FA 25 6513-C BIG DATA FINAL PRESENTATION

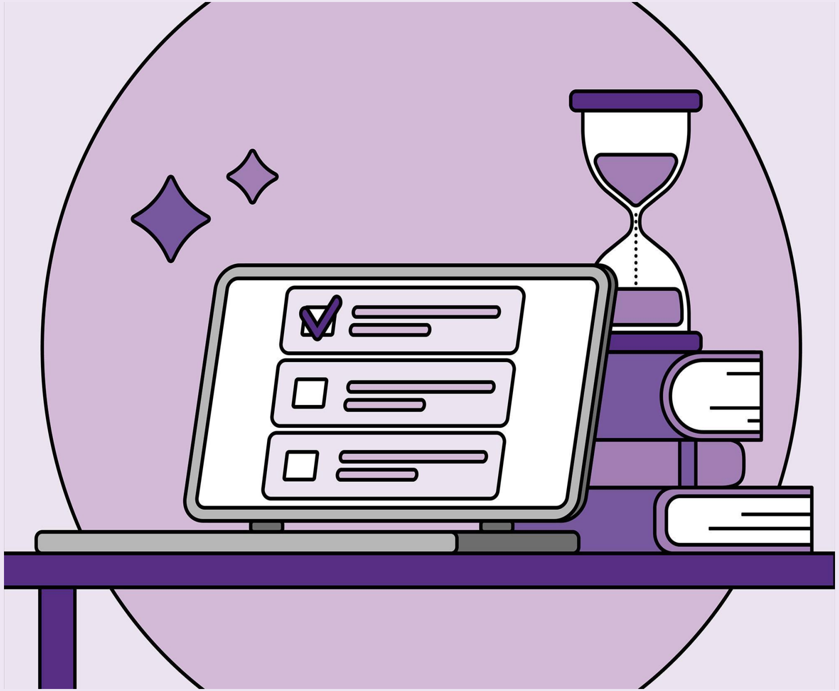
NYU

LLM DataPrep: Scalable Pipeline for Biomedical Fine-Tuning

Professor: Amit Patel

PROBLEM STATEMENT:

- Biomedical text (PubMed: **2.2M entries**) is **noisy, duplicated, and unstructured**
- Raw data **cannot** be used directly for LLM fine-tuning
- Need a **scalable pipeline** to transform big biomedical datasets into high-quality training data



NYU

DATA OVERVIEW

- Data Source Name: PubMed
- **Data Source Link:** https:/ /huggingface.co/datasets/MedRAG/pubmed
- Dataset File Size: 2.78 GB (Parquet) & ~70GB(JSONL)
- Approximate Number of Records: 2.21 Million
- **Columns Include:** id, title, content, contents, PMID

