

Ivan Belstov i.beltsov@innopolis.university

Methodology

Data Preparation

There are 3 files responsible for data preparation: `prepare_data.sh`, `prepare_data.py` and `prepare_index_data.py`

`prepare_data.sh` transfers a.parquet from local to hdfs, starts `prepare_data.py`, copies newly made data folder to hdfs and starts `prepare_index_data.py`

`prepare_data.py` is unchanged from initial file

`prepare_index_data.py` reads files from hdfs data folder and creates a new partition using rdd

Indexer tasks

`index.sh` – Pipeline Runner Script that orchestrates the full indexing pipeline.

Actions:

- Accepts input path (local or HDFS) and uploads local files if needed
- Runs first MapReduce job to compute raw term frequencies
- Runs second MapReduce job to generate document stats and build the inverted index
- Triggers the Python script (`app.py`) to load results into Cassandra

MapReduce Stage 1

`mapper1.py` – Tokenizes document text, emits one entry per word occurrence.

Takes tab-separated lines with `doc_id`, `title`, `text` as input.

Returns lines with `doc_id`, `term`, `1`, and `title`.

reducer1.py – Sums word counts per document to get term frequency.

Takes sorted output from mapper1.py.

Returns aggregated doc_id, term, tf, title.

MapReduce Stage 2

mapper2.py – Computes total document length and prepares structured input for indexing.

Takes output from reducer1 and returns:

DOCLEN lines with total doc length and title.

TERM lines with individual term frequencies per document.

reducer2.py – Groups all documents per term, computes document frequency (df) per term and finalizes all index entries.

Takes output from mapper2.

Output:

VOCAB entries: term → document frequency (df).

INDEX entries: term → document ID, tf.

DOCLEN entries: document ID → length, title.

app.py – Loads structured index output into Cassandra tables

Functions:

fetch_hdfs_output(): Reads reducer2 output from HDFS.

connect_cassandra(): Connects to Cassandra and ensures keyspace exists.

ensure_tables(): Creates tables for vocabulary, inverted_index, and documents.

parse_and_insert(): Parses output and performs batch inserts into Cassandra.

Stored Tables:

vocabulary(term, df)

inverted_index(term, doc_id, tf)

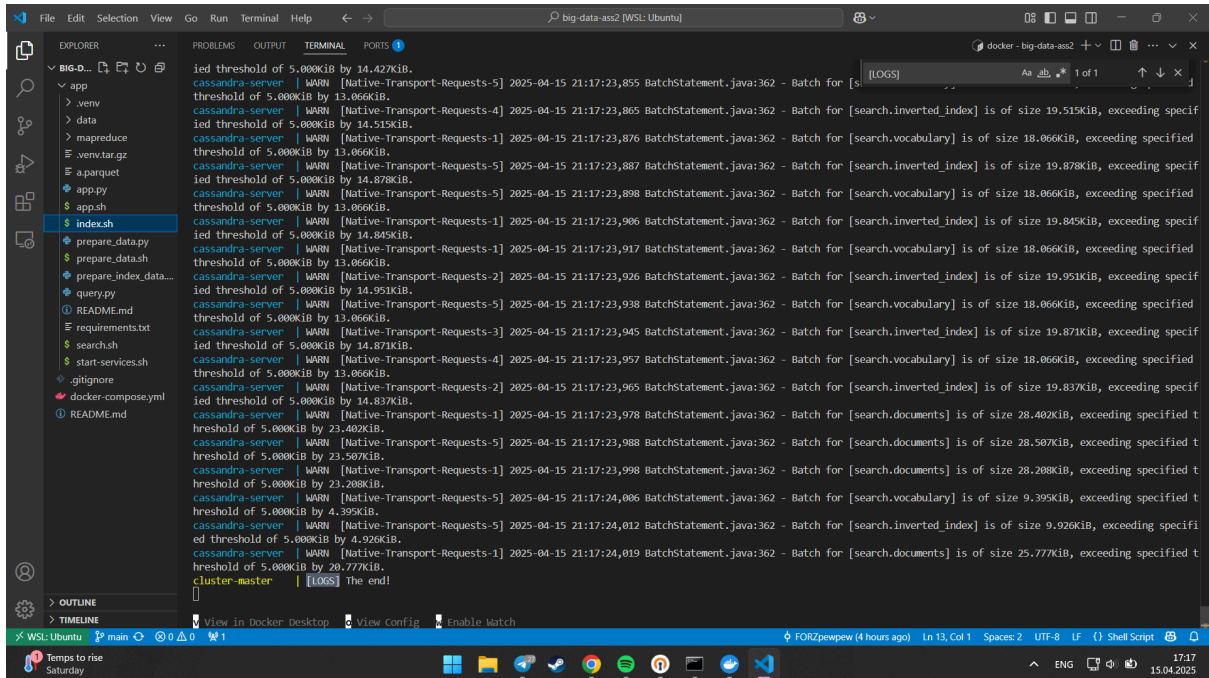
documents(doc_id, length, title)

