Sentiment Analysis of Amazon Product Reviews Using Machine Learning

23AID302-Big Data Analytics

Group No:B-1

S.Sandeep cb.ai.u4aid23140

G.Siva Sai kumar cb.ai.u4aid23156

D.Mahadev Naidu cb.ai.u4aid23111

K.Manoj kumar cb.ai.u4aid23161

Department: AIDS

Faculty In-charge: Dr. Sreeja

College, Academic Year: Amrita Vishwa Vidyapeetham, 2025

CONTENTS

INTRODUCTION

- Massive amounts of textual data are generated daily on platforms like Amazon.
- Handling and analyzing such large-scale data is difficult using traditional systems.
- Big Data technologies like Hadoop and Apache Spark enable distributed storage and parallel computation.
- The project demonstrates how Big Data frameworks can efficiently manage and analyze massive text datasets for real-world insights.

PROBLEM STATEMENT

- Manual review analysis is time-consuming, inefficient, and prone to bias.
- Customers often struggle to interpret the quality of products because of the overwhelming number of reviews.
- Companies need an automated, accurate, and scalable system to understand customer sentiments from vast review datasets.
- The absence of such systems can lead to:
  - Poor product improvement decisions.
  - Ineffective personalized recommendations.
  - Missed opportunities for market trend analysis.

OBJECTIVES

- To build an end-to-end Big Data pipeline for sentiment classification.
- To perform text preprocessing using Scala (Spark Core + Spark SQL).
- To utilize Hadoop HDFS for distributed data storage and retrieval.
- To train a Logistic Regression model using PySpark MLlib for binary sentiment analysis.
- To show the scalability, reliability, and efficiency of Big Data tools for large-scale text analytics.

DATASET OVERVIEW

- Contains 34,686,770 Amazon reviews spanning 18 years (up to March 2013).
- Reviews are from 6,643,669 users on 2,441,053 products.
- Polarity subset used for sentiment classification:
  - 1,800,000 training samples for each class (positive & negative).
  - 200,000 testing samples for each class.
  - Total: 4,000,000 samples (3.6M training + 0.4M testing).
- Collected by the Stanford Network Analysis Project (SNAP).

Size:-2.5 GB

https://www.kaggle.com/datasets/kritanjalijain/amazon-reviews?select=amazon_review_polarity_csv.tgz

Methodology

- **Step 1 Data Ingestion**
  - Load the Amazon reviews dataset from HDFS
  - Keep raw, intermediate, and processed datasets separate for reproducibility.

- **Step 2 — Data Preprocessing**
  - Remove or fill missing values in polarity, title, and text.
  - Combine textual fields (title + text) for a single column.
  - Clean text by converting to lowercase, removing punctuation, special characters, and URLs.

- **Step 3 — Data Loading**
  - Connected PySpark with Hadoop HDFS using SparkSession.
  - Loaded preprocessed CSV files for ML tasks.

