Overview

The search engine implementation is structured around a MapReduce pipeline, leveraging Hadoop for distributed processing and Cassandra for data storage.

Data Preparation

1. Data Ingestion: The input data is initially stored in a local directory and then uploaded to HDFS.
1. Data Formatting: Each document is processed to extract relevant metadata, such as document ID and title

MapReduce Pipeline (three reducers and mappers)

1. Document Indexing:
   - Mapper: Processes each document to extract terms and their occurrences.
   - Reducer: Aggregates term occurrences across documents to build an inverted index.
1. Document Statistics:
   - Mapper: Computes basic statistics for each document, such as term frequency.
   - Reducer: Aggregates these statistics to provide a comprehensive view of document characteristics.
1. Term Statistics:
   - Mapper: Extracts term-specific data, such as document frequency.
   - Reducer: Compiles term statistics to support relevance scoring.

Data Storage

- Cassandra Integration: The final indices and statistics are stored in Cassandra, enabling efficient retrieval and query execution. This choice supports the scalability and fault tolerance required for handling large datasets.

Query Execution

- Spark RDDs: Queries are executed using Spark, which retrieves data from Cassandra and computes relevance scores using the BM25 algorithm.

Demonstration

Execute the following command in your terminal:

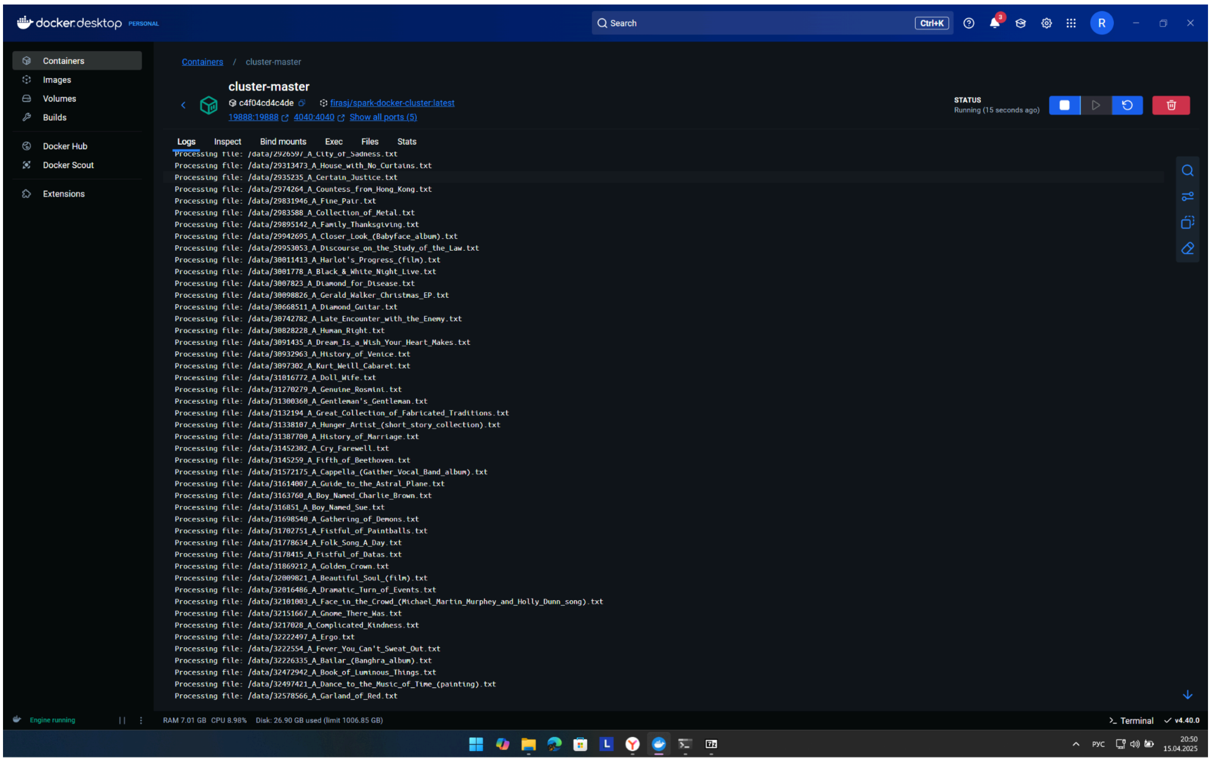
```
docker compose up --build
```

This command will build and launch all necessary services.

