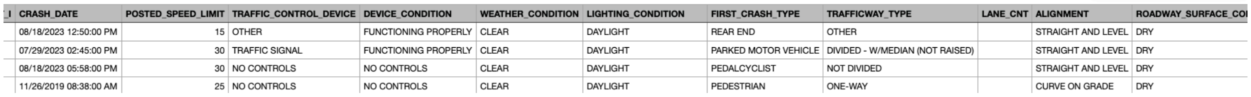
RBDA Data Ingestion - Yuxiang Huang - yh5047

There are four map reduce programs in total for data profiling, data cleaning on two different dataset separately.

1 Data Profiling

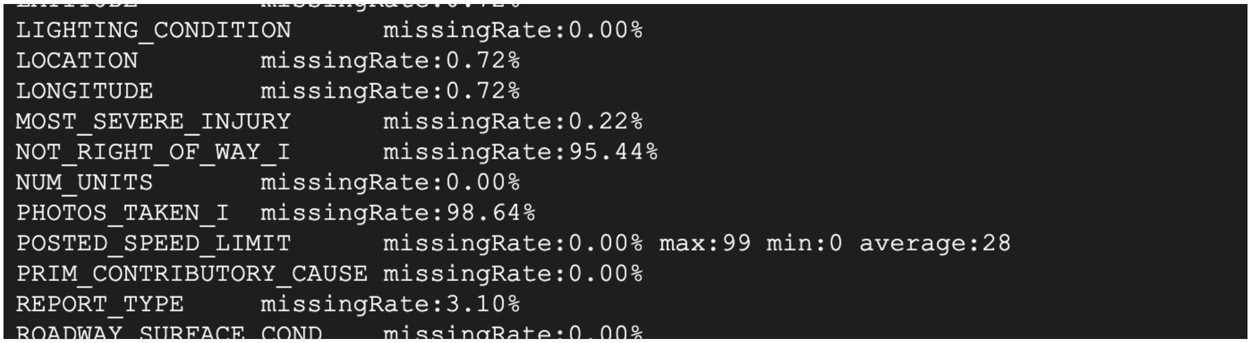
1.1 Chicago Traffic Accident Data

Download from: https://catalog.data.gov/dataset/traffic-crashes-crashes



_I	CRASH_DATE	POSTED_SPEED_LIMIT	TRAFFIC_CONTROL_DEVICE	DEVICE_CONDITION	WEATHER_CONDITION	LIGHTING_CONDITION	FIRST_CRASH_TYPE	TRAFFICWAY_TYPE	LANE_CNT	ALIGNMENT	ROADWAY_SURFACE_CO
	08/18/2023 12:50:00 PM	15	OTHER	FUNCTIONING PROPERLY	CLEAR	DAYLIGHT	REAR END	OTHER		STRAIGHT AND LEVEL	DRY
	07/29/2023 02:45:00 PM	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY	CLEAR	DAYLIGHT	PARKED MOTOR VEHICLE	DIVIDED - W/MEDIAN (NOT RAISED)		STRAIGHT AND LEVEL	DRY
	08/18/2023 05:58:00 PM	30	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	PEDALCYCLIST	NOT DIVIDED		STRAIGHT AND LEVEL	DRY
	11/26/2019 08:38:00 AM	25	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	PEDESTRIAN	ONE-WAY		CURVE ON GRADE	DRY

