Project Data Ingestion

Course Number	CSCI-GA-2436
Done	☐
Created time	@October 15, 2025 12:41 AM
Due Date	@11/28/2025
Relation	Realtime Big Data Analysis
Semester	2025-Fall

Data Source

Primary Dataset

Source: https://data.ny.gov/Transportation/MTA-Subway-Hourly-Ridership-2020-2024/wujg-7c2s/about_data

The data source we are interested is the MTA from: `ny.data.gov` . Here is the link to original dataset. Here is the first several lines of the dataset:

```
transit_timestamp,transit_mode,station_complex_id,station_complex,borough,payment_method,fare_class_category,ridership,transfers,latitude,longitude,Georeference
12/28/2024 08:00:00 PM,subway,475,96 St (Q),Manhattan,metrocard,Metrocard - Fair Fare,7,0,40.784317,-73.94715,POINT (-73.94715 40.784317)
12/28/2024 06:00:00 AM,subway,626,"Franklin Av (2,3,4,5)/Botanic Garden (S)",Brooklyn,metrocard,Metrocard - Unlimited 30-Day,7,0,40.67068,-73.95813,POINT (-73.95813 40.67068)
12/28/2024 12:00:00 PM,subway,607,"34 St-Herald Sq (B,D,F,M,N,Q,R,W)",Manhattan,omny,OMNY - Other,67,0,40.749718,-73.98782,POINT (-73.98782 40.749718)
12/28/2024 07:00:00 PM,subway,612,"Lexington Av-53 St (E,M)/51 St (6)",Manhattan,omny,OMNY - Students,24,0,40.757553,-73.969055,POINT (-73.9690
```

```
55 40.757553)
12/28/2024 05:00:00 PM,subway,123,Grand St (L),Brooklyn,metrocard,Metroc
ard - Seniors & Disability,2,1,40.711926,-73.94067,POINT (-73.94067 40.71192
6)
```

From the sample data above, we could get the schema of this dataset:

```
{
  "transit_timestamp": "datetime",
  "transit_mode": "string",
  "station_complex_id": "integer",
  "station_complex": "string",
  "borough": "string",
  "payment_method": "string",
  "fare_class_category": "string",
  "ridership": "integer",
  "transfers": "integer",
  "latitude": "float",
  "longitude": "float",
  "georeference": "string"
}
```

Our columns of interest would be:

```
transit_timestamp       - Essential for temporal joins with weather/crime datase
ts
station_complex_id      - Location identifier (foreign key to mapping dataset)
payment_method          - Payment type (OMNY or MetroCard)
fare_class_category     - Fare type (aggregated across in final output)
ridership               - Number of rides for a given station/hour/fare type
```

Note: borough, latitude, and longitude are functionally dependent on station_complex_id and will be joined from the mapping dataset downstream rather than stored in the processed ridership data.

Secondary Dataset

In addition to this main dataset, we also leveraged this MTA Subway Stations and Complexes dataset from `catalog.data.gov` for the mapping between station_complex_id to the corresponding station, borough, latitude and longitude.

Source: https://catalog.data.gov/dataset/mta-subway-stations-and-complexes

Key Columns:

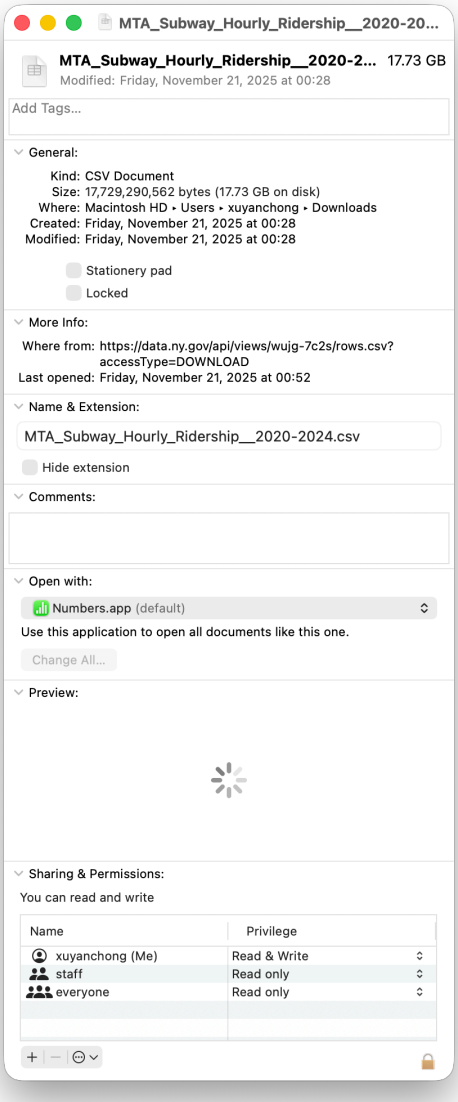
- Complex ID (Column 1) - Maps to station_complex_id
- Borough (Column 9) - Authoritative borough designation (M, Bx, Q, Bk, SI)
- Latitude (Column 13) - Authoritative latitude
- Longitude (Column 14) - Authoritative longitude

Ingestion

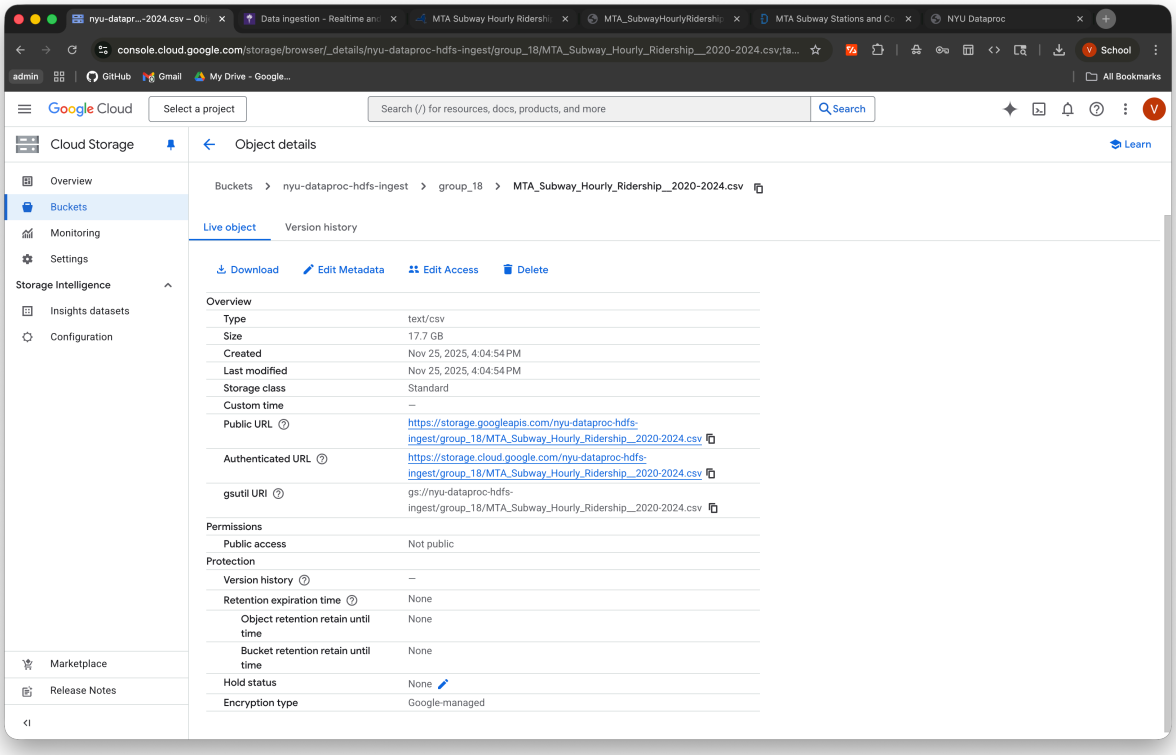
File Locations:

- Primary Dataset: `gs://nyu-dataproc-hdfs-ingest/group_18/MTA_Subway_Hourly_Ridership__2020-2024.csv`
- Mapping Dataset: `gs://nyu-dataproc-hdfs-ingest/group_18/MTA_Subway_Stations_and_Complexes.csv`

The primary data was downloaded from `ny.data.gov` in csv format. Its size is 17.73GB.

