Methodology

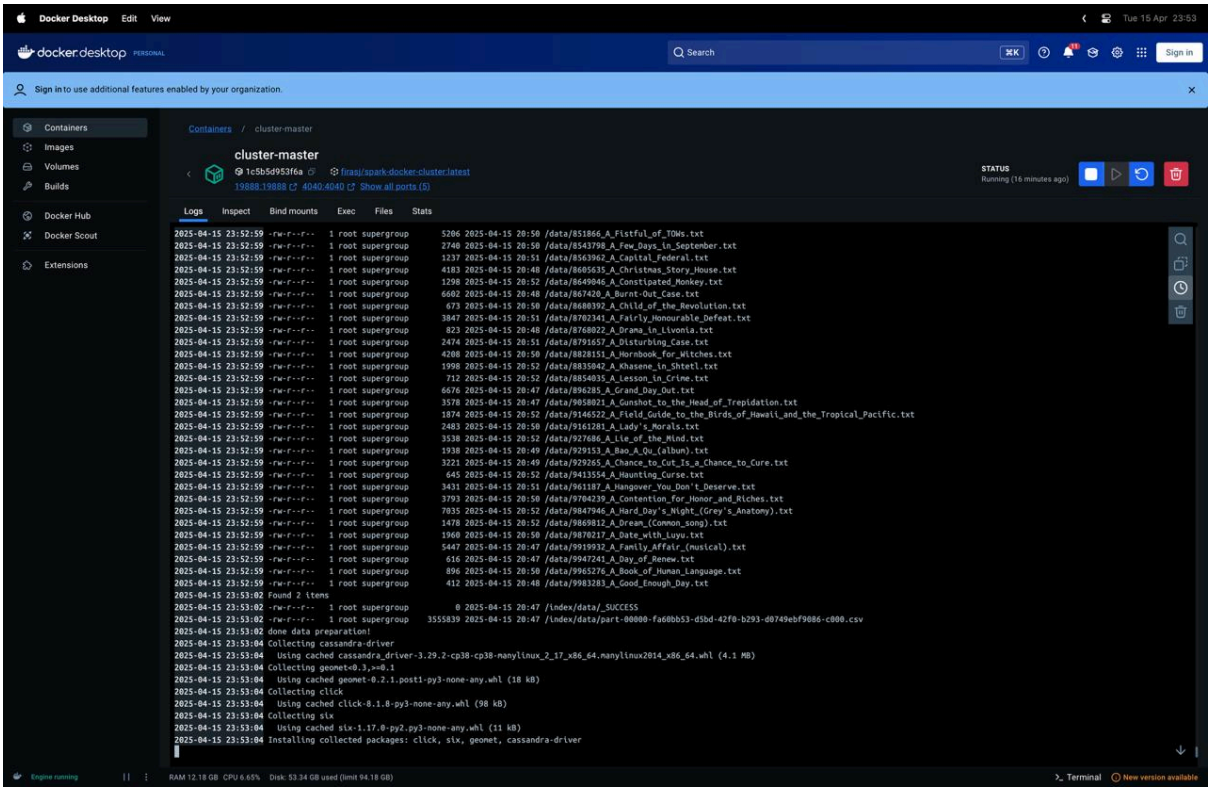
After the data preparation (prepared_data.py), the system builds the indices (index.sh). It begins by activating the virtual environment, ensures the Cassandra driver is bundled, and runs app.py to configure the database schema. This script defines four Cassandra tables: one for the inverted index to store word-document relationships with frequencies, one for storing document titles and lengths, a table for overall statistics like total document count, and another for vocabulary data including TF-IDF values used in scoring.

The main indexing logic is split between the mapper and reducer. The mapper reads through each document, breaks down the text into individual words, filters out common stopwords, and emits relevant information such as terms, document IDs, and frequencies, along with metadata like the title and document length. The reducer then gathers this data, calculates global statistics like IDF, and writes all the processed information to Cassandra in batches for efficiency.

