# Data Profiling, Cleaning, and Ingestion using MapReduce

CSCI-GA.2436 Realtime and Big Data Analytics

**Haardik Dharma (hd2585)**
**New York University - Courant Institute of Mathematical Sciences**
November 20, 2024

This report gives an overview of the data profiling and cleaning performed on the Fire Incidents dataset, which is publicly accessible from the City of San Francisco.

## Dataset overview:

Following are the columns of the dataset:

- **Incident Number:** A unique identifier for each fire incident.
- **Exposure Number:** Indicates if the incident is related to another incident (e.g., fire spreading to adjacent structures).
- **ID:** A unique identifier for each record in the dataset.
- **Address:** The location where the incident occurred.
- **Incident Date:** The date when the incident took place.
- **Call Number:** The unique identifier for the 911 call associated with the incident.
- **Alarm DtTm:** The date and time when the alarm was received.
- **Arrival DtTm:** The date and time when the first fire unit arrived at the scene.
- **Close DtTm:** The date and time when the incident was closed or resolved.
- **City:** The city where the incident occurred (likely San Francisco for most entries).
- **zipcode:** The postal code of the incident location.
- **Battalion:** The fire department battalion that is responsible for the area.
- **Station Area:** The fire station area where the incident occurred.
- **Box:** The fire box or call box associated with the incident location.
- **Suppression Units:** The number of fire suppression units that responded to the incident.
- **Suppression Personnel:** The number of firefighting personnel who responded to the incident.
- **Fire Fatalities:** The number of deaths caused by the fire.
- **Fire Injuries:** The number of injuries sustained by firefighters during the incident.
- **Civilian Fatalities:** The number of civilian deaths resulting from the incident.
- **Civilian Injuries:** The number of civilian injuries resulting from the incident.
- **Number of Alarms:** The number of alarms raised for the incident.
- **Primary Situation:** The main type or nature of the incident as observed by firefighters on the scene.
- **Mutual Aid:** Indicates if assistance was provided by or to other fire departments.
- **Action Taken Primary:** The main action taken by firefighters to address the incident.

- **Action Taken Secondary:** Additional actions taken by firefighters during the incident.
- **Action Taken Other:** Any other actions taken by firefighters during the incident.
- **Supervisor District:** The San Francisco supervisor district where the incident occurred.
- **neighborhood_district:** The neighborhood or district within San Francisco where the incident took place.
- **data_as_of:** The date when the data was last updated or current as of.
- **data_loaded_at:** The date and time when the data was loaded into the dataset.

## Dataset cleaning and filtering:

The original dataset available on the website contains 66 columns. I cleaned up 36 of these columns, leaving us with 30 columns (mentioned above) that we will use for our project. Some of the columns removed were -

- area_of_fire_origin
- ignition_cause
- ignition_factor_primary,
- ignition_factor_secondary,
- heat_source,
- item_first_ignited,
- human_factors_associated_with_ignition,
- floor_of_fire_origin

The primary objective of this project is to identify response times and efficiency patterns during emergencies. Therefore, we do not focus on how the fire was ignited or its origin. Most of the columns we removed were text-based, and many of them were empty, as the causes of the fires are often unknown or not reported.

Each row in the dataset represents a fire incident. Since the dataset is updated daily, we decided to limit our analysis to incidents from the last decade, specifically from January 1, 2014, to October 31, 2024, in order to maintain consistency with other data sources within our project.

**Java Classes used:**

1. **FireIncidentCleaning.java, FireIncidentCleaningMapper.java, FireIncidentCleaningReducer.java -** This MapReduce job was run on the original dataset containing 66 columns. The Mapper does the job of parsing each CSV line and validates the structure by checking the number of columns. It then removes the unnecessary columns by the index numbers specified. The Reducer removes the keys that were used in the mapping phase, as they are no longer needed. It ensures that all cleaned data lines are written to the output, maintaining the structure created by the mapper. The output of this MapReduce job is the cleaned dataset with 30 columns. Below is the preview of the dataset after running this job -

Example of 1 record in the dataset:

```
Incident Number,Exposure Number,ID,Address,Incident Date,Call Number,Alarm DtTm,Arrival DtTm,Close DtTm,City,zipcode,Battalion,Station Area,Box,
Suppression Units,Suppression Personnel,Fire Fatalities,Fire Injuries,Civilian Fatalities,Civilian Injuries,Number of Alarms,Primary Situation,Mutual Aid,
Action Taken Primary,Action Taken Secondary,Action Taken Other,Supervisor District,neighborhood_district,data_as_of,data_loaded_at
22077209,0,220772090,DODGE STREET,2022/06/17,221680293,2022/06/17 02:59:32 AM,2022/06/17 03:05:19 AM,2022/06/17 03:05:56 AM,San Francisco,94102,B02,03,1554,
1,4,0,0,0,1,"151 Outside rubbish, trash or waste fire",N None,
87 Investigate fire out on arrival,,,5,Tenderloin,2022/06/17 03:05:56 AM,2024/11/17 02:16:16 AM
```

