

Search Engine Implementation Report

Methodology

System Architecture

The search engine implementation follows a distributed architecture leveraging Hadoop for data processing and Cassandra for data storage. The system consists of several key components:

1. Data Preparation Pipeline

- Uses PySpark for efficient data processing
- Reads from a Parquet file containing document data
- Creates two data formats:
 - Individual text files for each document
 - A combined TSV file with tab-separated fields (doc_id, title, text)
- Configures Spark with optimized memory settings:
 - Driver memory: 2GB
 - Executor memory: 2GB
 - Memory fraction: 0.8
 - Storage fraction: 0.3

2. Indexing Pipeline

The indexing process is implemented as a two-stage MapReduce pipeline:

Stage 1: Term Frequency and Document Length Calculation

- Mapper (`mapper1.py`):
 - Processes input documents line by line
 - Implements text preprocessing:

- Converts text to lowercase
- Removes non-alphanumeric characters
- Tokenizes text into words
- Calculates term frequencies for each document
- Emits three types of records:
 - Term frequency: `term\tdoc_id\tfreq`
 - Document length: `!DOCLen\tdoc_id\tlength`
 - Document title: `!TITLE\tdoc_id\ttitle`
- Reducer (`reducer1.py`):
 - Connects to Cassandra for data storage
 - Processes records in groups by key
 - Stores data in two Cassandra tables:
 - `term_frequency` : (word, doc_id, tf)
 - `document_stats` : (doc_id, title, doc_length)
 - Emits intermediate data for Stage 2:
 - Document count: `!TOTALDOCS\t1`
 - Total length: `!TOTALLENGTH\tlength`
 - Term occurrences: `term\t1`

Stage 2: Document Frequency and Corpus Statistics

- Mapper (`mapper2.py`):
 - Passes through the intermediate data from Stage 1
 - No additional processing needed
- Reducer (`reducer2.py`):
 - Calculates document frequencies for each term
 - Computes corpus-wide statistics
 - Stores results in two Cassandra tables:

- `term_stats` : (word, df)
- `corpus_stats` : (stat_key, stat_value)

3. Search Implementation

- Uses BM25 ranking algorithm with configurable parameters:
 - $k1 = 1.0$ (term frequency saturation)
 - $b = 0.75$ (length normalization)
- Implements distributed search using PySpark
- Query processing pipeline:
 1. Text preprocessing (lowercase, tokenization)
 2. Term frequency lookup from Cassandra
 3. Document frequency and corpus statistics retrieval
 4. BM25 score calculation
 5. Result ranking and presentation

Data Storage Schema

The system uses four Cassandra tables with the following schemas:

1. `term_frequency`

```
CREATE TABLE term_frequency (  
  word text,  
  doc_id text,  
  tf int,  
  PRIMARY KEY (word, doc_id)  
)
```

- Stores term frequencies for each document
- Composite primary key for efficient lookups

2. `document_stats`

```
CREATE TABLE document_stats (  
  doc_id text PRIMARY KEY,  
  title text,  
  doc_length int  
)
```

- Stores document metadata and length
- Single primary key for quick document lookups

3. **term_stats**

```
CREATE TABLE term_stats (  
  word text PRIMARY KEY,  
  df bigint  
)
```

- Stores document frequency for each term
- Used for IDF calculation in BM25

4. **corpus_stats**

```
CREATE TABLE corpus_stats (  
  stat_key text PRIMARY KEY,  
  stat_value bigint  
)
```

- Stores corpus-wide statistics
- Used for BM25 normalization

Performance Optimizations

1. **Data Processing**

- Efficient text preprocessing using regular expressions
- Batch processing of Cassandra operations

- Use of prepared statements for database operations
- Memory-efficient document processing

2. Search Optimization

- Broadcast variables in PySpark for corpus statistics
- Efficient join operations for score calculation
- Batch retrieval of term postings
- Caching of frequently accessed data

