Introduction to Big Data. Assignment 2

Dzhavid Sadreddinov

April 2025

1 Methodology

1.1 System Architecture Overview

Our search engine implementation consists of three core components:

- **Indexing Pipeline**:
  - Batch processing using MapReduce
  - Cassandra for persistent storage
  - Hadoop for distributed computation
- **Search Components**:
  - Query processing and tokenization
  - BM25 ranking algorithm
  - Distributed scoring using PySpark
- **Infrastructure**:
  - Cassandra cluster for low-latency lookups
  - Spark for distributed computation
  - HDFS for document storage

1.2 Indexing Design Choices

1.2.1 Document Processing Pipeline

Implemented a two-phase MapReduce workflow:

- **Phase 1**: Document frequency counting (mapper1 → reducer1)
- **Phase 2**: Term frequency counting (mapper2 → reducer2)
- **Phase 3**: Document lengths evalution (mapper3 → reducer3)

1.2.2 Cassandra Schema Design

Optimized for read performance with denormalized tables:

```
CREATE TABLE term_frequencies (
    document_id bigint,
    term text,
    frequency counter,
    PRIMARY KEY ((document_id, term))
);
CREATE TABLE IF NOT EXISTS document_frequencies (
        term text PRIMARY KEY,
        count counter
);
CREATE TABLE IF NOT EXISTS document_lengths (
        document_id bigint PRIMARY KEY,
        length int,
        title text
);
```

- Store term frequencies at each document
- Store document frequencies for each term
- Store document lengths and their titles for search results

1.3 Search Implementation

1.3.1 BM25 Ranking Algorithm

Implemented the standard BM25 ranking function:

$$\text{score}(D, Q) = \sum_{q_i \in Q} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} \tag{1}$$

Where:

- $k_1 = 1.2$ (term frequency saturation)
- $b = 0.75$ (length normalization)

1.3.2 Distributed Scoring

1. **Query Processing**:
    - Case normalization
    - Tokenization using regex `\w+`
2. **Term Lookup**:

```
query_terms_bc = sc.broadcast(query_terms)
document_freqs_bc = sc.broadcast(document_freqs)
total_docs_bc = sc.broadcast(total_docs)
avg_doc_length_bc = sc.broadcast(avg_doc_length)
term_freqs_all_bc = sc.broadcast(term_freqs_all)
```

3. **Score Calculation**:
    - Mapped across matching documents
    - Reduce phase sums partial scores
    - Top-k selection using `takeOrdered`