WHAT TECH STACK WE USED

APACHE Spark™

Spark SQL

Spark MLlib

hadoop

Parquet

plotly

HIVE

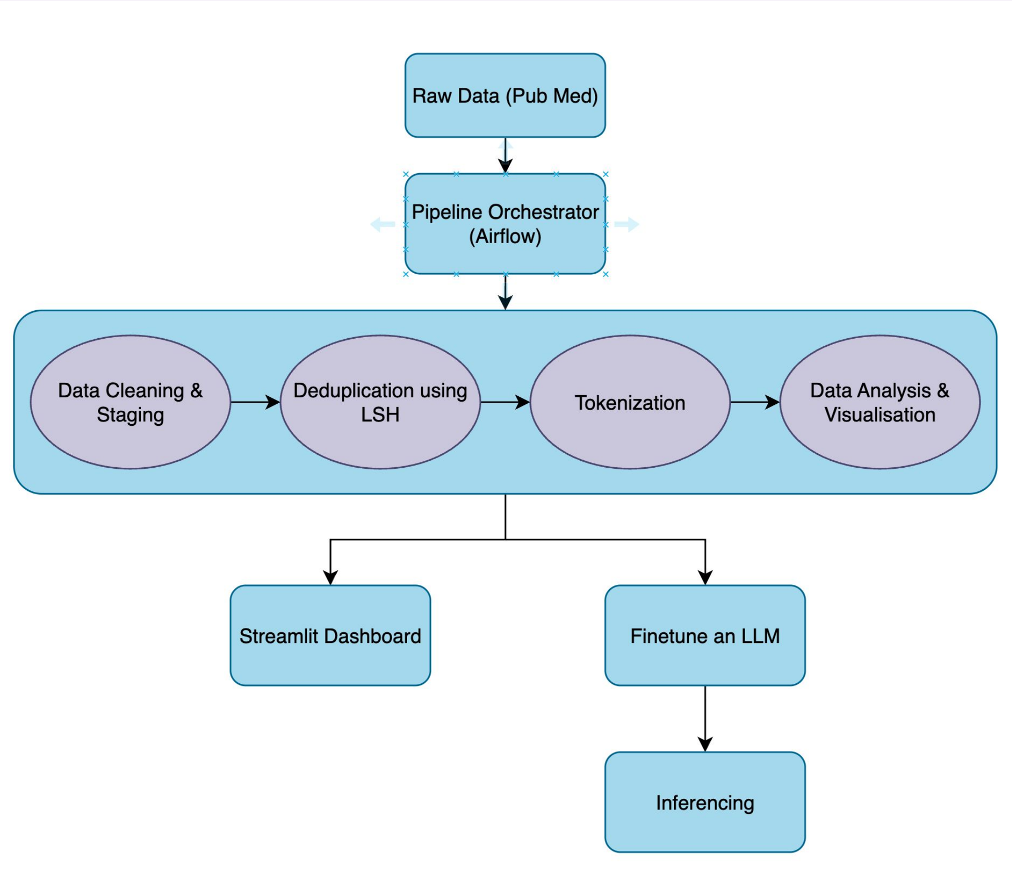
Streamlit

Apache Airflow

PROPOSED SOLUTION

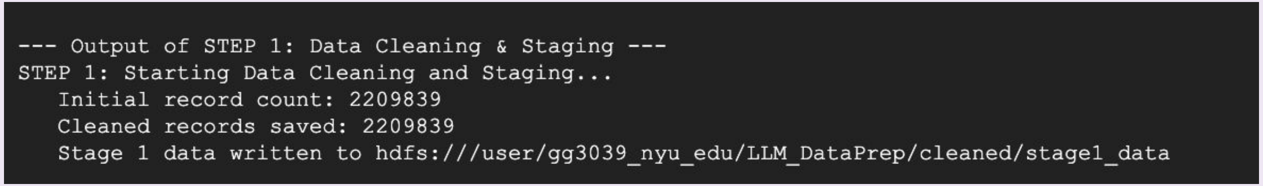
- Build a scalable end-to-end ETL pipeline to process millions of biomedical text records efficiently.
- Automatically clean, normalize, deduplicate, and tokenize PubMed data to produce high-quality training inputs.
- Orchestrate and automate all pipeline stages using Apache Airflow for reliability and repeatability.
- Generate a structured, LLM-ready dataset for domain-specific fine-tuning.

High Level Pipeline Architecture



NYU

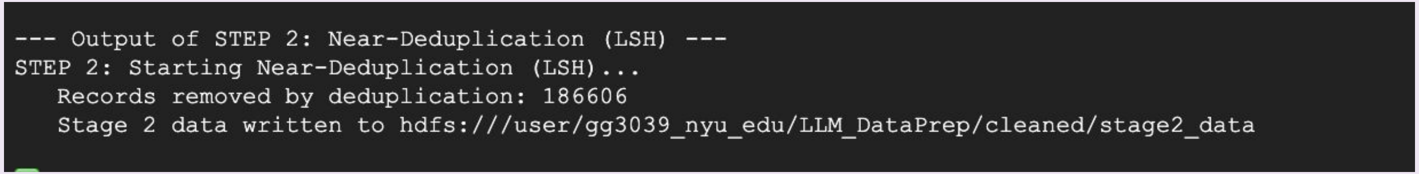
Step 1 - Data Cleaning



```
--- Output of STEP 1: Data Cleaning & Staging ---
STEP 1: Starting Data Cleaning and Staging...
   Initial record count: 2209839
   Cleaned records saved: 2209839
   Stage 1 data written to hdfs:///user/gg3039_nyu_edu/LLM_DataPrep/cleaned/stage1_data
```