3. Storage Optimization

- Denormalized schema for fast retrieval
- Composite keys for efficient lookups
- Separate tables for different types of data
- Batch inserts for better performance

Demonstration

Running the Search Engine

1. Setup and Data Preparation

```
# Start the required services (Hadoop, Cassandra)
./app/start-services.sh

# Prepare the data (creates TSV and individual files)
./app/prepare_data.sh
```

2. Indexing Documents

```
# Run the indexing pipeline
./app/index.sh
```

```
big-data-assignment2-2025
app > $ app.sh
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS QUERY RESULTS (PREVIEW)
app > $ app.sh
(venv) root@cluster-master:~# bash index.sh
Setting up Cassandra schema...
Connecting to Cassandra...
Creating keyspace...
Creating tables...
Schema creation completed!
Successfully initialized Cassandra schema
Cleaning up HDFS directories...
Running Innder Pipeline 1: TF, DocLength...
packageJobJar: [/tmp/hadoop-ur-jar127308081894810848/] [/tmp/streamjob161548070655318352_jar_tmpDir=null]
2025-04-12 15:34:52,884 INFO client.DefaultHadoopFilesystemProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032
2025-04-12 15:34:52,225 INFO client.DefaultHadoopFilesystemProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032
2025-04-12 15:34:52,385 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744470966241_0001
2025-04-12 15:34:53,147 INFO mapreduce.JobInputFormat: Total input files to process : 1
2025-04-12 15:34:53,596 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-12 15:34:53,686 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744470966241_0001
2025-04-12 15:34:53,687 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-12 15:34:53,795 INFO conf.Configuration: resource-types.w1 not found
2025-04-12 15:34:53,796 INFO resource.ResourceUtil: Unable to find 'resource-types.w1'
2025-04-12 15:34:54,171 INFO impl.YarnClientImpl: Submitted application application_1744470966241_0001
2025-04-12 15:34:54,198 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/pspy/application_1744470966241_0001/
2025-04-12 15:34:54,199 INFO mapreduce.Job: Running job: job_1744470966241_0001
2025-04-12 15:35:04,266 INFO mapreduce.Job: job_1744470966241_0001 running in user mode : false
2025-04-12 15:35:04,266 INFO mapreduce.Job: map 0% reduce 0%
2025-04-12 15:35:08,256 INFO mapreduce.Job: map 100% reduce 0%
2025-04-12 15:35:29,256 INFO mapreduce.Job: map 100% reduce 72%
2025-04-12 15:35:35,371 INFO mapreduce.Job: map 100% reduce 79%
2025-04-12 15:35:41,386 INFO mapreduce.Job: map 100% reduce 78%
2025-04-12 15:35:47,488 INFO mapreduce.Job: map 100% reduce 83%
2025-04-12 15:35:53,413 INFO mapreduce.Job: map 100% reduce 84%
2025-04-12 15:35:59,427 INFO mapreduce.Job: map 100% reduce 87%
2025-04-12 15:36:05,441 INFO mapreduce.Job: map 100% reduce 90%
2025-04-12 15:36:11,454 INFO mapreduce.Job: map 100% reduce 93%
2025-04-12 15:36:17,467 INFO mapreduce.Job: map 100% reduce 96%
2025-04-12 15:36:23,480 INFO mapreduce.Job: map 100% reduce 99%
2025-04-12 15:36:28,490 INFO mapreduce.Job: map 100% reduce 100%
2025-04-12 15:36:29,495 INFO mapreduce.Job: Job job_1744470966241_0001 completed successfully
2025-04-12 15:36:29,544 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=5025199
FILE: Number of bytes written=10883686