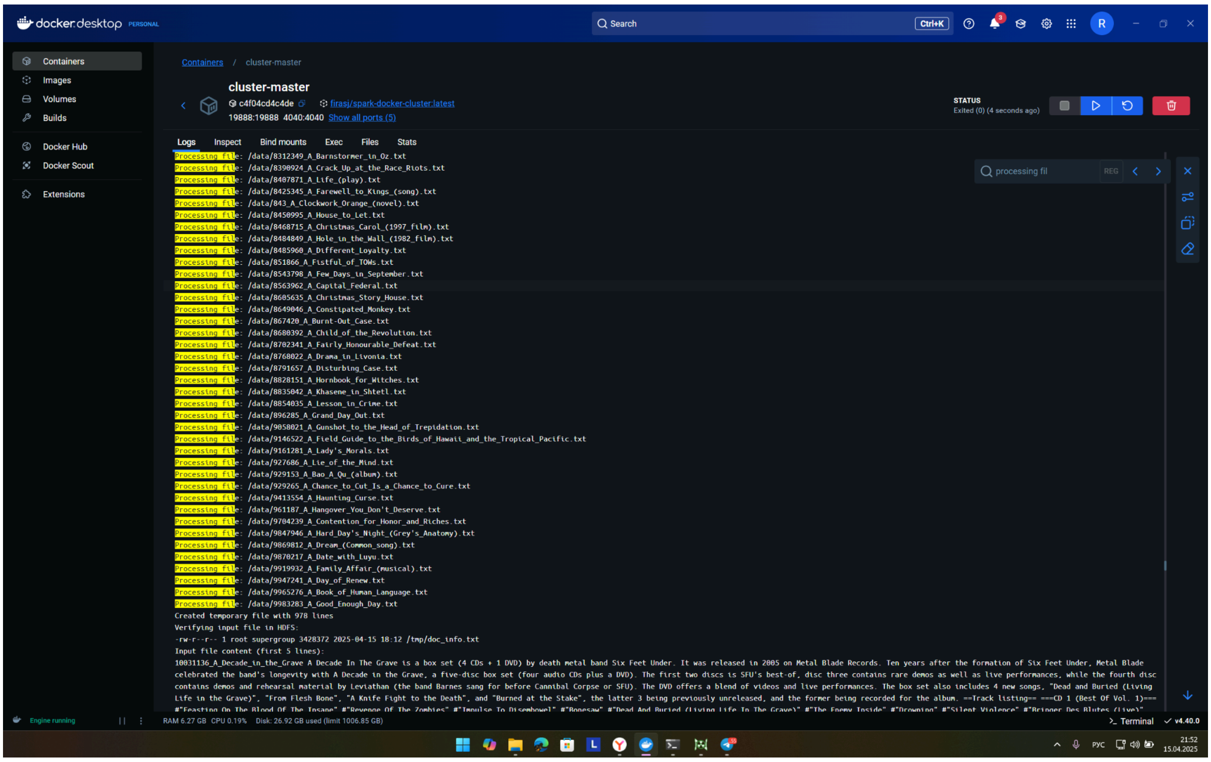
All files located in the `/data` folder will be automatically processed by the system.

The `app.sh` script will execute a predefined query.

All 1000 documents are indexed successfully:

