RBDA Project Part 1: Data Ingestion

Name: Aygun Najafova **NetID:** an4758 **Group:** 21

Idea

As a group, we are working on analysing important data points, including NYC restaurant inspection data, NYPD arrest data, NYC hotel reviews, and MTA transit data. From the above data sources, my dataset is the NYC Inspection data, and I have completed the data ingestion phase for the same. It is explained in more detail in the next sections.

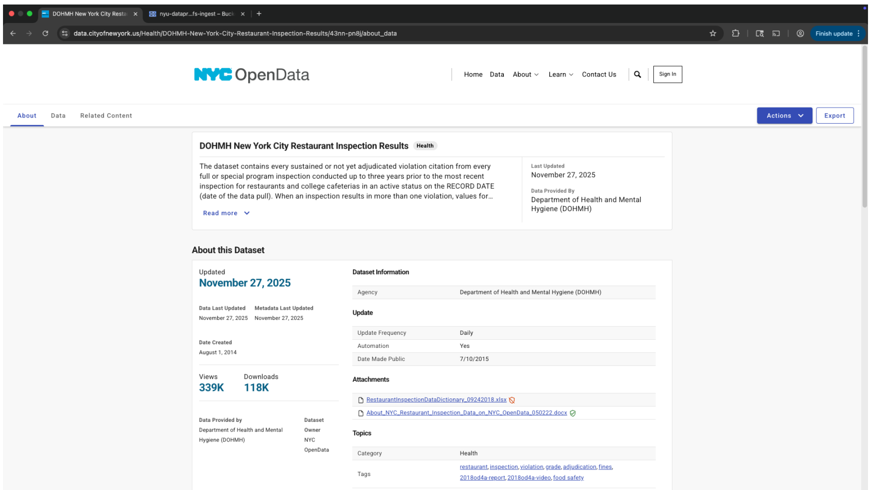
Data source

I am using the `DOHMH New York City Restaurant Inspection Results` dataset available on the NYC OpenData website.[1] As shown on the dataset's webpage, it contains detailed information about restaurant inspections in New York City, including violations and assigned grades, collected over more than ten years.

According to the website, the dataset contains over 291,000 rows and 27 columns. Each row is a single restaurant inspection record, and the columns have information on restaurant details, location, scoring and some additional information. These columns are explained below:

- **CAMIS**: Its a unique restaurant identifier stored as a text field.
- **DBA**: It is the restaurant name, represented as text.
- **BORO**: It is the borough name, stored as a text value.
- **BUILDING**: It is the number of the building in text format.
- **STREET**: It is the Street name recorded as text.
- **ZIPCODE**: The ZIP code of the restaurant as text.
- **PHONE**: The phone number of the place stored as text.
- **CUISINE DESCRIPTION**: The category of cuisine represented as text.
- **INSPECTION DATE**: The inspection date stored as a timestamp.
- **ACTION**: The outcome of the inspection stored as text.
- **VIOLATION CODE**: The code to identify violation, stored as text.
- **VIOLATION DESCRIPTION**: The description of the violation, stored as text.
- **CRITICAL FLAG**: Severity of the violation, represented as text.
- **SCORE**: The score of the inspection stored as an integer.
- **GRADE**: Grades for the inspection (A, B, C) represented as text.
- **GRADE DATE**: The date the grade was issued, as a timestamp.
- **RECORD DATE**: The date the record was entered or updated as a timestamp.
- **INSPECTION TYPE**: The type of inspection stored as text.
- **Latitude**: The latitude coordinate stored as a decimal number.
- **Longitude**: The longitude coordinate stored as a decimal number.
- Other Metadata Fields like `Community Board`, `Council District`, `Census Tract`, `BIN`, `BBL`, `NTA`, and `Location` that represent geographic identifiers and are not important, so I will remove these columns in my preprocessing phase.