- **Loads the raw PubMed text and builds a clean main_text field by merging and normalizing the content.**
- **Removes noise such as citations, extra spaces, and short/invalid entries.**
- **Saves a cleaned, standardized dataset to HDFS for the next stage.**

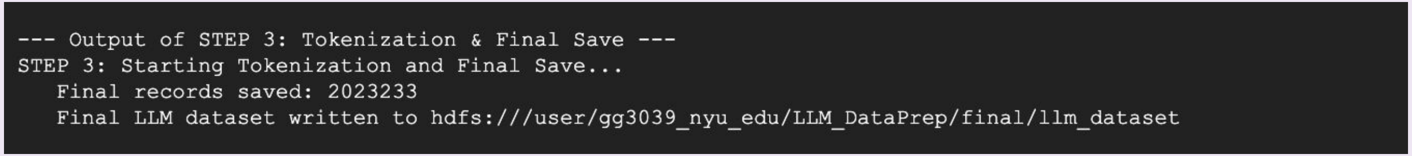
Step 2: Deduplication Using MinHash LSH



```
--- Output of STEP 2: Near-Deduplication (LSH) ---
STEP 2: Starting Near-Deduplication (LSH)...
   Records removed by deduplication: 186606
   Stage 2 data written to hdfs:///user/gg3039_nyu_edu/LLM_DataPrep/cleaned/stage2_data
```

- **Tokenizes each document and converts it into a hashed feature vector using HashingTF, enabling efficient similarity checks.**
- **Uses MinHash LSH to group together texts that are highly similar and flags them as potential duplicates.**
- **Removes these near-duplicate records and saves a cleaned, deduplicated dataset to HDFS while reporting how many were dropped.**

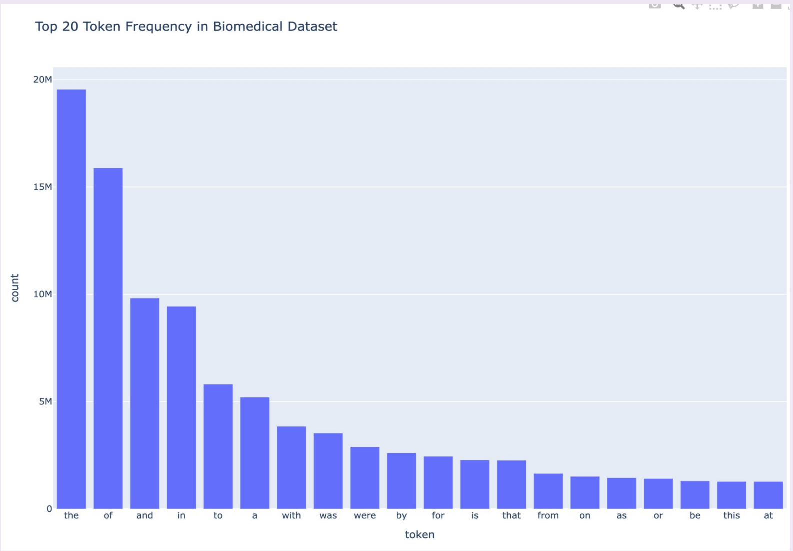
Step 3: Distributed Tokenization



```
--- Output of STEP 3: Tokenization & Final Save ---
STEP 3: Starting Tokenization and Final Save...
   Final records saved: 2023233
   Final LLM dataset written to hdfs:///user/gg3039_nyu_edu/LLM_DataPrep/final/llm_dataset
```