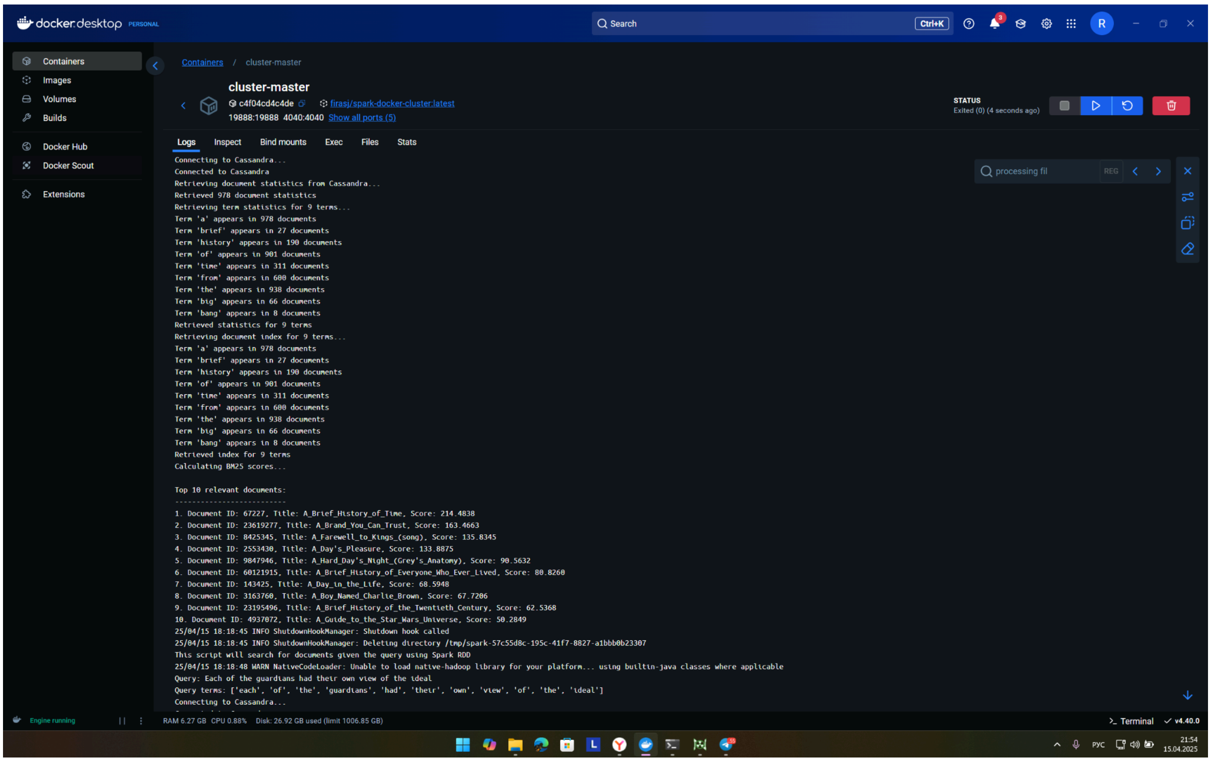


```
Processing file: /data/2926597_A_City_of_Sadness.txt
Processing file: /data/29313473_A_House_with_No_Curtains.txt
Processing file: /data/2935235_A_Certain_Justice.txt
Processing file: /data/2974264_A_Countess_from_Hong_Kong.txt
Processing file: /data/29831946_A_Fine_Pair.txt
Processing file: /data/2983588_A_Collection_of_Metal.txt
Processing file: /data/29895142_A_Family_Thanksgiving.txt
Processing file: /data/29942695_A_Closer_Look_(Babyface_album).txt
Processing file: /data/29953053_A_Discourse_on_the_Study_of_the_Law.txt
Processing file: /data/30011413_A_Harlot's_Progress_(film).txt
Processing file: /data/3001778_A_Black_&_White_Night_Live.txt
Processing file: /data/3007823_A_Diamond_for_Disease.txt
Processing file: /data/30098826_A_Gerald_Walker_Christmas_EP.txt
Processing file: /data/30668511_A_Diamond_Guitar.txt
Processing file: /data/30742782_A_Late_Encounter_with_the_Enemy.txt
Processing file: /data/30828228_A_Human_Right.txt
Processing file: /data/3091435_A_Dream_Is_a_Wish_Your_Heart_Makes.txt
Processing file: /data/30932963_A_History_of_Venice.txt
Processing file: /data/3097302_A_Kurt_Weill_Cabaret.txt
Processing file: /data/31016772_A_Doll_Wife.txt
Processing file: /data/31270279_A_Genuine_Rosmini.txt
Processing file: /data/31300360_A_Gentleman's_Gentleman.txt
Processing file: /data/3132194_A_Great_Collection_of_Fabricated_Traditions.txt
Processing file: /data/31338107_A_Hunger_Artist_(short_story_collection).txt
Processing file: /data/31387700_A_History_of_Marriage.txt
Processing file: /data/31452302_A_Cry_Farewell.txt
Processing file: /data/3145259_A_Fifth_of_Beethoven.txt
Processing file: /data/31572175_A_Cappella_(Gaither_Vocal_Band_album).txt
Processing file: /data/31614007_A_Guide_to_the_Astral_Plane.txt
Processing file: /data/3163760_A_Boy_Named_Charlie_Brown.txt
Processing file: /data/316851_A_Boy_Named_Sue.txt
Processing file: /data/31698540_A_Gathering_of_Demons.txt
Processing file: /data/31702751_A_Fistful_of_Paintballs.txt
Processing file: /data/31778634_A_Folk_Song_A_Day.txt
Processing file: /data/3178415_A_Fistful_of_Datas.txt
Processing file: /data/31869212_A_Golden_Crown.txt
