

Big Data. Assignment 2

Vabnits Alexandra

a.vabnits@innopolis.university

DS-01

Repository with solution: <https://github.com/sashhhaka/BigData-Assignment-2>

1. Methodology

1.1 System Architecture

The search engine employs a distributed architecture leveraging several technologies:

- **Hadoop/YARN:** For distributed processing of document data
- **PySpark:** For efficient data processing and search operations
- **Cassandra:** For storing the inverted index and document statistics
- **BM25 Algorithm:** For relevance ranking of search results

The system is composed of three main components:

1. **Document Indexing Pipeline:** A MapReduce workflow that processes documents and builds the inverted index
2. **Data Storage Layer:** Cassandra tables that store document statistics and the inverted index
3. **Query Processing Engine:** A PySpark application that handles search queries and ranks documents

1.2 Application Workflow

The complete workflow includes:

1. **Environment Setup**

- Starting SSH and Hadoop services (`service ssh restart` , `start-services.sh`)
- Creating and configuring Python virtual environment (`python3 -m venv .venv`)
- Installing dependencies and packaging the environment (`venv-pack`)
- Initializing Cassandra tables

2. Data Preparation (`prepare_data.sh`)

- Reading from Parquet files using PySpark
- Cleaning and sampling data (limiting to 1000 or 100 documents)
- Writing document metadata to HDFS for indexing

3. Indexing Process (`index.sh`)

- Running the MapReduce pipeline to create inverted index
- Storing results in Cassandra

4. Search Operations (`search.sh`)

- Loading data from Cassandra
- Processing user queries with BM25 ranking

1.3 Indexing Methodology in Detail

The indexing process follows a two-stage MapReduce pipeline:

Pipeline 1: Document Processing

- `mapper1.py` : Reads document data from HDFS, tokenizes text, and emits:
 - Document lengths: `DOCLEN_{doc_id}\t{length}`
 - Term occurrences: `{term}:{doc_id}\t1`
- `reducer1.py` : Aggregates the outputs to produce:
 - Total document lengths: `DOCLEN_{doc_id}\t{total_length}`
 - Term frequencies per document: `{term}\t{doc_id}\t{tf}`

Pipeline 2: Vocabulary Building

- `mapper2.py` : For each term-document pair, emits `{term}\t1` to count documents containing each term

- `reducer2.py` : Aggregates to produce document frequencies: `{term}\t{df}`

1.4 Data Storage and Flow

Cassandra Schema Design:

The data is stored in three Cassandra tables:

1. inverted_index:

```
CREATE TABLE inverted_index (
    term text,
    doc_id text,
    tf int,
    PRIMARY KEY (term, doc_id)
)
```

Stores term frequency for each term-document pair

2. vocabulary:

```
CREATE TABLE vocabulary (
    term text PRIMARY KEY,
    doc_freq int
)
```

Contains document frequency for IDF calculations

3. doc_stats:

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

Stores document lengths for normalization in BM25

Data Flow:

1. MapReduce outputs downloaded from HDFS
2. Data is loaded into Cassandra tables
3. During search, PySpark loads data from Cassandra:

```
inv_index_df = spark.read.format("org.apache.spark.sql.cassandra")
    .options(table="inverted_index", keyspace="bigdata").load()
```

1.5 BM25 Ranking Implementation

BM25 algorithm implementation:

```
def bm25_score(tf, df, dl, avg_dl, N, k1=1.2, b=0.75):
    # IDF component
    idf = math.log((N - df + 0.5) / (df + 0.5) + 1.0)
    # TF saturation component
    numerator = tf * (k1 + 1)
    denominator = tf + k1 * ((1 - b) + b * (dl / avg_dl))
    return idf * (numerator / denominator)
```

This implements the Robertson-Sparck Jones formula with:

- Proper IDF calculation
- Term frequency saturation
- Document length normalization
- Standard tuning parameters ($k1=1.2$, $b=0.75$)

2. Demonstration

2.1 Running the Search Engine

To run this search engine:

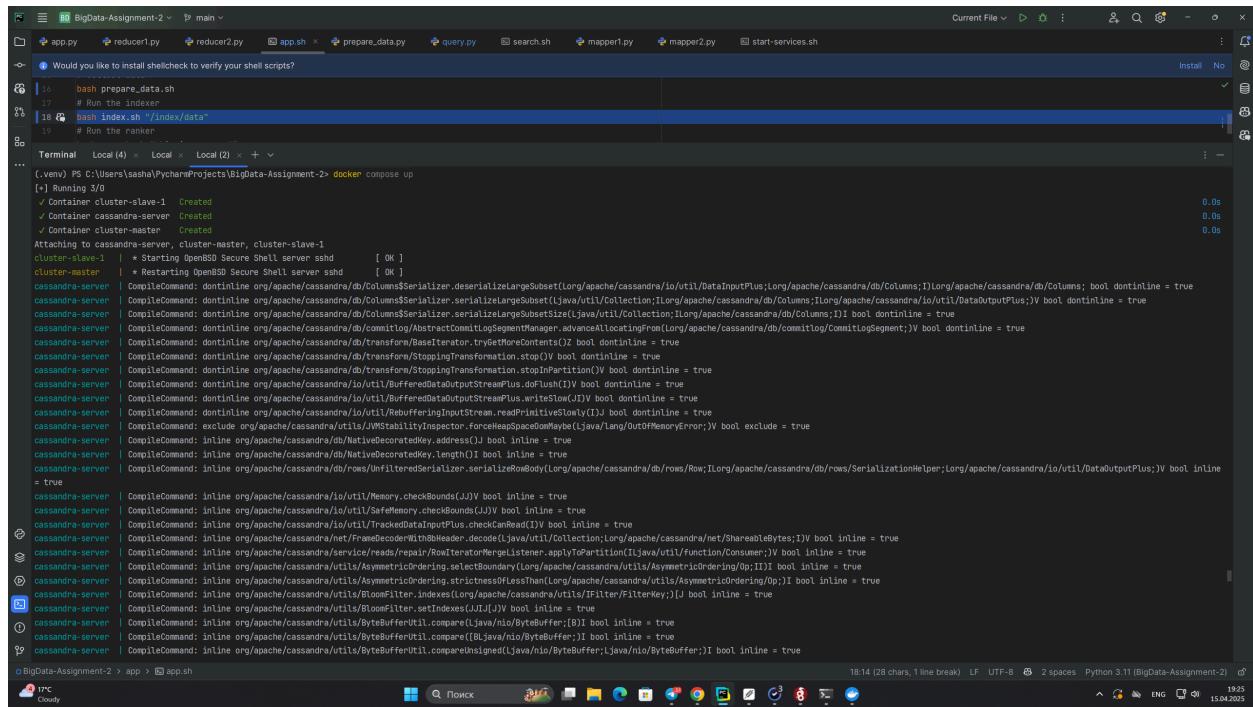
1. Clone the repository
2. Start the environment using Docker Compose:

```
docker compose up
```

This starts three containers:

- Hadoop master node
- Hadoop worker node
- Cassandra server

Starting the application:



```
PS C:\Users\saahal\PycharmProjects\BigData-Assignment-2> docker compose up
[*] Running 3/0
  ⬤ Container cluster-slave-1  Created
  ⬤ Container cassandra-server  Created
  ⬤ Container cluster-master  Created

Attaching to cassandra-server, cluster-master, cluster-slave-1
cluster-slave-1 |  * Starting OpenBSD Secure Shell server sshd      [ OK ]
cluster-master |  * Restarting OpenBSD Secure Shell server sshd      [ OK ]
cassandra-server |  CompileCommand: dontInline org/apache/cassandra/db/Column$Serializer.deserializeLargeSubset(lorg/apache/cassandra/io/util/DataInputPlus;lorg/apache/cassandra/db/Columns; boolean dontInline = true
cassandra-server |  CompileCommand: dontInline org/apache/cassandra/db/Column$Serializer.serializeLargeSubset(ljava/util/Collection;lorg/apache/cassandra/db/Columns;lorg/apache/cassandra/io/util/DataOutputPlus;)V boolean dontInline = true
cassandra-server |  CompileCommand: dontInline org/apache/cassandra/db/Column$Serializer.serializeLargeSubsetSize(ljava/util/Collection;lorg/apache/cassandra/db/columns;I)I boolean dontInline = true
cassandra-server |  CompileCommand: dontInline org/apache/cassandra/db/Commitlog/AbstractCommitlogSegmentManager.advanceAllocatingFrom(lorg/apache/cassandra/db/commitlog/CommitlogSegment;)V boolean dontInline = true
cassandra-server |  CompileCommand: dontInline org/apache/cassandra/db/Commitlog/AbstractCommitlogSegmentManager.advanceAllocatingFrom(lorg/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/StoppingTransformation.stop(O)V boolean dontInline = true
cassandra-server |  CompileCommand: dontInline org/apache/cassandra/db/transform/StoppingTransformation.stopInPartition(W)V boolean dontInline = true
cassandra-server |  CompileCommand: dontInline org/apache/cassandra/io/util/BufferedDataOutputStreamPlus.doFlush(J)V 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.readPrintableSlowly(I)V boolean dontInline = true
cassandra-server |  CompileCommand: exclude org/apache/cassandra/utils/JVMStabilityInspector.forceHeadSpaceOnMaybe(java/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/rows/UnfilteredSerializer.serializeRowBody(lorg/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(J)V boolean inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/io/util/SafeMemory.checkBounds(J)V boolean inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/io/util/TrackedDataInputPlus.checkCanRead(D)V boolean inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/net/FreadDecoderWithHeader.decode(Ljava/util/Collection;lorg/apache/cassandra/net/ShareableBytes;I)V boolean inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/service/ServiceListener.applyToPartition(Ljava/util/function/Consumer;)V boolean inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/AsymmetricOrdering.selectBoundary(lorg/apache/cassandra/utils/AsymmetricOrdering;I)I boolean inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/AsymmetricOrdering.strictnessOfLessThan(lorg/apache/cassandra/utils/AsymmetricOrdering;O)I boolean inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/RowFilter.indexes(lorg/apache/cassandra/utils/FilterKey;)J boolean inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/RowFilter.setIndexes(JJJ)J boolean inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/ByteBufferUtil.compare(Ljava/nio/ByteBuffer;[B)I boolean inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/ByteBufferUtil.compare([B;java/nio/ByteBuffer;)I boolean inline = true
cassandra-server |  CompileCommand: inline org/apache/cassandra/utils/ByteBufferUtil.compareUnsigned(Ljava/nio/ByteBuffer;Ljava/nio/ByteBuffer;)I boolean inline = true
18:14 (28 chars, 1 line break) LF UTF-8 2 spaces Python 3.11 (BigData-Assignment-2) 19:25
BigData-Assignment-2 > app > app.sh
```