[1] Dataset available at: `https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j`


DOHMH New York City Resta
nyu-datapr...fs-ingest – Buck
data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j/about_data
Finish update
NYC OpenData
Home
Data
About
Learn
Contact Us
Sign In
About
Data
Related Content
Actions
Export
Read more

DOHMH New York City Restaurant Inspection Results Health

The dataset contains every sustained or not yet adjudicated violation citation from every full or special program inspection conducted up to three years prior to the most recent inspection for restaurants and college cafeterias in an active status on the RECORD DATE (date of the data pull). When an inspection results in more than one violation, values for...

Last Updated
November 27, 2025

Data Provided By
Department of Health and Mental Hygiene (DOHMH)

About this Dataset

Updated
November 27, 2025

Data Last Updated
November 27, 2025

Metadata Last Updated
November 27, 2025

Date Created
August 1, 2014

Views
339K

Downloads
118K

Data Provided by
Department of Health and Mental Hygiene (DOHMH)

Dataset Owner
NYC OpenData

Dataset Information

Agency	Department of Health and Mental Hygiene (DOHMH)

Update

Update Frequency	Daily
Automation	Yes
Date Made Public	7/10/2015

Attachments

- RestaurantInspectionDataDictionary_09242018.xlsx
- About_NYC_Restaurant_Inspection_Data_on_NYC_OpenData_050222.docx

Topics

Category	Health
Tags	restaurant, inspection, violation, grade, adjudication, fines, 2018od4a-report, 2018od4a-video, food safety

