

Big Data Assignment 2 Report

Arsenii Pavlov ars.pavlov@innopolis.university

Methodology

Data Preparation

- 1) Used a Parquet file containing Wikipedia articles. Selected 4000 documents using PySpark's `sample()` and `limit()` functions.
- 2) Each document is saved as `<doc_id>_<doc_title>.txt` in HDFS. Spaces in titles are replaced with `_`.

pipeline:

- 1) Read the Parquet file into a Spark DataFrame.
- 2) Extracted id, title, and text columns.
- 3) Saved documents to HDFS to `/data` using tab-separated format `<doc_id>\t<doc_title>\t<doc_text>`.

Indexer Tasks

The indexer uses two Hadoop MapReduce pipelines and save results in Cassandra.

1) TF Calculation

Mapper1

- 1) Read input documents from HDFS.
- 2) Tokenizes text cast to lowercase and splits on non-alphanumeric characters.
- 3) Convert to key-value pairs: `<term>#<doc_id>` to 1.

Reducer1

- 1) Aggregates counts for `<term>#<doc_id>` to compute TF.
- 2) Output: `<term> <doc_id> <tf>` to `/tmp/index/pipeline1`

2) DF Calculation

Mapper2

- 1) Reads TF data from Pipeline 1.
- 2) Convert term to 1 for each unique term-document pair.

Reducer2:

- 1) Sums counts to compute DF.
- 2) Output `<term>\t<df>` to `/tmp/index/pipeline2`

3) Saving to Cassandra

Tables:

documents: map document IDs and titles

Ranker Tasks

BM25 Calculation

- 1) Fetch DF for query terms from vocabulary table.
- 2) Fetch TF and document lengths from inverted_index and documents.
- 3) Compute BM25 score using provided formula
- 4) Sum scores across all query terms.
- 5) Rank documents by score and return top 10.

- 1) Join vocabulary and inverted_index.
- 2) Apply BM25 formula using Spark SQL functions.
- 3) Join results with `documents` table to get titles.

Demonstration

[illegible]