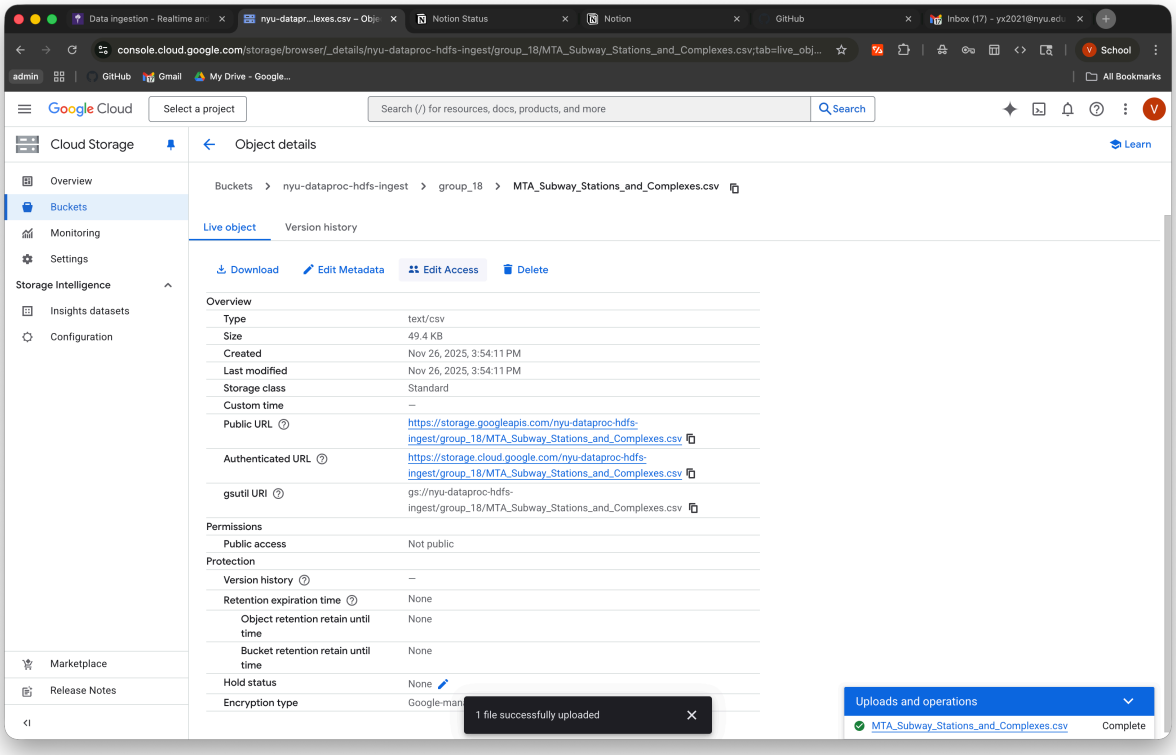


Then, the csv file is uploaded to https://dataproc.hpc.nyu.edu/ingest .



The file path could be seen from the screenshot above.

The secondary dataset was downloaded from source website and uploaded to the same bucket as shown below:



Processing

MR Job1: Cleaning

We are going to run one-map reduce job to clean the raw dataset. The map reduce job is going to achieve below goals:

1. Filter out records that is not in year 2024
2. Drop unneeded columns: `transit_mode` , `station_complex` , `transfers` , `borough` , `latitude` , `longitude` , and `georeference` .
3. Reformat the timestamp from `MM/dd/yyyy hh:mm:ss aa` to `yyyy-MM-dd HH:00:00` .
4. Fill in invalid values such as none, null, and 0.

Output Schema (5 fields):

> (timestamp, station_complex_id, payment_method, fare_class_category, ridership)

Output Location: `project/preprocessing/mta_processed/cleaned/`

MR Job 2: Aggregation

Another MapReduce job will be run to aggregate the cleaned data to produce hourly ridership totals:

```
Aggregation Logic:

transit_timestamp       - GROUP BY (part of composite key)
station_complex_id      - GROUP BY (part of composite key)
payment_method          - GROUP BY (part of composite key, OMNY or MetroCard)
fare_class_category     - Dropped (sum across all fare types)
ridership               - SUM of all fare types
```

Unique composite key would be:

```
(transit_timestamp, station_complex_id, payment_method)
```

Output Schema:

```
(timestamp, station_complex_id, payment_method, total_ridership)
```

Output Location: `project/preprocessing/mta_processed/station_hourly/`

Granularity: One row per station per hour per payment method

Example Output:

```
2024-12-28 20:00:00,475,metrocard,145
2024-12-28 20:00:00,475,omny,78
```

Script

In order to run the Map Reduce job in an organized way, the `Makefile` and script file is leveraged:

```
# Makefile for MTA MapReduce Pipeline

# Configuration
HADOOP_CLASSPATH = $(shell hadoop classpath)
CLASSES_DIR = classes
JAVA_FILES = MTAFilterClean.java MTAStationHourly.java
JAR_FILES = mta-filter-clean.jar mta-station-hourly.jar

# HDFS paths
OUTPUT_BASE = project/preprocessing/mta_processed
CLEANED_OUTPUT = $(OUTPUT_BASE)/cleaned
STATION_OUTPUT = $(OUTPUT_BASE)/station_hourly

# Default target
.PHONY: all
all: compile

# Compile all Java files and create JARs
.PHONY: compile
compile:
	@echo "====================================="
	@echo "Compiling MTA MapReduce jobs..."
	@echo "====================================="
	@mkdir -p $(CLASSES_DIR)
	@echo "Compiling MTAFilterClean.java (Cleaning Job)..."
	@javac -classpath $(HADOOP_CLASSPATH) -d $(CLASSES_DIR) MTAFilter
Clean.java
	@echo "Compiling MTAStationHourly.java (Aggregation Job)..."
	@javac -classpath $(HADOOP_CLASSPATH) -d $(CLASSES_DIR) MTAStatio
nHourly.java
	@echo ""
	@echo "Creating JAR files..."
	@cd $(CLASSES_DIR) && jar -cvf ../mta-filter-clean.jar MTAFilterClean*.clas
s
	@cd $(CLASSES_DIR) && jar -cvf ../mta-station-hourly.jar MTAStationHourl
```

```
y*.class
	@echo ""
	@echo "✓ Compilation complete!"
	@echo "Generated JAR files:"
	@ls -lh *.jar 2>/dev/null || echo "  No JAR files found"

# Run the complete pipeline
.PHONY: run
run: compile
	@echo ""
	@echo "========================================"
	@echo "Running MTA Preprocessing Pipeline"
	@echo "========================================"
	@./run_pipeline.sh
```