- **Converts main_text and title into numeric token sequences that LLMs use for learning**
- **Standardizes the dataset to keep only the fields required for downstream LLM training: id, PMID, title, main_text, tokens**
- **Processes tokenization in parallel across the Spark cluster, enabling millions of records to be handled efficiently**
- **Writes the final structured dataset as Parquet to HDFS for fast retrieval during model fine-tuning**
- **Produces a terminal summary with the total number of tokenized records**

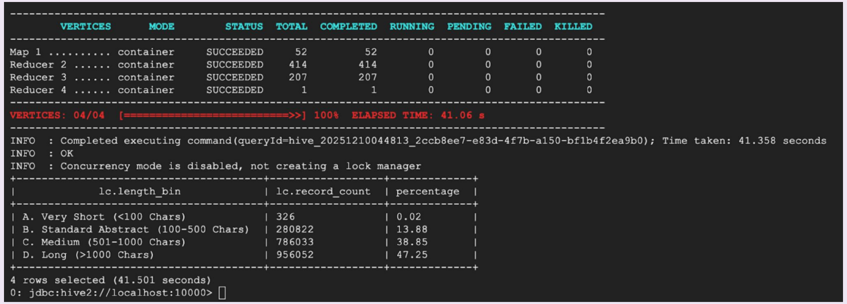
Step 4: Data Analysis & Visualization



- Analyzes the cleaned and tokenized dataset to compute basic statistics like total records and average text length.
- Identifies the most common tokens in the corpus to understand vocabulary patterns.
- Generates an interactive Plotly chart and saves it for visualization and reporting.

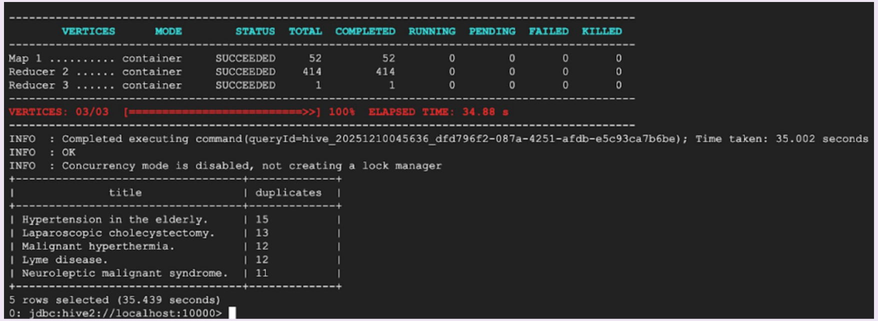
NYU

Data Analysis - Hive Queries



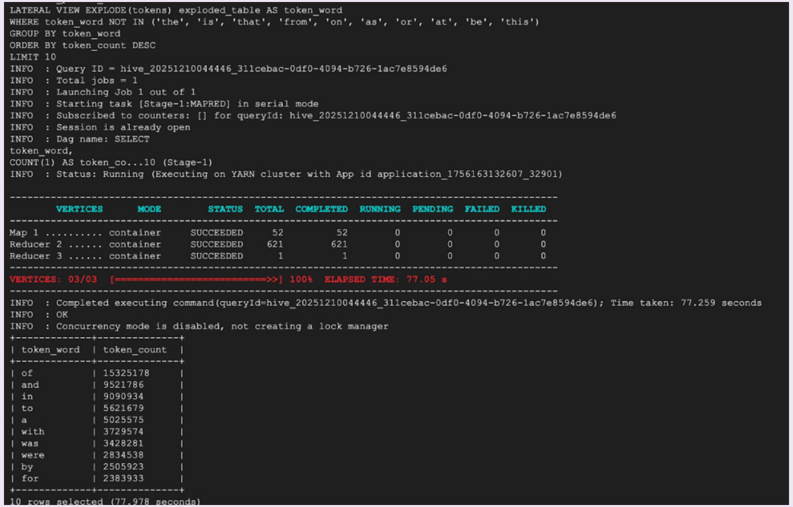
```
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     52         52        0        0       0       0
Reducer 2 ...... container     SUCCEEDED    414        414        0        0       0       0
Reducer 3 ...... container     SUCCEEDED    207        207        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 41.06 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20251210044813_2ccb8ee7-e83d-4f7b-a150-bf1b4f2ea9b0); Time taken: 41.358 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+----------------------------------------+-------------------+-------------+
|             lc.length_bin              | lc.record_count   | percentage  |
+----------------------------------------+-------------------+-------------+
| A. Very Short (<100 Chars)             | 326               | 0.02        |
| B. Standard Abstract (100-500 Chars)   | 280822            | 13.88       |
| C. Medium (501-1000 Chars)             | 786033            | 38.85       |
| D. Long (>1000 Chars)                  | 956052            | 47.25       |
+----------------------------------------+-------------------+-------------+
4 rows selected (41.501 seconds)
0: jdbc:hive2://localhost:10000>
```

