

Report

Big Data: Assignment 2

Simple Search Engine using Hadoop MapReduce

Mariia Shmakova

April 14, 2025

I Methodology

A. *Data Collection and Preparation*

- Initially, a parquet file containing the ID, title, and text fields for each document was downloaded. Using PySpark, data was extracted from 1000 documents and saved in a specified file format.
- Each document was named using its id and title, with spaces replaced by underscores.
- The processed documents were stored in HDFS in the data folder and an intermediate RDD was created to store the document metadata in `/index/data`.

B. *Indexer task*

Documents are stored in HDFS (`/index/data`) as plain text files. Each document is processed line-by-line to extract words. Indexing implemented using Hadoop Streaming with Python-based mapper and reducer scripts.

Hadoop Streaming was chosen for its flexibility in integrating custom mapper and reducer scripts written in Python. This allows us to process large datasets efficiently while maintaining simplicity in implementation.

The mapper tokenizes the text into words and emits key-value pairs (word, document ID). The reducer aggregates these pairs to build the inverted index.

C. *Ranker task*

- **BM25 Implementation** : The ranker script (`query.py`) calculates BM25 scores for all documents based on the user query. It retrieves data from Cassandra, computes scores, and ranks the top 10 documents.
- **Broadcast Variables** : To optimize performance, we used Spark broadcast variables to share BM25 statistics across all nodes.
- **RDD API** : The implementation relies on PySpark's RDD API for distributed computation.

II Demonstration

A. Data Preparation section

```
MINGW64/c/Users/shmak/Downloads/big-data-assignment2-2025-main
cluster-master | 322487_B...J...Arms Strong.txt
cluster-master | 32721476_B...Radhabai_Ananda_Rao.txt
cluster-master | 33573982_BBC_Radio_Extra.txt
cluster-master | 3395088_B-Dienst.txt
cluster-master | 3685183_B...Gabor.txt
cluster-master | 37627_BBC_World_Service.txt
cluster-master | 38918043_B_Street_Theatre.txt
cluster-master | 4082240_BBC_Allied_Expeditionary_Forces_Programme.txt
cluster-master | 409812_B_U.G._Mafia.txt
cluster-master | 41496195_BCOR.txt
cluster-master | 41603178_B...Rajam_Iyer.txt
cluster-master | 41800947_BBC_Orchies.txt
cluster-master | 42019425_B_Delta.txt
cluster-master | 42353931_B...Fornosa.txt
cluster-master | 43474572_B...Mario_Pinto.txt
cluster-master | 46782668_BAE_Chadderton.txt
cluster-master | 4756452_B90_nuclear_bomb.txt
cluster-master | 47848393_B-Scada.txt
cluster-master | 48232968_BBGM.txt
cluster-master | 48566810_B...Gunn.txt
cluster-master | 49273124_BIG_Super_Saturday.txt
cluster-master | 49535168_BBC_Cymru_Fyw.txt
cluster-master | 49997913_BBC_Arantia_Larochette.txt
cluster-master | 51081249_B...Wayne_Hughes_Jr...txt
cluster-master | 51132536_B...Barry_Shapiro.txt
cluster-master | 52200322_BC_Liquor_Stores.txt
cluster-master | 52574706_B...Elmore_High_School.txt
cluster-master | 5270789_B...Munusamy_Naidu.txt
cluster-master | 5277959_BAWN.txt
cluster-master | '53190751_BC_Sirius_Mure'$'\310\231''u_l_t'$'\303\242''rgu_Mure'$'\310\231''.txt'
cluster-master | 5360891_B...Snowden.txt
cluster-master | 53695501_BBWM.txt
cluster-master | 54709981_BCCIV.txt
cluster-master | '55493656_B10_road_(Namibia).txt'
cluster-master | 55714217_BC_Grand_Sport.txt
cluster-master | 56605120_BBC_News_Pidgin.txt
cluster-master | 569394_BBC_Radio_Northampton.txt
cluster-master | 570090_BBC_Radio_Somerset.txt
cluster-master | '57257820_B...J...Hill_(American_Football).txt'
cluster-master | 58951011_B...M...Muzamel_Haque.txt
cluster-master | '62026514_BERT_(language_model).txt'
cluster-master | 63169256_B...Mifflyn_Hood_Brick_Company_Building.txt
cluster-master | 63479899_BBQ_Brawl.txt
cluster-master | 63582857_B...Codanayaguy.txt
cluster-master | 65136137_B...Akber_Pasha.txt
cluster-master | '66409490_B_nai_b'rith_Cuba.txt'
cluster-master | 67487344_B-NL_Challenge_Trophy.txt
cluster-master | 67872949_BAL_Defensive_Player_of_the_Year.txt
cluster-master | 6819034_B...Sandhya.txt
cluster-master | 72916065_B_I_Videoigraphy.txt
cluster-master | 7744423_BCW_Can-Am_Heavyweight_Championship.txt
cluster-master | 7942202_BBC_East_Midlands.txt
cluster-master | 8273753_B_C...roll.txt
cluster-master | 882295_B.E.S...Publishing.txt
cluster-master | '8847663_B33_(New_York_City_bus).txt'
cluster-master | 9597512_B_notation.txt
cluster-master | a.parquet
cluster-master | Data preparation completed successfully!
```