Search tasks

For some reason docker crashes during search query (though it shouldn't), so i commented out search.sh. Basically the whole task is missing :(

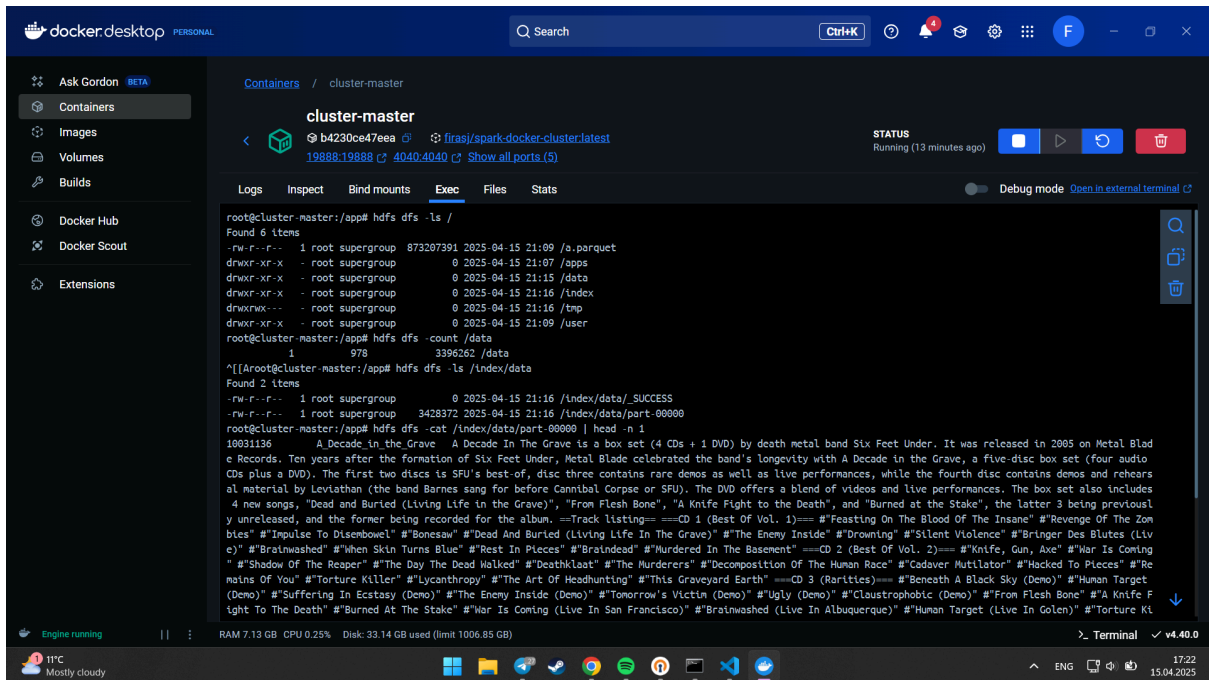
Demonstration

Warn messages from cassandra due to batch size is higher than threshold by a couple of KiB(but files are still added to cassandra)



```
ied threshold of 5.000KiB by 14.427KiB.
cassandra-server | WARN | [Native-Transport-Requests-5] 2025-04-15 21:17:23,855 BatchStatement.java:362 - Batch for [
threshold of 5.000KiB by 13.066KiB.
cassandra-server | WARN | [Native-Transport-Requests-4] 2025-04-15 21:17:23,865 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 19.515KiB, exceeding specif
ied threshold of 5.000KiB by 14.515KiB.
cassandra-server | WARN | [Native-Transport-Requests-1] 2025-04-15 21:17:23,876 BatchStatement.java:362 - Batch for [search.vocabulary] is of size 18.066KiB, exceeding specified
threshold of 5.000KiB by 13.066KiB.
cassandra-server | WARN | [Native-Transport-Requests-5] 2025-04-15 21:17:23,887 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 19.878KiB, exceeding specif
ied threshold of 5.000KiB by 14.878KiB.
cassandra-server | WARN | [Native-Transport-Requests-5] 2025-04-15 21:17:23,898 BatchStatement.java:362 - Batch for [search.vocabulary] is of size 18.066KiB, exceeding specified
threshold of 5.000KiB by 13.066KiB.
cassandra-server | WARN | [Native-Transport-Requests-1] 2025-04-15 21:17:23,906 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 19.845KiB, exceeding specif
ied threshold of 5.000KiB by 14.845KiB.
cassandra-server | WARN | [Native-Transport-Requests-1] 2025-04-15 21:17:23,917 BatchStatement.java:362 - Batch for [search.vocabulary] is of size 18.066KiB, exceeding specified
threshold of 5.000KiB by 13.066KiB.
cassandra-server | WARN | [Native-Transport-Requests-2] 2025-04-15 21:17:23,926 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 19.951KiB, exceeding specif
ied threshold of 5.000KiB by 14.951KiB.