Distribution of text lengths



```
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     52         52        0        0       0       0
Reducer 2 ...... container     SUCCEEDED    414        414        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 34.88 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20251210045636_dfd796f2-087a-4251-afdb-e5c93ca7b6be); Time taken: 35.002 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------------------------------------+-------------+
|               title                | duplicates  |
+------------------------------------+-------------+
| Hypertension in the elderly.       | 15          |
| Laparoscopic cholecystectomy.      | 13          |
| Malignant hyperthermia.            | 12          |
| Lyme disease.                      | 12          |
| Neuroleptic malignant syndrome.    | 11          |
+------------------------------------+-------------+
5 rows selected (35.439 seconds)
0: jdbc:hive2://localhost:10000>
```

Top 5 Most repeated titles



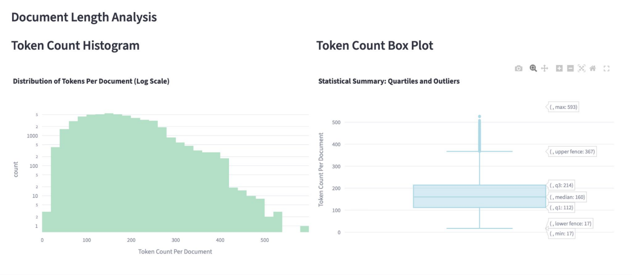
```
LATERAL VIEW EXPLODE(tokens) exploded_table AS token_word
WHERE token_word NOT IN ('the', 'is', 'that', 'from', 'on', 'as', 'or', 'at', 'be', 'this')
GROUP BY token_word
ORDER BY token_count DESC
LIMIT 10
INFO  : Query ID = hive_20251210044446_311cebac-0df0-4094-b726-1ac7e8594de6
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20251210044446_311cebac-0df0-4094-b726-1ac7e8594de6
INFO  : Session is already open
INFO  : Dag name: SELECT
token_word,
COUNT(1) AS token_co...10 (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1756163132607_32901)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     52         52        0        0       0       0
Reducer 2 ...... container     SUCCEEDED    621        621        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 77.05 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20251210044446_311cebac-0df0-4094-b726-1ac7e8594de6); Time taken: 77.259 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------+--------------+
| token_word  | token_count  |
+-------------+--------------+
| of          | 15325178     |
| and         | 9521786      |
| in          | 9090934      |
| to          | 5621679      |
| a           | 5025575      |
| with        | 3729574      |
| was         | 3428281      |
| were        | 2834538      |
| by          | 2505923      |
| for         | 2383933      |
+-------------+--------------+
10 rows selected (77.978 seconds)
```

