Student: Kuzmich Aleksandra

Big Data Assignment 2 report

The indexing process is split into:

Mapper1.py

Parses each input document (doc_id, title, text).

Combines title and text into a single string.

Tokenizes the content into lowercase words using regex.

Emits key-value pairs: word<TAB>doc_id<TAB>1.

Reducer1.py

Receives sorted mapper output as input.

Aggregates term frequencies for each (term, doc_id) pair.

Calculates document lengths by summing frequencies.

Connects to Cassandra and stores: inverted_index(term, doc_id, freq), doc_stats(doc_id, length)

The reducer writes two types of information into Cassandra: inverted_index stores the frequency of each term in each document. doc_stats stores the length (total number of words) of each document.

Index.sh

This bash script automates the indexing pipeline:

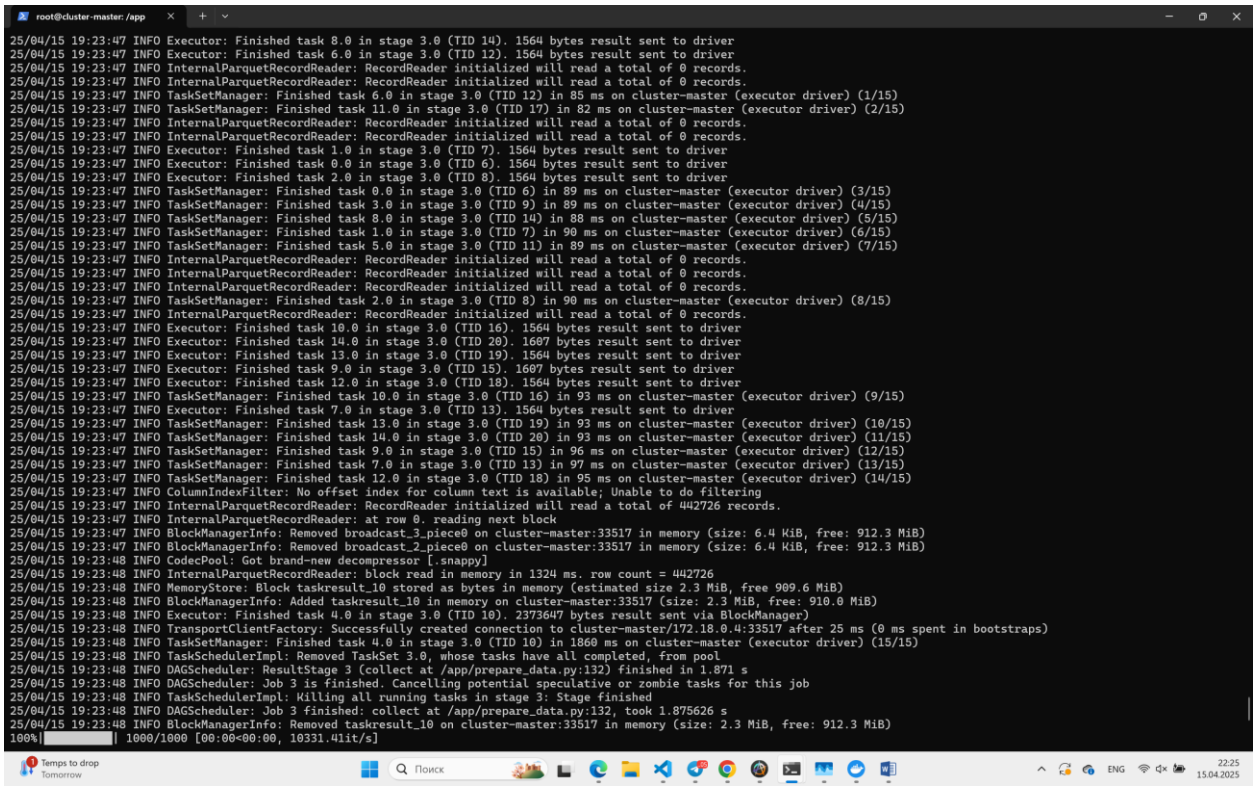
Accepts a file path to documents as input (default: /index/data).

Executes a Hadoop streaming job with mapper1.py.

Collects the output from HDFS and feeds it into reducer1.py via standard input.

reducer1.py writes the final index data into Cassandra.

Prepare data



Mapper+Reducer run