Processing file: /data/32009821_A_Beautiful_Soul_(film).txt
Processing file: /data/32016486_A_Dramatic_Turn_of_Events.txt
Processing file: /data/32101003_A_Face_in_the_Crowd_(Michael_Martin_Murphey_and_Holly_Dunn_song).txt
Processing file: /data/32151667_A_Gnome_There_Was.txt
Processing file: /data/3217028_A_Complicated_Kindness.txt
Processing file: /data/32222497_A_Ergo.txt
Processing file: /data/3222554_A_Fever_You_Can't_Sweat_Out.txt
Processing file: /data/32226335_A_Bailar_(Banghra_album).txt
Processing file: /data/32472942_A_Book_of_Luminous_Things.txt
Processing file: /data/32497421_A_Dance_to_the_Music_of_Time_(painting).txt
Processing file: /data/32578566_A_Garland_of_Red.txt
```



```
Processing file: /data/8312349_A_Barnstormer_in_Oz.txt
Processing file: /data/8390924_A_Crack_Up_at_the_Race_Riots.txt
Processing file: /data/8407871_A_Life_(play).txt
Processing file: /data/8425345_A_Farewell_to_Kings_(song).txt
Processing file: /data/843_A_Clockwork_Orange_(novel).txt
Processing file: /data/8450995_A_House_to_Let.txt
Processing file: /data/8468715_A_Christmas_Carol_(1997_film).txt
Processing file: /data/8484849_A_Hole_in_the_Wall_(1982_film).txt
Processing file: /data/8485960_A_Different_Loyalty.txt
Processing file: /data/851866_A_Fistful_of_TOWs.txt
Processing file: /data/8543798_A_Few_Days_in_September.txt
Processing file: /data/8563962_A_Capital_Federal.txt
Processing file: /data/8605635_A_Christmas_Story_House.txt
Processing file: /data/8649046_A_Constipated_Monkey.txt
Processing file: /data/867420_A_Burnt-Out_Case.txt
Processing file: /data/8680392_A_Child_of_the_Revolution.txt
Processing file: /data/8702341_A_Fairly_Honourable_Defeat.txt
Processing file: /data/8768022_A_Drama_in_Livonia.txt
Processing file: /data/8791657_A_Disturbing_Case.txt
Processing file: /data/8828151_A_Hornbook_for_Witches.txt
Processing file: /data/8835042_A_Khasene_in_Shtetl.txt
Processing file: /data/8854035_A_Lesson_in_Crime.txt
Processing file: /data/896285_A_Grand_Day_Out.txt
Processing file: /data/9058021_A_Gunshot_to_the_Head_of_Trepidation.txt
Processing file: /data/9146522_A_Field_Guide_to_the_Birds_of_Hawaii_and_the_Tropical_Pacific.txt