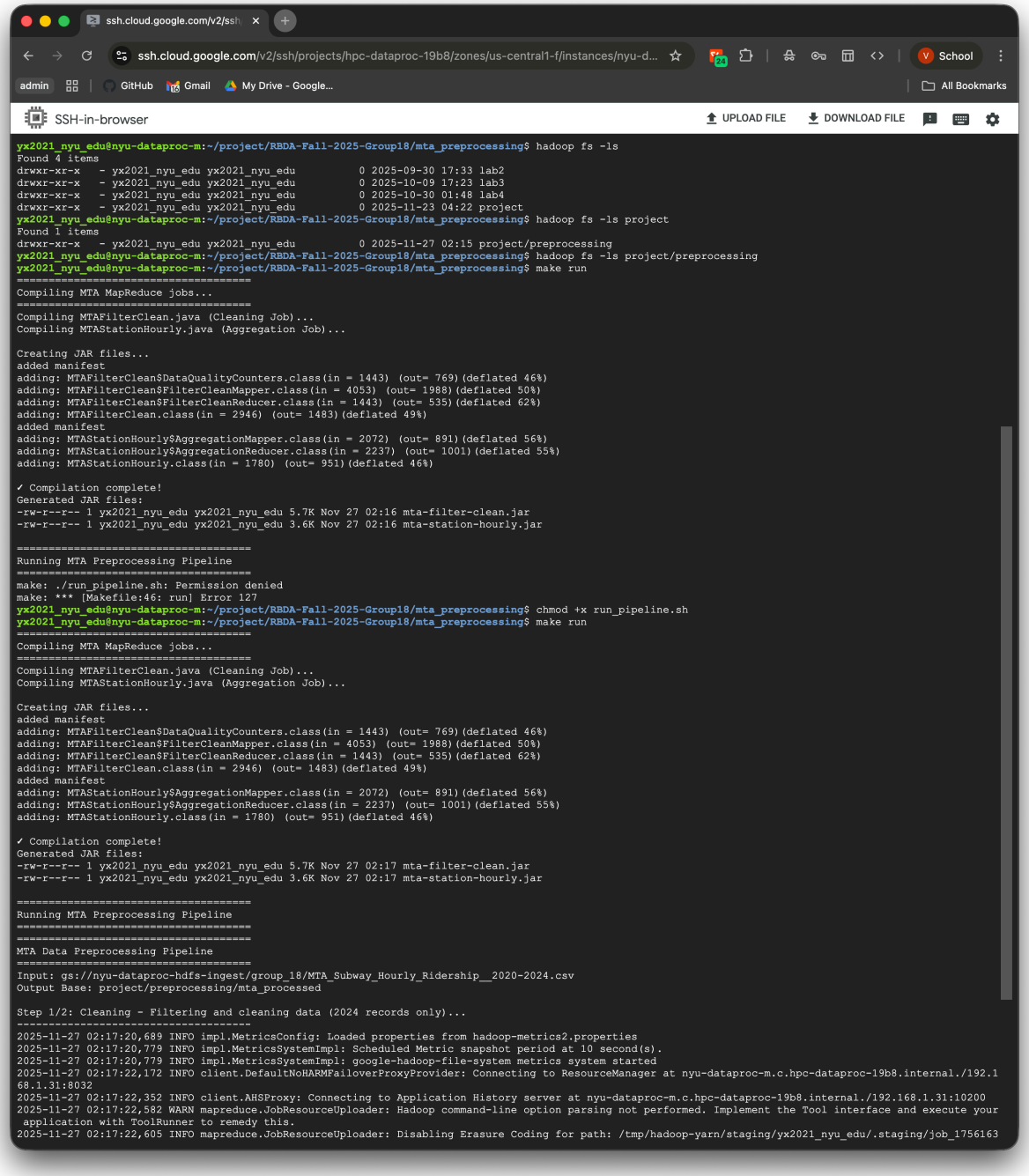
run_pipeline.sh

```
#!/bin/bash

# Complete MTA Data Preprocessing Pipeline
# This script runs both MapReduce jobs in sequence

# Configuration
INPUT_DATA="gs://nyu-dataproc-hdfs-ingest/group_18/MTA_Subway_Hourly_Ridership__2020-2024.csv"

# Output location - storing in HDFS under your project directory
OUTPUT_BASE="project/preprocessing/mta_processed"

# Output paths for each job
CLEANED_OUTPUT="${OUTPUT_BASE}/cleaned"
STATION_OUTPUT="${OUTPUT_BASE}/station_hourly"

echo "========================================"
echo "MTA Data Preprocessing Pipeline"
echo "========================================"
```

```
echo "Input: ${INPUT_DATA}"
echo "Output Base: ${OUTPUT_BASE}"
echo ""

# Cleaning Job: Filter and Clean (2024 data only)
echo "Step 1/2: Cleaning - Filtering and cleaning data (2024 records only)..."
echo "-------------------------------------"
hadoop jar mta-filter-clean.jar MTAFilterClean \
    ${INPUT_DATA} \
    ${CLEANED_OUTPUT}

if [ $? -ne 0 ]; then
    echo "ERROR: Cleaning job failed!"
    exit 1
fi
echo "✓ Step 1 complete: Cleaned data saved to ${CLEANED_OUTPUT}"
echo ""

# Aggregation Job: Station-Level Hourly Aggregation
echo "Step 2/2: Aggregation - Aggregating by station, hour, and payment met
hod..."
echo "-------------------------------------"
hadoop jar mta-station-hourly.jar MTAStationHourly \
    ${CLEANED_OUTPUT}/part-* \
    ${STATION_OUTPUT}

if [ $? -ne 0 ]; then
    echo "ERROR: Aggregation job failed!"
    exit 1
fi
echo "✓ Step 2 complete: Aggregated data saved to ${STATION_OUTPUT}"
echo ""

echo "====================================="
echo "Pipeline completed successfully!"
echo "====================================="
```

```
echo ""
echo "Output locations:"
echo "  1. Cleaned data (2024):     ${CLEANED_OUTPUT}"
echo "  2. Aggregated hourly data:  ${STATION_OUTPUT}"
echo ""
echo "Output schema:"
echo "  - Cleaned: timestamp, station_complex_id, payment_method, fare_class_category, ridership"
echo "  - Aggregated: timestamp, station_complex_id, payment_method, total_ridership"
echo ""
echo "Next steps:"
echo "  - Load aggregated output into Trino for querying"
echo "  - Join with MTA_Subway_Stations_and_Complexes.csv to get borough/lat/lon"
echo "  - Join with weather and crime data for analysis"
```