Initializing Cassandra tables and starting data preparation:

```

BigData-Assignment-2 > main >
app.py reducer1.py reducer2.py app.sh prepare_data.py query.py search.sh mapper.py mapper2.py start-services.sh
Terminal Local(4) < Local > Local(2) + ~
cluster-master | Requirement already satisfied: pytz>=2020.1 in ./venv/lib/python3.8/site-packages (from pandas>1.5.0->fastparquet->r requirements.txt (line 4)) (2025.2)
cluster-master | Collecting packages...
cluster-master | VenvPackError: File 'venv.tar.gz' already exists
cluster-master | hello app
cluster-master | Initializing Cassandra schema...
cluster-master | Cassandra schema initialized successfully
cluster-master | Available keyspaces:
cluster-master | RowKeyspace_name='bigdata', type='keyspace', name='bigdata')
cluster-master | RowKeyspace_name='system', type='keyspace', name='system')
cluster-master | RowKeyspace_name='system_auth', type='keyspace', name='system_auth')
cluster-master | RowKeyspace_name='system_distributed', type='keyspace', name='system_distributed')
cluster-master | RowKeyspace_name='system_schema', type='keyspace', name='system_schema')
cluster-master | RowKeyspace_name='system_traces', type='keyspace', name='system_traces')
cluster-master | RowKeyspace_name='system_virtual_schema', type='keyspace', name='system_virtual_schema')
cluster-master | Checking Cassandra tables...
cluster-master | Existing tables in keyspace 'bigdata':
cluster-master | - doc_stats
cluster-master | - inverted_index
cluster-master | - vocabulary
cluster-master | All required tables successfully created!
cluster-master | 25/04/15 16:26:38 INFO SparkContext: Running Spark version 3.5.4
cluster-master | 25/04/15 16:26:38 INFO SparkContext: OS info Linux, 5.15.167.4-microsoft-standard-WSL2, amd64
cluster-master | 25/04/15 16:26:39 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
cluster-master | 25/04/15 16:26:39 INFO ResourceUtils: =====
cluster-master | 25/04/15 16:26:39 INFO ResourceUtils: No custom resources configured for spark-driver.
cluster-master | 25/04/15 16:26:39 INFO ResourceUtils: =====
cluster-master | 25/04/15 16:26:39 INFO ResourceUtils: Submitted application: data preparation
cluster-master | 25/04/15 16:26:39 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: -, vendor: -, memory -> name: memory, amount: 8192, script: -, vendor: -, offHeap -> name: offHeap, amount: 4096, script: -, vendor: -, task resources: Map(cpus -> name: cpus, amount: 1.0)
cluster-master | 25/04/15 16:26:39 INFO ResourceProfile: Limiting resource is cpu
cluster-master | 25/04/15 16:26:39 INFO ResourceProfileManager: Added ResourceProfile id: 0
cluster-master | 25/04/15 16:26:39 INFO SecurityManager: Changing view acls to: root
cluster-master | 25/04/15 16:26:39 INFO SecurityManager: Changing modify acls to: root
cluster-master | 25/04/15 16:26:39 INFO SecurityManager: Changing view acls groups to:
cluster-master | 25/04/15 16:26:39 INFO SecurityManager: Changing modify acls groups to:
cluster-master | 25/04/15 16:26:39 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: root; groups with view permissions: EMPTY; users with modify permissions: root; groups with modify permissions: EMPTY
Icons: EMPTY
cluster-master | 25/04/15 16:26:39 INFO Utils: Successfully started service 'sparkDriver' on port 34583.
cluster-master | 25/04/15 16:26:39 INFO SparkEnv: Registering MapOutputTracker
cluster-master | 25/04/15 16:26:39 INFO SparkEnv: Registering BlockManagerMaster
BigData-Assignment-2 > app > app.sh
18:14 (28 chars, 1 line break) LF UTF-8 ⚡ 2 spaces Python 3.11(BigData-Assignment-2) 19:26 15.04.2023
```

