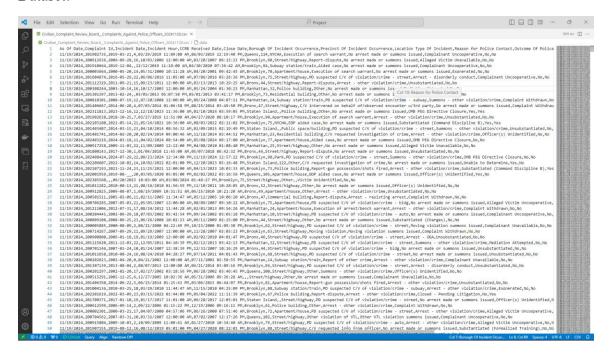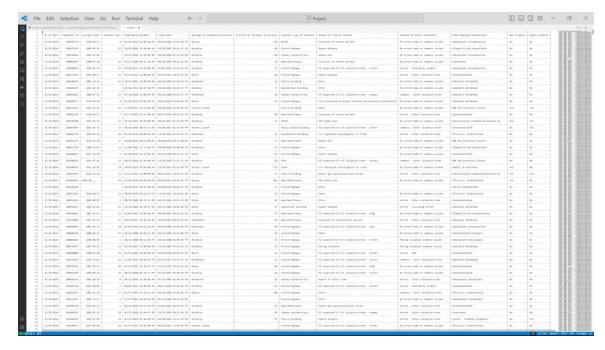Dataset used: Civilian Complaint Review Board: Complaints Against Police Officers [data.cityofnewyork.us/Public-Safety/Civilian-Complaint-Review-Board-Complaints-Against/2mby-ccnw/about_data]

"The primary table for all public data on complaints, including dates, locations and the outcomes of closed complaints received since the year 2000.

The dataset is part of a database of all public police misconduct records the Civilian Complaint Review Board (CCRB) maintains on complaints against New York Police Department uniformed members of service received in CCRB's jurisdiction since the year 2000, when CCRB's database was first built."

Dataset:

Fields:

Field Name	Description
As Of Date	The date when this data record was last updated or verified in the system
Complaint Id	Unique identifier assigned to each complaint filed with the CCRB
Incident Date	The date when the reported incident occurred
Incident Hour	The hour of the day when the incident took place
CCRB Received Date	The date when the CCRB officially received and logged the complaint
Close Date	The date when the investigation or case was closed by CCRB
Borough Of Incident Occurrence	The New York City borough where the incident took place
Precinct Of Incident Occurrence	The specific police precinct where the incident occurred
Location Type Of Incident	The type of location where the incident took place (e.g., street, residence, public building)

Field Name	Description
Reason for Police Contact	The initial reason why police officers interacted with the civilian
Outcome Of Police Encounter	The result or conclusion of the interaction between police and civilian
CCRB Complaint Disposition	The final determination or outcome of the CCRB investigation
BWC Evidence	Indicates whether Body-Worn Camera footage is available for this incident (Yes/No)
Video Evidence	Indicates whether any other video evidence is available for this incident (Yes/No)

The dataset was small, so I downloaded it to my local and uploaded it using the "upload file" function of "ssh-in-browser" to NYU DataProc.

As the data was in .csv format, I used the default text input which loaded one line at a time, skipped the header line, and split each line with commas in the mapper. I omitted useless columns like "as of date." I checked if there were missing values or malformed values (e.g., hours less than 0 or greater than 24, ending time earlier than starting time, etc.), and logged an analysis using the MapReduce counter:

```
Data Quality
        Invalid Date Format=2286
        Invalid Date Sequence=13
        Invalid Hour Format=3650
        Missing Borough=451
        Missing Location Type=2702
        Missing Outcome=616
        Missing Police Contact Reason=1893
```

Also, in the mapper, I aligned the date format.

I then used a self-defined data structure to hold useful data, with the original id being the key, and let the reducer write the key-value pair to output files:

```
rm5327_nyu_edu@nyu-dataproc-m:~$ hadoop fs -tail output/part-r-00000
ROOKLYN,Street/highway,Moving violation,Arrest - other violation/crime
202409313       09/10/2024 10,09/18/2024 11:29:00 AM,10/16/2024 10:28:10 AM,BRONX,Street/highway,PD suspected C/V of violation/crime - subway,Arrest - other violation/crime
202409321       09/17/2024 20,09/18/2024 01:40:00 PM,10/16/2024 10:28:11 AM,BROOKLYN,Street/highway,Demonstration/protest,UNKNOWN
202409538       09/22/2024 9,09/24/2024 09:14:00 AM,10/16/2024 10:28:52 AM,BROOKLYN,Street/highway,C/V requested investigation of crime,No arrest made or summons issued
202409546       09/20/2024 19,09/23/2024 05:14:00 PM,10/16/2024 10:28:56 AM,MANHATTAN,Street/highway,PD suspected C/V of violation/crime - auto,Arrest - other violation/crime
202410042       08/19/2023 10,10/08/2024 03:03:00 PM,11/07/2024 11:12:40 AM,BROOKLYN,Street/highway,C/V intervened on behalf of/observed encounter w/3rd party,No arrest made or summons issued
202410857       10/28/2023 20,10/31/2024 11:20:00 AM,11/07/2024 11:12:37 AM,BROOKLYN,Street/highway,Traffic accident,No arrest made or summons issued
rm5327_nyu_edu@nyu-dataproc-m:~$
```

Update:

After the data profiling phase, I found that my data profiling and cleaning didn't match the need to thoroughly analyze the dataset, so I made changes to my codes:

1. In the Reducer, I replaced the id as keys with NullWritable, because the original ids were generated by the publisher of the dataset and did not convey any information, while keeping them makes loading data in to Hive complicated.
2. In the Mapper, I added the "disposition" information which means how NYPD dealt with the civilian complaints.