cassandra-server | WARN | [Native-Transport-Requests-5] 2025-04-15 21:17:23,938 BatchStatement.java:362 - Batch for [search.vocabulary] is of size 18.066KiB, exceeding specified
threshold of 5.000KiB by 13.066KiB.
cassandra-server | WARN | [Native-Transport-Requests-3] 2025-04-15 21:17:23,945 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 19.871KiB, exceeding specif
ied threshold of 5.000KiB by 14.871KiB.
cassandra-server | WARN | [Native-Transport-Requests-4] 2025-04-15 21:17:23,957 BatchStatement.java:362 - Batch for [search.vocabulary] is of size 18.066KiB, exceeding specified
threshold of 5.000KiB by 13.066KiB.
cassandra-server | WARN | [Native-Transport-Requests-2] 2025-04-15 21:17:23,965 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 19.837KiB, exceeding specif
ied threshold of 5.000KiB by 14.837KiB.
cassandra-server | WARN | [Native-Transport-Requests-1] 2025-04-15 21:17:23,978 BatchStatement.java:362 - Batch for [search.documents] is of size 28.402KiB, exceeding specified t
hreshold of 5.000KiB by 23.402KiB.
cassandra-server | WARN | [Native-Transport-Requests-5] 2025-04-15 21:17:23,988 BatchStatement.java:362 - Batch for [search.documents] is of size 28.507KiB, exceeding specified t
hreshold of 5.000KiB by 23.507KiB.
cassandra-server | WARN | [Native-Transport-Requests-1] 2025-04-15 21:17:23,998 BatchStatement.java:362 - Batch for [search.documents] is of size 28.208KiB, exceeding specified t
hreshold of 5.000KiB by 23.208KiB.
cassandra-server | WARN | [Native-Transport-Requests-5] 2025-04-15 21:17:24,006 BatchStatement.java:362 - Batch for [search.vocabulary] is of size 9.399KiB, exceeding specified t
hreshold of 5.000KiB by 4.399KiB.
cassandra-server | WARN | [Native-Transport-Requests-5] 2025-04-15 21:17:24,012 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 9.926KiB, exceeding specifi
ed threshold of 5.000KiB by 4.926KiB.
cassandra-server | WARN | [Native-Transport-Requests-1] 2025-04-15 21:17:24,019 BatchStatement.java:362 - Batch for [search.documents] is of size 25.777KiB, exceeding specified t
hreshold of 5.000KiB by 20.777KiB.
cluster-master | [LOGS] The end!
```