Putting data to hdfs:

```

BigData-Assignment-2 > main >
Project < Local(2) + ~
Terminal Local(4) < Local > Local(2) + ~
cluster-master | 25/04/15 16:26:57 INFO ShutdownHookManager: Shutdown hook called
cluster-master | 25/04/15 16:26:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-e8dc7b5-17d1-42bc-af92-248a5c2121bf
cluster-master | 25/04/15 16:26:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-9080fd7-58aa-4ab0-bff1-816fc4c7c3d/pyspark-15db2aab-82e7-49e6-ba3c-d398e2e1a62d
cluster-master | Putting data to hdfs...
cluster-master | put: /data/10931136.A.Decade_in_the_Grave.txt': File exists
cluster-master | put: /data/10907832.A_Case_for_the_Court.txt': File exists
cluster-master | put: /data/10999975.A_Different_Light_(album).txt': File exists
cluster-master | put: /data/10137549.A_Good_Thief_Tips_His_Way.txt': File exists
cluster-master | put: /data/10174652.A_History_of_Money_and_Banking_in_the_United_States.txt': File exists
cluster-master | put: /data/1023157_A_Balinese_Trance_Seance.txt': File exists
cluster-master | put: /data/1022871_A_Death_in_the_Family_(comics).txt': File exists
cluster-master | put: /data/10230485.A_Deut_Sinking_Stroy.txt': File exists
cluster-master | put: /data/10254092.A_Flat_Man.txt': File exists
cluster-master | put: /data/10381993.A_Hell's_House_(1973_Iomey_file).txt': File exists
cluster-master | put: /data/1039311.A_Hero_of_Our_Time.txt': File exists
cluster-master | put: /data/10399316.A_Flowering_Tree.txt': File exists
cluster-master | put: /data/10534798.A_Black_and_White_World.txt': File exists
cluster-master | put: /data/10570204.A_Ban_Called_Tension.txt': File exists
cluster-master | put: /data/1057891.A_Hard_Day's_Night_(song).txt': File exists
cluster-master | put: /data/1083462_A_Hillbilly_Tribute_to_ACDC.txt': File exists
cluster-master | put: /data/10849480.A_Day_in_the_Death_of_Danny_B.txt': File exists
cluster-master | put: /data/1085897.A_Dangerous_Path.txt': File exists
cluster-master | put: /data/1090070_A_Dictionary_of_Canadianism_on_Historical_Principles.txt': File exists
cluster-master | put: /data/1107793.A_Bad_Spell_in_Vurt.txt': File exists
cluster-master | put: /data/1107589.A_Doctor's_Report_on_Diabetics.txt': File exists
cluster-master | put: /data/11140461.A_Blueprint_of_the_World.txt': File exists
cluster-master | put: /data/1115810_A_Hanging.txt': File exists
cluster-master | put: /data/1121270_A_Lesson_in_Romantics.txt': File exists
cluster-master | put: /data/1131587_A_Go_Go_Potshot_album.txt': File exists
cluster-master | put: /data/11490217_A_Guitar_to_Groovy_lovin'.txt': File exists
cluster-master | put: /data/11528779.A_brewer's_Tales.txt': File exists
cluster-master | put: /data/11631735.A_Ballad_of_the_West.txt': File exists
cluster-master | put: /data/11753653.A_Journal_of_the_Plague_Year_(album).txt': File exists
cluster-master | put: /data/11971020_A_Lifetime_on_Mars.txt': File exists
cluster-master | put: /data/1199274_A_Cold_Night's_Death.txt': File exists
cluster-master | put: /data/1199321.A_Frogule_Hope.txt': File exists
cluster-master | put: /data/11984610_A_Catalogue_of_Crime.txt': File exists
cluster-master | put: /data/12000597.A_King_and_No_King.txt': File exists
cluster-master | put: /data/12132506.A_Crystal_Christmas.txt': File exists
cluster-master | put: /data/12212352.A_Flintstones_Christmas_Carol.txt': File exists
cluster-master | put: /data/12403122.A_Giant_Alien_Force_More_Violent_&_Sick_Than_Anything_You_Can_Imagine.txt': File exists
BigData-Assignment-2 > app > prepare_data.sh
4:1 LF UTF-8 ⚡ 2 spaces Python 3.11(BigData-Assignment-2) 19:29 15.04.2023
```

2.2 Indexing Results

Running MapReduce pipeline 1:

```

Project ▾ BigData-Assignment-2 ▾ main ▾
Terminal Local (4) × Local × Local (2) × + ×
cluster-master | Running Pipeline 1: Aggregating term frequencies and document statistics...
cluster-master | 2025-04-15 16:27:22,645 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
cluster-master | package为之器: [mapreduce|appender.py, mapreduce|reducer1.py, /tmp/hadoop-unjar397005515207676397/] []
cluster-master | 2025-04-15 16:27:23,372 INFO client.DefaultHttpMMapFileoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-15 16:27:23,350 INFO client.DefaultHttpMMapFileoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-15 16:27:24,148 INFO mapred.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/job_1744734319027_0001
cluster-master | 2025-04-15 16:27:25,301 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master | 2025-04-15 16:27:25,749 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master | 2025-04-15 16:27:25,880 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744734319027_0001
cluster-master | 2025-04-15 16:27:25,880 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master | 2025-04-15 16:27:26,029 INFO conf.Configuration: resource-types.xml not found
cluster-master | 2025-04-15 16:27:26,029 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master | 2025-04-15 16:27:26,957 INFO mapreduce.YarnClientImpl: Submitted application application_1744734319027_0001
cluster-master | 2025-04-15 16:27:27,015 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744734319027_0001/
cluster-master | 2025-04-15 16:27:27,017 INFO mapreduce.Job: Running job: job_1744734319027_0001
cluster-master | 2025-04-15 16:27:34,119 INFO mapreduce.Job: job_1744734319027_0001 running in uber mode : false
cluster-master | 2025-04-15 16:27:34,120 INFO mapreduce.Job: map ON reduce 0X
cluster-master | 2025-04-15 16:27:38,158 INFO mapreduce.Job: map 100% reduce 0N
cluster-master | 2025-04-15 16:27:43,180 INFO mapreduce.Job: map 100% reduce 100%
cluster-master | 2025-04-15 16:27:44,190 INFO mapreduce.Job: Job job_1744734319027_0001 completed successfully
cluster-master | 2025-04-15 16:27:44,248 INFO mapreduce.Job: Counters: 54
cluster-master |     File System Counters
cluster-master |         FILE: Number of bytes read:11187916
cluster-master |         FILE: Number of bytes written:23205349
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:350221
cluster-master |         HDFS: Number of bytes written:4379500
cluster-master |         HDFS: Number of read operations=11
cluster-master |         HDFS: Number of large read operations=0
cluster-master |         HDFS: Number of write operations=2
cluster-master |         HDFS: Number of bytes read erasure-coded=0
cluster-master |     Job Counters
cluster-master |         Launched map tasks=2
cluster-master |         Launched reduce tasks=1
cluster-master |         Data-local map tasks=2
cluster-master |         Total time spent by all maps in occupied slots (ms)=4586
cluster-master |         Total time spent by all reduces in occupied slots (ms)=2261
cluster-master |         Total time spent by all map tasks (ms)=4586
cluster-master |         Total time spent by all reduce tasks (ms)=2261
cluster-master |         Total vcore-milliseconds taken by all map tasks=4586
Project ▾ BigData-Assignment-2 ▾ app ▾ prepare_data.sh