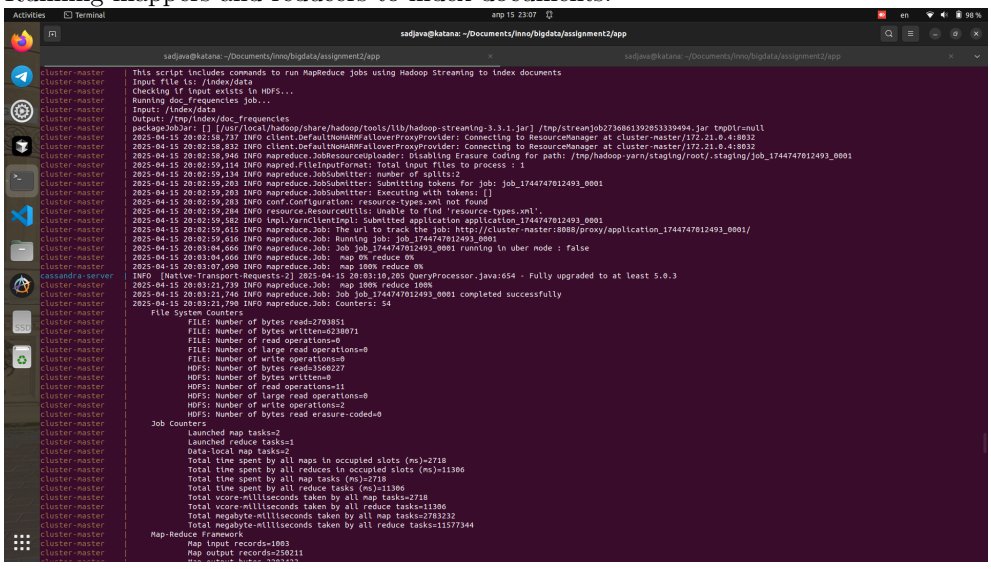
2 Guide on running

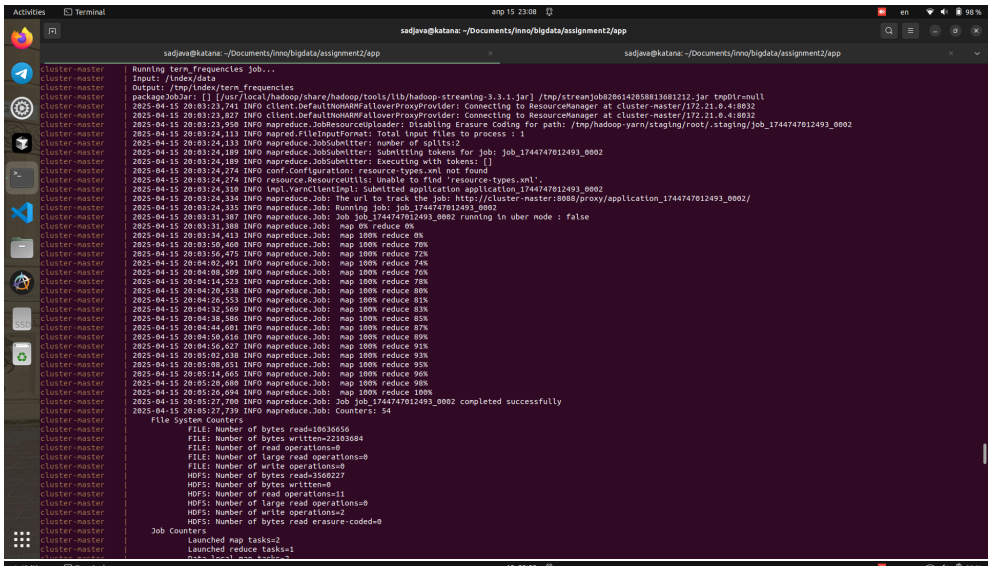
Clone the repo and run docker compose. Download zipped cassandra package to /app/cassandra.zip from here

```
git clone https://github.com/sadjava/big-data-assignment2.git
cd big-data-assignment2
docker compose up
```

3 Screenshots

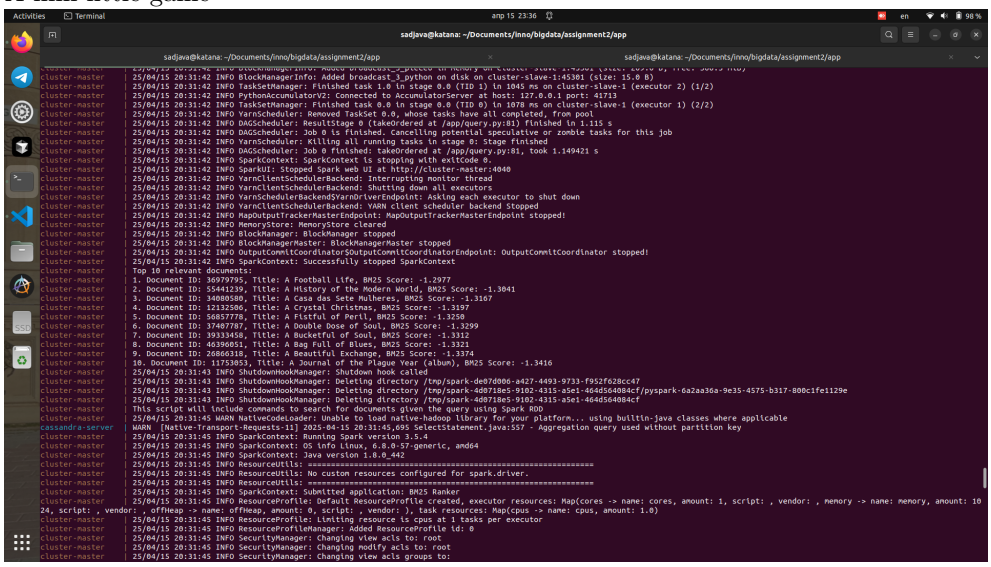
Running mappers and reducers to index documents:



