
25/04/15 19:23:47 INFO Executor: Finished task 8.0 in stage 3.0 (TID 14). 1564 bytes result sent to driver
25/04/15 19:23:47 INFO Executor: Finished task 6.0 in stage 3.0 (TID 12). 1564 bytes result sent to driver
25/04/15 19:23:47 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 0 records.
25/04/15 19:23:47 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 0 records.
25/04/15 19:23:47 INFO TaskSetManager: Finished task 6.0 in stage 3.0 (TID 12) in 85 ms on cluster-master (executor driver) (1/15)
25/04/15 19:23:47 INFO TaskSetManager: Finished task 11.0 in stage 3.0 (TID 17) in 82 ms on cluster-master (executor driver) (2/15)
25/04/15 19:23:47 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 0 records.
25/04/15 19:23:47 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 0 records.
25/04/15 19:23:47 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 0 records.
25/04/15 19:23:47 INFO Executor: Finished task 0.0 in stage 3.0 (TID 7). 1564 bytes result sent to driver
25/04/15 19:23:47 INFO Executor: Finished task 2.0 in stage 3.0 (TID 8). 1564 bytes result sent to driver
25/04/15 19:23:47 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 6) in 89 ms on cluster-master (executor driver) (3/15)
25/04/15 19:23:47 INFO TaskSetManager: Finished task 3.0 in stage 3.0 (TID 9) in 89 ms on cluster-master (executor driver) (4/15)
25/04/15 19:23:47 INFO TaskSetManager: Finished task 8.0 in stage 3.0 (TID 14) in 88 ms on cluster-master (executor driver) (5/15)
25/04/15 19:23:47 INFO TaskSetManager: Finished task 1.0 in stage 3.0 (TID 7) in 90 ms on cluster-master (executor driver) (6/15)
25/04/15 19:23:47 INFO TaskSetManager: Finished task 5.0 in stage 3.0 (TID 11) in 89 ms on cluster-master (executor driver) (7/15)
25/04/15 19:23:47 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 0 records.
25/04/15 19:23:47 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 0 records.
25/04/15 19:23:47 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 0 records.
25/04/15 19:23:47 INFO TaskSetManager: Finished task 2.0 in stage 3.0 (TID 8) in 90 ms on cluster-master (executor driver) (8/15)
25/04/15 19:23:47 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 0 records.
25/04/15 19:23:47 INFO Executor: Finished task 10.0 in stage 3.0 (TID 16). 1564 bytes result sent to driver
25/04/15 19:23:47 INFO Executor: Finished task 14.0 in stage 3.0 (TID 20). 1607 bytes result sent to driver
25/04/15 19:23:47 INFO Executor: Finished task 13.0 in stage 3.0 (TID 19). 1564 bytes result sent to driver
25/04/15 19:23:47 INFO Executor: Finished task 9.0 in stage 3.0 (TID 15). 1607 bytes result sent to driver
25/04/15 19:23:47 INFO Executor: Finished task 12.0 in stage 3.0 (TID 18). 1564 bytes result sent to driver
25/04/15 19:23:47 INFO TaskSetManager: Finished task 10.0 in stage 3.0 (TID 16) in 93 ms on cluster-master (executor driver) (9/15)
25/04/15 19:23:47 INFO Executor: Finished task 7.0 in stage 3.0 (TID 13). 1564 bytes result sent to driver
25/04/15 19:23:47 INFO TaskSetManager: Finished task 13.0 in stage 3.0 (TID 19) in 93 ms on cluster-master (executor driver) (10/15)
25/04/15 19:23:47 INFO TaskSetManager: Finished task 14.0 in stage 3.0 (TID 20) in 92 ms on cluster-master (executor driver) (11/15)
25/04/15 19:23:47 INFO TaskSetManager: Finished task 9.0 in stage 3.0 (TID 15) in 96 ms on cluster-master (executor driver) (12/15)
25/04/15 19:23:47 INFO TaskSetManager: Finished task 7.0 in stage 3.0 (TID 13) in 97 ms on cluster-master (executor driver) (13/15)
25/04/15 19:23:47 INFO TaskSetManager: Finished task 12.0 in stage 3.0 (TID 18) in 95 ms on cluster-master (executor driver) (14/15)
25/04/15 19:23:47 INFO ColumnIndexFilter: No offset index for column text is available; Unable to do filtering
25/04/15 19:23:47 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 442726 records.
25/04/15 19:23:47 INFO InternalParquetRecordReader: at row 0, reading next block
25/04/15 19:23:47 INFO BlockManagerInfo: Removed broadcast_2_piece0 on cluster-master:33517 in memory (size: 6.4 MiB, free: 912.3 MiB)
25/04/15 19:23:47 INFO BlockManagerInfo: Removed broadcast_2_piece0 on cluster-master:33517 in memory (size: 6.4 MiB, free: 912.3 MiB)
25/04/15 19:23:48 INFO CodecPool: Got brand-new decompressor [.snappy]
25/04/15 19:23:48 INFO InternalParquetRecordReader: block read in memory in 1324 ms. row count = 442726
25/04/15 19:23:48 INFO MemoryStore: Block taskresult_10 stored as bytes in memory (estimated size 2.3 MiB, free 909.6 MiB)
25/04/15 19:23:48 INFO BlockManagerInfo: Added taskresult_10 in memory on cluster-master:33517 (size: 2.3 MiB, free: 910.0 MiB)
25/04/15 19:23:48 INFO Executor: Finished task 4.0 in stage 3.0 (TID 10). 2372647 bytes result sent via BlockManager
25/04/15 19:23:48 INFO TransportClientFactory: Successfully created connection to cluster-master/172.18.0.4:33517 after 25 ms (0 ms spent in bootstraps)
25/04/15 19:23:48 INFO TaskSetManager: Finished task 4.0 in stage 3.0 (TID 10) in 1860 ms on cluster-master (executor driver) (15/15)
25/04/15 19:23:48 INFO DAGSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
25/04/15 19:23:48 INFO DAGScheduler: ResultStage 3 (collect at /app/prepare_data.py:132) finished in 1.871 s
25/04/15 19:23:48 INFO DAGScheduler: Job 3 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/15 19:23:48 INFO TaskSchedulerImpl: Killing all running tasks in stage 3: Stage finished
25/04/15 19:23:48 INFO DAGScheduler: Job 3 finished: collect at /app/prepare_data.py:132, took 1.875626 s
25/04/15 19:23:48 INFO BlockManagerInfo: Removed taskresult_10 on cluster-master:33517 in memory (size: 2.3 MiB, free: 912.3 MiB)
100%|#####| 1000/1000 [00:00<00:00, 10331.41it/s]
```

## Mapper+Reducer run