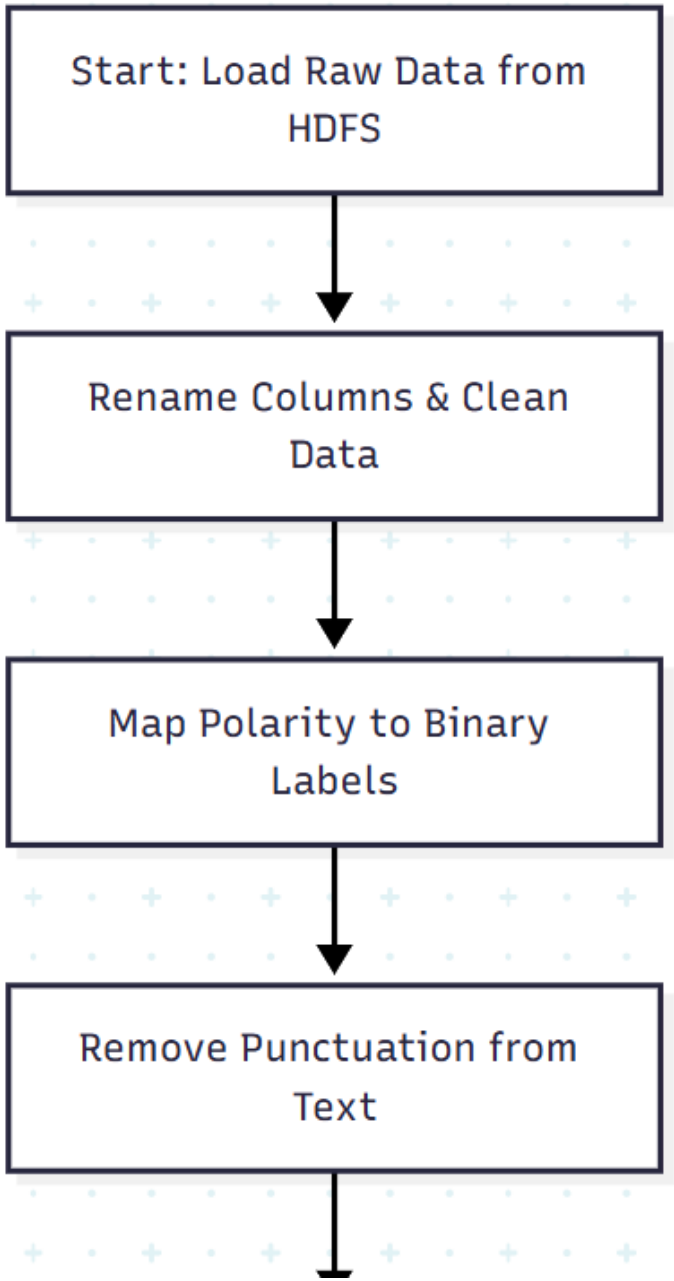
- **Step 4 — Feature Engineering**
  - Applied Tokenization, StopWords Removal, and TF-IDF Vectorization.
  - Converted text into numerical feature vectors suitable for machine learning.

- **Step 5 — Model Training**
  - Trained a Logistic Regression model using PySpark MLlib.
  - Saved the trained model back into HDFS.

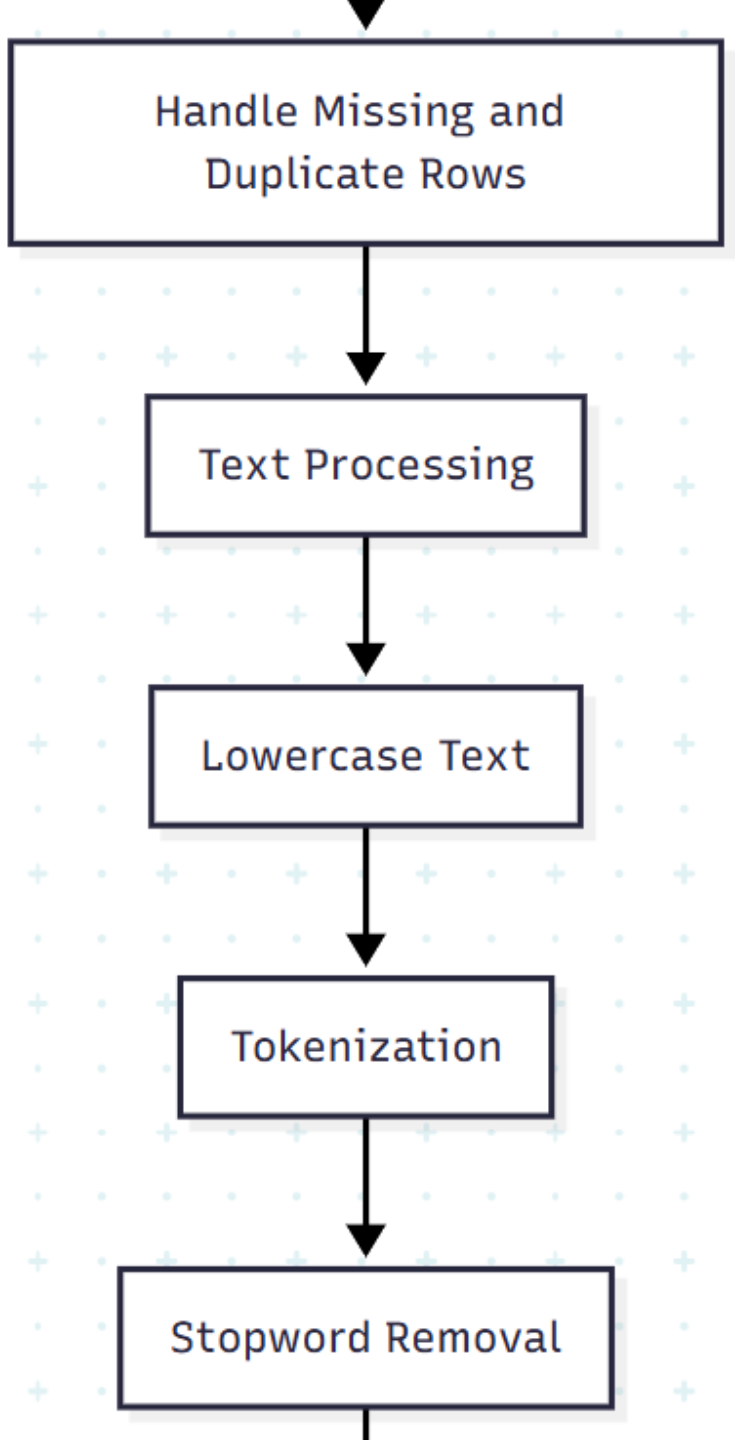
- **Step 6 — Model Evaluation**
  - Evaluated model on test data for **accuracy, precision, recall, F1, and AUC**.
  - Stored metrics as **JSON output** for reporting.