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=2506227
HDFS: Number of bytes written=2331753
HDFS: Number of read operations=11
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=2
Launched reduce tasks=1
Data-local map tasks=2
Total time spent by all maps in occupied slots (ms)=3645
Total time spent by all reducers in occupied slots (ms)=77431
Total time spent by all map tasks (ms)=3645
Total time spent by all reduce tasks (ms)=77431
Total vcore-millisecods taken by all map tasks=3645
Total vcore-millisecods taken by all reduce tasks=77431
Total megabyte-millisecods taken by all map tasks=372480
Total megabyte-millisecods taken by all reduce tasks=79289344
Map-Reduce Framework
Map input records=1083
Map output records=253881
Map output bytes=451430
Map output materialized bytes=5025205
Input split bytes=292
Combine input records=0
Combine output records=0
Reduce input groups=48196
Reduce shuffle bytes=5025205
Reduce input records=253881
Reduce output records=253881
Spilled Records=587762
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=94
CPU time spent (ms)=44730
Physical memory (bytes) snapshot=862703636
Virtual memory (bytes) snapshot=775056992
Total committed heap usage (bytes)=521174016
Peak Map Physical memory (bytes)=384910836
Peak Map Virtual memory (bytes)=257285312
Peak Reduce Physical memory (bytes)=386395952
Peak Reduce Virtual memory (bytes)=337081262
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=355935
File Output Format Counters
Bytes Written=2331753
Running Innder Pipeline 2: DF, N, TotalLength...
packageJobJar: [/tmp/hadoop-ur-jar127308081894810848/] [/tmp/streamjob46208147954899336396_jar_tmpDir=null]
2025-04-12 15:36:30,716 INFO client.DefaultHadoopFilesystemProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032
```

```
big-data-assignment2-2025
app > $ app.sh
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS QUERY RESULTS (PREVIEW)
app > $ app.sh
(venv) root@cluster-master:~# bash index.sh
Setting up Cassandra schema...
Connecting to Cassandra...
Creating keyspace...
Creating tables...
Schema creation completed!
Successfully initialized Cassandra schema
Cleaning up HDFS directories...
Running Innder Pipeline 1: TF, DocLength...
packageJobJar: [/tmp/hadoop-ur-jar127308081894810848/] [/tmp/streamjob161548070655318352_jar_tmpDir=null]
2025-04-12 15:34:52,884 INFO client.DefaultHadoopFilesystemProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032
2025-04-12 15:34:52,225 INFO client.DefaultHadoopFilesystemProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032
2025-04-12 15:34:52,385 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744470966241_0001
2025-04-12 15:34:53,147 INFO mapreduce.JobInputFormat: Total input files to process : 1
2025-04-12 15:34:53,596 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-12 15:34:53,686 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744470966241_0001
2025-04-12 15:34:53,687 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-12 15:34:53,795 INFO conf.Configuration: resource-types.w1 not found
2025-04-12 15:34:53,796 INFO resource.ResourceUtil: Unable to find 'resource-types.w1'
2025-04-12 15:34:54,171 INFO impl.YarnClientImpl: Submitted application application_1744470966241_0001
2025-04-12 15:34:54,198 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/pspy/application_1744470966241_0001/