Demonstration

For running need to start with: **docker compose up -d**

Indexing the documents


Docker Desktop Edit View
Tue 15 Apr 23:53
docker desktop PERSONAL
Search
Sign in
Sign in to use additional features enabled by your organization.
Containers
Images
Volumes
Builds
Docker Hub
Docker Scout
Extensions
Containers / cluster-master
cluster-master
1c5b5d953f6a
firasj/spark-docker-cluster:latest
19888:19888 4040:4040 Show all ports (5)
STATUS
Running (16 minutes ago)
Logs Inspect Bind mounts Exec Files Stats
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 5206 2025-04-15 20:50 /data/851866_A_Fistful_of_TOWs.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 2740 2025-04-15 20:50 /data/8543798_A_Few_Days_in_September.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 1237 2025-04-15 20:51 /data/8563962_A_Capital_Federal.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 4103 2025-04-15 20:48 /data/8605635_A_Christmas_Story_House.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 1298 2025-04-15 20:52 /data/8649046_A_Constipated_Monkey.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 6602 2025-04-15 20:48 /data/867420_A_Burnt-Out_Case.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 673 2025-04-15 20:50 /data/8680392_A_Child_of_the_Revolution.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 3847 2025-04-15 20:51 /data/8702341_A_Fairly_Honourable_Defeat.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 823 2025-04-15 20:48 /data/8768022_A_Drama_in_Livonia.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 2474 2025-04-15 20:51 /data/8791657_A_Disturbing_Case.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 4208 2025-04-15 20:50 /data/8828151_A_Hornbook_for_Witches.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 1998 2025-04-15 20:52 /data/8835042_A_Khasene_Un_Shtetl.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 712 2025-04-15 20:52 /data/8854035_A_Lesson_in_Crime.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 6676 2025-04-15 20:47 /data/896285_A_Grand_Day_Out.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 3578 2025-04-15 20:47 /data/9058021_A_Gunshot_to_the_Head_of_Trepidation.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 1874 2025-04-15 20:52 /data/9146522_A_Field_Guide_to_the_Birds_of_Hawaii_and_the_Tropical_Pacific.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 2483 2025-04-15 20:50 /data/9161281_A_Lady's_Morals.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 3538 2025-04-15 20:52 /data/927686_A_Lie_of_the_Mind.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 1938 2025-04-15 20:49 /data/929153_A_Bao_A_Qu_(album).txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 3221 2025-04-15 20:49 /data/929265_A_Chance_to_Cut_Is_a_Chance_to_Cure.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 645 2025-04-15 20:52 /data/9413554_A_Haunting_Curse.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 3431 2025-04-15 20:51 /data/961187_A_Hangover_You_Don't_Deserve.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 3793 2025-04-15 20:50 /data/9704239_A_Contention_for_Honor_and_Riches.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 7035 2025-04-15 20:52 /data/9847946_A_Hard_Day's_Night_(Grey's_Anatomy).txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 1478 2025-04-15 20:52 /data/9869812_A_Dream_(Common_song).txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 1960 2025-04-15 20:50 /data/9870217_A_Date_with_Luyu.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 5447 2025-04-15 20:47 /data/9919932_A_Family_Affair_(musical).txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 616 2025-04-15 20:47 /data/9947241_A_Day_of_Renew.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 896 2025-04-15 20:50 /data/9965276_A_Book_of_Human_Language.txt
2025-04-15 23:52:59 -rw-r--r-- 1 root supergroup 412 2025-04-15 20:48 /data/9983283_A_Good_Enough_Day.txt
2025-04-15 23:53:02 Found 2 items
2025-04-15 23:53:02 -rw-r--r-- 1 root supergroup 0 2025-04-15 20:47 /index/data/_SUCCESS
2025-04-15 23:53:02 -rw-r--r-- 1 root supergroup 3555839 2025-04-15 20:47 /index/data/part-00000-fa60bb53-d5bd-42f0-b293-d0749ebf9086-c000.csv
2025-04-15 23:53:02 done data preparation!
2025-04-15 23:53:04 Collecting cassandra-driver
2025-04-15 23:53:04 Using cached cassandra_driver-3.29.2-cp38-cp38-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (4.1 MB)
2025-04-15 23:53:04 Collecting geomet<0.3,>=0.1
2025-04-15 23:53:04 Using cached geomet-0.2.1.post1-py3-none-any.whl (18 kB)
2025-04-15 23:53:04 Collecting click
2025-04-15 23:53:04 Using cached click-8.1.8-py3-none-any.whl (98 kB)
2025-04-15 23:53:04 Collecting six
2025-04-15 23:53:04 Using cached six-1.17.0-py2.py3-none-any.whl (11 kB)
2025-04-15 23:53:04 Installing collected packages: click, six, geomet, cassandra-driver
Engine running
RAM 12.18 GB CPU 6.65% Disk: 53.34 GB used (limit 94.18 GB)
Terminal
New version available