Screenshots


ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-d...
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$ hadoop fs -ls
Found 4 items
drwxr-xr-x - yx2021_nyu_edu yx2021_nyu_edu 0 2025-09-30 17:33 lab2
drwxr-xr-x - yx2021_nyu_edu yx2021_nyu_edu 0 2025-10-09 17:23 lab3
drwxr-xr-x - yx2021_nyu_edu yx2021_nyu_edu 0 2025-10-30 01:48 lab4
drwxr-xr-x - yx2021_nyu_edu yx2021_nyu_edu 0 2025-11-23 04:22 project
yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$ hadoop fs -ls project
Found 1 items
drwxr-xr-x - yx2021_nyu_edu yx2021_nyu_edu 0 2025-11-27 02:15 project/preprocessing
yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$ hadoop fs -ls project/preprocessing
yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$ make run
Compiling MTA MapReduce jobs...
Compiling MTAFilterClean.java (Cleaning Job)...
Compiling MTAStationHourly.java (Aggregation Job)...
Creating JAR files...
added manifest
adding: MTAFilterClean$DataQualityCounters.class(in = 1443) (out= 769)(deflated 46%)
adding: MTAFilterClean$FilterCleanMapper.class(in = 4053) (out= 1988)(deflated 50%)
adding: MTAFilterClean$FilterCleanReducer.class(in = 1443) (out= 535)(deflated 62%)
adding: MTAFilterClean.class(in = 2946) (out= 1483)(deflated 49%)
added manifest
adding: MTAStationHourly$AggregationMapper.class(in = 2072) (out= 891)(deflated 56%)
adding: MTAStationHourly$AggregationReducer.class(in = 2237) (out= 1001)(deflated 55%)
adding: MTAStationHourly.class(in = 1780) (out= 951)(deflated 46%)
✓ Compilation complete!
Generated JAR files:
-rw-r--r-- 1 yx2021_nyu_edu yx2021_nyu_edu 5.7K Nov 27 02:16 mta-filter-clean.jar
-rw-r--r-- 1 yx2021_nyu_edu yx2021_nyu_edu 3.6K Nov 27 02:16 mta-station-hourly.jar
Running MTA Preprocessing Pipeline
make: ./run_pipeline.sh: Permission denied
make: *** [Makefile:46: run] Error 127
yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$ chmod +x run_pipeline.sh
yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$ make run
Compiling MTA MapReduce jobs...
Compiling MTAFilterClean.java (Cleaning Job)...
Compiling MTAStationHourly.java (Aggregation Job)...
Creating JAR files...
added manifest
adding: MTAFilterClean$DataQualityCounters.class(in = 1443) (out= 769)(deflated 46%)
adding: MTAFilterClean$FilterCleanMapper.class(in = 4053) (out= 1988)(deflated 50%)
adding: MTAFilterClean$FilterCleanReducer.class(in = 1443) (out= 535)(deflated 62%)
adding: MTAFilterClean.class(in = 2946) (out= 1483)(deflated 49%)
added manifest
adding: MTAStationHourly$AggregationMapper.class(in = 2072) (out= 891)(deflated 56%)
adding: MTAStationHourly$AggregationReducer.class(in = 2237) (out= 1001)(deflated 55%)
adding: MTAStationHourly.class(in = 1780) (out= 951)(deflated 46%)
✓ Compilation complete!
Generated JAR files:
-rw-r--r-- 1 yx2021_nyu_edu yx2021_nyu_edu 5.7K Nov 27 02:17 mta-filter-clean.jar
-rw-r--r-- 1 yx2021_nyu_edu yx2021_nyu_edu 3.6K Nov 27 02:17 mta-station-hourly.jar
Running MTA Preprocessing Pipeline
MTA Data Preprocessing Pipeline
Input: gs://nyu-dataproc-hdfs-ingest/group_18/MTA_Subway_Hourly_Ridership__2020-2024.csv
Output Base: project/preprocessing/mta_processed
Step 1/2: Cleaning - Filtering and cleaning data (2024 records only)...
2025-11-27 02:17:20,689 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-11-27 02:17:20,779 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-11-27 02:17:20,779 INFO impl.MetricsSystemImpl: google-hadoop-file-system metrics system started
2025-11-27 02:17:22,172 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.31:8032
2025-11-27 02:17:22,352 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.31:10200
2025-11-27 02:17:22,582 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-11-27 02:17:22,605 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/yx2021_nyu_edu/.staging/job_1756163