2025-04-12 15:34:54,199 INFO mapreduce.Job: Running job: job_1744470966241_0001
2025-04-12 15:35:04,266 INFO mapreduce.Job: job_1744470966241_0001 running in user mode : false
2025-04-12 15:35:04,266 INFO mapreduce.Job: map 0% reduce 0%
2025-04-12 15:35:08,256 INFO mapreduce.Job: map 100% reduce 0%
2025-04-12 15:35:29,256 INFO mapreduce.Job: map 100% reduce 72%
2025-04-12 15:35:35,371 INFO mapreduce.Job: map 100% reduce 79%
2025-04-12 15:35:41,386 INFO mapreduce.Job: map 100% reduce 78%
2025-04-12 15:35:47,488 INFO mapreduce.Job: map 100% reduce 83%
2025-04-12 15:35:53,413 INFO mapreduce.Job: map 100% reduce 84%
2025-04-12 15:35:59,427 INFO mapreduce.Job: map 100% reduce 87%
2025-04-12 15:36:05,441 INFO mapreduce.Job: map 100% reduce 90%
2025-04-12 15:36:11,454 INFO mapreduce.Job: map 100% reduce 93%
2025-04-12 15:36:17,467 INFO mapreduce.Job: map 100% reduce 96%
2025-04-12 15:36:23,480 INFO mapreduce.Job: map 100% reduce 99%
2025-04-12 15:36:28,490 INFO mapreduce.Job: map 100% reduce 100%
2025-04-12 15:36:29,495 INFO mapreduce.Job: Job job_1744470966241_0001 completed successfully
2025-04-12 15:36:29,544 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=5025199
FILE: Number of bytes written=10883686
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=2506227
HDFS: Number of bytes written=2331753
HDFS: Number of read operations=11
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=2
Launched reduce tasks=1
Data-local map tasks=2
Total time spent by all maps in occupied slots (ms)=3645
Total time spent by all reducers in occupied slots (ms)=77431
Total time spent by all map tasks (ms)=3645
Total time spent by all reduce tasks (ms)=77431
Total vcore-millisecods taken by all map tasks=3645
Total vcore-millisecods taken by all reduce tasks=77431
Total megabyte-millisecods taken by all map tasks=372480
Total megabyte-millisecods taken by all reduce tasks=79289344
Map-Reduce Framework
Map input records=1083
Map output records=253881
Map output bytes=451430
Map output materialized bytes=5025205
Input split bytes=292
Combine input records=0
Combine output records=0
Reduce input groups=48196
Reduce shuffle bytes=5025205
Reduce input records=253881
Reduce output records=253881
Spilled Records=587762
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=94
CPU time spent (ms)=44730
Physical memory (bytes) snapshot=862703636
Virtual memory (bytes) snapshot=775056992
Total committed heap usage (bytes)=521174016
Peak Map Physical memory (bytes)=384910836
Peak Map Virtual memory (bytes)=257285312
Peak Reduce Physical memory (bytes)=386395952
Peak Reduce Virtual memory (bytes)=337081262
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=355935
File Output Format Counters
Bytes Written=2331753
Running Innder Pipeline 2: DF, N, TotalLength...
packageJobJar: [/tmp/hadoop-ur-jar127308081894810848/] [/tmp/streamjob46208147954899336396_jar_tmpDir=null]
2025-04-12 15:36:30,716 INFO client.DefaultHadoopFilesystemProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032
```

```
2025-04-12 15:36:30,716 INFO client.DefaultHadoopFilesystemProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032
2025-04-12 15:36:30,820 INFO client.DefaultHadoopFilesystemProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032