Processing file: /data/9161281_A_Lady's_Morals.txt
Processing file: /data/927686_A_Lie_of_the_Mind.txt
Processing file: /data/929153_A_Bao_A_Qu_(album).txt
Processing file: /data/929265_A_Chance_to_Cut_Is_a_Chance_to_Cure.txt
Processing file: /data/9413554_A_Haunting_Curse.txt
Processing file: /data/961187_A_Hangover_You_Don't_Deserve.txt
Processing file: /data/9704239_A_Contention_for_Honor_and_Riches.txt
Processing file: /data/9847946_A_Hard_Day's_Night_(Grey's_Anatomy).txt
Processing file: /data/9869812_A_Dream_(Common_song).txt
Processing file: /data/9870217_A_Date_with_Luyu.txt
Processing file: /data/9919932_A_Family_Affair_(musical).txt
Processing file: /data/9947241_A_Day_of_Renew.txt
Processing file: /data/9965276_A_Book_of_Human_Language.txt
Processing file: /data/9983283_A_Good_Enough_Day.txt
Created temporary file with 978 lines
Verifying input file in HDFS:
-rw-r--r-- 1 root supergroup 3428372 2025-04-15 18:12 /tmp/doc_info.txt
Input file content (first 5 lines):
10031136_A_Decade_in_the_Grave A Decade In The Grave is a box set (4 CDs + 1 DVD) by death metal band Six Feet Under. It was released in 2005 on Metal Blade Records. Ten years after the formation of Six Feet Under, Metal Blade
celebrated the band's longevity with A Decade in the Grave, a five-disc box set (four audio CDs plus a DVD). The first two discs is SFU's best-of, disc three contains rare demos as well as live performances, while the fourth disc
contains demos and rehearsal material by Leviathan (the band Barnes sang for before Cannibal Corpse or SFU). The DVD offers a blend of videos and live performances. The box set also includes 4 new songs, "Dead and Buried (Living
Life in the Grave)", "From Flesh Bone", "A Knife Fight to the Death", and "Burned at the Stake", the latter 3 being previously unreleased, and the former being recorded for the album. ==Track listing== ===CD 1 (Best Of Vol. 1)===
#"Feasting On The Blood Of The Insane" #"Revenge Of The Zombies" #"Impulse To Disembowel" #"Bonesaw" #"Dead And Buried (Living Life In The Grave)" #"The Enemy Inside" #"Drowning" #"Silent Violence" #"Bringer Des Blutes (Live)"
```