```

Rinning MapReduce pipeline 2:

```

Project ▾ BigData-Assignment-2 ▾ main ▾
Terminal Local (4) × Local × Local (2) × + ×
cluster-master | Bytes Read:355929
cluster-master | File Output Format Counters
cluster-master | Bytes Written:379500
cluster-master | 2025-04-15 16:27:44,248 INFO streaming.StreamJob: Output directory: /tmp/mapreduce_pipeline1_output
cluster-master | Pipeline 1 completed. Output is stored in HDFS directory: /tmp/mapreduce_pipeline1_output.
cluster-master | Running Pipeline 2: Calculating vocabulary document frequencies...
cluster-master | 2025-04-15 16:27:45,069 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
cluster-master | package为之器: [mapreduce|appender2.py, mapreduce|reducer2.py, /tmp/hadoop-unjar4951299031051839454/]
cluster-master | 2025-04-15 16:27:45,835 INFO client.DefaultHttpMMapFileoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-15 16:27:45,835 INFO client.DefaultHttpMMapFileoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master | 2025-04-15 16:27:46,957 INFO mapreduce.YarnClientImpl: Submitted application application_1744734319027_0002
cluster-master | 2025-04-15 16:27:46,957 INFO mapreduce.JobSubmitter: number of splits: 1
cluster-master | 2025-04-15 16:27:47,507 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744734319027_0002
cluster-master | 2025-04-15 16:27:48,002 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master | 2025-04-15 16:27:48,125 INFO conf.Configuration: resource-types.xml not found
cluster-master | 2025-04-15 16:27:48,125 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master | 2025-04-15 16:27:48,171 INFO mapreduce.YarnClientImpl: Submitted application application_1744734319027_0002
cluster-master | 2025-04-15 16:27:48,201 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744734319027_0002/
cluster-master | 2025-04-15 16:27:48,203 INFO mapreduce.Job: Running job: job_1744734319027_0002
cluster-master | 2025-04-15 16:27:48,236 INFO mapreduce.Job: Job job_1744734319027_0002 running in uber mode : false
cluster-master | 2025-04-15 16:27:48,287 INFO mapreduce.Job: map ON reduce 0X
cluster-master | 2025-04-15 16:27:54,324 INFO mapreduce.Job: map 100% reduce 0N
cluster-master | 2025-04-15 16:28:03,346 INFO mapreduce.Job: map 100% reduce 100%
cluster-master | 2025-04-15 16:28:04,356 INFO mapreduce.Job: Job job_1744734319027_0002 completed successfully
cluster-master | 2025-04-15 16:28:04,412 INFO mapreduce.Job: Counters: 54
cluster-master |     File System Counters
cluster-master |         FILE: Number of bytes read:2711877
cluster-master |         FILE: Number of bytes written:6253325
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:4383836
cluster-master |         HDFS: Number of bytes written:417954
cluster-master |         HDFS: Number of read operations=11
cluster-master |         HDFS: Number of large read operations=0
cluster-master |         HDFS: Number of write operations=2
cluster-master |         HDFS: Number of bytes read erasure-coded=0
cluster-master |     Job Counters
cluster-master |         Launched map tasks=2
cluster-master |         Launched reduce tasks=1
cluster-master |         Data-local map tasks=2
cluster-master |         Total time spent by all maps in occupied slots (ms)=4586
cluster-master |         Total time spent by all reduces in occupied slots (ms)=2261
cluster-master |         Total time spent by all map tasks (ms)=4586
cluster-master |         Total vcore-milliseconds taken by all map tasks=4586
Project ▾ BigData-Assignment-2 ▾ app ▾ prepare_data.sh