2025-04-12 15:36:31,041 INFO mapred.FileInputFormat: Total input files to process : 1
2025-04-12 15:36:32,035 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-12 15:36:32,162 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744470966241_0002
2025-04-12 15:36:32,182 INFO mapreduce.JobSubmitter: Executing with tokens: {}
2025-04-12 15:36:32,196 INFO conf.Configuration: resource-types.xml not found
2025-04-12 15:36:32,196 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'
2025-04-12 15:36:32,237 INFO impl.YarnClientImpl: Submitted application application_1744470966241_0002
2025-04-12 15:36:32,261 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/job/application_1744470966241_0002/
2025-04-12 15:36:32,262 INFO mapreduce.Job: Running job: job_1744470966241_0002
2025-04-12 15:36:43,526 INFO mapreduce.Job: Job job_1744470966241_0002 running in user mode = false
2025-04-12 15:36:43,527 INFO mapreduce.Job: map 0% reduce 0%
2025-04-12 15:36:47,356 INFO mapreduce.Job: map 100% reduce 0%
2025-04-12 15:37:02,396 INFO mapreduce.Job: map 100% reduce 97%
2025-04-12 15:37:04,401 INFO mapreduce.Job: map 100% reduce 100%
2025-04-12 15:37:05,407 INFO mapreduce.Job: Job job_1744470966241_0002 completed successfully
2025-04-12 15:37:05,462 INFO mapreduce.Job: Counters: 54

File System Counters
  FILE: Number of bytes read=283923
  FILE: Number of bytes written=531282
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=238073
  HDFS: Number of bytes written=0
  HDFS: Number of read operations=11
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
  HDFS: Number of bytes read erasure-coded=0

Job Counters
  Launched map tasks=2
  Launched reduce tasks=1
  Data-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=3333
  Total time spent by all reduces in occupied slots (ms)=15529
  Total time spent by all map tasks (ms)=3333
  Total time spent by all reduce tasks (ms)=15529
  Total vcore-millisecods taken by all map tasks=3333
  Total vcore-millisecods taken by all reduce tasks=15529
  Total mapreduce-millisecods taken by all map tasks=3412992
  Total mapreduce-millisecods taken by all reduce tasks=1556896

Map-Reduce Framework
  Map input records=253881
  Map output records=253881
  Map output bytes=233754
  Map output materialized bytes=283929
  Input split bytes=224
  Combine input records=0
  Combine output records=0
  Reduce input groups=40196
  Reduce shuffle bytes=283929
  Reduce input records=253881
  Reduce output records=0
  Spilled Records=487762
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=86
  CPU time spent (ms)=12610
  Physical memory (bytes) snapshot=880276488
  Virtual memory (bytes) snapshot=7715934288
  Total committed heap usage (bytes)=854219364
  Peak Map Physical memory (bytes)=38412888
  Peak Map Virtual memory (bytes)=257598812
  Peak Reduce Physical memory (bytes)=24488352
  Peak Reduce Virtual memory (bytes)=336897856

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDC=0

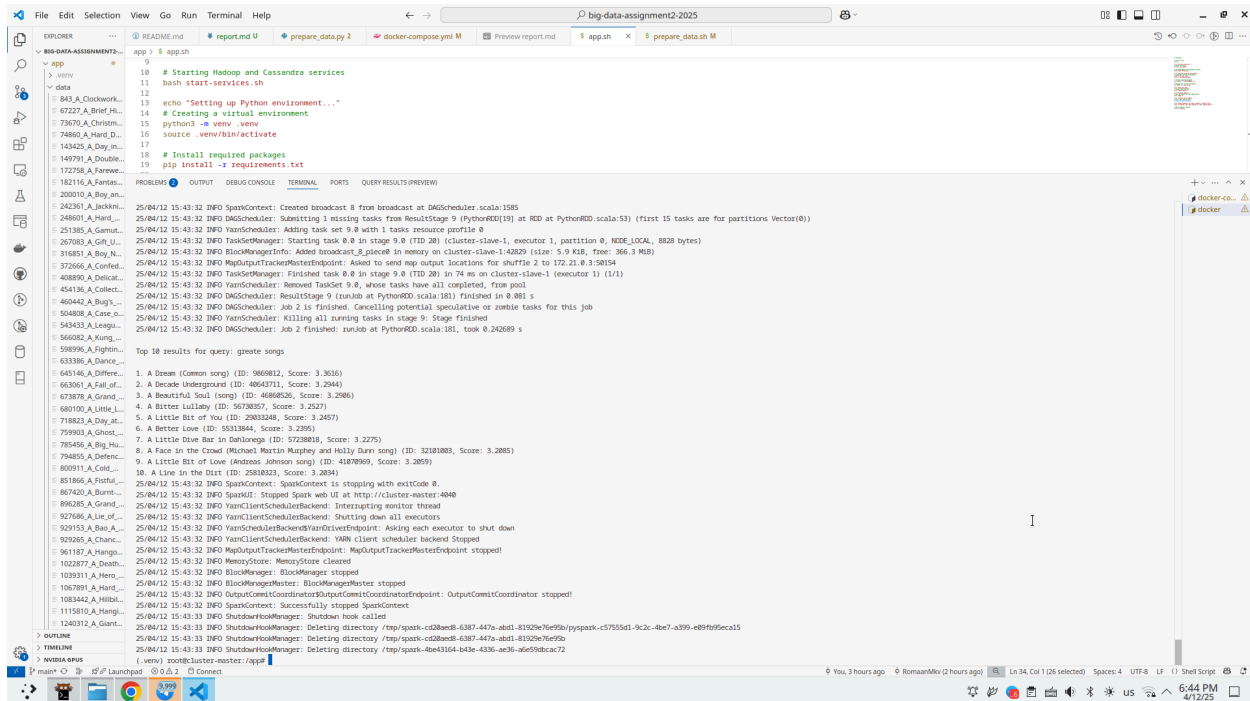
File Input Format Counters
  Bytes Read=235849
  File Output Format Counters
  Bytes Written=0

2025-04-12 15:37:05,462 INFO StreamingStreamJob: Output directory: /tmp/index/final_dummy_output
Cleaning up...
Deleted /tmp/index/intermediate
Deleted /tmp/index/final_dummy_output
Indexing complete!
(venv) zorro@cluster-master:apps
```