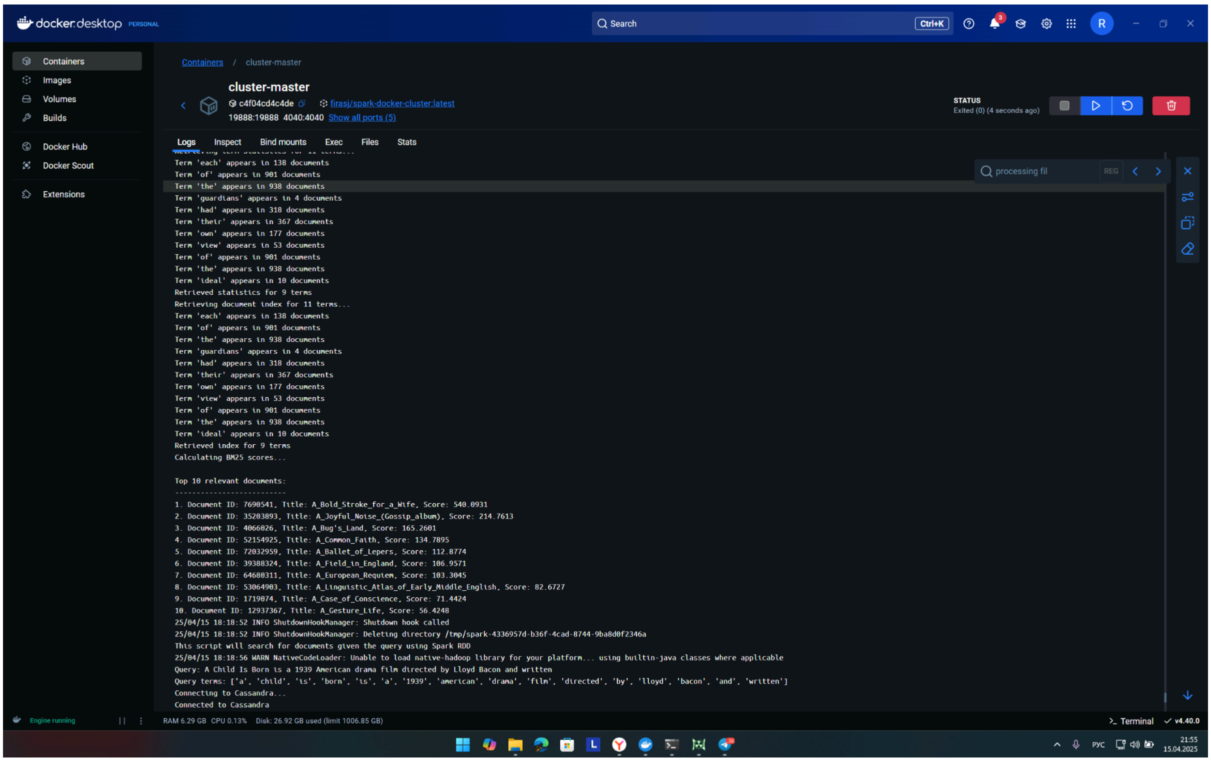
The first query is "A Brief History of Time: From the Big Bang" from book “67227_A_Brief_History_of_Time”, as we can see stats are calculated correctly and the expected document ranked 1st:



```
Connecting to Cassandra...
Connected to Cassandra
Retrieving document statistics from Cassandra...
Retrieved 978 document statistics
Retrieving term statistics for 9 terms...
Term 'a' appears in 978 documents
Term 'brief' appears in 27 documents
Term 'history' appears in 190 documents
Term 'of' appears in 901 documents
Term 'time' appears in 311 documents
Term 'from' appears in 600 documents
Term 'the' appears in 938 documents
Term 'big' appears in 66 documents
Term 'bang' appears in 8 documents
Retrieved statistics for 9 terms
Retrieving document index for 9 terms...
Term 'a' appears in 978 documents
Term 'brief' appears in 27 documents
Term 'history' appears in 190 documents
Term 'of' appears in 901 documents
Term 'time' appears in 311 documents
Term 'from' appears in 600 documents
Term 'the' appears in 938 documents
Term 'big' appears in 66 documents
Term 'bang' appears in 8 documents
Retrieved index for 9 terms
Calculating BM25 scores...

Top 10 relevant documents:
------------------------
1. Document ID: 67227, Title: A_Brief_History_of_Time, Score: 214.4838
2. Document ID: 23619277, Title: A_Brand_You_Can_Trust, Score: 163.4663
3. Document ID: 8425345, Title: A_Farewell_to_Kings_(song), Score: 135.8345
4. Document ID: 2553430, Title: A_Day's_Pleasure, Score: 133.8875
5. Document ID: 9847946, Title: A_Hard_Day's_Night_(Grey's_Anatomy), Score: 90.5632
6. Document ID: 60121915, Title: A_Brief_History_of_Everyone_Who_Ever_Lived, Score: 80.8260
7. Document ID: 143425, Title: A_Day_in_the_Life, Score: 68.5948
8. Document ID: 3163760, Title: A_Boy_Named_Charlie_Brown, Score: 67.7206
9. Document ID: 23195496, Title: A_Brief_History_of_the_Twentieth_Century, Score: 62.5368
10. Document ID: 4937072, Title: A_Guide_to_the_Star_Wars_Universe, Score: 50.2849
25/04/15 18:18:45 INFO ShutdownHookManager: Shutdown hook called
25/04/15 18:18:45 INFO ShutdownHookManager: Deleting directory /tmp/spark-57c55d8c-195c-41f7-8827-a1bbb0b23307
This script will search for documents given the query using Spark RDD
25/04/15 18:18:48 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Query: Each of the guardians had their own view of the ideal
Query terms: ['each', 'of', 'the', 'guardians', 'had', 'their', 'own', 'view', 'of', 'the', 'ideal']
Connecting to Cassandra...
```