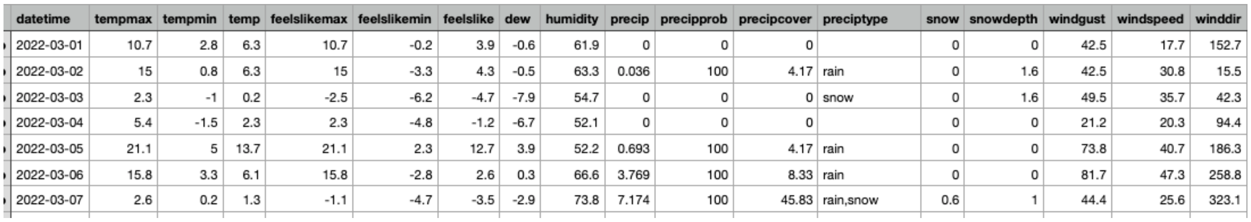
The data contains 48 columns, but most of them are strings, and some of them contain missing values. Thus, I wrote a MapReduce program to perform data profiling, where the percentage of missing value is counted. For numeric fields, the maximum, minimum and average values are calculated.



```
LATITUDE        missingRate:0.72%
LIGHTING_CONDITION      missingRate:0.00%
LOCATION        missingRate:0.72%
LONGITUDE       missingRate:0.72%
MOST_SEVERE_INJURY      missingRate:0.22%
NOT_RIGHT_OF_WAY_I      missingRate:95.44%
NUM_UNITS       missingRate:0.00%
PHOTOS_TAKEN_I  missingRate:98.64%
POSTED_SPEED_LIMIT      missingRate:0.00% max:99 min:0 average:28
PRIM_CONTRIBUTORY_CAUSE missingRate:0.00%
REPORT_TYPE     missingRate:3.10%
ROADWAY_SURFACE_COND    missingRate:0.00%
```

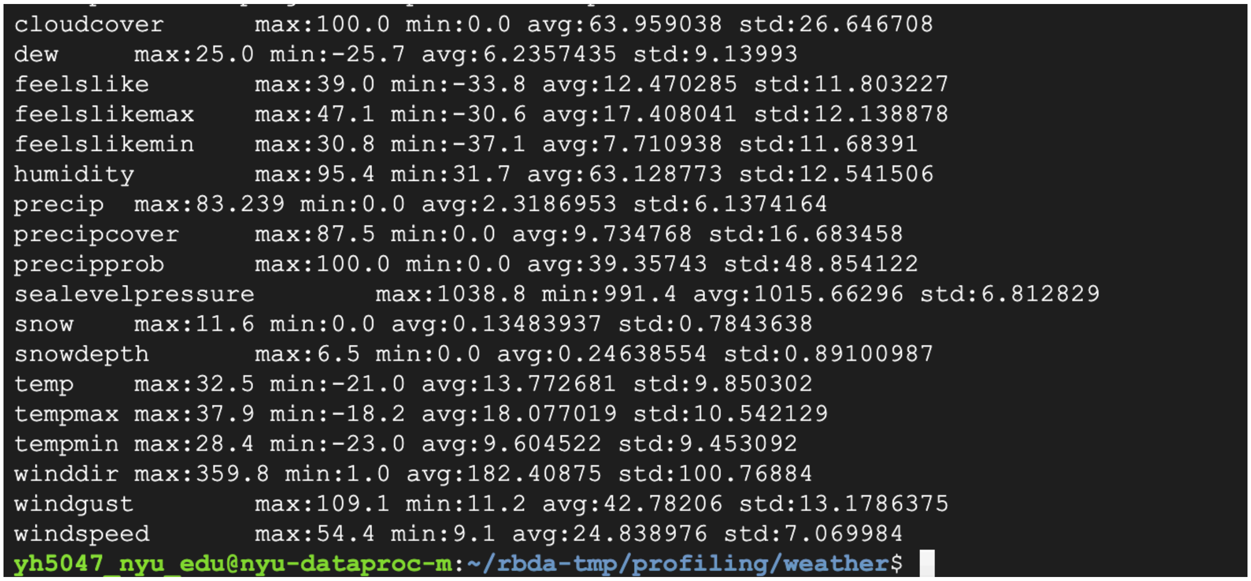
1.2 Chicago Weather Data

Download from: https://www.visualcrossing.com/weather/weather-data-services/Chicago/metric/last15days