Figure 1: Data source on NYC OpenData

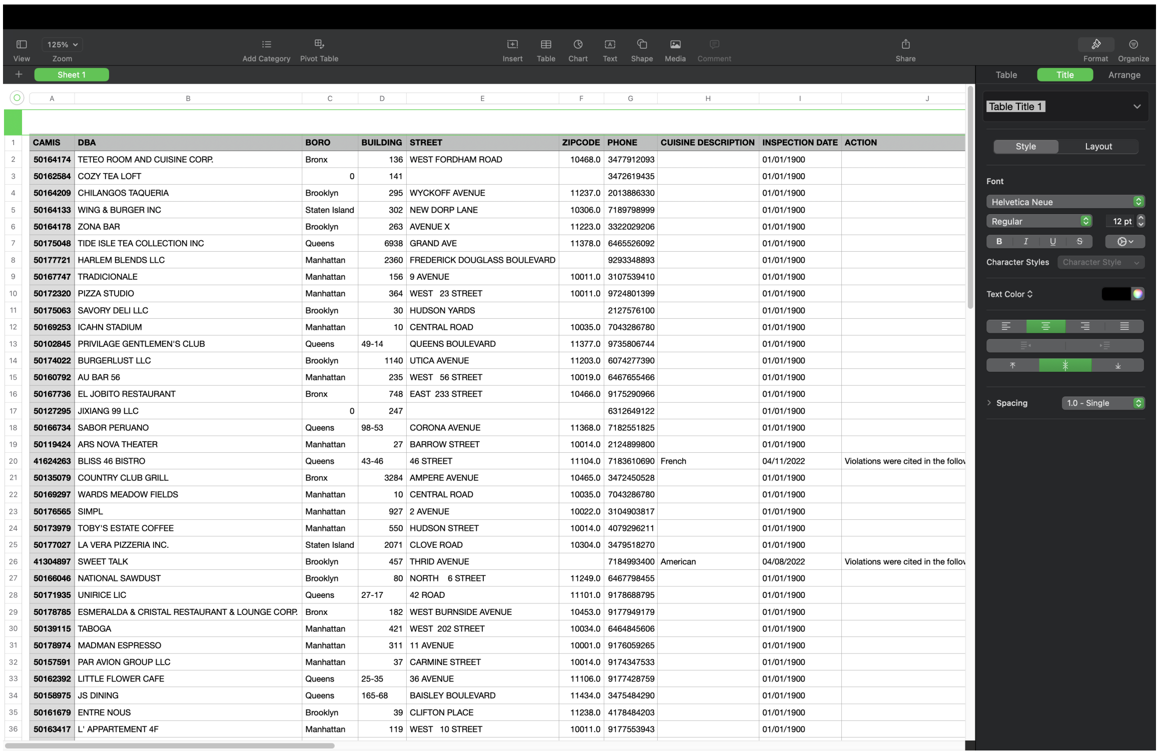
Data analysis

To begin my analysis, I selected the first 100 rows of the dataset for a quick visual inspection. This was done using the following command:

```
head -n 101 NYC_Restaurent_Inscp_Data.csv > NYC_Restaurent_Inscp_Data_First100.csv
```

After opening the file in a spreadsheet, I observed several missing values in different columns. Some records had empty `CUISINE DESCRIPTION` fields, missing ZIPCODE, and even missing borough information. Also, important columns such as `ZIPCODE` and `INSPECTION DATE` were present which makes it suitable to index the data for future study of trends in the subsequent step of the project.

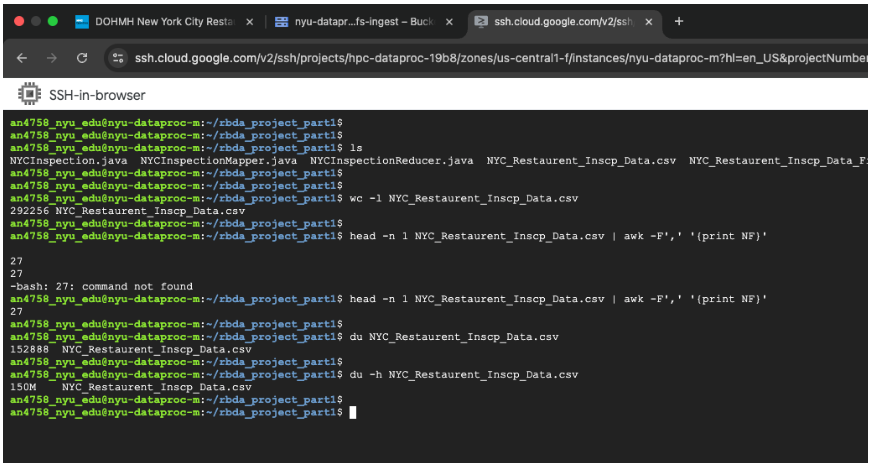
Additionally, I noticed that the dataset contains very old inspection dates, with some records dating back to the early 1900s. These entries are not useful to analyse recent trends. I am interested in recent and frequently monitored restaurants. Thus, I will only consider records from the year 2015 onwards.