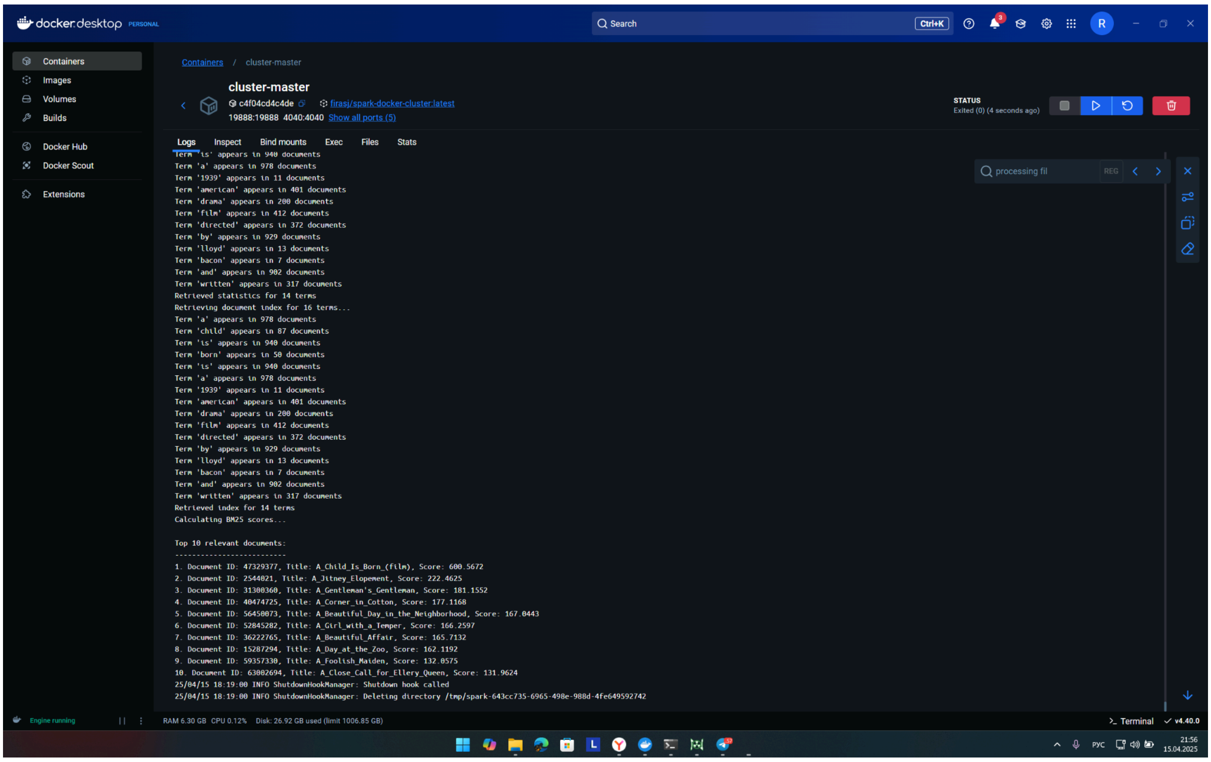
Second query from random book "Each of the guardians had their own view of the ideal", the same, stats are calculated:



```
Term 'each' appears in 138 documents
Term 'of' appears in 901 documents
Term 'the' appears in 938 documents
Term 'guardians' appears in 4 documents
Term 'had' appears in 318 documents
Term 'their' appears in 367 documents
Term 'own' appears in 177 documents
Term 'view' appears in 53 documents
Term 'of' appears in 901 documents
Term 'the' appears in 938 documents
Term 'ideal' appears in 10 documents
Retrieved statistics for 9 terms
Retrieving document index for 11 terms...
Term 'each' appears in 138 documents
Term 'of' appears in 901 documents
Term 'the' appears in 938 documents
Term 'guardians' appears in 4 documents
Term 'had' appears in 318 documents
Term 'their' appears in 367 documents
Term 'own' appears in 177 documents
Term 'view' appears in 53 documents
Term 'of' appears in 901 documents
Term 'the' appears in 938 documents
Term 'ideal' appears in 10 documents
Retrieved index for 9 terms
Calculating BM25 scores...

Top 10 relevant documents:
-------------------------
1. Document ID: 7690541, Title: A_Bold_Stroke_for_a_Wife, Score: 540.0931
2. Document ID: 35203893, Title: A_Joyful_Noise_(Gossip_album), Score: 214.7613
3. Document ID: 4066026, Title: A_Bug's_Land, Score: 165.2601
4. Document ID: 52154925, Title: A_Common_Faith, Score: 134.7895
5. Document ID: 72032959, Title: A_Ballet_of_Lepers, Score: 112.8774
6. Document ID: 39388324, Title: A_Field_in_England, Score: 106.9571
7. Document ID: 64680311, Title: A_European_Requiem, Score: 103.3045
8. Document ID: 53064903, Title: A_Linguistic_Atlas_of_Early_Middle_English, Score: 82.6727
9. Document ID: 1719074, Title: A_Case_of_Conscience, Score: 71.4424
10. Document ID: 12937367, Title: A_Gesture_Life, Score: 56.4248
25/04/15 18:18:52 INFO ShutdownHookManager: Shutdown hook called
25/04/15 18:18:52 INFO ShutdownHookManager: Deleting directory /tmp/spark-4336957d-b36f-4cad-8744-9ba8d0f2346a
This script will search for documents given the query using Spark RDD
25/04/15 18:18:56 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Query: A Child Is Born is a 1939 American drama film directed by Lloyd Bacon and written
Query terms: ['a', 'child', 'is', 'born', 'is', 'a', '1939', 'american', 'drama', 'film', 'directed', 'by', 'lloyd', 'bacon', 'and', 'written']
Connecting to Cassandra...
Connected to Cassandra
```