```
root@cluster-master: /app
2025-04-15 20:01:37,676 INFO impl.YarnClientImpl: Submitted application application_1744744490562_0019
2025-04-15 20:01:37,705 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744744490562_0019/
2025-04-15 20:01:37,706 INFO mapreduce.Job: Running job: job_1744744490562_0019
2025-04-15 20:01:39,170 INFO mapreduce.Job: Job job_1744744490562_0019 running in uber mode : false
2025-04-15 20:01:39,171 INFO mapreduce.Job: map 0% reduce 0%
2025-04-15 20:01:40,262 INFO mapreduce.Job: map 25% reduce 0%
2025-04-15 20:01:45,268 INFO mapreduce.Job: map 30% reduce 0%
2025-04-15 20:01:46,277 INFO mapreduce.Job: map 35% reduce 0%
2025-04-15 20:01:47,285 INFO mapreduce.Job: map 60% reduce 0%
2025-04-15 20:01:49,297 INFO mapreduce.Job: map 65% reduce 0%
2025-04-15 20:01:50,303 INFO mapreduce.Job: map 90% reduce 0%
2025-04-15 20:01:51,311 INFO mapreduce.Job: map 95% reduce 0%
2025-04-15 20:01:53,333 INFO mapreduce.Job: map 100% reduce 0%
2025-04-15 20:01:54,439 INFO mapreduce.Job: Job job_1744744490562_0019 completed successfully
2025-04-15 20:01:54,439 INFO mapreduce.Job: Counters: 33

File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=5516450
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=61267
  HDFS: Number of bytes written=0
  HDFS: Number of read operations=140
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=40
  HDFS: Number of bytes read erasure-coded=0

Job Counters
  Launched map tasks=20
  Data-local map tasks=20
  Total time spent by all maps in occupied slots (ms)=42776
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=42776
  Total vcore-seconds taken by all map tasks=42776
  Total megabyte-milliseconds taken by all map tasks=43882624

Map-Reduce Framework
  Map input records=20
  Map output records=0
  Input split bytes=2479
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=1500
  CPU time spent (ms)=7390
  Physical memory (bytes) snapshot=4655169536
  Virtual memory (bytes) snapshot=5199188864
  Total committed heap usage (bytes)=3672637440
  Peak Map Physical memory (bytes)=247832576
  Peak Map Virtual memory (bytes)=2603143168

File Input Format Counters
  Bytes Read=58788
  File Output Format Counters
  Bytes Written=0
2025-04-15 20:01:54,439 INFO streaming.StreamJob: Output directory: /tmp/index-output
Mapper end
Reducer
```