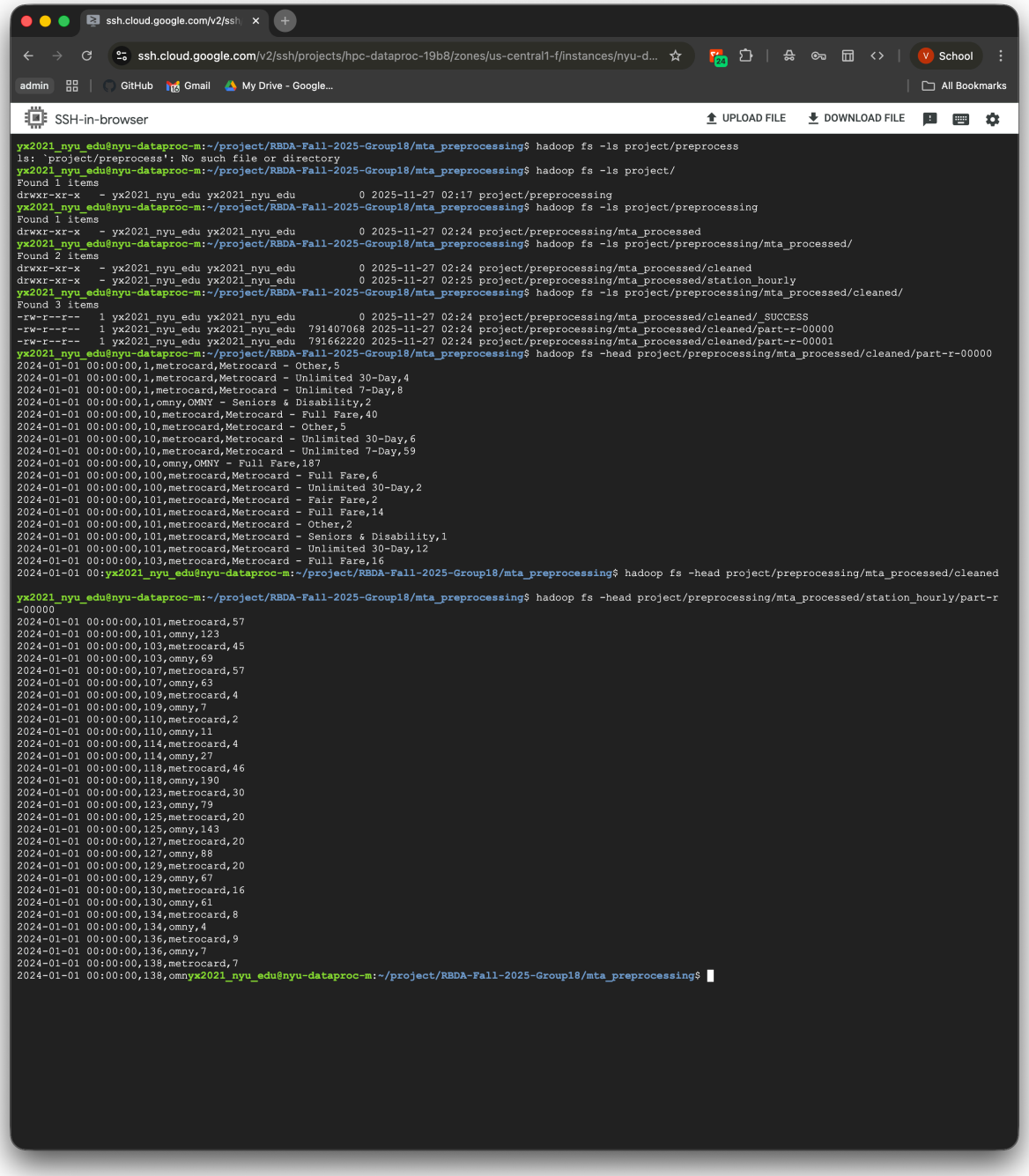

ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-d...
School
admin
GitHub
Gmail
My Drive - Google...
All Bookmarks
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
Step 1/2: Cleaning - Filtering and cleaning data (2024 records only)...
2025-11-27 02:17:20,689 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-11-27 02:17:20,779 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-11-27 02:17:20,779 INFO impl.MetricsSystemImpl: google-hadoop-file-system metrics system started
2025-11-27 02:17:22,172 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.31:8032
2025-11-27 02:17:22,352 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.31:10200
2025-11-27 02:17:22,582 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-11-27 02:17:22,605 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/yx2021_nyu_edu/.staging/job_1756163132607_14208
Nov 27, 2025 2:17:23 AM com.google.cloud.hadoop.fs.gcs.GhfsGlobalStorageStatistics trackDuration
INFO: periodic connector metrics: {gcs_api_time=360, gcs_api_total_request_count=2, gcs_connector_time=498, gcs_list_file_request=1, gcs_list_file_request_duration=179, gcs_list_file_request_max=179, gcs_list_file_request_mean=179, gcs_list_file_request_min=179, gcs_metadata_request=1, gcs_metadata_request_duration=181, gcs_metadata_request_max=181, gcs_metadata_request_mean=181, gcs_metadata_request_min=181, gs_filesystem_create=2, gs_filesystem_initialize=1, op_get_file_status=1, op_get_file_status_duration=498, op_get_file_status_max=498, op_get_file_status_mean=498, op_get_file_status_min=498, op_glob_status=1, uptimeSeconds=2} [CONTEXT ratelimit_period="5 MINUTES"]
2025-11-27 02:17:23,467 INFO input.FileInputFormat: Total input files to process : 1
2025-11-27 02:17:23,546 INFO mapreduce.JobSubmitter: number of splits:132
2025-11-27 02:17:23,742 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1756163132607_14208
2025-11-27 02:17:23,742 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-11-27 02:17:23,946 INFO conf.Configuration: resource-types.xml not found
2025-11-27 02:17:23,946 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-11-27 02:17:24,060 INFO impl.YarnClientImpl: Submitted application application_1756163132607_14208
2025-11-27 02:17:24,098 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m.c.hpc-dataproc-19b8.internal.:8088/proxy/application_1756163132607_14208/
2025-11-27 02:17:24,099 INFO mapreduce.Job: Running job: job_1756163132607_14208
2025-11-27 02:17:33,262 INFO mapreduce.Job: Job job_1756163132607_14208 running in uber mode : false
2025-11-27 02:17:33,264 INFO mapreduce.Job: map 0% reduce 0%
2025-11-27 02:18:10,641 INFO mapreduce.Job: map 1% reduce 0%
2025-11-27 02:18:14,669 INFO mapreduce.Job: map 2% reduce 0%
2025-11-27 02:18:20,704 INFO mapreduce.Job: map 3% reduce 0%
2025-11-27 02:18:23,722 INFO mapreduce.Job: map 4% reduce 0%
2025-11-27 02:18:28,750 INFO mapreduce.Job: map 5% reduce 0%
2025-11-27 02:18:32,772 INFO mapreduce.Job: map 6% reduce 0%
2025-11-27 02:18:35,794 INFO mapreduce.Job: map 7% reduce 0%
2025-11-27 02:18:40,824 INFO mapreduce.Job: map 8% reduce 0%
2025-11-27 02:18:44,845 INFO mapreduce.Job: map 9% reduce 0%
2025-11-27 02:18:47,861 INFO mapreduce.Job: map 10% reduce 0%
2025-11-27 02:18:52,891 INFO mapreduce.Job: map 11% reduce 0%
2025-11-27 02:18:56,914 INFO mapreduce.Job: map 12% reduce 0%
2025-11-27 02:18:59,941 INFO mapreduce.Job: map 13% reduce 0%
2025-11-27 02:19:04,967 INFO mapreduce.Job: map 14% reduce 0%
2025-11-27 02:19:08,987 INFO mapreduce.Job: map 15% reduce 0%
2025-11-27 02:19:12,026 INFO mapreduce.Job: map 16% reduce 0%
2025-11-27 02:19:14,036 INFO mapreduce.Job: map 17% reduce 0%
2025-11-27 02:19:15,042 INFO mapreduce.Job: map 18% reduce 0%
2025-11-27 02:19:17,060 INFO mapreduce.Job: map 19% reduce 0%
2025-11-27 02:19:20,076 INFO mapreduce.Job: map 20% reduce 0%
2025-11-27 02:19:21,086 INFO mapreduce.Job: map 21% reduce 0%
2025-11-27 02:19:22,097 INFO mapreduce.Job: map 22% reduce 0%
2025-11-27 02:19:23,103 INFO mapreduce.Job: map 23% reduce 0%
2025-11-27 02:19:25,113 INFO mapreduce.Job: map 24% reduce 0%
2025-11-27 02:19:31,149 INFO mapreduce.Job: map 25% reduce 0%
2025-11-27 02:19:36,173 INFO mapreduce.Job: map 26% reduce 0%
2025-11-27 02:19:39,187 INFO mapreduce.Job: map 27% reduce 0%
2025-11-27 02:19:41,207 INFO mapreduce.Job: map 28% reduce 0%
2025-11-27 02:19:48,266 INFO mapreduce.Job: map 29% reduce 0%
2025-11-27 02:19:59,346 INFO mapreduce.Job: map 30% reduce 0%
2025-11-27 02:20:04,370 INFO mapreduce.Job: map 31% reduce 0%
2025-11-27 02:20:08,387 INFO mapreduce.Job: map 32% reduce 0%
2025-11-27 02:20:13,407 INFO mapreduce.Job: map 33% reduce 0%
2025-11-27 02:20:18,431 INFO mapreduce.Job: map 34% reduce 0%
2025-11-27 02:20:22,451 INFO mapreduce.Job: map 35% reduce 0%
2025-11-27 02:20:25,463 INFO mapreduce.Job: map 36% reduce 0%
2025-11-27 02:20:30,483 INFO mapreduce.Job: map 37% reduce 0%
2025-11-27 02:20:33,494 INFO mapreduce.Job: map 38% reduce 0%
2025-11-27 02:20:37,510 INFO mapreduce.Job: map 39% reduce 0%
2025-11-27 02:20:42,530 INFO mapreduce.Job: map 40% reduce 0%
2025-11-27 02:20:45,542 INFO mapreduce.Job: map 41% reduce 0%
2025-11-27 02:20:49,558 INFO mapreduce.Job: map 42% reduce 0%
2025-11-27 02:20:52,570 INFO mapreduce.Job: map 43% reduce 0%
2025-11-27 02:20:53,579 INFO mapreduce.Job: map 44% reduce 0%
2025-11-27 02:20:56,590 INFO mapreduce.Job: map 45% reduce 0%
2025-11-27 02:20:58,598 INFO mapreduce.Job: map 46% reduce 0%
2025-11-27 02:21:01,614 INFO mapreduce.Job: map 47% reduce 0%
2025-11-27 02:21:02,618 INFO mapreduce.Job: map 48% reduce 0%
2025-11-27 02:21:06,634 INFO mapreduce.Job: map 49% reduce 0%
2025-11-27 02:21:08,641 INFO mapreduce.Job: map 50% reduce 0%
2025-11-27 02:21:11,653 INFO mapreduce.Job: map 51% reduce 0%
2025-11-27 02:21:14,663 INFO mapreduce.Job: map 52% reduce 0%