CAMIS	DBA	BORO	BUILDING	STREET	ZIPCODE	PHONE	CUISINE DESCRIPTION	INSPECTION DATE	ACTION
50164174	TETEO ROOM AND CUISINE CORP.	Bronx	136	WEST FORDHAM ROAD	10468.0	3477912093		01/01/1900	
50162584	COZY TEA LOFT	0	141			3472619435		01/01/1900	
50164209	CHILANGOS TAQUERIA	Brooklyn	295	WYCKOFF AVENUE	11237.0	2013886330		01/01/1900	
50164133	WING & BURGER INC	Staten Island	302	NEW DORP LANE	10306.0	7189798999		01/01/1900	
50164178	ZONA BAR	Brooklyn	263	AVENUE X	11223.0	3322029206		01/01/1900	
50175048	TIDE ISLE TEA COLLECTION INC	Queens	6938	GRAND AVE	11378.0	6465526092		01/01/1900	
50177721	HARLEM BLENDS LLC	Manhattan	2360	FREDERICK DOUGLASS BOULEVARD		9293348893		01/01/1900	
50167747	TRADICIONALE	Manhattan	156	9 AVENUE	10011.0	3107539410		01/01/1900	
50172320	PIZZA STUDIO	Manhattan	364	WEST 23 STREET	10011.0	9724801399		01/01/1900	
50175063	SAVORY DELI LLC	Brooklyn	30	HUDSON YARDS		2127576100		01/01/1900	
50169253	ICAHN STADIUM	Manhattan	10	CENTRAL ROAD	10035.0	7043286780		01/01/1900	
50102845	PRIVILAGE GENTLEMEN'S CLUB	Queens	49-14	QUEENS BOULEVARD	11377.0	9735806744		01/01/1900	
50174022	BURGERLUST LLC	Brooklyn	1140	UTICA AVENUE	11203.0	6074277390		01/01/1900	
50160792	AU BAR 56	Manhattan	235	WEST 56 STREET	10019.0	6467655466		01/01/1900	
50167736	EL JOBITO RESTAURANT	Bronx	748	EAST 233 STREET	10466.0	9175290966		01/01/1900	
50127295	JIXIANG 99 LLC	0	247			6312649122		01/01/1900	
50166734	SABOR PERUANO	Queens	98-53	CORONA AVENUE	11368.0	7182551825		01/01/1900	
50119424	ARS NOVA THEATER	Manhattan	27	BARROW STREET	10014.0	2124899800		01/01/1900	
41624263	BLISS 46 BISTRO	Queens	43-46	46 STREET	11104.0	7183610690	French	04/11/2022	Violations were cited in the follov
50135079	COUNTRY CLUB GRILL	Bronx	3284	AMPERE AVENUE	10465.0	3472450528		01/01/1900	
50169297	WARDS MEADOW FIELDS	Manhattan	10	CENTRAL ROAD	10035.0	7043286780		01/01/1900	
50176565	SIMPL	Manhattan	927	2 AVENUE	10022.0	3104903817		01/01/1900	
50173979	TOBY'S ESTATE COFFEE	Manhattan	550	HUDSON STREET	10014.0	4079296211		01/01/1900	
50177027	LA VERA PIZZERIA INC.	Staten Island	2071	CLOVE ROAD	10304.0	3479518270		01/01/1900	
41304897	SWEET TALK	Brooklyn	457	THRID AVENUE		7184993400	American	04/08/2022	Violations were cited in the follov
50166046	NATIONAL SAWDUST	Brooklyn	80	NORTH 6 STREET	11249.0	6467798455		01/01/1900	
50171935	UNIRICE LIC	Queens	27-17	42 ROAD	11101.0	9178688795		01/01/1900	
50178785	ESMERALDA & CRISTAL RESTAURANT & LOUNGE CORP.	Bronx	182	WEST BURNSIDE AVENUE	10453.0	9177949179		01/01/1900	
50139115	TABOGA	Manhattan	421	WEST 202 STREET	10034.0	6464845606		01/01/1900	
50178974	MADMAN ESPRESSO	Manhattan	311	11 AVENUE	10001.0	9176059265		01/01/1900	
50157591	PAR AVION GROUP LLC	Manhattan	37	CARMINE STREET	10014.0	9174347533		01/01/1900	
50162392	LITTLE FLOWER CAFE	Queens	25-35	36 AVENUE	11106.0	9177428759		01/01/1900	
50158975	JS DINING	Queens	165-68	BAISLEY BOULEVARD	11434.0	3475484290		01/01/1900	
50161679	ENTRE NOUS	Brooklyn	39	CLIFTON PLACE	11238.0	4178484203		01/01/1900	
50163417	L' APPARTEMENT 4F	Manhattan	119	WEST 10 STREET	10011.0	9177553943		01/01/1900	

Figure 2: Opening a subset of data in a spreadsheet

One more part of my analysis was actually counting the content and analysing the columns and the actual size of the data. I can see that it is a decently big dataset with over 292,256 rows and it is around 150 MB in size.