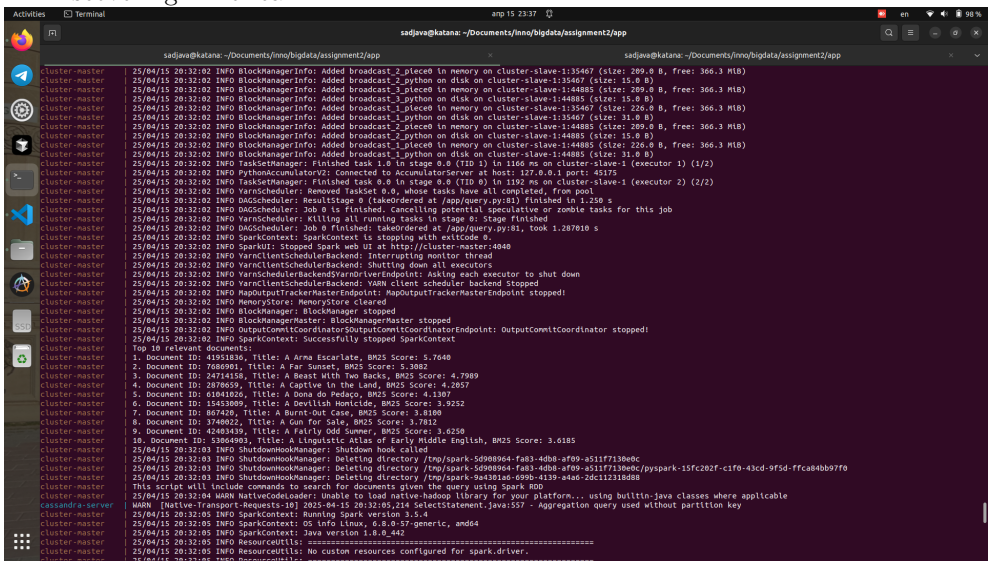

Activities
Terminal
anp 15 23:08
en
98 %
sadjava@katana: ~/Documents/inno/bigdata/assignment2/app
sadjava@katana: ~/Documents/inno/bigdata/assignment2/app
sadjava@katana: ~/Documents/inno/bigdata/assignment2/app
cluster-master | Running term_frequencies job...
cluster-master | Input: /index/data
cluster-master | Output: /tmp/index/term_frequencies
cluster-master | packageJobJar: [] [/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar] /tmp/streamjob8286142058813681212.jar tmpDir=null
cluster-master | 2025-04-15 20:03:23,741 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032
cluster-master | 2025-04-15 20:03:23,827 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032
cluster-master | 2025-04-15 20:03:23,950 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744747012493_0002
cluster-master | 2025-04-15 20:03:24,113 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master | 2025-04-15 20:03:24,133 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master | 2025-04-15 20:03:24,189 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744747012493_0002
cluster-master | 2025-04-15 20:03:24,189 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master | 2025-04-15 20:03:24,274 INFO conf.Configuration: resource-types.xml not found
cluster-master | 2025-04-15 20:03:24,274 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master | 2025-04-15 20:03:24,310 INFO impl.YarnClientImpl: Submitted application application_1744747012493_0002
cluster-master | 2025-04-15 20:03:24,334 INFO mapreduce.Job: The url to track the job: http://cluster-naster:8088/proxy/application_1744747012493_0002/
cluster-master | 2025-04-15 20:03:24,335 INFO mapreduce.Job: Running job: job_1744747012493_0002
cluster-master | 2025-04-15 20:03:31,387 INFO mapreduce.Job: Job job_1744747012493_0002 running in uber mode : false
cluster-master | 2025-04-15 20:03:31,388 INFO mapreduce.Job: map 0% reduce 0%
cluster-master | 2025-04-15 20:03:34,413 INFO mapreduce.Job: map 100% reduce 0%
cluster-master | 2025-04-15 20:03:50,460 INFO mapreduce.Job: map 100% reduce 70%
cluster-master | 2025-04-15 20:03:56,475 INFO mapreduce.Job: map 100% reduce 72%
cluster-master | 2025-04-15 20:04:02,491 INFO mapreduce.Job: map 100% reduce 74%
cluster-master | 2025-04-15 20:04:08,509 INFO mapreduce.Job: map 100% reduce 76%
cluster-master | 2025-04-15 20:04:14,523 INFO mapreduce.Job: map 100% reduce 78%
cluster-master | 2025-04-15 20:04:20,538 INFO mapreduce.Job: map 100% reduce 80%
cluster-master | 2025-04-15 20:04:26,553 INFO mapreduce.Job: map 100% reduce 81%
cluster-master | 2025-04-15 20:04:32,569 INFO mapreduce.Job: map 100% reduce 83%
cluster-master | 2025-04-15 20:04:38,586 INFO mapreduce.Job: map 100% reduce 85%
cluster-master | 2025-04-15 20:04:44,601 INFO mapreduce.Job: map 100% reduce 87%
cluster-master | 2025-04-15 20:04:50,616 INFO mapreduce.Job: map 100% reduce 89%
cluster-master | 2025-04-15 20:04:56,627 INFO mapreduce.Job: map 100% reduce 91%
cluster-master | 2025-04-15 20:05:02,638 INFO mapreduce.Job: map 100% reduce 93%
cluster-master | 2025-04-15 20:05:08,651 INFO mapreduce.Job: map 100% reduce 95%
cluster-master | 2025-04-15 20:05:14,665 INFO mapreduce.Job: map 100% reduce 96%
cluster-master | 2025-04-15 20:05:20,680 INFO mapreduce.Job: map 100% reduce 98%
cluster-master | 2025-04-15 20:05:26,694 INFO mapreduce.Job: map 100% reduce 100%
cluster-master | 2025-04-15 20:05:27,700 INFO mapreduce.Job: Job job_1744747012493_0002 completed successfully
cluster-master | 2025-04-15 20:05:27,739 INFO mapreduce.Job: Counters: 54
cluster-master | File System Counters