datetime	tempmax	tempmin	temp	feelslikemax	feelslikemin	feelslike	dew	humidity	precip	precipprob	precipcover	preciptype	snow	snowdepth	windgust	windspeed	winddir
2022-03-01	10.7	2.8	6.3	10.7	-0.2	3.9	-0.6	61.9	0	0	0		0	0	42.5	17.7	152.7
2022-03-02	15	0.8	6.3	15	-3.3	4.3	-0.5	63.3	0.036	100	4.17	rain	0	1.6	42.5	30.8	15.5
2022-03-03	2.3	-1	0.2	-2.5	-6.2	-4.7	-7.9	54.7	0	0	0	snow	0	1.6	49.5	35.7	42.3
2022-03-04	5.4	-1.5	2.3	2.3	-4.8	-1.2	-6.7	52.1	0	0	0		0	0	21.2	20.3	94.4
2022-03-05	21.1	5	13.7	21.1	2.3	12.7	3.9	52.2	0.693	100	4.17	rain	0	0	73.8	40.7	186.3
2022-03-06	15.8	3.3	6.1	15.8	-2.8	2.6	0.3	66.6	3.769	100	8.33	rain	0	0	81.7	47.3	258.8
2022-03-07	2.6	0.2	1.3	-1.1	-4.7	-3.5	-2.9	73.8	7.174	100	45.83	rain,snow	0.6	1	44.4	25.6	323.1

The weather data contains no missing value, and most of columns are string type. In this case, we can calculate the maximum, minimum , average and variant values for all numeric fields.



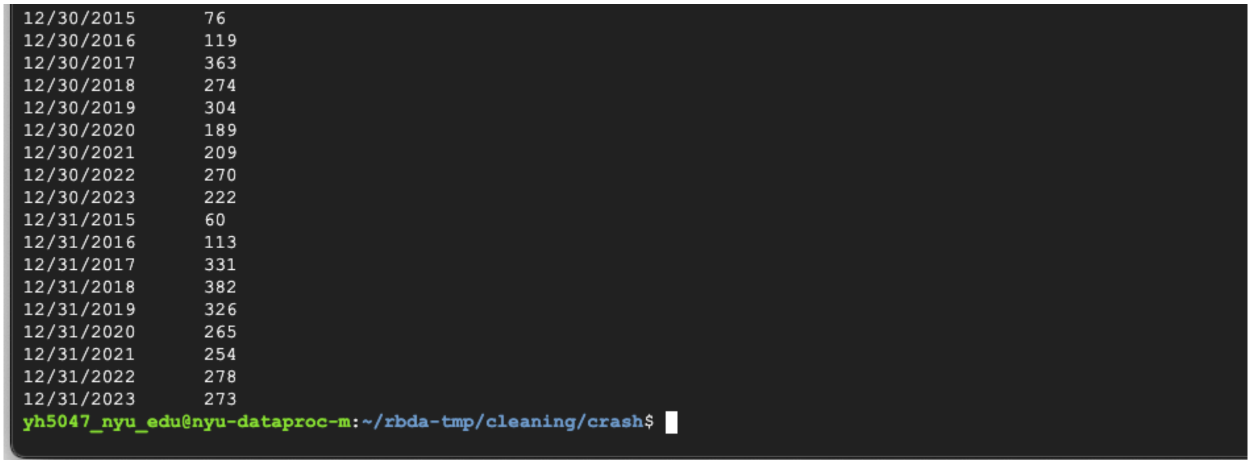
```
cloudcover      max:100.0 min:0.0 avg:63.959038 std:26.646708
dew     max:25.0 min:-25.7 avg:6.2357435 std:9.13993
feelslike       max:39.0 min:-33.8 avg:12.470285 std:11.803227
feelslikemax    max:47.1 min:-30.6 avg:17.408041 std:12.138878
feelslikemin    max:30.8 min:-37.1 avg:7.710938 std:11.68391
humidity        max:95.4 min:31.7 avg:63.128773 std:12.541506
precip  max:83.239 min:0.0 avg:2.3186953 std:6.1374164
precipcover     max:87.5 min:0.0 avg:9.734768 std:16.683458
precipprob      max:100.0 min:0.0 avg:39.35743 std:48.854122
sealevelpressure        max:1038.8 min:991.4 avg:1015.66296 std:6.812829
snow    max:11.6 min:0.0 avg:0.13483937 std:0.7843638
snowdepth       max:6.5 min:0.0 avg:0.24638554 std:0.89100987
temp    max:32.5 min:-21.0 avg:13.772681 std:9.850302
tempmax max:37.9 min:-18.2 avg:18.077019 std:10.542129
tempmin max:28.4 min:-23.0 avg:9.604522 std:9.453092
winddir max:359.8 min:1.0 avg:182.40875 std:100.76884
windgust        max:109.1 min:11.2 avg:42.78206 std:13.1786375
windspeed       max:54.4 min:9.1 avg:24.838976 std:7.069984
yh5047_nyu_edu@nyu-dataproc-m:~/rbda-tmp/profiling/weather$
```