DOHMH New York City Resta
nyu-datapr...fs-ingest – Buck
ssh.cloud.google.com/v2/ssh
ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumbe
SSH-in-browser

```
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls
NYCInspection.java  NYCInspectionMapper.java  NYCInspectionReducer.java  NYC_Restaurent_Inscp_Data.csv  NYC_Restaurent_Inscp_Data_F
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ wc -l NYC_Restaurent_Inscp_Data.csv
292256 NYC_Restaurent_Inscp_Data.csv
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ head -n 1 NYC_Restaurent_Inscp_Data.csv | awk -F',' '{print NF}'

27
27
-bash: 27: command not found
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ head -n 1 NYC_Restaurent_Inscp_Data.csv | awk -F',' '{print NF}'
27
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ du NYC_Restaurent_Inscp_Data.csv
152888  NYC_Restaurent_Inscp_Data.csv
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ du -h NYC_Restaurent_Inscp_Data.csv
150M    NYC_Restaurent_Inscp_Data.csv
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ 
```

Figure 3: Basic analysis using shell commands

Data Ingestion using MapReduce

Now that I have analysed the dataset, the next step is to write a MapReduce program to process the data and ingest it into the Hadoop system. This step is essential because the processed data will be used by Hive or Trino for further analysis in the next phase of the project, and it can also be loaded into Tableau for interactive visual analysis.

Mapper

The mapper performs a number of operations, including checking for any missing values in the CSV and also transforming the missing values for some fields with valid data.

- **Input Processing**
  Each input in Mapper is the CSV row. It is received as a `Text` value. I am parsing using Apache Commons CSV with default comma-delimited, quote-aware settings.

- **Filtering and Validation**
  I am processing each row to perform validation on the schema of the data and only considering specific INPECTION DATE records.

  - The first entry of the Mapper is the CSV header row, which is skipped since it is of no use as it is already added at the Reducer side.

    I am also dropping the entire record from the dataset when:

    - The **SCORE** column is null or empty. This should be dropped since there is no suitable replacement for it.
    - The records with **INSPECTION DATE** before 2015 are dropped since I am only interested in the analysis of data from January 2015.
    - The **ZIPCODE** is an essential field, and thus I am dropping the records that have missing ZIPCODEs.
    - I am also dropping the records where **VIOLATION CODE** is null or empty.

- **Transformations**

  - All the values of the column fields in the records are trimmed and the nulls or missing values are replaced with empty strings.
  - For the case where **PHONE** is empty or missing, I am replacing it with 0 as a default number case.
  - The **SCORE** is a double value, which is converted to an Integer and stored in the format of Text string.
  - The **GRADE** is an important column but most of the records do not provide a grade or contain missing values. To fix this the missing grades, I calculated it using the official NYC Department of Health scoring guidelines using the below lookup.

    $$\text{GRADE} = \begin{cases} \text{A}, & 0 \leq \text{SCORE} \leq 13, \\ \text{B}, & 14 \leq \text{SCORE} \leq 27, \\ \text{C}, & \text{SCORE} \geq 28. \end{cases}$$

    If a GRADE is already present, I will just remove the whitespace and process it. This ensures I always have a GRADE value in my records.

- **Column Selection**
  After processing all the columns and validating their schema, now I am only considering the following columns in my final output CSV to the reducer side.
  Allowed Columns:

  CAMIS, INSPECTION DATE, DBA, BORO, BUILDING, STREET, ZIPCODE, PHONE, CUISINE DESCRIPTION, ACTION, VIOLATION CODE, VIOLATION DESCRIPTION, CRITICAL FLAG, SCORE, GRADE, INSPECTION TYPE

  Dropped Columns

  GRADE DATE, RECORD DATE, Latitude, Longitude,
  Community Board, Council District, Census Tract, BIN, BBL, NTA, Location

- At the end, I am also formatting the Mapper output values by properly escaping the special characters like new line or intermediate ',' in column values.