- **Step 7 — Visualization**
  - used **Spark UI (localhost:4040)** to monitor job stages, tasks, and performance metrics.

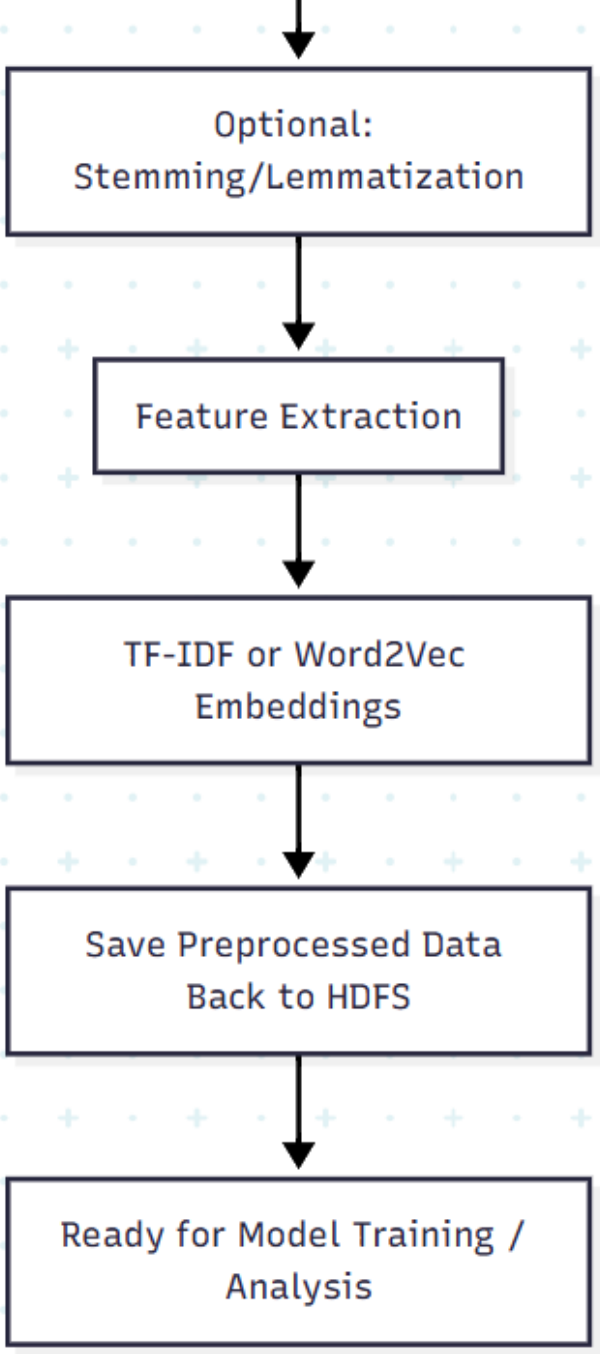
WORKFLOW


Start: Load Raw Data from HDFS
Rename Columns & Clean Data
Map Polarity to Binary Labels
Remove Punctuation from Text