It will run all containers and run data preparation indexing and 1 query

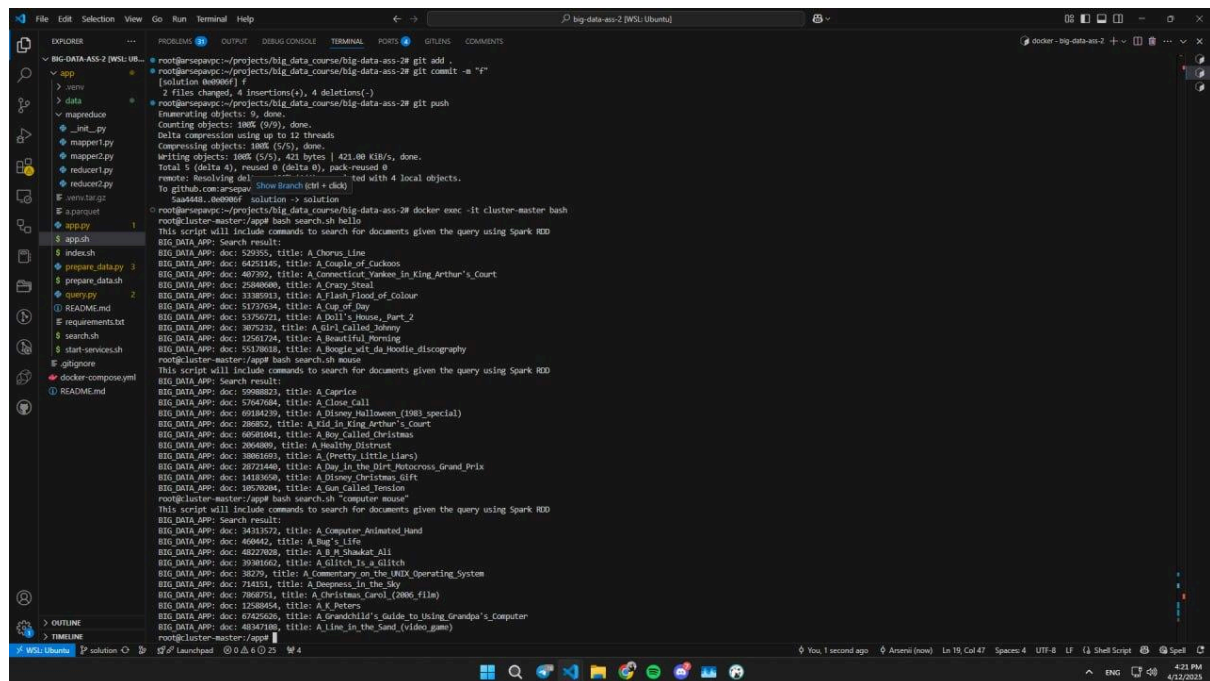
```
File Edit Selection View Go Run Terminal Help
docker - big-data-ast-2 [WSL: Ubuntu]

EXPLORER
▼ BIG-Data-AST-2 [WSL: Ubuntu]
  ▼ src
    data
  mapreduce
  init_py
  mapper1.py
  mapper2.py
  reducer1.py
  reducer2.py
  venv.txt
  sparknet
  app.sh
  index.sh
  prepare_data.sh
  prepare_index_data_3
  query.py
  README.md
  requirements.txt
  search
  start-services.sh
  upload_to_cassandra_4
  ghpages
  README.md

PROBLEMS
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS

25/04/15 20:40:15 INFO FileCacheReader: Reading File path: hdfs://cluster-master:9000/a/parquet, range: 536870172-681088640, partition values: [empty row]
25/04/15 20:40:18 INFO Executor: Finished task 4.0 in stage 4.0 (TID 13). 348 bytes result sent to driver
25/04/15 20:40:18 INFO TaskScheduler: Starting task 5.0 in stage 4.0 (TID 14) (Cluster-master, executor driver, partition 5, INM, 9587 bytes)
25/04/15 20:40:18 INFO Executor: Running task 5.0 in stage 4.0 (TID 14)
25/04/15 20:40:18 INFO TaskScheduler: Finished task 4.0 in stage 4.0 (TID 13) in 33 ms on cluster-master (executor driver) (5/7)
25/04/15 20:40:18 INFO FileCacheReader: Reading File path: hdfs://cluster-master:9000/a/parquet, range: 671868868-8630368, partition values: [empty row]
25/04/15 20:40:18 INFO Executor: Finished task 5.0 in stage 4.0 (TID 14). 3441 bytes result sent to driver
25/04/15 20:40:18 INFO TaskScheduler: Starting task 6.0 in stage 4.0 (TID 15) (Cluster-master, executor driver, partition 6, INM, 9587 bytes)
25/04/15 20:40:18 INFO Executor: Running task 6.0 in stage 4.0 (TID 15)
25/04/15 20:40:18 INFO TaskScheduler: Finished task 5.0 in stage 4.0 (TID 14) in 17 ms on cluster-master (executor driver) (6/7)
25/04/15 20:40:18 INFO FileCacheReader: Reading File path: hdfs://cluster-master:9000/a/parquet, range: 8630368-873287391, partition values: [empty row]
25/04/15 20:40:18 INFO Executor: Finished task 6.0 in stage 4.0 (TID 15). 3441 bytes result sent to driver
25/04/15 20:40:18 INFO TaskScheduler: Finished task 6.0 in stage 4.0 (TID 15) in 21 ms on cluster-master (executor driver) (7/7)
25/04/15 20:40:18 INFO TaskScheduler: Removed TaskSet 4.0, whose tasks have all completed, from pool
25/04/15 20:40:18 INFO DAGScheduler: ShuffleMapStage 4 (JavaSayPython at NativeMethodAccessorImpl.java:3) finished in 2.197 s
25/04/15 20:40:18 INFO DAGScheduler: Scheduling for newly running stages
25/04/15 20:40:18 INFO DAGScheduler: running: Set()
25/04/15 20:40:18 INFO DAGScheduler: waiting: Set(ResultStage 5)
25/04/15 20:40:18 INFO DAGScheduler: failed: Set()
25/04/15 20:40:18 INFO DAGScheduler: Submitting ResultStage 5 (PythonIO[10] at foreach at /app/prepare_data.py:25), which has no missing parents
25/04/15 20:40:18 INFO MemoryStore: Block broadcast 6 stored as values in memory (estimated size 24.0 KiB, free 365.9 MiB)