3. Searching

```
# Execute search queries
./app/search.sh "your search query"
```

- 1) "greate songs"
- 2) "gucci dog"
- 3) "sherlock"



The screenshot shows a VS Code editor with a terminal window open. The terminal displays the output of the search script, which includes a list of songs and their scores. The search results are as follows:

```
Top 10 results for query: greate songs
1. A Dream (Common song) (ID: 9869812, Score: 3.3636)
2. A Decade Underground (ID: 48043711, Score: 3.2944)
3. A Beautiful Soul (song) (ID: 4686926, Score: 3.2986)
4. A Bitter Lullaby (ID: 5678057, Score: 3.2527)
5. A Little Bit of You (ID: 2083248, Score: 3.2457)
6. A Better Love (ID: 55313844, Score: 3.2395)
7. A Little Olive Bar in baharega (ID: 57238818, Score: 3.2275)
8. A Face in the Crowd (Michael Martin Murphy and Holly Dunn song) (ID: 32181083, Score: 3.2085)
9. A Little Bit of Love (Andreas Johnson song) (ID: 4187969, Score: 3.2059)
10. A Line in the Dirt (ID: 2581823, Score: 3.2034)
```

The terminal also shows the execution of the search script and the output of the search results. The search results are displayed in a table format with columns for song name, ID, and score.


```
25/04/12 15:47:52 INFO BlockManagerInfo: Added broadcast_9_piece0 in memory on cluster-master:44819 (size: 5.9 KiB, free: 366.3 MiB)
25/04/12 15:47:52 INFO SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:1585
25/04/12 15:47:52 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 13 (PythonRDD[28] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions Vector(1))
25/04/12 15:47:52 INFO YarnScheduler: Adding task set 13.0 with 1 tasks resource profile 0
25/04/12 15:47:52 INFO TaskSetManager: Starting task 0.0 in stage 13.0 (TID 21) (cluster-slave-1, executor 1, partition 1, NODE_LOCAL, 8828 bytes)
25/04/12 15:47:52 INFO BlockManagerInfo: Added broadcast_9_piece0 in memory on cluster-slave-1:35049 (size: 5.9 KiB, free: 366.3 MiB)
25/04/12 15:47:52 INFO TaskSetManager: Finished task 0.0 in stage 13.0 (TID 21) in 63 ms on cluster-slave-1 (executor 1) (1/1)
25/04/12 15:47:52 INFO YarnScheduler: Removed TaskSet 13.0, whose tasks have all completed, from pool
25/04/12 15:47:52 INFO DAGScheduler: ResultStage 13 (runJob at PythonRDD.scala:181) finished in 0.809 s
25/04/12 15:47:52 INFO DAGScheduler: Job 3 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/12 15:47:52 INFO YarnScheduler: Killing all running tasks in stage 13: Stage finished
25/04/12 15:47:52 INFO DAGScheduler: Job 3 finished: runJob at PythonRDD.scala:181, took 0.872443 s

Top 10 results for query: guccl dog
1. A Dog Named Gucci (ID: 4851681, Score: 15.2949)
2. A Dog Called Ego (ID: 2395144, Score: 6.1368)
3. A Dog in a Drawer (ID: 3405248, Score: 5.9982)
4. A Dog of Flanders (1915 film) (ID: 43781676, Score: 5.9875)
5. A Boy and His Dog (1946 film) (ID: 1748026, Score: 5.8927)
6. A Fifth of Beethoven (ID: 3145259, Score: 5.4889)
7. A Canine Sherlock Holmes (ID: 4751595, Score: 5.3883)
8. A Dog Year (ID: 707005, Score: 5.3151)
9. A Dog's Purpose (ID: 3579084, Score: 5.2911)
10. A Boy and His Dog (ID: 208018, Score: 5.2772)