ssh.cloud.google.com/v2/ssh
ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-d...
School
admin
GitHub
Gmail
My Drive - Google...
All Bookmarks
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
2025-11-27 02:21:11,653 INFO mapreduce.Job: map 51% reduce 0%
2025-11-27 02:21:14,663 INFO mapreduce.Job: map 52% reduce 0%
2025-11-27 02:21:18,683 INFO mapreduce.Job: map 53% reduce 0%
2025-11-27 02:21:23,711 INFO mapreduce.Job: map 54% reduce 0%
2025-11-27 02:21:27,729 INFO mapreduce.Job: map 55% reduce 0%
2025-11-27 02:21:29,736 INFO mapreduce.Job: map 56% reduce 0%
2025-11-27 02:21:32,759 INFO mapreduce.Job: map 57% reduce 0%
2025-11-27 02:21:36,773 INFO mapreduce.Job: map 58% reduce 0%
2025-11-27 02:21:44,815 INFO mapreduce.Job: map 59% reduce 0%
2025-11-27 02:21:48,830 INFO mapreduce.Job: map 60% reduce 0%
2025-11-27 02:21:53,857 INFO mapreduce.Job: map 61% reduce 0%
2025-11-27 02:21:58,875 INFO mapreduce.Job: map 62% reduce 0%
2025-11-27 02:22:02,890 INFO mapreduce.Job: map 63% reduce 0%
2025-11-27 02:22:06,906 INFO mapreduce.Job: map 64% reduce 0%
2025-11-27 02:22:10,919 INFO mapreduce.Job: map 65% reduce 0%
2025-11-27 02:22:15,936 INFO mapreduce.Job: map 66% reduce 0%
2025-11-27 02:22:18,947 INFO mapreduce.Job: map 67% reduce 0%
2025-11-27 02:22:22,960 INFO mapreduce.Job: map 68% reduce 0%
2025-11-27 02:22:27,978 INFO mapreduce.Job: map 69% reduce 0%
2025-11-27 02:22:29,989 INFO mapreduce.Job: map 70% reduce 0%
2025-11-27 02:22:31,998 INFO mapreduce.Job: map 71% reduce 0%
2025-11-27 02:22:35,010 INFO mapreduce.Job: map 73% reduce 0%
2025-11-27 02:22:39,025 INFO mapreduce.Job: map 74% reduce 0%
2025-11-27 02:22:43,039 INFO mapreduce.Job: map 75% reduce 0%
2025-11-27 02:22:46,051 INFO mapreduce.Job: map 76% reduce 0%
2025-11-27 02:22:51,069 INFO mapreduce.Job: map 77% reduce 0%
2025-11-27 02:22:54,101 INFO mapreduce.Job: map 78% reduce 0%
2025-11-27 02:22:56,110 INFO mapreduce.Job: map 79% reduce 0%
2025-11-27 02:22:59,130 INFO mapreduce.Job: map 80% reduce 0%
2025-11-27 02:23:02,142 INFO mapreduce.Job: map 81% reduce 0%
2025-11-27 02:23:06,159 INFO mapreduce.Job: map 82% reduce 0%
2025-11-27 02:23:10,172 INFO mapreduce.Job: map 83% reduce 0%
2025-11-27 02:23:14,187 INFO mapreduce.Job: map 84% reduce 0%
2025-11-27 02:23:17,197 INFO mapreduce.Job: map 85% reduce 0%
2025-11-27 02:23:19,204 INFO mapreduce.Job: map 86% reduce 0%
2025-11-27 02:23:22,215 INFO mapreduce.Job: map 87% reduce 0%
2025-11-27 02:23:27,232 INFO mapreduce.Job: map 88% reduce 0%
2025-11-27 02:23:32,249 INFO mapreduce.Job: map 89% reduce 0%
2025-11-27 02:23:38,269 INFO mapreduce.Job: map 90% reduce 0%
2025-11-27 02:23:43,287 INFO mapreduce.Job: map 91% reduce 0%
2025-11-27 02:23:47,302 INFO mapreduce.Job: map 92% reduce 0%
2025-11-27 02:23:54,325 INFO mapreduce.Job: map 93% reduce 0%
2025-11-27 02:23:56,333 INFO mapreduce.Job: map 94% reduce 0%
2025-11-27 02:24:01,351 INFO mapreduce.Job: map 95% reduce 0%
2025-11-27 02:24:03,358 INFO mapreduce.Job: map 96% reduce 0%
2025-11-27 02:24:07,371 INFO mapreduce.Job: map 97% reduce 0%
2025-11-27 02:24:10,382 INFO mapreduce.Job: map 98% reduce 0%
2025-11-27 02:24:13,392 INFO mapreduce.Job: map 99% reduce 0%
2025-11-27 02:24:19,413 INFO mapreduce.Job: map 100% reduce 0%
2025-11-27 02:24:29,445 INFO mapreduce.Job: map 100% reduce 35%
2025-11-27 02:24:30,448 INFO mapreduce.Job: map 100% reduce 74%
2025-11-27 02:24:35,465 INFO mapreduce.Job: map 100% reduce 85%
2025-11-27 02:24:36,468 INFO mapreduce.Job: map 100% reduce 96%
2025-11-27 02:24:37,472 INFO mapreduce.Job: map 100% reduce 100%
2025-11-27 02:24:38,482 INFO mapreduce.Job: Job job_1756163132607_14208 completed successfully
2025-11-27 02:24:38,594 INFO mapreduce.Job: Counters: 63
File System Counters
FILE: Number of bytes read=1637094326
FILE: Number of bytes written=3312922108
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
GS: Number of bytes read=17729827138
GS: Number of bytes written=0
GS: Number of read operations=4328572
GS: Number of large read operations=0
GS: Number of write operations=0
HDFS: Number of bytes read=19272
HDFS: Number of bytes written=1583069288
HDFS: Number of read operations=274
HDFS: Number of large read operations=0
HDFS: Number of write operations=6
HDFS: Number of bytes read erasure-coded=0
Job Counters
Killed map tasks=4
Launched map tasks=136
Launched reduce tasks=2
Rack-local map tasks=136
Total time spent by all maps in occupied slots (ms)=54623724
Total time spent by all reduces in occupied slots (ms)=193264
Total time spent by all map tasks (ms)=13655931
Total time spent by all reduce tasks (ms)=48316