```

After finishing both pipelines, statistics about indexed documents:

```

Project ▾ BigData-Assignment-2 ▾ main ▾
Terminal Local (4) × Local ▾ Local (2) × +
cluster-master | Shuffle Errors
cluster-master | BAD_ID=0
cluster-master | CONNECTION=0
cluster-master | ID_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=438598
cluster-master | File Output Format Counters
cluster-master | Bytes Written=47954
cluster-master | 2025-04-15 16:28:04,412 INFO streaming.StreamJob: Output directory: /tmp/mapreduce_pipeline2_output
cluster-master | Pipeline 2 completed. Vocabulary output is stored in HDFS directory: /tmp/mapreduce_pipeline2_output.
cluster-master | Loading MapReduce output into Cassandra...
cluster-master | hello app
cluster-master | Initializing Cassandra schema...
cluster-master | Cassandra schema initialized successfully
cluster-master | Available keyspaces:
cluster-master | RowKeyspace_name='bigdata', type='keyspace', name='bigdata'
cluster-master | RowKeyspace_name='system', type='keyspace', name='system'
cluster-master | RowKeyspace_name='system_auth', type='keyspace', name='system_auth'
cluster-master | RowKeyspace_name='system_distributed', type='keyspace', name='system_distributed'
cluster-master | RowKeyspace_name='system_schema', type='keyspace', name='system_schema'
cluster-master | RowKeyspace_name='system_traces', type='keyspace', name='system_traces'
cluster-master | RowKeyspace_name='system_views', type='keyspace', name='system_views'
cluster-master | RowKeyspace_name='system_virtual_schema', type='keyspace', name='system_virtual_schema'
cluster-master | Checking Cassandra tables...
cluster-master | Existing tables in keyspace 'bigdata':
cluster-master | - doc_stats
cluster-master | - inverted_index
cluster-master | - vocabulary
cluster-master | All required tables successfully created!
cluster-master | Loading data from MapReduce outputs into Cassandra...
cluster-master | Data successfully loaded into Cassandra:
cluster-master | - 1000 document statistics entries
cluster-master | - 25952 inverted index entries
cluster-master | - 4084 vocabulary entries
cluster-master | Indexing tasks complete. Data has been loaded into Cassandra.
cluster-master | Search for query: this is a query!
cluster-master | :: Loading settings :: url = jar:file:/usr/local/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
o BigData-Assignment-2 > app > prepare_data.sh
PC Cloudy 4:1 LF UTF-8 2 spaces Python 3.11 (BigData-Assignment-2) 19:34 15.04.2025

```

2.3 Search Results

When executing search queries:

. /search.sh "this is a query!"

Output:

```

Project ▾ BigData-Assignment-2 ▾ main ▾
Terminal Local (4) × Local × Local (2) + ×
cluster-master | 25/04/15 16:37:39 INFO TaskSetManager: Starting task 13.0 in stage 7.0 (TID 64) (cluster-slave-1, executor 2, partition 13, NODE_LOCAL, 8828 bytes)
cluster-master | 25/04/15 16:37:39 INFO TaskSetManager: Finished task 11.0 in stage 7.0 (TID 62) in 42 ms on cluster-slave-1 (executor 2) (12/14)
cluster-master | 25/04/15 16:37:39 INFO TaskSetManager: Finished task 12.0 in stage 7.0 (TID 63) in 48 ms on cluster-slave-1 (executor 1) (13/14)
cluster-master | 25/04/15 16:37:39 INFO TaskSetManager: Finished task 13.0 in stage 7.0 (TID 64) in 59 ms on cluster-slave-1 (executor 2) (14/14)
cluster-master | 25/04/15 16:37:39 INFO DAGScheduler: Removed TaskSet 7.0, whose tasks have all completed, from pool
cluster-master | 25/04/15 16:37:39 INFO DAGScheduler: ResultStage 7 (takeOrdered at /app/query.py:98) finished in 0.446 s
cluster-master | 25/04/15 16:37:39 INFO DAGScheduler: Job 5 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master | 25/04/15 16:37:39 INFO YarnScheduler: Killing all running tasks in stage 7: Stage finished
cluster-master | 25/04/15 16:37:39 INFO DAGScheduler: Job 5 finished: takeOrdered at /app/query.py:98, took 4.206215 s
cluster-master |
cluster-master | Top 10 relevant documents:
cluster-master | ID: 47515959, Title: A Canine Sherlock Holmes, BM25: 8.6579
cluster-master | ID: 30828220, Title: A Human Right, BM25: 1.7356
cluster-master | ID: 18171042, Title: A Chrestomathy, BM25: 1.7137
cluster-master | ID: 7068624, Title: A Black Mass, BM25: 1.7039
cluster-master | ID: S7279816, Title: A Book of American Martyrs, BM25: 1.6964
cluster-master | ID: 41801556, Title: A (The Walking Dead), BM25: 1.6794
cluster-master | ID: 45303168, Title: A Bearded Man, BM25: 1.6719
cluster-master | ID: 16379814, Title: A Breathhtaking Guy, BM25: 1.6470
cluster-master | ID: 6248566, Title: A Calf for Christmas, BM25: 1.6414
cluster-master | ID: 2761148, Title: A History of Philosophy (Coppleston), BM25: 1.6406
cluster-master | 25/04/15 16:37:39 INFO SparkContext: SparkContext is stopping with exitcode 0.
cluster-master | 25/04/15 16:37:39 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master | 25/04/15 16:37:39 INFO YarnClientSchedulerBackend: Interrupting monitor thread
cluster-master | 25/04/15 16:37:39 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster-master | 25/04/15 16:37:39 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
cluster-master | 25/04/15 16:37:39 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
cluster-master | 25/04/15 16:37:39 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master | 25/04/15 16:37:39 INFO MemoryStore: MemoryStore cleared
cluster-master | 25/04/15 16:37:39 INFO BlockManager: BlockManager stopped
cluster-master | 25/04/15 16:37:39 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master | 25/04/15 16:37:39 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master | 25/04/15 16:37:40 INFO SparkContext: Successfully stopped SparkContext
cluster-master | 25/04/15 16:37:40 INFO ShutdownHookManager: Shutdown hook called
cluster-master | 25/04/15 16:37:40 INFO ShutdownHookManager: Deleting directory /tmp/spark-065418c8-bbb8-440f-a479-86f2018a94/pyspark-1d8a87b3-8bb4-4328-a85a-51820cbba624
cluster-master | 25/04/15 16:37:40 INFO ShutdownHookManager: Deleting directory /tmp/spark-16fe0d6e-864a-4c9c-8873-643e10400957
cluster-master | 25/04/15 16:37:40 INFO ShutdownHookManager: Deleting directory /tmp/spark-065418c8-bbb8-440f-a479-86f2018a94
cluster-master | 25/04/15 16:37:40 INFO CassandraConnector: Disconnected from Cassandra cluster.
cluster-master | 25/04/15 16:37:40 INFO SerialIShutdownHooks: Successfully executed shutdown hook: Clearing session cache for C* connector
cluster-master exited with code 0

View in Docker Desktop View Config Enable Watch
BigData-Assignment-2 > app > app.sh
PC Cloud 14:15 LF UTF-8 2 spaces Python 3.11(BigData-Assignment-2) 19:38 15.04.2025