cluster-master | FILE: Number of bytes read=10636656
cluster-master | FILE: Number of bytes written=22103684
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=3560227
cluster-master | HDFS: Number of bytes written=0
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
Activities
Terminal
anp 15 23:08
en
98 %
sadjava@katana: ~/Documents/inno/bigdata/assignment2/app
sadjava@katana: ~/Documents/inno/bigdata/assignment2/app
sadjava@katana: ~/Documents/inno/bigdata/assignment2/app
cluster-master | Running doc_lengths job...
cluster-master | Input: /index/data
cluster-master | Output: /tmp/index/doc_lengths
cluster-master | packageJobJar: [] [/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar] /tmp/streamjob3956875152162072854.jar tmpDir=null
cluster-master | 2025-04-15 20:05:29,677 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032
cluster-master | 2025-04-15 20:05:29,763 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032
cluster-master | 2025-04-15 20:05:29,875 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744747012493_0003
cluster-master | 2025-04-15 20:05:30,033 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master | 2025-04-15 20:05:30,053 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master | 2025-04-15 20:05:30,104 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744747012493_0003
cluster-master | 2025-04-15 20:05:30,104 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master | 2025-04-15 20:05:30,183 INFO conf.Configuration: resource-types.xml not found
cluster-master | 2025-04-15 20:05:30,184 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master | 2025-04-15 20:05:30,214 INFO impl.YarnClientImpl: Submitted application application_1744747012493_0003
cluster-master | 2025-04-15 20:05:30,237 INFO mapreduce.Job: The url to track the job: http://cluster-naster:8088/proxy/application_1744747012493_0003/
cluster-master | 2025-04-15 20:05:30,238 INFO mapreduce.Job: Running job: job_1744747012493_0003
cluster-master | 2025-04-15 20:05:37,296 INFO mapreduce.Job: Job job_1744747012493_0003 running in uber mode : false
cluster-master | 2025-04-15 20:05:37,296 INFO mapreduce.Job: map 0% reduce 0%
cluster-master | 2025-04-15 20:05:40,321 INFO mapreduce.Job: map 100% reduce 0%
cluster-master | 2025-04-15 20:05:44,336 INFO mapreduce.Job: map 100% reduce 100%
cluster-master | 2025-04-15 20:05:45,345 INFO mapreduce.Job: Job job_1744747012493_0003 completed successfully
cluster-master | 2025-04-15 20:05:45,380 INFO mapreduce.Job: Counters: 54
cluster-master | File System Counters
cluster-master | FILE: Number of bytes read=37656
cluster-master | FILE: Number of bytes written=905669
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=3560227
cluster-master | HDFS: Number of bytes written=0
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)=2436
cluster-master | Total time spent by all reduces in occupied slots (ms)=1461
cluster-master | Total time spent by all map tasks (ms)=2436
cluster-master | Total time spent by all reduce tasks (ms)=1461
cluster-master | Total vcore-milliseconds taken by all map tasks=2436
cluster-master | Total vcore-milliseconds taken by all reduce tasks=1461
cluster-master | Total megabyte-milliseconds taken by all map tasks=2494464
cluster-master | Total megabyte-milliseconds taken by all reduce tasks=1496064
cluster-master | Map-Reduce Framework
cluster-master | Map input records=1003
cluster-master | Map output records=997
cluster-master | Map output bytes=35656
cluster-master | Map output materialized bytes=37662
cluster-master | Input split bytes=292
cluster-master | Combine input records=0