ssh.cloud.google.com/v2/ssh
ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-d...
School
admin
GitHub
Gmail
My Drive - Google...
All Bookmarks
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
Job Counters
Killed map tasks=4
Launched map tasks=136
Launched reduce tasks=2
Rack-local map tasks=136
Total time spent by all maps in occupied slots (ms)=54623724
Total time spent by all reduces in occupied slots (ms)=193264
Total time spent by all map tasks (ms)=13655931
Total time spent by all reduce tasks (ms)=48316
Total vcore-milliseconds taken by all map tasks=13655931
Total vcore-milliseconds taken by all reduce tasks=48316
Total megabyte-milliseconds taken by all map tasks=55934693376
Total megabyte-milliseconds taken by all reduce tasks=197902336
Map-Reduce Framework
Map input records=120855568
Map output records=27012513
Map output bytes=1583069288
Map output materialized bytes=1637095898
Input split bytes=19272
Combine input records=0
Combine output records=0
Reduce input groups=27012513
Reduce shuffle bytes=1637095898
Reduce input records=27012513
Reduce output records=27012513
Spilled Records=54025026
Shuffled Maps =264
Failed Shuffles=0
Merged Map outputs=264
GC time elapsed (ms)=125291
CPU time spent (ms)=12967750
Physical memory (bytes) snapshot=158393806848
Virtual memory (bytes) snapshot=675374002176
Total committed heap usage (bytes)=133131403264
Peak Map Physical memory (bytes)=1342824448
Peak Map Virtual memory (bytes)=5117521920
Peak Reduce Physical memory (bytes)=1255378944
Peak Reduce Virtual memory (bytes)=5022560256
MTAFilterClean$DataQualityCounters
FILTERED_WRONG_YEAR=93843054
TOTAL_RECORDS=120855567
VALID_RECORDS=27012513
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=17729827138
File Output Format Counters
Bytes Written=1583069288
=== DATA QUALITY REPORT ===
FILTERED_WRONG_YEAR: 93843054
TOTAL_RECORDS: 120855567
VALID_RECORDS: 27012513
✓ Step 1 complete: Cleaned data saved to project/preprocessing/mta_processed/cleaned
Step 2/2: Aggregation - Aggregating by station, hour, and payment method...
2025-11-27 02:24:41,923 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.31:8032
2025-11-27 02:24:42,270 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.31:10200
2025-11-27 02:24:42,606 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-11-27 02:24:42,623 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/yx2021_nyu_edu/.staging/job_1756163132607_14209
2025-11-27 02:24:43,000 INFO input.FileInputFormat: Total input files to process : 2
2025-11-27 02:24:43,084 INFO mapreduce.JobSubmitter: number of splits:12
2025-11-27 02:24:43,289 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1756163132607_14209
2025-11-27 02:24:43,289 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-11-27 02:24:43,492 INFO conf.Configuration: resource-types.xml not found
2025-11-27 02:24:43,492 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-11-27 02:24:43,603 INFO impl.YarnClientImpl: Submitted application application_1756163132607_14209
2025-11-27 02:24:43,639 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m.c.hpc-dataproc-19b8.internal.:8088/proxy/application_1756163132607_14209/
2025-11-27 02:24:43,640 INFO mapreduce.Job: Running job: job_1756163132607_14209
2025-11-27 02:24:52,786 INFO mapreduce.Job: Job job_1756163132607_14209 running in uber mode : false
2025-11-27 02:24:52,788 INFO mapreduce.Job: map 0% reduce 0%
2025-11-27 02:25:23,049 INFO mapreduce.Job: map 36% reduce 0%


ssh.cloud.google.com/v2/ssh
ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-d...
School
admin
GitHub
Gmail
My Drive - Google...
All Bookmarks
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
Step 2/2: Aggregation - Aggregating by station, hour, and payment method...
2025-11-27 02:24:41,923 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.31:8032
2025-11-27 02:24:42,270 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.31:10200
2025-11-27 02:24:42,606 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-11-27 02:24:42,623 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/yx2021_nyu_edu/.staging/job_1756163132607_14209
2025-11-27 02:24:43,000 INFO input.FileInputFormat: Total input files to process : 2
2025-11-27 02:24:43,084 INFO mapreduce.JobSubmitter: number of splits:12
2025-11-27 02:24:43,289 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1756163132607_14209
2025-11-27 02:24:43,289 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-11-27 02:24:43,492 INFO conf.Configuration: resource-types.xml not found
2025-11-27 02:24:43,492 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-11-27 02:24:43,603 INFO impl.YarnClientImpl: Submitted application application_1756163132607_14209
2025-11-27 02:24:43,639 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m.c.hpc-dataproc-19b8.internal.:8088/proxy/application_1756163132607_14209/
2025-11-27 02:24:43,640 INFO mapreduce.Job: Running job: job_1756163132607_14209
2025-11-27 02:24:52,786 INFO mapreduce.Job: Job job_1756163132607_14209 running in uber mode : false
2025-11-27 02:24:52,788 INFO mapreduce.Job: map 0% reduce 0%
2025-11-27 02:25:23,049 INFO mapreduce.Job: map 36% reduce 0%
2025-11-27 02:25:24,056 INFO mapreduce.Job: map 81% reduce 0%
2025-11-27 02:25:25,062 INFO mapreduce.Job: map 94% reduce 0%
2025-11-27 02:25:26,068 INFO mapreduce.Job: map 100% reduce 0%
2025-11-27 02:25:46,178 INFO mapreduce.Job: map 100% reduce 74%
2025-11-27 02:25:52,211 INFO mapreduce.Job: map 100% reduce 92%
2025-11-27 02:25:54,222 INFO mapreduce.Job: map 100% reduce 95%
2025-11-27 02:25:55,228 INFO mapreduce.Job: map 100% reduce 100%
2025-11-27 02:25:56,243 INFO mapreduce.Job: Job job_1756163132607_14209 completed successfully
2025-11-27 02:25:56,364 INFO mapreduce.Job: Counters: 56
File System Counters
FILE: Number of bytes read=1005242440
FILE: Number of bytes written=2014537114
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1583112204
HDFS: Number of bytes written=250669519
HDFS: Number of read operations=46
HDFS: Number of large read operations=0
HDFS: Number of write operations=6
HDFS: Number of bytes read erasure-coded=0
Job Counters
Killed map tasks=1
Launched map tasks=12
Launched reduce tasks=2
Data-local map tasks=5
Rack-local map tasks=7
Total time spent by all maps in occupied slots (ms)=1400488
Total time spent by all reduces in occupied slots (ms)=209176
Total time spent by all map tasks (ms)=350122
Total time spent by all reduce tasks (ms)=52294
Total vcore-milliseconds taken by all map tasks=350122
Total vcore-milliseconds taken by all reduce tasks=52294
Total megabyte-milliseconds taken by all map tasks=1434099712
Total megabyte-milliseconds taken by all reduce tasks=214196224
Map-Reduce Framework
Map input records=27012513
Map output records=27012513
Map output bytes=951217402
Map output materialized bytes=1005242572
Input split bytes=1956
Combine input records=0
Combine output records=0
Reduce input groups=7255913
Reduce shuffle bytes=1005242572
Reduce input records=27012513
Reduce output records=7255913
Spilled Records=54025026
Shuffled Maps =24
Failed Shuffles=0
Merged Map outputs=24
GC time elapsed (ms)=6008
CPU time spent (ms)=236970
Physical memory (bytes) snapshot=15578419200
Virtual memory (bytes) snapshot=69968683008
Total committed heap usage (bytes)=13828620288
Peak Map Physical memory (bytes)=1126318080
Peak Map Virtual memory (bytes)=5009588224
Peak Reduce Physical memory (bytes)=1225146368
Peak Reduce Virtual memory (bytes)=4998754304