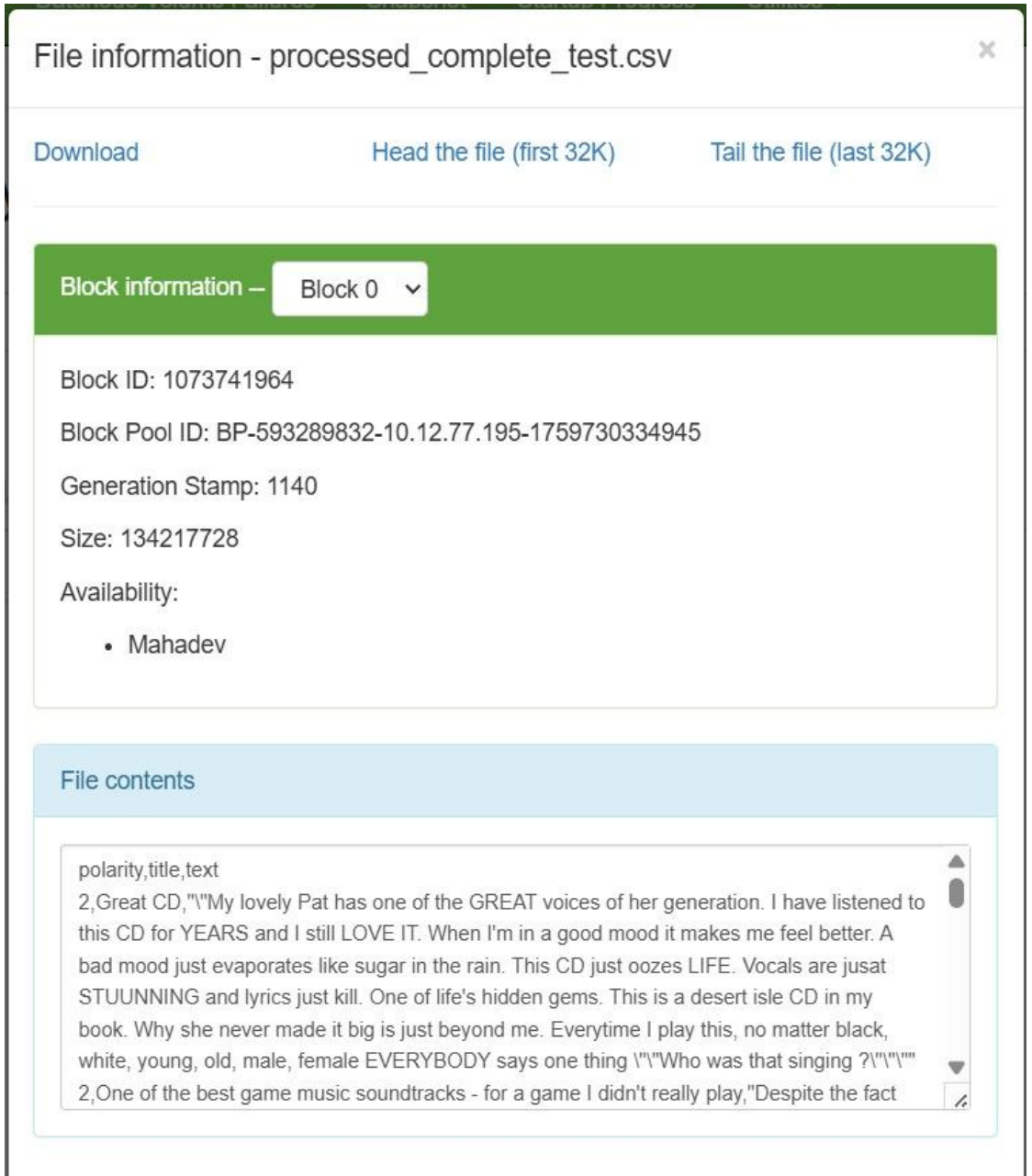
WORKFLOW


Handle Missing and Duplicate Rows
Text Processing
Lowercase Text
Tokenization
Stopword Removal

WORKFLOW


Optional:
Stemming/Lemmatization
Feature Extraction
TF-IDF or Word2Vec
Embeddings
Save Preprocessed Data
Back to HDFS
Ready for Model Training /
Analysis