3.1 Results of search

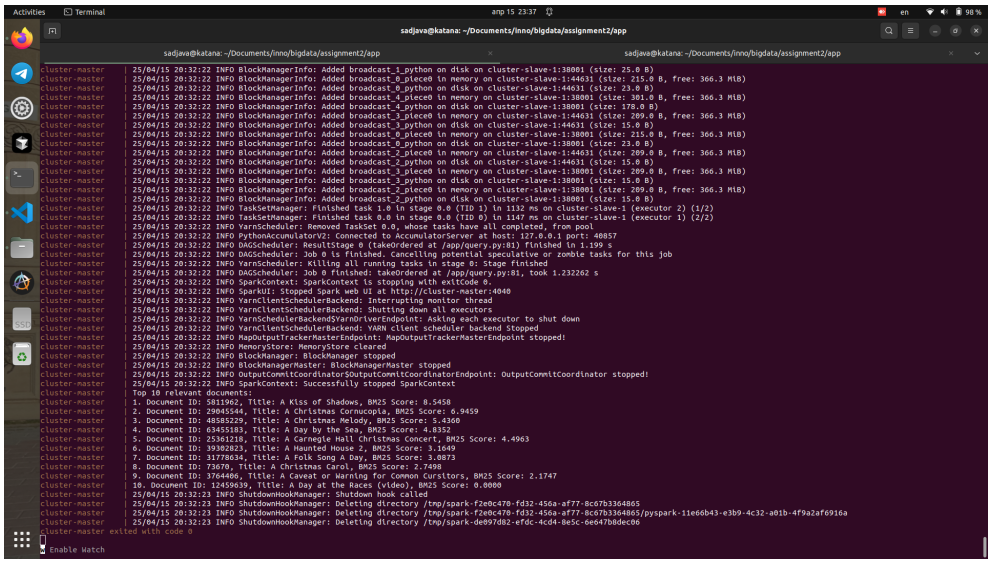
A film little game



Discovering America



Merry Christmas



Here's a detailed analysis of the search results with explanations and reflections:

4 Search Results Analysis

4.1 Query: "a film little game"

- **Top Result**: "A Football Life" (Score: -1.2977)
- **Observation**: The negative BM25 scores suggest:
  - Poor term matching between query and documents
  - Possible IDF (Inverse Document Frequency) issues where terms are too common
  - Mismatch between query intent and indexed content
- **Analysis**:
  - The query terms ("film", "little", "game") appear too generic
  - Many documents contain these common words without semantic relevance
  - Scores cluster closely (-1.29 to -1.33) indicating low discrimination

4.2 Query: "america discovering"

- **Top Result**: "A Arma Escarlate" (Score: 5.7640)
- **Key Findings**:

  - Positive scores indicate better term specificity
  - Top scoring documents likely contain both "america" and "discovering"
  - Score distribution shows better discrimination (5.76 to 3.61)

- **Anomalies**:
  - Portuguese title "A Arma Escarlate" ranking highest suggests:
    * Possible metadata issues in indexing
    * May contain English text about discovery despite Portuguese title
  - Linguistic Atlas appears relevant but ranks last in top 10

4.3 Query: "Merry christmas"

- **Top Result**: "A Kiss of Shadows" (Score: 8.5458)
- **Notable Patterns**:
  - Highest absolute scores among all queries
  - Christmas-themed titles rank appropriately:
    * "A Christmas Cornucopia" (2nd)
    * "A Christmas Melody" (3rd)
    * "A Christmas Carol" (8th)
- **Surprises**:
  - Non-Christmas title "A Kiss of Shadows" ranking first suggests:
    * Possible term frequency dominance in document text
    * Metadata/text discrepancy in indexing
    * "merry" may match other contexts (e.g., "merry making")