

Isai Tinoco Gutierrez, Russell Ferrall

Dr. Laila

CSC 364

26 August 2025

## Lab 3 Report – Analyzing Steam Game Ratings with Hadoop MapReduce

### Data Description

We selected the Steam Games Review 2024 dataset from Kaggle.

- **Size:** ~128 million reviews, >80,000 games, >30 million unique users. ~14 GB
- **Format:** CSV files
- **Key attributes used in our analysis:**
  - *appid* (game identifier, derived from file name)
  - *language* (review language)
  - *voted\_up* (positive or negative review)
  - *early\_access* (flag for early access reviews)

#### Example record (simplified):

appid, language, voted_up, early_access, review_text
730, english, true, false, "Great shooting mechanics..."

The dataset exceeds the 1 GB requirement and contains diverse attributes suitable for filtering and aggregation.

### Data Cleaning Process

To ensure data quality and relevance, we applied the following filters in our **Mapper** code:

- **Removed headers and malformed lines.**
- **Filtered by language:** kept only English reviews (*language == english*).
- **Excluded early access reviews** (to avoid bias from unfinished games).
- **Checked for corruption:** incremented counters for corrupt or filtered rows.

The cleaning ensured that our analysis focused only on high-quality, comparable reviews across games.

## Data Analysis Process

We implemented the analysis using **Hadoop MapReduce** with the following pipeline:

### 1. Mapper (**ReviewMapper.java**):

- Parsed each CSV line using *CsvParser*.
- Extracted *appid*, *language*, *voted\_up*, and *early\_access*.
- Emitted (*appid*, (*positive*=1/0, *total*=1)).

### 2. Combiner (**ReviewCombiner.java**):

- Locally aggregated partial counts of positive and total reviews per game.
- Reduced network overhead by sending fewer intermediate results.

### 3. Reducer (**ReviewReducer.java**):

- Summed all (*positive*, *total*) values per *appid*.
- Output format:

appid positive total
----------------------

### 4. Driver (**Main.java**):

- Configured job with input splits (128 MB for efficiency).
- Set Mapper, Combiner, Reducer, and custom *PairWritable* class for intermediate values.
- Stored results in HDFS output directory.

This pipeline was run on **Hadoop 3.4.1** using a Maven-built JAR (*pom.xml* specifies dependencies).

## Results of the Analysis

We collected review statistics per game and visualized the results:

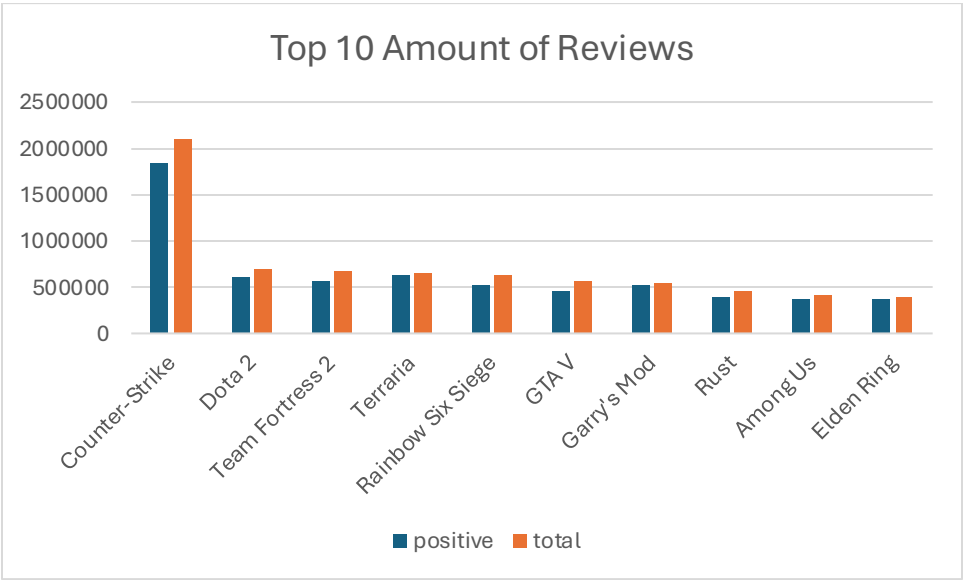
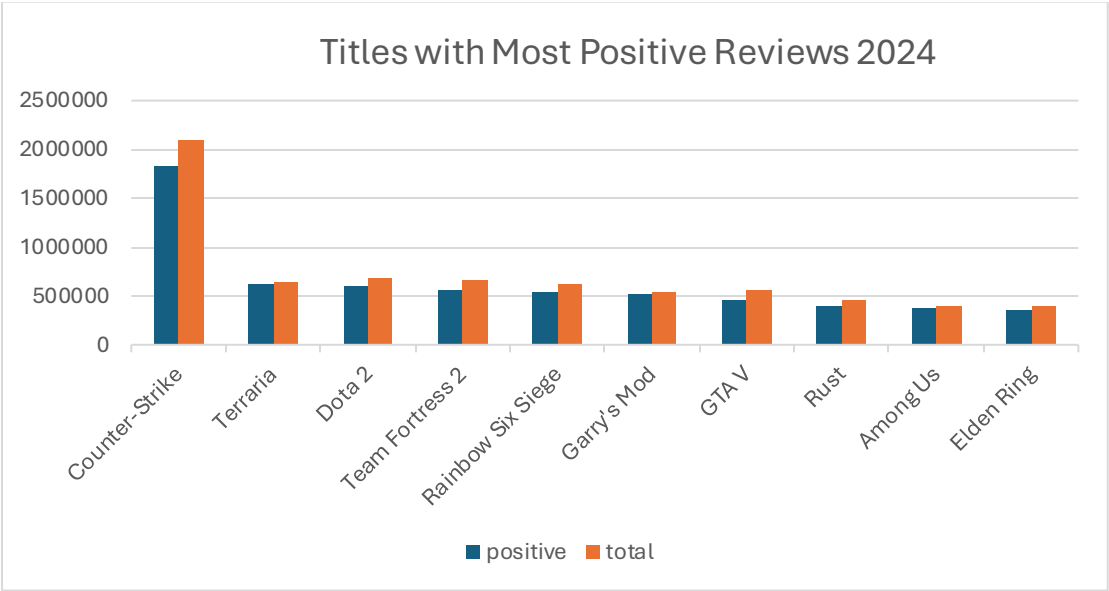
### Top Games by Positive Reviews (2024)

- Counter-Strike dominates with ~1.9 million positive reviews (out of ~2.1 million total).
- Other high-ranking games: *Terraria*, *Dota 2*, *Team Fortress 2*, *Rainbow Six Siege*.

### Top 10 Total Reviews

- Same trend: *Counter-Strike*, *Dota 2*, and *Team Fortress 2* remain at the top, confirming their large active communities.
- Smaller but strong performers include *Among Us*, *Rust*, and *Elden Ring*.

(Charts shown below)



**Conclusion**

This project demonstrated:

- The ability to clean and process a **multi-gigabyte dataset** with Hadoop.
- A custom MapReduce pipeline to count and filter Steam reviews efficiently.
- Clear insights into which games dominate both in volume and positivity of reviews.

**Key points:**

Older, community-driven games (*Counter-Strike*, *Dota 2*) have massive reviews, while new hits (*Elden Ring*, *Among Us*) are competitive but smaller in scale.