This screenshot shows contents of hdfs / folder, number of documents in /data folder which is 978, contents of /index/data folder and the first line of /index/data/part-00000



```
root@cluster-master:/app# df -ls /
Found 6 items
-rw-r--r-- 1 root supergroup 873297391 2025-04-15 21:09 /a.parquet
drwxr-xr-x 1 root supergroup 0 2025-04-15 21:07 /apps
drwxr-xr-x 1 root supergroup 0 2025-04-15 21:15 /data
drwxr-xr-x 1 root supergroup 0 2025-04-15 21:16 /index
drwxr-xr-x 1 root supergroup 0 2025-04-15 21:16 /tmp
drwxr-xr-x 1 root supergroup 0 2025-04-15 21:09 /user
root@cluster-master:/app# df -ls -count /data
1 978
3396262 /data
^[[Aroot@cluster-master:/app# df -ls /index/data
Found 2 items
-rw-r--r-- 1 root supergroup 0 2025-04-15 21:16 /index/data/_SUCCESS
-rw-r--r-- 1 root supergroup 3428372 2025-04-15 21:16 /index/data/part-00000
root@cluster-master:/app# df -ls -cat /index/data/part-00000 | head -n 1
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 Zom bies" #"Impulse To Disembowel" #"Bonesaw" #"Dead And Buried (Living Life In The Grave)" #"The Enemy Inside" #"Drowning" #"Silent Violence" #"Bringer Des Blues (Liv e)" #"Brainwashed" #"When Skin Turns Blue" #"Rest In Pieces" #"Branded" #"Murdered In The Basement" ==CD 2 (Best Of Vol. 2)== #"Knife, Gun, Axe" #"War Is Coming #"Shadow Of The Reaper" #"The Day The Dead Walked" #"Deathklast" #"The Murderers" #"Decomposition Of The Human Race" #"Cadaver Mutilator" #"Hacked To Pieces" #"Re mains Of You" #"Torture Killer" #"Lycanthropy" #"The Art Of Headhunting" #"This Graveyard Earth" ==CD 3 (Rarities)== #"Beneath A Black Sky (Demo)" #"Human Target (Demo)" #"Suffering In Ecstasy (Demo)" #"The Enemy Inside (Demo)" #"Tomorrow's Victim (Demo)" #"Ugly (Demo)" #"Claustrophobic (Demo)" #"From Flesh Bone" #"A Knife F ight To The Death" #"Burned At The Stake" #"War Is Coming (Live In San Francisco)" #"Brainwashed (Live In Albuquerque)" #"Human Target (Live In Golen)" #"Torture Ki
```

Next screenshots show cassandra db tables, their contents and count of rows (global count is expensive without partitions, but there's no workaround :c)

```
pepew@pepew: /mnt/c/Users/gachi
cqlsh:search> describe tables;

documents inverted_index vocabulary

cqlsh:search> select count(*) from documents;

count
-----
978

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:search> select * from documents limit 5;

doc_id | length | title
-----+-----+-----
10230685 | 115 | A_Dead_Sinking_Story
27568194 | 567 | A_Hero_Ain't_Nothin'_but_a_Sandwich_(film)
39710446 | 330 | A_Little_Bit_of_Luck
38294693 | 338 | A_Change_Is_Gonna_Come_(Jack_McDuff_album)
51794980 | 306 | A_Family_Secret_(Upstairs,_Downstairs)

(5 rows)

cqlsh:search> select count(*) from inverted_index;

count
-----
244375

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:search> _
```

```
pepew@pepew: /mnt/c/Users/gachi
(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:search> select * from inverted_index limit 5;

term | doc_id | tf
-----+-----+-----
dobson | 13633480 | 1
bessus | 12000397 | 3
ix | 19789501 | 1
ix | 32497421 | 1
ix | 67078438 | 2

(5 rows)

cqlsh:search> select count(*) from vocabulary;

count
-----
39630

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:search> select * from vocabulary limit 5;

term | df
-----+-----
dobson | 1
bessus | 1
ix | 4
await | 2
libertad | 1

(5 rows)

cqlsh:search> _
```