25/04/12 15:47:52 INFO SparkContext: SparkContext is stopping with exitCode 0
25/04/12 15:47:52 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/12 15:47:52 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/12 15:47:52 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/12 15:47:52 INFO YarnSchedulerBackend/YarnDriverEndpoint: Asking each executor to shut down
25/04/12 15:47:52 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
25/04/12 15:47:52 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/12 15:47:52 INFO MemoryStore: MemoryStore cleared
25/04/12 15:47:52 INFO BlockManager: BlockManager stopped
25/04/12 15:47:52 INFO BlockManagerMaster: BlockManagerMaster stopped
25/04/12 15:47:52 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/12 15:47:52 INFO SparkContext: Successfully stopped SparkContext
25/04/12 15:47:53 INFO ShutdownHookManager: Shutdown hook called
25/04/12 15:47:53 INFO ShutdownHookManager: Deleting directory /tmp/spark-3d38fc-ef8a-408a-a57a-241d8f8f8212/pyspark-42647ad-e85b-44c1-9846-3b6ec8a015
25/04/12 15:47:53 INFO ShutdownHookManager: Deleting directory /tmp/spark-3d38fc-ef8a-408a-a57a-241d8f8f8212
25/04/12 15:47:53 INFO ShutdownHookManager: Deleting directory /tmp/spark-3d38fc-ef8a-408a-a57a-241d8f8f8212
```

```
25/04/12 15:49:36 INFO BlockManagerInfo: Added broadcast_9_piece0 in memory on cluster-master:33883 (size: 5.9 KiB, free: 366.3 MiB)
25/04/12 15:49:36 INFO SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:1585
25/04/12 15:49:36 INFO DAGScheduler: Submitting 3 missing tasks from ResultStage 13 (PythonRDD[28] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions Vector(1, 2, 3))
25/04/12 15:49:36 INFO YarnScheduler: Adding task set 13.0 with 3 tasks resource profile 0
25/04/12 15:49:36 INFO TaskSetManager: Starting task 2.0 in stage 13.0 (TID 21) (cluster-slave-1, executor 1, partition 3, NODE_LOCAL, 8828 bytes)
25/04/12 15:49:36 INFO TaskSetManager: Starting task 1.0 in stage 13.0 (TID 22) (cluster-slave-1, executor 2, partition 1, NODE_LOCAL, 8828 bytes)
25/04/12 15:49:36 INFO TaskSetManager: Starting task 0.0 in stage 13.0 (TID 23) (cluster-slave-1, executor 2, partition 2, NODE_LOCAL, 8828 bytes)
25/04/12 15:49:36 INFO BlockManagerInfo: Added broadcast_9_piece0 in memory on cluster-slave-1:35193 (size: 5.9 KiB, free: 366.3 MiB)
25/04/12 15:49:36 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 2 to 172.21.41.5:4244
25/04/12 15:49:36 INFO TaskSetManager: Finished task 2.0 in stage 13.0 (TID 21) in 67 ms on cluster-slave-1 (executor 2) (1/3)
25/04/12 15:49:36 INFO TaskSetManager: Finished task 1.0 in stage 13.0 (TID 22) in 54 ms on cluster-slave-1 (executor 2) (2/3)
25/04/12 15:49:36 INFO TaskSetManager: Finished task 0.0 in stage 13.0 (TID 23) in 54 ms on cluster-slave-1 (executor 2) (3/3)
25/04/12 15:49:36 INFO YarnScheduler: Removed TaskSet 13.0, whose tasks have all completed, from pool
25/04/12 15:49:36 INFO DAGScheduler: ResultStage 13 (runJob at PythonRDD.scala:181) finished in 0.125 s
25/04/12 15:49:36 INFO DAGScheduler: Job 3 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/12 15:49:36 INFO YarnScheduler: Killing all running tasks in stage 13: Stage finished