RESULTS


File information - processed_complete_test.csv
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information –
Block 0
Block ID: 1073741964
Block Pool ID: BP-593289832-10.12.77.195-1759730334945
Generation Stamp: 1140
Size: 134217728
Availability:
Mahadev
File contents
polarity,title,text
2,Great CD,"\"My lovely Pat has one of the GREAT voices of her generation. I have listened to this CD for YEARS and I still LOVE IT. When I'm in a good mood it makes me feel better. A bad mood just evaporates like sugar in the rain. This CD just oozes LIFE. Vocals are jusat STUUNNING and lyrics just kill. One of life's hidden gems. This is a desert isle CD in my book. Why she never made it big is just beyond me. Everytime I play this, no matter black, white, young, old, male, female EVERYBODY says one thing \"\"Who was that singing ?\"\"\"\"
2,One of the best game music soundtracks - for a game I didn't really play,"Despite the fact

File information - processed_complete_train.csv

Download | Head the file (first 32K) | Tail the file (last 32K)

Block information – Block 0

Block ID: 1073741953

Block Pool ID: BP-593289832-10.12.77.195-1759730334945

Generation Stamp: 1129

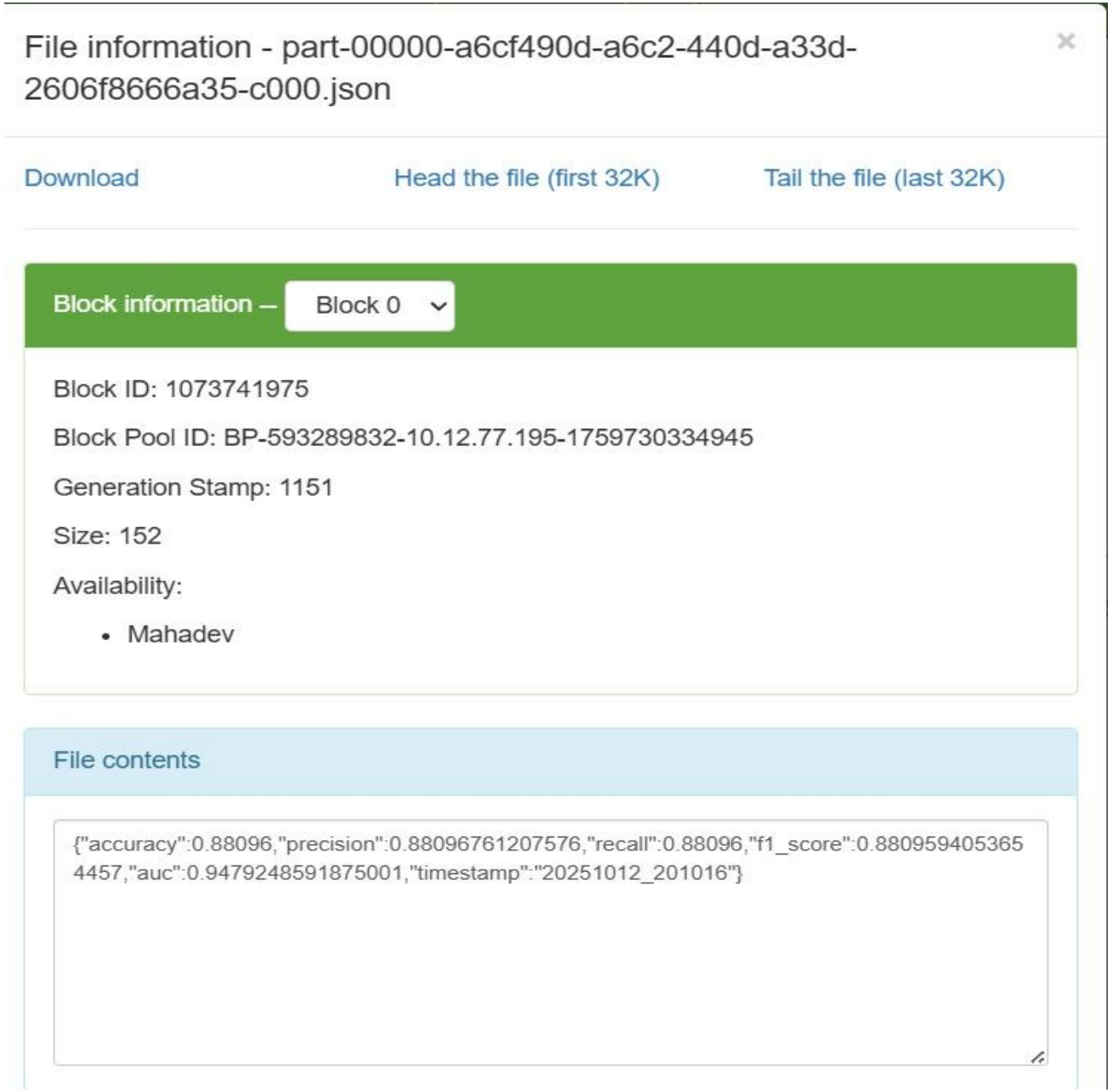
Size: 134217728

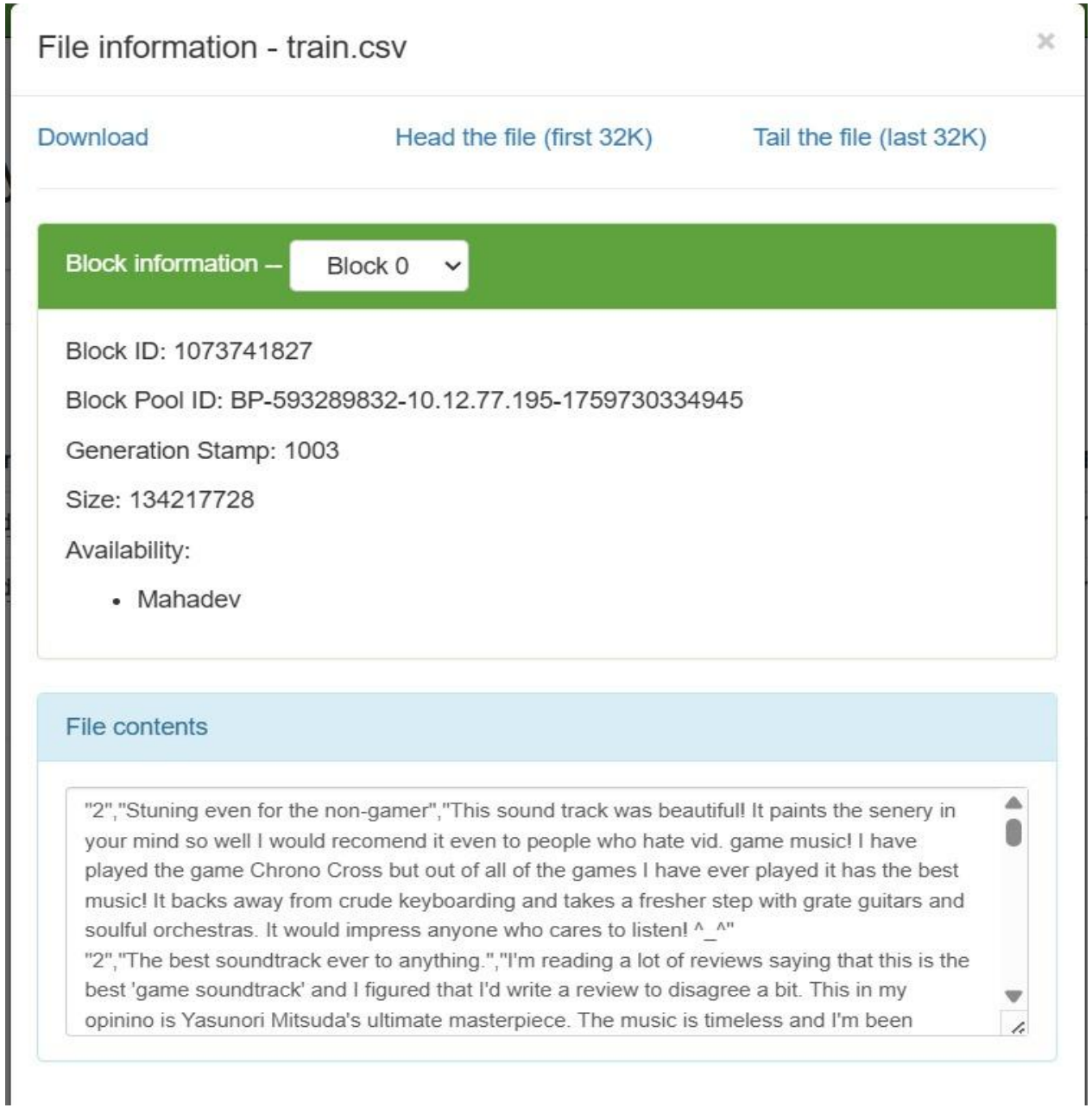
Availability:

- Mahadev

File contents

```
polarity,title,text
2,Great CD,"\"My lovely Pat has one of the GREAT voices of her generation. I have listened to this CD for YEARS and I still LOVE IT. When I'm in a good mood it makes me feel better. A bad mood just evaporates like sugar in the rain. This CD just oozes LIFE. Vocals are jusat STUUNNING and lyrics just kill. One of life's hidden gems. This is a desert isle CD in my book. Why she never made it big is just beyond me. Everytime I play this, no matter black, white, young, old, male, female EVERYBODY says one thing \"\"Who was that singing ?\"\"\""
2,One of the best game music soundtracks - for a game I didn't really play,"Despite the fact
```


File information - part-00000-a6cf490d-a6c2-440d-a33d-2606f8666a35-c000.json
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information –
Block 0
Block ID: 1073741975
Block Pool ID: BP-593289832-10.12.77.195-1759730334945
Generation Stamp: 1151
Size: 152
Availability:
Mahadev
File contents
{"accuracy":0.88096,"precision":0.88096761207576,"recall":0.88096,"f1_score":0.880959405365
4457,"auc":0.9479248591875001,"timestamp":"20251012_201016"}


File information - train.csv
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information –
Block 0
Block ID: 1073741827
Block Pool ID: BP-593289832-10.12.77.195-1759730334945
Generation Stamp: 1003
Size: 134217728
Availability:
Mahadev
File contents
"2","Stuning even for the non-gamer","This sound track was beautiful! It paints the senery in your mind so well I would recomend it even to people who hate vid. game music! I have played the game Chrono Cross but out of all of the games I have ever played it has the best music! It backs away from crude keyboarding and takes a fresher step with grate guitars and soulful orchestras. It would impress anyone who cares to listen! ^_^"
"2","The best soundtrack ever to anything.","I'm reading a lot of reviews saying that this is the best 'game soundtrack' and I figured that I'd write a review to disagree a bit. This in my opinino is Yasunori Mitsuda's ultimate masterpiece. The music is timeless and I'm been

Browse Directory

/user/mahad Go!

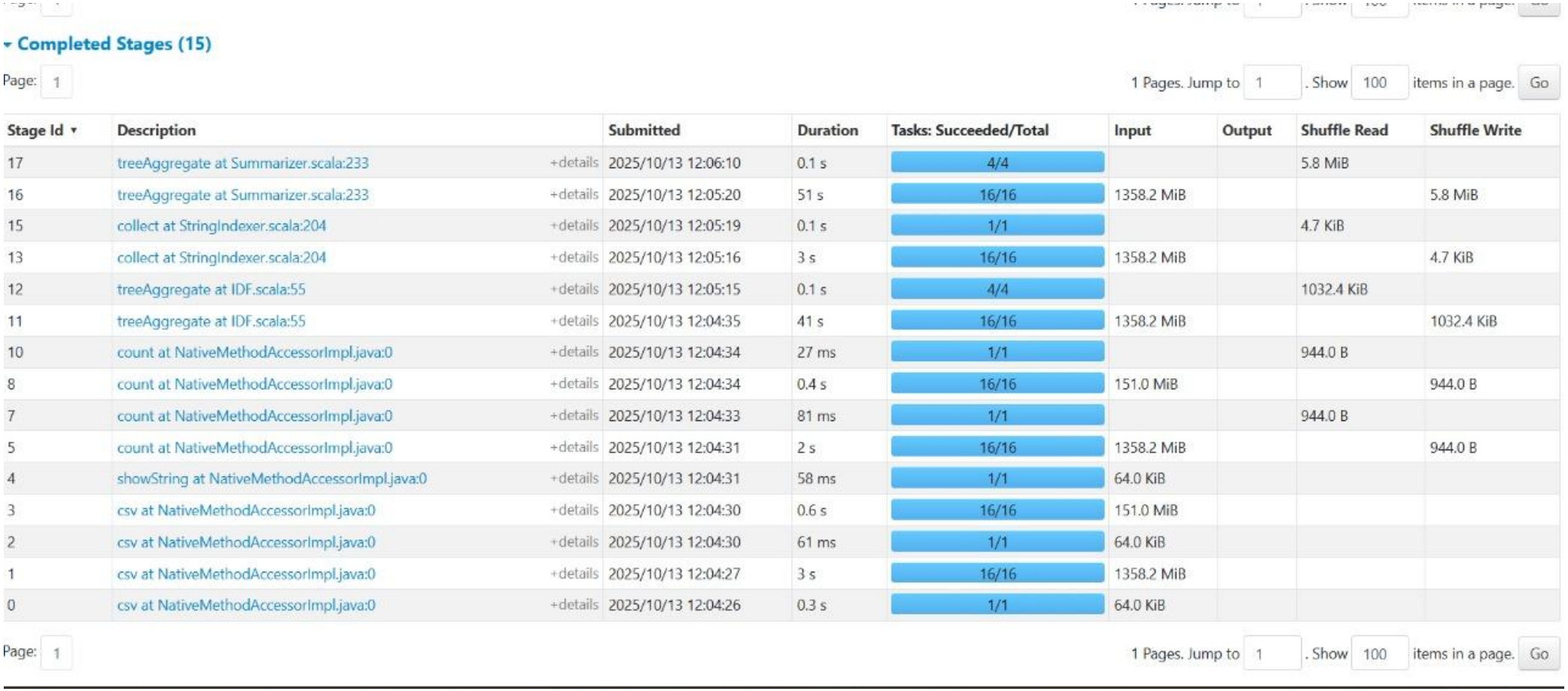
Show 25 entries Search:

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
☐	drwxr-xr-x	mahad	supergroup	0 B	Oct 12 21:09	0	0 B	amazon_metrics
☐	drwxr-xr-x	mahad	supergroup	0 B	Oct 06 12:02	0	0 B	input
☐	drwxr-xr-x	mahad	supergroup	0 B	Oct 06 17:33	0	0 B	models
☐	drwxr-xr-x	mahad	supergroup	0 B	Oct 06 15:26	0	0 B	processed

Showing 1 to 4 of 4 entries Previous 1 Next

Hadoop, 2022.

Completed Stages



▾ Completed Stages (15)

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id ▾	Description		Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
17	treeAggregate at Summarizer.scala:233	+details	2025/10/13 12:06:10	0.1 s	4/4			5.8 MiB	
16	treeAggregate at Summarizer.scala:233	+details	2025/10/13 12:05:20	51 s	16/16	1358.2 MiB			5.8 MiB
15	collect at StringIndexer.scala:204	+details	2025/10/13 12:05:19	0.1 s	1/1			4.7 KiB	
13	collect at StringIndexer.scala:204	+details	2025/10/13 12:05:16	3 s	16/16	1358.2 MiB			4.7 KiB
12	treeAggregate at IDF.scala:55	+details	2025/10/13 12:05:15	0.1 s	4/4			1032.4 KiB	
11	treeAggregate at IDF.scala:55	+details	2025/10/13 12:04:35	41 s	16/16	1358.2 MiB			1032.4 KiB
10	count at NativeMethodAccessorImpl.java:0	+details	2025/10/13 12:04:34	27 ms	1/1			944.0 B	
8	count at NativeMethodAccessorImpl.java:0	+details	2025/10/13 12:04:34	0.4 s	16/16	151.0 MiB			944.0 B
7	count at NativeMethodAccessorImpl.java:0	+details	2025/10/13 12:04:33	81 ms	1/1			944.0 B	
5	count at NativeMethodAccessorImpl.java:0	+details	2025/10/13 12:04:31	2 s	16/16	1358.2 MiB			944.0 B
4	showString at NativeMethodAccessorImpl.java:0	+details	2025/10/13 12:04:31	58 ms	1/1	64.0 KiB			
3	csv at NativeMethodAccessorImpl.java:0	+details	2025/10/13 12:04:30	0.6 s	16/16	151.0 MiB			
2	csv at NativeMethodAccessorImpl.java:0	+details	2025/10/13 12:04:30	61 ms	1/1	64.0 KiB			
1	csv at NativeMethodAccessorImpl.java:0	+details	2025/10/13 12:04:27	3 s	16/16	1358.2 MiB			
0	csv at NativeMethodAccessorImpl.java:0	+details	2025/10/13 12:04:26	0.3 s	1/1	64.0 KiB			

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Conclusion

- Successfully **cleaned, formatted, and preprocessed raw data** for use in Hadoop.
- Ensured **data consistency, accuracy, and reliability** by handling missing, duplicate, and noisy values.
- Converted data into **Hadoop-compatible formats** for smooth storage and processing.
- improved the **efficiency of Hadoop jobs** by reducing redundancy and optimizing input data.

THANK YOU