25/04/15 20:40:18 INFO MemoryStore: Block broadcast 6 placed into blocks in memory (estimated size 12.1 KiB, free 365.8 MiB)
25/04/15 20:40:18 INFO BlockManager: Added broadcast 6, placed in memory on cluster-master-32363 (size: 12.1 KiB, free: 366.2 MiB)
25/04/15 20:40:18 INFO SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:1585
25/04/15 20:40:18 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 5 (PythonIO[10] at foreach at /app/prepare_data.py:25) (first 15 tasks are for partitions Vector(0))
25/04/15 20:40:18 INFO TaskScheduler: Starting task 8.0 in stage 5.0 (TID 16) (Cluster-master, executor driver, partition 8, MODE_LOCAL, 8999 bytes)
25/04/15 20:40:18 INFO Executor: Running task 8.0 in stage 5.0 (TID 16)
25/04/15 20:40:18 INFO ShuffleBlockFetcherIterator: Getting 1 (4.0 MiB) non-empty blocks including 1 (4.0 MiB) local and 0 (0.0) host-local and 0 (0.0) push-merged-local and 0 (0.0) remote blocks
25/04/15 20:40:18 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
25/04/15 20:40:21 INFO PythonRunner: Times: total = 3810, boot = 375, init = 21, finish = 2622
25/04/15 20:40:21 INFO Executor: Finished task 8.0 in stage 5.0 (TID 16). 3788 bytes result sent to driver
25/04/15 20:40:21 INFO TaskScheduler: Finished task 8.0 in stage 5.0 (TID 16) in 3868 ms on cluster-master (executor driver) (1/1)
25/04/15 20:40:21 INFO TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool.
25/04/15 20:40:21 INFO PythonContextMapperV2: Connected to AccumulatorServer at host: 127.0.0.1 port: 48699
25/04/15 20:40:21 INFO DAGScheduler: ResultStage 5 (foreach at /app/prepare_data.py:25) finished in 3.077 s
25/04/15 20:40:21 INFO DAGScheduler: Job 3 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/15 20:40:21 INFO TaskSchedulerImpl: Killing all running tasks in stage 5: Stage finished
25/04/15 20:40:21 INFO DAGScheduler: Job 3 finished: foreach at /app/prepare_data.py:25, took 5.232212 s
BIG_DATA_APP: data prepared
25/04/15 20:40:21 INFO SparkContext: Invoking stop() from Shutdown hook
25/04/15 20:40:21 INFO SparkContext: SparkContext is stopping with exitcode 0.
25/04/15 20:40:21 INFO SparkUI: Stopped Spark web UI at http://cluster-master:6840
25/04/15 20:40:21 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/15 20:40:21 INFO MemoryStore: MemoryStore cleared
25/04/15 20:40:21 INFO BlockManager: BlockManager stopped
25/04/15 20:40:21 INFO BlockManagerMaster: BlockManagerMaster stopped
25/04/15 20:40:21 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/15 20:40:21 INFO SparkContext: Successfully stopped SparkContext
25/04/15 20:40:21 INFO ShutdownHookManager: Shutdown hook called
25/04/15 20:40:21 INFO ShutdownHookManager: Deleting directory /tmp/spark-ca7553-3980-4dfc-4abf-73795647b047
25/04/15 20:40:21 INFO ShutdownHookManager: Deleting directory /tmp/spark-ca7553-3980-4dfc-4abf-73795647b047
25/04/15 20:40:21 INFO ShutdownHookManager: Deleting directory /tmp/spark-b8d7a3f5-ca16-4448-906a-568989483c05
BIG_DATA_APP: Putting data to hdfs
```