ssh.cloud.google.com/v2/ssh
ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-d...
School
admin
GitHub
Gmail
My Drive - Google...
All Bookmarks
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
2025-11-27 02:25:56,364 INFO mapreduce.Job: Counters: 56
File System Counters
FILE: Number of bytes read=1005242440
FILE: Number of bytes written=2014537114
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1583112204
HDFS: Number of bytes written=250669519
HDFS: Number of read operations=46
HDFS: Number of large read operations=0
HDFS: Number of write operations=6
HDFS: Number of bytes read erasure-coded=0
Job Counters
Killed map tasks=1
Launched map tasks=12
Launched reduce tasks=2
Data-local map tasks=5
Rack-local map tasks=7
Total time spent by all maps in occupied slots (ms)=1400488
Total time spent by all reduces in occupied slots (ms)=209176
Total time spent by all map tasks (ms)=350122
Total time spent by all reduce tasks (ms)=52294
Total vcore-milliseconds taken by all map tasks=350122
Total vcore-milliseconds taken by all reduce tasks=52294
Total megabyte-milliseconds taken by all map tasks=1434099712
Total megabyte-milliseconds taken by all reduce tasks=214196224
Map-Reduce Framework
Map input records=27012513
Map output records=27012513
Map output bytes=951217402
Map output materialized bytes=1005242572
Input split bytes=1956
Combine input records=0
Combine output records=0
Reduce input groups=7255913
Reduce shuffle bytes=1005242572
Reduce input records=27012513
Reduce output records=7255913
Spilled Records=54025026
Shuffled Maps =24
Failed Shuffles=0
Merged Map outputs=24
GC time elapsed (ms)=6008
CPU time spent (ms)=236970
Physical memory (bytes) snapshot=15578419200
Virtual memory (bytes) snapshot=69968683008
Total committed heap usage (bytes)=13828620288
Peak Map Physical memory (bytes)=1126318080
Peak Map Virtual memory (bytes)=5009588224
Peak Reduce Physical memory (bytes)=1225146368
Peak Reduce Virtual memory (bytes)=4998754304
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=1583110248
File Output Format Counters
Bytes Written=250669519
✓ Step 2 complete: Aggregated data saved to project/preprocessing/mta_processed/station_hourly
==================================
Pipeline completed successfully!
==================================
Output locations:
1. Cleaned data (2024): project/preprocessing/mta_processed/cleaned
2. Aggregated hourly data: project/preprocessing/mta_processed/station_hourly
Output schema:
- Cleaned: timestamp, station_complex_id, payment_method, fare_class_category, ridership
- Aggregated: timestamp, station_complex_id, payment_method, total_ridership
Next steps:
- Load aggregated output into Trino for querying
- Join with MTA_Subway_Stations_and_Complexes.csv to get borough/lat/lon
- Join with weather and crime data for analysis
yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$


ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-d...
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$ hadoop fs -ls project/preprocess
ls: `project/preprocess': No such file or directory
yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$ hadoop fs -ls project/
Found 1 items
drwxr-xr-x - yx2021_nyu_edu yx2021_nyu_edu 0 2025-11-27 02:17 project/preprocessing
yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$ hadoop fs -ls project/preprocessing
Found 1 items
drwxr-xr-x - yx2021_nyu_edu yx2021_nyu_edu 0 2025-11-27 02:24 project/preprocessing/mta_processed
yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$ hadoop fs -ls project/preprocessing/mta_processed/
Found 2 items
drwxr-xr-x - yx2021_nyu_edu yx2021_nyu_edu 0 2025-11-27 02:24 project/preprocessing/mta_processed/cleaned
drwxr-xr-x - yx2021_nyu_edu yx2021_nyu_edu 0 2025-11-27 02:25 project/preprocessing/mta_processed/station_hourly
yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$ hadoop fs -ls project/preprocessing/mta_processed/cleaned/
Found 3 items
-rw-r--r-- 1 yx2021_nyu_edu yx2021_nyu_edu 0 2025-11-27 02:24 project/preprocessing/mta_processed/cleaned/_SUCCESS
-rw-r--r-- 1 yx2021_nyu_edu yx2021_nyu_edu 791407068 2025-11-27 02:24 project/preprocessing/mta_processed/cleaned/part-r-00000
-rw-r--r-- 1 yx2021_nyu_edu yx2021_nyu_edu 791662220 2025-11-27 02:24 project/preprocessing/mta_processed/cleaned/part-r-00001
yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$ hadoop fs -head project/preprocessing/mta_processed/cleaned/part-r-00000
2024-01-01 00:00:00,1,metrocard,Metrocard - Other,5
2024-01-01 00:00:00,1,metrocard,Metrocard - Unlimited 30-Day,4
2024-01-01 00:00:00,1,metrocard,Metrocard - Unlimited 7-Day,8
2024-01-01 00:00:00,1,omny,OMNY - Seniors & Disability,2
2024-01-01 00:00:00,10,metrocard,Metrocard - Full Fare,40
2024-01-01 00:00:00,10,metrocard,Metrocard - Other,5
2024-01-01 00:00:00,10,metrocard,Metrocard - Unlimited 30-Day,6
2024-01-01 00:00:00,10,metrocard,Metrocard - Unlimited 7-Day,59
2024-01-01 00:00:00,10,omny,OMNY - Full Fare,187
2024-01-01 00:00:00,100,metrocard,Metrocard - Full Fare,6
2024-01-01 00:00:00,100,metrocard,Metrocard - Unlimited 30-Day,2
2024-01-01 00:00:00,101,metrocard,Metrocard - Fair Fare,2
2024-01-01 00:00:00,101,metrocard,Metrocard - Full Fare,14
2024-01-01 00:00:00,101,metrocard,Metrocard - Other,2
2024-01-01 00:00:00,101,metrocard,Metrocard - Seniors & Disability,1
2024-01-01 00:00:00,101,metrocard,Metrocard - Unlimited 30-Day,12
2024-01-01 00:00:00,103,metrocard,Metrocard - Full Fare,16
2024-01-01 00:yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$ hadoop fs -head project/preprocessing/mta_processed/cleaned
yx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$ hadoop fs -head project/preprocessing/mta_processed/station_hourly/part-r-00000
2024-01-01 00:00:00,101,metrocard,57
2024-01-01 00:00:00,101,omny,123
2024-01-01 00:00:00,103,metrocard,45
2024-01-01 00:00:00,103,omny,69
2024-01-01 00:00:00,107,metrocard,57
2024-01-01 00:00:00,107,omny,63
2024-01-01 00:00:00,109,metrocard,4
2024-01-01 00:00:00,109,omny,7
2024-01-01 00:00:00,110,metrocard,2
2024-01-01 00:00:00,110,omny,11
2024-01-01 00:00:00,114,metrocard,4
2024-01-01 00:00:00,114,omny,27
2024-01-01 00:00:00,118,metrocard,46
2024-01-01 00:00:00,118,omny,190
2024-01-01 00:00:00,123,metrocard,30
2024-01-01 00:00:00,123,omny,79
2024-01-01 00:00:00,125,metrocard,20
2024-01-01 00:00:00,125,omny,143
2024-01-01 00:00:00,127,metrocard,20
2024-01-01 00:00:00,127,omny,88
2024-01-01 00:00:00,129,metrocard,20
2024-01-01 00:00:00,129,omny,67
2024-01-01 00:00:00,130,metrocard,16
2024-01-01 00:00:00,130,omny,61
2024-01-01 00:00:00,134,metrocard,8
2024-01-01 00:00:00,134,omny,4
2024-01-01 00:00:00,136,metrocard,9
2024-01-01 00:00:00,136,omny,7
2024-01-01 00:00:00,138,metrocard,7
2024-01-01 00:00:00,138,omnyx2021_nyu_edu@nyu-dataproc-m:~/project/RBDA-Fall-2025-Group18/mta_preprocessing$