2. **FireIncidentDateFiltering.java -** This is a Map-only job that filters fire incident records, keeping only those that occurred between January 1, 2014, and October 31, 2024. It maintains the original structure of the data, including the header row which contains the column names. It brings down the number of rows from **684k** to **349k.**

## Dataset Profiling:

For data profiling, I followed these 2 steps:

1. Calculated the Sum, Average, Min, Max for the below columns:

- **Fire Fatalities:** The number of deaths caused by the fire to the personnel.
- **Fire Injuries:** The number of injuries sustained by firefighters during the incident.
- **Civilian Fatalities:** The number of civilian deaths resulting from the incident.
- **Civilian Injuries:** The number of civilian injuries resulting from the incident.

**Java Classes used:**

**FireIncidentsInjuredDeathProfiling.java, FireIncidentsInjuredDeathMapper.java, FireIncidentsInjuredDeathReducer.java -** The Mapper processes each input line, except the header row, which contains the column names. It then extracts values for the columns - Fire Fatalities, Fire Injuries, Civilian Fatalities, and Civilian Injuries and emits each of these categories as a key with its corresponding value to the reducer. The Reducer receives grouped values for each category and calculates the count (this is the total number of rows), sum (total number of injuries/fatalities), average, minimum, and maximum for each category. It finally outputs a summary string for each category with these calculated statistics.

**Output:**

outputprofiling > merged_profiling_output

```
1  Civilian Fatalities Count: 348939, Sum: 19, Avg: 0.00, Min: 0, Max: 2
2  Fire Injuries   Count: 348939, Sum: 33, Avg: 0.00, Min: 0, Max: 4
3  Civilian Injuries   Count: 348939, Sum: 327, Avg: 0.00, Min: 0, Max: 9
4  Fire Fatalities Count: 348939, Sum: 0, Avg: 0.00, Min: 0, Max: 0
5
```

2. Calculated Mean, Std Deviation, Min, Max, Median response and Turnaround times, using the below columns:

- **Alarm DtTm:** The date and time when the alarm was received.
- **Arrival DtTm:** The date and time when the first fire unit arrived at the scene.
- **Close DtTm:** The date and time when the incident was closed or resolved.

**Java Classes used:**

**FireIncidentTimeProfiling.java,FireIncidentTimeMapper.java, FireIncidentTimeReducer.java -** The Mapper parses the alarm, arrival, and close timestamps from the input data. It then checks if the timestamps are in chronological order, skipping records with inconsistent time sequences. For valid records, it calculates the response time (arrival time - alarm time) and turnaround time (close time - alarm time) in milliseconds. Finally, it emits two key-value pairs: one for "Response Time" and another for "Turnaround Time", with the calculated times as values. The reducer sorts the collected times in ascending order. It then calculates the following statistics and emits them as the output: Minimum, Maximum, Mean, Standard Deviation, and Median.

**Output:**

outputtimeprofiling > time_profiling_output.txt

```
1   Response Time   Mean: 563251.86, StdDev: 4826612.46, Min: 0, Max: 518594000, Median: 279000.00
2   Turnaround Time  Mean: 3384385.42, StdDev: 16631893.88, Min: 0, Max: 1475694000, Median: 1050000.00
```

The above times are in milliseconds. Converting them to minutes, we get:

**Response Time:**

- **Mean:** 9.39 minutes
- **Median:** 4.65 minutes
- **Standard Deviation:** 80.44 minutes
- **Minimum:** 0 minutes
- **Maximum:** 8,643.23 minutes

**Turnaround Time:**

- **Mean:** 56.41 minutes
- **Median:** 17.50 minutes
- **Standard Deviation:** 277.20 minutes
- **Minimum:** 0 minutes
- **Maximum:** 24,594.90 minutes

The significant difference between the mean and median values for both response and turnaround times indicates the presence of outliers in the dataset, particularly skewing towards longer durations. For response times, the mean is 9.39 minutes, while the median is only 4.65 minutes. Similarly, for turnaround times, the mean is 56.41 minutes, compared to a median of just 17.50 minutes. This difference between the mean and median is a classic sign of a right-skewed distribution, where a small number of extremely high values are pulling the mean upwards, while the median remains a more accurate representation of the typical case.

The presence of outliers is also evident from the high maximum values: 8,643.23 minutes for response time and 24,594.90 minutes for turnaround time. These large values, along with large standard deviations (80.44 minutes for response time and

277.20 minutes for turnaround time), are also indicating a long right tail in the distribution. This right skew suggests that while most incidents are managed within a reasonable timeframe, there are rare cases with extraordinarily long response or resolution times.