Most Common Non-stop word tokens

NYU

All the HIVE analysis Screenshots

Visualization - Streamlit



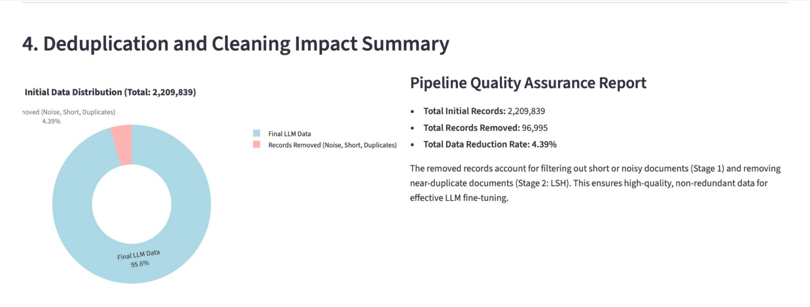
Document Length Analysis

Token Count Histogram

Token Count Box Plot

Distribution of Tokens Per Document (Log Scale)

Statistical Summary: Quartiles and Outliers



4. Deduplication and Cleaning Impact Summary

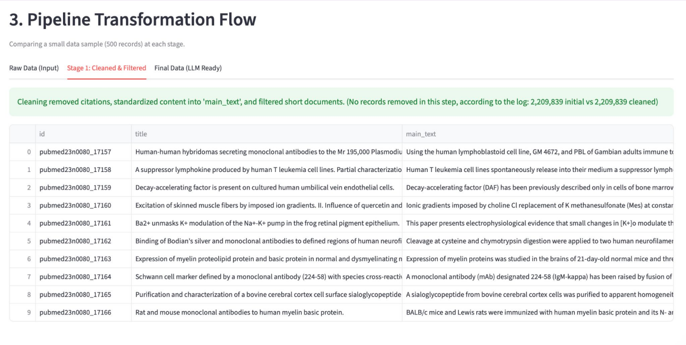
Initial Data Distribution (Total: 2,209,839)

noved (Noise, Short, Duplicates)
4.39%
Final LLM Data
95.6%
Final LLM Data
Records Removed (Noise, Short, Duplicates)

Pipeline Quality Assurance Report

- **Total Initial Records:** 2,209,839
- **Total Records Removed:** 96,995
- **Total Data Reduction Rate: 4.39%**

The removed records account for filtering out short or noisy documents (Stage 1) and removing near-duplicate documents (Stage 2: LSH). This ensures high-quality, non-redundant data for effective LLM fine-tuning.

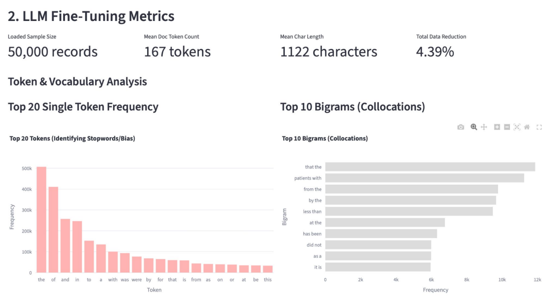


3. Pipeline Transformation Flow

Comparing a small data sample (500 records) at each stage.

Cleaning removed citations, standardized content into 'main_text', and filtered short documents. (No records removed in this step, according to the log: 2,209,839 initial vs 2,209,839 cleaned)