B. Indexer tasks section

Running indexer:

```
MINGW64/c/Users/shmak/Downloads/big-data-assignment2-2025-main
cluster-master 25/04/14 16:23:15 INFO ShutdownHookManager: Deleting directory /tmp/spark-d6b4c9b0-9ecb-4392-878f-7c58a574db4a
cluster-master 25/04/14 16:23:15 INFO ShutdownHookManager: Deleting directory /tmp/spark-836d317e-0b28-4d65-b65e-451c9337a6dc
cluster-master Verifying output files...
cluster-master 15141813.BEE_Japan.txt
cluster-master 26062673.BFAST.txt
cluster-master 27751114.B_Monkey.txt
cluster-master 322487.B_J._Armstrong.txt
cluster-master 37627.BBC_World_Service.txt
cluster-master 4082240.BBC_Allied_Expeditionary_Forces_Programme.txt
cluster-master 42755508.BET_Awards_2014.txt
cluster-master 43474572.B_Mario_Pinto.txt
cluster-master 5360891.B._J._Snowden.txt
cluster-master 67872949.BAL_Defensive_Player_of_the_Year.txt
cluster-master 882295.B.E.S._Publishing.txt
cluster-master a.parquet
cluster-master Data preparation completed successfully!
cluster-master Verifying input path '/index/data'...
cluster-master Cleaning up previous output directory...
cluster-master Activating virtual environment...
cluster-master Downloading Hadoop Streaming JAR...
cluster-master --2025-04-14 16:23:28-- https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-streaming/3.4.1/hadoop-streaming-3.4.1.jar
cluster-master Connecting to repo1.maven.org (repo1.maven.org)... 146.75.116.209, 2a04:4e42:7d::209
cluster-master HTTP request sent, awaiting response... 200 OK
cluster-master Length: 141777 (138K) [application/java-archive]
cluster-master Saving to: '/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.4.1.jar'
cluster-master /usr/local/hadoop/s 100%[=====] 138.45K 732KB/s in 0.2s
cluster-master
cluster-master 2025-04-14 16:23:29 (732 KB/s) - '/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.4.1.jar' saved [141777/141777]
cluster-master
cluster-master Updating commons-cli to version 1.4...
cluster-master --2025-04-14 16:23:29-- https://repo1.maven.org/maven2/commons-cli/commons-cli/1.4/commons-cli-1.4.jar
cluster-master Resolving repo1.maven.org (repo1.maven.org)... 146.75.116.209, 2a04:4e42:7d::209
cluster-master Connecting to repo1.maven.org (repo1.maven.org)... 146.75.116.209:443... connected.
cluster-master HTTP request sent, awaiting response... 200 OK
cluster-master Length: 53820 (53K) [application/java-archive]
cluster-master Saving to: '/tmp/commons-cli-1.4.jar'
cluster-master /tmp/commons-cl-1 100%[=====] 52.56K --.-KB/s in 0.07s
cluster-master
cluster-master 2025-04-14 16:23:30 (715 KB/s) - '/tmp/commons-cli-1.4.jar' saved [53820/53820]
cluster-master
cluster-master Running Hadoop Streaming job...
cluster-master packageJobJar: [/tmp/hadoop-unjar5086396067056269659/] [] /tmp/streamjob4232527493433692582.jar tmpDir=null
cluster-master 2025-04-14 16:23:34.994 INFO client.DefaultHARMPFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.20.0.3:8032
cluster-master 2025-04-14 16:23:35.413 INFO client.DefaultHARMPFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.20.0.3:8032
cluster-master 2025-04-14 16:23:36.073 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744647324269_0001
cluster-master 2025-04-14 16:23:36.579 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master 2025-04-14 16:23:36.780 INFO mapreduce.JobSubmitter: number of splits:5
cluster-master 2025-04-14 16:23:37.107 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744647324269_0001
cluster-master 2025-04-14 16:23:37.107 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master 2025-04-14 16:23:37.444 INFO conf.Configuration: resource-types.xml not found
cluster-master 2025-04-14 16:23:37.445 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'
cluster-master 2025-04-14 16:23:39.025 INFO impl.YarnClientImpl: Submitted application application_1744647324269_0001
cluster-master 2025-04-14 16:23:39.075 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744647324269_0001/
cluster-master 2025-04-14 16:23:39.075 INFO mapreduce.Job: Running job: job_1744647324269_0001
```