HackerNews Data Ingestion Analysis Project Report

Junhao Fu (jf4519)

Executive Summary

This project implements a large-scale dataset processing pipeline to analyze comments data from HackerNews. Using Hadoop MapReduce, I completed jobs of data profiling, cleaning, and data ingestion of a 15GB dataset containing user comments from HackerNews. The pipeline enables efficient processing of large-scale dataset while ensuring data quality and providing insights into technical discussions.

Data Source Description

Dataset Overview

- **Source**: HackerNews Comments Dataset
- **Size**: 15GB
- **Format**: CSV file
- **Location**: HDFS directory (/user/jf4519_nyu_edu/project/comments.csv)
- **Schema**:
    - id: Unique identifier for each comment
    - title: Title of the related post
    - text: Comment content
    - by: Username of the commenter
    - score: Comment score/rating
    - time: Timestamp of the comment
    - type: Type of the entry (comment, story, etc.)

Data Characteristics

- Historical data spanning multiple years
- Rich text comments content containing technical discussions and other topics
- Hierarchical structure with comments linked to stories
- User engagement through scoring application

Implementation

1. Data Pipeline Architecture

The implementation follows a 3-stage pipeline:

- Data Profiling
- Data Cleaning
- Data Integration

2. Data Profiling Component

Key Features:

- Completeness checking for all columns
- Data type checking
- Value distribution analysis
- Missing value detection

3. Data Cleaning Component

Cleaning Operations:

- Text standardization
- HTML entity and extra whitespace removal
- Timestamp normalization
- Duplicate removal

4. Data Ingestion

Running these commands in terminal to get the pipeline result

```
# Install mrjob package
pip install mrjob

# Configure mrjob for Hadoop (create ~/.mrjob.conf file)
cat > ~/.mrjob.conf << EOL
runners:
  hadoop:
    hadoop_home: /usr/lib/hadoop
    hadoop_streaming_jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
EOL

# Run the data profiling job
python data_profiling_mr.py \
    -r hadoop \
    project/comments.csv \
    --output-dir project/profiling_output

# Run the data cleaning job
python data_cleaning_mr.py \
    -r hadoop \
    project/comments.csv \
    --output-dir project/cleaned_output

# Check the results
echo "Checking profiling results:"
hadoop fs -cat project/profiling_output/part-* | head -n 10

echo "Checking cleaned data:"
hadoop fs -cat project/cleaned_output/part-* | head -n 10
```

Conclusions and Future Work

Achievements

1. Processed and cleaned 15GB of HackerNews data successfully.
2. Implemented MapReduce pipeline to complete Large-scale data processing efficiently.

Future Improvements

We will enhance text analysis to complete sentiment analysis and topic modeling for social media dataset in the future.

Appendix: Source Code

get_dataset.py: Source code for downloading the dataset from hugging face source.

data_profiling_mr.py: Source code for *data profiling job.*

data_cleaning_mr.py: Souce code for *data cleaning job.*