	id	title	main_text
0	pubmed23n0080_17157	Human-human hybridomas secreting monoclonal antibodies to the Mr 195,000 Plasmodiu	Using the human lymphoblastoid cell line, GM 4672, and PBL of Gambian adults immune tc
1	pubmed23n0080_17158	A suppressor lymphokine produced by human T leukemia cell lines. Partial characterizatio	Human T leukemia cell lines spontaneously release into their medium a suppressor lymph
2	pubmed23n0080_17159	Decay-accelerating factor is present on cultured human umbilical vein endothelial cells.	Decay-accelerating factor (DAF) has been previously described only in cells of bone marrov
3	pubmed23n0080_17160	Excitation of skinned muscle fibers by imposed ion gradients. II. Influence of quercetin and	Ionic gradients imposed by choline Cl replacement of K methanesulfonate (Mes) at constar
4	pubmed23n0080_17161	Ba2+ unmasks K+ modulation of the Na+-K+ pump in the frog retinal pigment epithelium.	This paper presents electrophysiological evidence that small changes in [K+]o modulate th
5	pubmed23n0080_17162	Binding of Bodian's silver and monoclonal antibodies to defined regions of human neurofi	Cleavage at cysteine and chymotrypsin digestion were applied to two human neurofilamer
6	pubmed23n0080_17163	Expression of myelin proteolipid protein and basic protein in normal and dysmyelinating n	Expression of myelin proteins was studied in the brains of 21-day-old normal mice and thre
7	pubmed23n0080_17164	Schwann cell marker defined by a monoclonal antibody (224-58) with species cross-reacti	A monoclonal antibody (mAb) designated 224-58 (IgM-kappa) has been raised by fusion of
8	pubmed23n0080_17165	Purification and characterization of a bovine cerebral cortex cell surface sialoglycopeptide	A sialoglycopeptide from bovine cerebral cortex cells was purified to apparent homogeneit
9	pubmed23n0080_17166	Rat and mouse monoclonal antibodies to human myelin basic protein.	BALB/c mice and Lewis rats were immunized with human myelin basic protein and its N- ar



2. LLM Fine-Tuning Metrics

Loaded Sample Size	Mean Doc Token Count	Mean Char Length	Total Data Reduction
50,000 records	167 tokens	1122 characters	4.39%

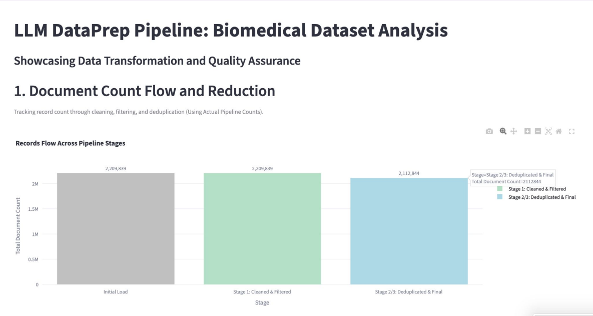
Token & Vocabulary Analysis

Top 20 Single Token Frequency

Top 10 Bigrams (Collocations)

Top 20 Tokens (Identifying Stopwords/Bias)

Top 10 Bigrams (Collocations)