[illegible]



```
root@arsenavpc:~/projects/big_data_course/big-data-ass-2# git add .
root@arsenavpc:~/projects/big_data_course/big-data-ass-2# git commit -m "f"
[initial commit] 1
2 files changed, 4 insertions(+), 4 deletions(-)
root@arsenavpc:~/projects/big_data_course/big-data-ass-2# git push
Enumerating objects: 9, done.
Counting objects: 100% (9/9), done.
Delta compression using up to 32 threads
Compressing objects: 100% (5/5), done.
Writing objects: 100% (5/5), 421 bytes | 421.00 KiB/s, done.
Total 5 (delta 4), reused 0 (delta 0), pack-reused 0
remote: Resolving url: https://github.com:root@arsenavpc:~/projects/big_data_course/big-data-ass-2# docker exec -it cluster-master bash
root@cluster-master:/app# bash search.sh hello
This script will include commands to search for documents given the query using Spark RDD
BIG_DATA_APP: doc: 529355, title: A Chorus Line
BIG_DATA_APP: doc: 64251145, title: A Couple of Cuckoos
BIG_DATA_APP: doc: 407262, title: A Connecticut Yankee in King Arthur's Court
BIG_DATA_APP: doc: 25840600, title: A Crazy Steal
BIG_DATA_APP: doc: 31989913, title: A Flash Flood of Colour
BIG_DATA_APP: doc: 5372634, title: A Gun of Day
BIG_DATA_APP: doc: 53756723, title: A Doll's House, Part 2
BIG_DATA_APP: doc: 3075232, title: A Girl Called Sonny
BIG_DATA_APP: doc: 32583726, title: A Beautiful Morning
BIG_DATA_APP: doc: 55178018, title: A Boogie wit da Hoodie discography
root@cluster-master:/app# bash search.sh mouse
This script will include commands to search for documents given the query using Spark RDD
BIG_DATA_APP: Search result:
BIG_DATA_APP: doc: 10988823, title: A Caprice
BIG_DATA_APP: doc: 57647084, title: A Close Call
BIG_DATA_APP: doc: 69184239, title: A Disney Halloween (1983 special)
BIG_DATA_APP: doc: 2668452, title: A Kid in King Arthur's Court
BIG_DATA_APP: doc: 80541045, title: A Boy Called Christmas
BIG_DATA_APP: doc: 20648809, title: A Healthy Distrust
BIG_DATA_APP: doc: 38881693, title: A (Pretty Little) Liar(s)
BIG_DATA_APP: doc: 28721446, title: A Day in the Dirt: Patrons Grand Prix
BIG_DATA_APP: doc: 14183608, title: A Disney Christmas Gift
BIG_DATA_APP: doc: 10576284, title: A Gun Called Tension
root@cluster-master:/app# bash search.sh "computer mouse"
This script will include commands to search for documents given the query using Spark RDD
BIG_DATA_APP: Search result:
BIG_DATA_APP: doc: 14131572, title: A Computer Animated Hand
BIG_DATA_APP: doc: 4698442, title: A Bug's Life
BIG_DATA_APP: doc: 48227628, title: A B.P. Shocker All
BIG_DATA_APP: doc: 39381662, title: A Glitch Is a Glitch
BIG_DATA_APP: doc: 38279, title: A Commentary on the UNIX Operating System
BIG_DATA_APP: doc: 714151, title: A Despair in the Sky
BIG_DATA_APP: doc: 7862753, title: A Christmas Carol (2009 film)
BIG_DATA_APP: doc: 12588454, title: A.K. Peters
BIG_DATA_APP: doc: 67425628, title: A Grandchild's Guide to Using Grandpa's Computer
BIG_DATA_APP: doc: 40347186, title: A Line in the Sand (video game)
```

The results are actually good it can be seen in the example with computer mouse titles (they all connected to computer)