Third query "A Child Is Born is a 1939 American drama film directed by Lloyd Bacon and written" from the book named the smae “"A Child Is Born…”, and its being ranked 1st.



```
Term 'is' appears in 940 documents
Term 'a' appears in 978 documents
Term '1939' appears in 11 documents
Term 'american' appears in 401 documents
Term 'drama' appears in 200 documents
Term 'film' appears in 412 documents
Term 'directed' appears in 372 documents
Term 'by' appears in 929 documents
Term 'lloyd' appears in 13 documents
Term 'bacon' appears in 7 documents
Term 'and' appears in 902 documents
Term 'written' appears in 317 documents
Retrieved statistics for 14 terms
Retrieving document index for 16 terms...
Term 'a' appears in 978 documents
Term 'child' appears in 87 documents
Term 'is' appears in 940 documents
Term 'born' appears in 50 documents
Term 'is' appears in 940 documents
Term 'a' appears in 978 documents
Term '1939' appears in 11 documents
Term 'american' appears in 401 documents
Term 'drama' appears in 200 documents
Term 'film' appears in 412 documents
Term 'directed' appears in 372 documents
Term 'by' appears in 929 documents
Term 'lloyd' appears in 13 documents
Term 'bacon' appears in 7 documents
Term 'and' appears in 902 documents
Term 'written' appears in 317 documents
Retrieved index for 14 terms
Calculating BM25 scores...

Top 10 relevant documents:
-------------------------
1. Document ID: 47329377, Title: A_Child_Is_Born_(film), Score: 600.5672
2. Document ID: 2544021, Title: A_Jitney_Elopement, Score: 222.4625
3. Document ID: 31300360, Title: A_Gentleman's_Gentleman, Score: 181.1552
4. Document ID: 40474725, Title: A_Corner_in_Cotton, Score: 177.1168
5. Document ID: 56450073, Title: A_Beautiful_Day_in_the_Neighborhood, Score: 167.0443
6. Document ID: 52845282, Title: A_Girl_with_a_Temper, Score: 166.2597
7. Document ID: 36222765, Title: A_Beautiful_Affair, Score: 165.7132
8. Document ID: 15287294, Title: A_Day_at_the_Zoo, Score: 162.1192
9. Document ID: 59357330, Title: A_Foolish_Maiden, Score: 132.0575
10. Document ID: 63002694, Title: A_Close_Call_for_Ellery_Queen, Score: 131.9624
25/04/15 18:19:00 INFO ShutdownHookManager: Shutdown hook called
25/04/15 18:19:00 INFO ShutdownHookManager: Deleting directory /tmp/spark-643cc735-6965-498e-988d-4fe649592742
```