```

`./search.sh "big data"`

Output:

```

Project ▾ BigData-Assignment-2 ▾ main ▾
Terminal Local (4) × Local × Local (2) + ×
cluster-master | 25/04/15 16:41:48 INFO TaskSetManager: Finished task 11.0 in stage 7.0 (TID 59) in 59 ms on cluster-slave-1 (executor 1) (12/12)
cluster-master | 25/04/15 16:41:48 INFO YarnScheduler: Removed TaskSet 7.0, whose tasks have all completed, from pool
cluster-master | 25/04/15 16:41:48 INFO DAGScheduler: ResultStage 7 (takeOrdered at /app/query.py:98) finished in 0.404 s
cluster-master | 25/04/15 16:41:48 INFO YarnScheduler: Job 5 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master | 25/04/15 16:41:48 INFO DAGScheduler: Job 5 finished: takeOrdered at /app/query.py:98, took 4.162782 s
cluster-master |
cluster-master | Top 10 relevant documents:
cluster-master | ID: 3178415, Title: A Fistful of Datas, BM25: 8.4109
cluster-master | ID: 10909703, Title: A Dictionary of Canadianism on Historical Principles, BM25: 6.5209
cluster-master | ID: 2323594, Title: A History of the Arab Peoples, BM25: 6.0997
cluster-master | ID: 880911, Title: A Cold Wind Blows (game), BM25: 5.1832
cluster-master | ID: 65926201, Title: A Franklin kész leírása, BM25: 5.1694
cluster-master | ID: 9146522, Title: A Field Guide to the Birds of Hawaii and the Tropical Pacific, BM25: 5.1128
cluster-master | ID: 67425426, Title: A Grandchild's Guide to Using Grandpa's Computer, BM25: 4.9294
cluster-master | ID: 7854052, Title: A Big Heart of Love, BM25: 4.8951
cluster-master | ID: 67573012, Title: A Huge Eye Gazing Pulseating Brain That Rules from the Centre of the Ultraworld, BM25: 4.7726
cluster-master | ID: 65774444, Title: A Flower Above the Clouds, BM25: 4.7399
cluster-master | 25/04/15 16:41:48 INFO SparkContext: SparkContext is stopping with exitcode 0.
cluster-master | 25/04/15 16:41:48 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master | 25/04/15 16:41:48 INFO YarnClientSchedulerBackend: Interrupting monitor thread
cluster-master | 25/04/15 16:41:48 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster-master | 25/04/15 16:41:48 INFO YarnClientSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
cluster-master | 25/04/15 16:41:48 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
cluster-master | 25/04/15 16:41:48 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master | 25/04/15 16:41:48 INFO MemoryStore: MemoryStore cleared
cluster-master | 25/04/15 16:41:48 INFO BlockManager: BlockManager stopped
cluster-master | 25/04/15 16:41:48 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master | 25/04/15 16:41:48 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master | 25/04/15 16:41:48 INFO SparkContext: Successfully stopped SparkContext
cluster-master | 25/04/15 16:41:49 INFO ShutdownHookManager: Deleting directory /tmp/spark-54df0ed7-2h99-4e55-9fab-3325bf1cf151
cluster-master | 25/04/15 16:41:49 INFO ShutdownHookManager: Deleting directory /tmp/spark-12cf89e7-83b0-43b5-98a7-15c59996ba4/pyspark-565%e0d550-4a55-b0c5-04336af5dd11
cluster-master | 25/04/15 16:41:49 INFO ShutdownHookManager: Deleting directory /tmp/spark-12cf89e7-83b0-43b5-98a7-15c59996ba4
cluster-master | 25/04/15 16:41:49 INFO CassandraConnector: Disconnected from Cassandra cluster.
cluster-master | 25/04/15 16:41:49 INFO SerialIShutdownHooks: Successfully executed shutdown hook: Clearing session cache for C* connector
cluster-master exited with code 0

View in Docker Desktop View Config Enable Watch
BigData-Assignment-2 > app > app.sh
PC Cloud 20:25 LF UTF-8 2 spaces Python 3.11(BigData-Assignment-2) 19:42 15.04.2025