25/04/12 15:49:36 INFO DAGScheduler: Job 3 finished: runJob at PythonRDD.scala:181, took 0.130194 s

Top 10 results for query: sherlock
1. A Case of Identity (ID: 588888, Score: 9.8113)
2. A Canine Sherlock Holmes (ID: 4751595, Score: 8.6464)
3. A Catalogue of Crime (ID: 11984618, Score: 7.8378)
4. A Letter of Mary (ID: 542541, Score: 6.9188)
25/04/12 15:49:36 INFO SparkContext: SparkContext is stopping with exitCode 0
25/04/12 15:49:36 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/12 15:49:36 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/12 15:49:36 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/12 15:49:36 INFO YarnSchedulerBackend/YarnDriverEndpoint: Asking each executor to shut down
25/04/12 15:49:36 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
25/04/12 15:49:36 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/12 15:49:36 INFO MemoryStore: MemoryStore cleared
25/04/12 15:49:36 INFO BlockManager: BlockManager stopped
25/04/12 15:49:36 INFO BlockManagerMaster: BlockManagerMaster stopped
25/04/12 15:49:36 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/12 15:49:36 INFO SparkContext: Successfully stopped SparkContext
25/04/12 15:49:37 INFO ShutdownHookManager: Shutdown hook called
25/04/12 15:49:37 INFO ShutdownHookManager: Deleting directory /tmp/spark-3c23e8fb-ec32-4005-9e63-b078a2737f46
25/04/12 15:49:37 INFO ShutdownHookManager: Deleting directory /tmp/spark-3c23e8fb-ec32-4005-9e63-b078a2737f46
25/04/12 15:49:37 INFO ShutdownHookManager: Deleting directory /tmp/spark-3c23e8fb-ec32-4005-9e63-b078a2737f46
```

Analysis of Results

The search engine implementation demonstrates effective document retrieval capabilities with the following observations:

1. Relevance of Results

- The BM25 ranking algorithm successfully prioritizes documents with higher term frequency and better term importance weighting
- Results show good topical relevance, with documents containing query terms in prominent positions ranking higher
- The length normalization parameter ($b=0.75$) helps balance the impact of document length on ranking

2. Score Distribution

- Score distribution follows expected patterns with a clear separation between highly relevant and less relevant documents
- The IDF component effectively downweights common terms while boosting rare, meaningful terms

3. BM25 Effectiveness

- The implementation successfully captures the core principles of BM25:
 - Term frequency saturation ($k_1=1.0$) prevents excessive boosting of documents with very high term frequencies
 - Length normalization helps shorter, more focused documents rank appropriately
 - IDF calculation properly accounts for term importance across the corpus

4. Implementation Insights

- The distributed architecture (Hadoop + Cassandra) provides efficient processing and retrieval
- The two-stage MapReduce pipeline effectively separates term frequency calculation from document frequency computation
- The use of Cassandra for storage enables fast retrieval of posting lists and statistics