2 Data Cleaning and Ingestion

2.1 Chicago Traffic Accident Data

As for the traffic accident dataset, what we care about is how many crashes happened in one day. However, the dataset is just a list of all crash records. Thus, we need to merge all crashes that happened in one day, and output something like a key-value pair <date, number of crashes>.

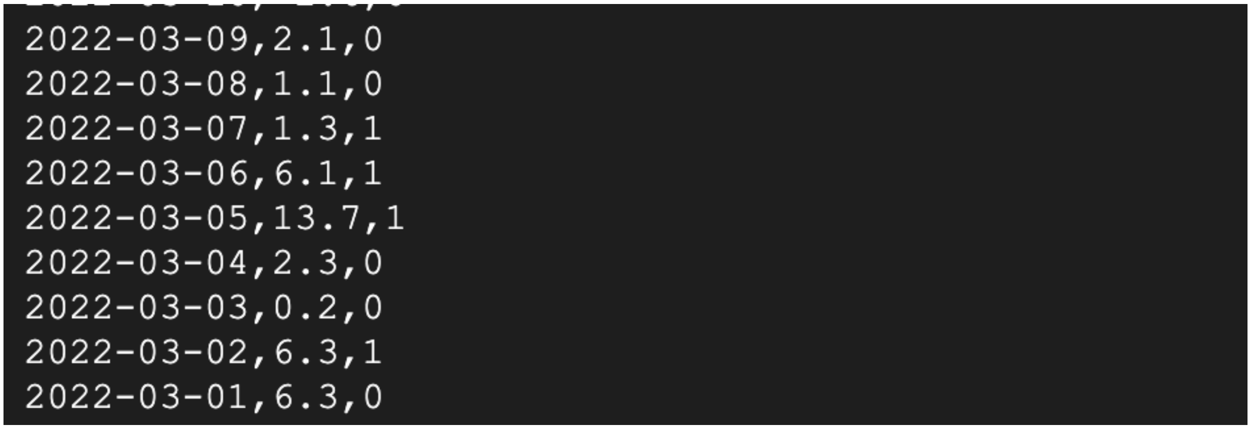
In this case, we only need to use the column "CRASH_DATE" and generate the result output from this single column.



```
12/30/2015      76
12/30/2016      119
12/30/2017      363
12/30/2018      274
12/30/2019      304
12/30/2020      189
12/30/2021      209
12/30/2022      270
12/30/2023      222
12/31/2015      60
12/31/2016      113
12/31/2017      331
12/31/2018      382
12/31/2019      326
12/31/2020      265
12/31/2021      254
12/31/2022      278
12/31/2023      273
yh5047_nyu_edu@nyu-dataproc-m:~/rbda-tmp/cleaning/crash$
```

2.2 Chicago Weather Data

In our project, we only need "datetime", "temp", and "condition". Thus we can drop all other columns. The datetime and temperature columns are perfect which do not need to normalize. While the "condition" column contains complicated strings such as "Rain, Overcast", "Partially cloudy", and "Clear". For normalization, I decide to make the column represent whether it rained or not, that is, if the string contains "rain", it will become "1", otherwise it will be "0".



```
2022-03-09,2.1,0
2022-03-08,1.1,0
2022-03-07,1.3,1
2022-03-06,6.1,1
2022-03-05,13.7,1
2022-03-04,2.3,0
2022-03-03,0.2,0
2022-03-02,6.3,1
2022-03-01,6.3,0
```