```

```
./search.sh "frog princess"
```

Output:

```
BigData-Assignment-2 > ./main.py
Project
Terminal Local (4) x Local (2) x + ~
cluster-master | 25/04/15 16:45:58 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID 65) in 69 ms on cluster-slave-1 (executor 1) (13/14)
cluster-master | 25/04/15 16:45:58 INFO TaskSetManager: Finished task 2.0 in stage 7.0 (TID 64) in 59 ms on cluster-slave-1 (executor 2) (14/14)
cluster-master | 25/04/15 16:45:58 INFO YarnScheduler: Removed TaskSet 7.0, whose tasks have all completed, from pool
cluster-master | 25/04/15 16:45:58 INFO DAGScheduler: ResultStage 7 (takeOrdered at /app/query.py:98) finished in 0.491 s
cluster-master | 25/04/15 16:45:58 INFO YarnScheduler: Cancelling potential speculative or zombie tasks for this job
cluster-master | 25/04/15 16:45:58 INFO DAGScheduler: Killing all running tasks in stage 7: Stage finished
cluster-master | 25/04/15 16:45:58 INFO DAGScheduler: Job 5 finished: takeOrdered at /app/query.py:98, took 4.036349 s
cluster-master |
cluster-master | Top 10 relevant documents:
cluster-master | ID: 6877746, Title: A Land Remembered, BM25: 8.6825
cluster-master | ID: 45631030, Title: A Fallen Idol, BM25: 7.5696
cluster-master | ID: 21789335, Title: A Kind of Magic (TV series), BM25: 5.9785
cluster-master | ID: 37742844, Title: A Dark Lantern, BM25: 5.6982
cluster-master | ID: 5811942, Title: A Kiss of Shadows, BM25: 5.3806
cluster-master | ID: 1039311, Title: A Hero of Our Time, BM25: 5.0404
cluster-master | ID: 7013864, Title: A Coach for Cinderella, BM25: 4.8369
cluster-master | ID: 52359819, Title: A Life Time Love, BM25: 4.6780
cluster-master | ID: 39571443, Title: A Bunch of Nonsense, BM25: 4.5331
cluster-master | ID: 12008397, Title: A King and No King, BM25: 4.0258
cluster-master | 25/04/15 16:45:58 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master | 25/04/15 16:45:58 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master | 25/04/15 16:45:58 INFO YarnClientSchedulerBackend: Interrupting monitor thread
cluster-master | 25/04/15 16:45:58 INFO YarnClientSchedulerBackend: Shutting down all executors
cluster-master | 25/04/15 16:45:58 INFO YarnClientSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
cluster-master | 25/04/15 16:45:58 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
cluster-master | 25/04/15 16:45:58 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master | 25/04/15 16:45:58 INFO MemoryStore: MemoryStore cleared
cluster-master | 25/04/15 16:45:58 INFO BlockManager: BlockManager stopped
cluster-master | 25/04/15 16:45:58 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master | 25/04/15 16:45:58 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master | 25/04/15 16:45:59 INFO SparkContext: Successfully stopped SparkContext
cluster-master | 25/04/15 16:45:59 INFO ShutdownHookManager: Shutdown hook called
cluster-master | 25/04/15 16:45:59 INFO ShutdownHookManager: Deleting directory /tmp/spark-9ff282fa-f850-43fd-8b4b-sea2d588506
cluster-master | 25/04/15 16:45:59 INFO ShutdownHookManager: Deleting directory /tmp/spark-5eb7496-58b8-4ebc-b4dd-64dc9c0e1af
cluster-master | 25/04/15 16:45:59 INFO ShutdownHookManager: Deleting directory /tmp/spark-9ff282fa-f850-43fd-8b4b-sea2d588506/pyspark-61a400e6-a14f-4697-ad7c-10407483d039
cluster-master | 25/04/15 16:45:59 INFO CassandraConnector: Disconnected from Cassandra cluster.
cluster-master | 25/04/15 16:45:59 INFO SerialShutdownHooks: Successfully executed shutdown hook: Clearing session cache for C* connector
cluster-master exited with code 0
```

2.4 Reflections

The top-ranked documents for each query are relevant, indicating that the ranking algorithm works as expected.

For the "big data" query, documents with guides or researches and something big are prioritized. For "frog princess", we can see mostly fantastic and fairy tale books are recommended.

The implementation successfully demonstrates how distributed computing technologies can be combined to create a scalable search engine. The separation of indexing and query processing allows for efficient usage and future updates.

Potential improvements can be implementing relevance feedback and using more advanced text preprocessing.

This search engine provides a solid foundation that can be scaled to handle larger document collections or implementing more complex algorithms.