LLM DataPrep Pipeline: Biomedical Dataset Analysis

Showcasing Data Transformation and Quality Assurance

1. Document Count Flow and Reduction

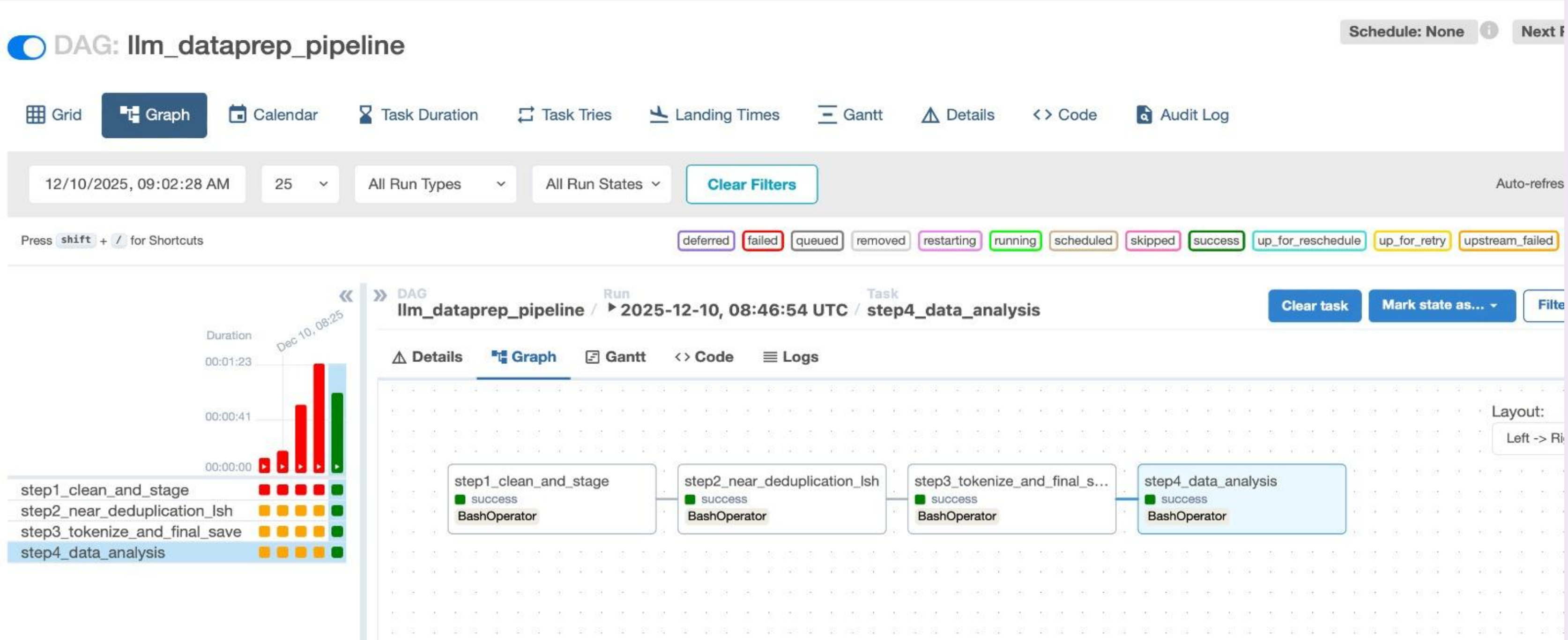
Tracking record count through cleaning, filtering, and deduplication (Using Actual Pipeline Counts).

Records Flow Across Pipeline Stages

Total Document Count
2M
1.5M
1M
0.5M
0
2,209,839
2,209,839
2,112,844
Initial Load
Stage 1: Cleaned & Filtered
Stage 2/3: Deduplicated & Final
Stage
Stage=Stage 2/3: Deduplicated & Final
Total Document Count=2112844
Stage 1: Cleaned & Filtered
Stage 2/3: Deduplicated & Final

NYU

Pipeline orchestration through Airflow



DAG: llm_dataprep_pipeline

Schedule: None

12/10/2025, 09:02:28 AM

25

All Run Types

All Run States

Press shift + / for Shortcuts

llm_dataprep_pipeline / 2025-12-10, 08:46:54 UTC / step4_data_analysis

Layout: Left -> Ri

NYU

Big Data Aspects We Focused On

- **Distributed processing:** Spark jobs across Dataproc cluster
- **Scalability:** Pipeline works on millions of biomedical records
- **Fault tolerance:** HDFS storage + Spark retry mechanisms
- **Optimized storage:** Parquet format for compression & speed
- **Parallel ETL stages:** Cleaning, dedupe, tokenization handled at scale
- **Workflow Orchestration at Scale (Airflow DAGs)** : Automated all pipeline stages using Apache Airflow, ensuring reproducibility, scheduling, and monitoring of large-scale ETL workflows.
- **Data Analysis:** Large-scale SQL analytics performed through Hive for dataset insights

Challenges faced

- Setting up Airflow in NYU Dataproc
- Unavailability of GPU resources
- Preparing LLM fine-tuning-ready format

Lessons Learned

- Preprocessing quality drives LLM quality
- Distributed processing is essential at this scale
- Modular ETL design reduces debugging time
- Parquet + Spark = fast and scalable

THANK YOU

Student Names and Net IDs:
Geetha Krishna Guruju (Net Id: gg3039)
Tejdeep Chippa (Net Id: tc4263)
Isha Kookanapalli Math (Net Id: ik2688)

ANY QUESTIONS