Data Profiling, Cleaning, and Ingestion using MapReduce

Rhea Chandok (rc5397)

Objective

To perform data profiling, cleaning, and ingestion on a dataset containing incident-related information. Specifically, we aimed to:

1. Filter and clean data based on specific criteria.
2. Compute profiling statistics to understand data quality and characteristics.
3. Save the cleaned dataset and profiling statistics for further analysis.

Dataset Overview

The dataset consisted of the following columns:

- **rowid**: Unique identifier for each row.
- **response_type**: Type of response to the incident.
- **incident_number**: Incident identification number.
- **call_date**: Date of the call.
- **final_priority**: Priority level assigned to the incident.
- **month_name**: Month of the call.
- **month_no**: Numerical representation of the month.
- **response_time_min**: Response time in minutes.
- **data_as_of**: Timestamp indicating the data's freshness.
- **data_loaded_at**: Timestamp when the data was ingested.

Steps Performed

1. Data Cleaning and Filtering

- Dropped unnecessary columns: `rowid`, `data_as_of`, and `data_loaded_at`, which were not required for analysis.
- Filtered rows to include only data from January 1, 2014, to October 31, 2024.
- Excluded rows where `response_time_min` contained invalid or missing values.

2. Profiling Statistics

Computed key statistics for the `response_time_min` column:

- **Minimum Response Time**: The fastest recorded response time.
- **Maximum Response Time**: The slowest recorded response time.
- **Average Response Time**: The average of all valid response times.
- **Standard Deviation of Response Time**: The standard deviation in response times.
- **Total Cleaned Rows**: The count of rows retained after cleaning.

3. MapReduce Implementation

- **Data Cleaning Mapper**:
  - Read input records and parsed the relevant fields.
  - Dropped rows with invalid values, such as missing `response_time_min` or invalid date formats.
  - Forwarded cleaned `response_time_min` values to the Reducer.
  - Incremented counters to track dropped rows due to missing or invalid data.
- **Data Cleaning Reducer**:
  - Aggregated data from the Mapper.
  - Outputted the cleaned data to be used in subsequent processing steps.
- **Data Profiling Mapper**:
  - Read cleaned data from the previous job (output of the Data Cleaning step).
  - Parsed the `response_time_min` field and forwarded it to the Reducer for statistical aggregation.
- **Data Profiling Reducer**:
  - Aggregated the `response_time_min` values.
  - Calculated profiling statistics, including `min`, `max`, `average`, and `standard deviation`.
  - Outputted the computed statistics along with the total count of records processed.

Output

1. Cleaned Data

- Contains rows with valid data from January 1, 2014, to October 31, 2024.

```
BLS,14013538,2014-02-09,3,Feb,2,18.166667
BLS,14013101,2014-02-08,3,Feb,2,4.466667
BLS,14012808,2014-02-07,3,Feb,2,1.566667
BLS,14011345,2014-02-03,3,Feb,2,2.316667
BLS,19011181,2019-01-27,3,Jan,1,2.866667
BLS,19008998,2019-01-21,3,Jan,1,3.883333
BLS,19006544,2019-01-16,3,Jan,1,2.733333
BLS,23035796,2023-03-14,3,Mar,3,4.05
BLS,23036385,2023-03-16,3,Mar,3,6.4
BLS,23036246,2023-03-15,3,Mar,3,1.616667
BLS,23035980,2023-03-15,3,Mar,3,3.666667
BLS,19009093,2019-01-22,3,Jan,1,3.75
BLS,19008469,2019-01-20,3,Jan,1,2.166667
BLS,19006727,2019-01-16,3,Jan,1,3
BLS,19006213,2019-01-15,3,Jan,1,2.9
BLS,19007492,2019-01-18,3,Jan,1,4.666667
BLS,19006374,2019-01-16,3,Jan,1,4.633333
BLS,19005149,2019-01-12,3,Jan,1,4.4
BLS,19080529,2019-07-08,3,Jul,7,3.15
BLS,19078858,2019-07-04,3,Jul,7,3.916667
BLS,19079602,2019-07-05,3,Jul,7,3.516667
BLS,19078955,2019-07-04,3,Jul,7,1.733333
```

Figure 1: Cleaned Data

2. Counters

- **Rows Dropped Due to Invalid Dates**: Count of rows with dates outside the range.
- **Rows Dropped Due to Missing/Invalid Response Times**: Count of rows with missing or invalid `response_time_min`.

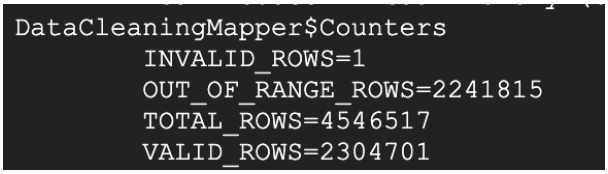

Figure 2: Counters

3. Profiling Statistics

- Includes:
  - **Minimum Response Time**
  - **Maximum Response Time**
  - **Average Response Time**
  - **Standard Deviation of Response Time**
  - **Total Cleaned Rows**

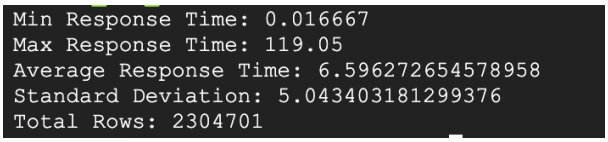

Figure 3: Profiling Statistics