  Output Record Structure:

  - **Key:** `NullWritable`
  - **Value:** CSV row string with all 16 selected columns.

- Additionally, I am also collecting a set of Mapper-level statistics to understand data quality, study some distributions, and validate the filtering logic using counters. It helps validate the correctness of the preprocessing step and provides insights into how the dataset is processed after cleaning. The counters are as follows:

  - **Input Counters:** These counters track the total number of raw input rows read by the mapper. It provides understanding of how many records were present in the input.
  - **Output Counters:** These counters count the number of records that were successfully cleaned and emitted as final processed output.
  - **Borough Counters:** These counters aggregate valid records by borough to provide some interesting insights into patterns in the dataset. For each borough, I am tracking:
    - Total number of valid inspection records per borough.
    - Number of records with a Grade A rating per borough.
    - Number of restaurants serving American cuisine per borough.

Reducer

The reducer is responsible for producing the final cleaned output. The `NYCInspectionReducer` performs two tasks that is writing the header row and emitting all valid data records passed from the mapper.

Writting Header row

During the `setup()` phase, the reducer writes the CSV header as the very first output line. This header lists the 16 selected columns in the exact order defined by the output schema:

INSPECTION DATE, CAMIS, DBA, BORO, BUILDING, STREET, ZIPCODE,
PHONE, CUISINE DESCRIPTION, ACTION, VIOLATION CODE, VIOLATION DESCRIPTION,
CRITICAL FLAG, SCORE, GRADE, INSPECTION TYPE

This ensures that the output CSV can be directly imported into tools like Hive or Trino.

Data Record Writing

The `reduce()` method receives grouped values from the mapper and writes them directly to the output.

- The reducer receives all mapper outputs grouped as an `Iterable` of records for each key. Since the mapper uses an empty key, each group effectively contains a sequence of fully formatted CSV rows. The reducer simply iterates through this iterable and writes each record directly to the final output without performing any additional transformations.

- **Output Format:**

  - **Key:** `NullWritable`, this ensures no key field appears in the final CSV output.
  - **Value:** The cleaned and processed CSV row generated by the mapper.

Output location

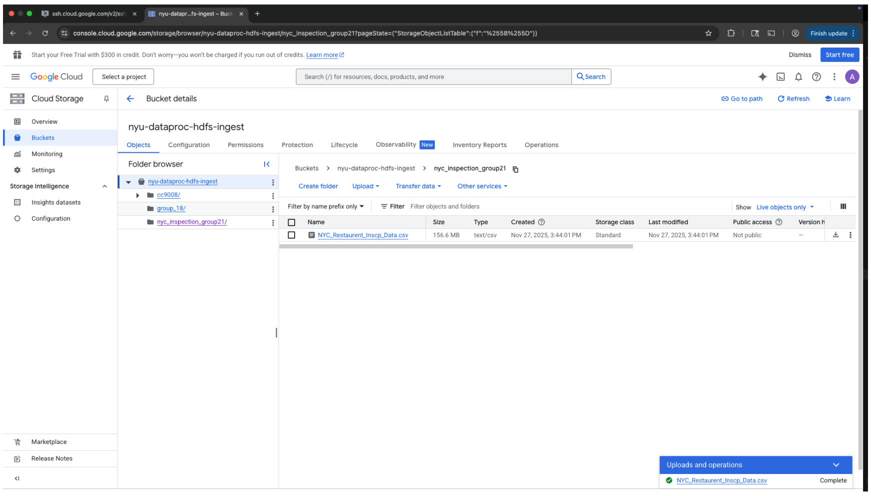
The output is stored in a single file because `setNumReduceTasks(1)` in my driver code collects all cleaned records into one reducer, producing a consolidated CSV output (`part-r-00000`) which is easy to load and analyse.

Running MapReduce code

In this part now I am actually running the MapReduce code on the Dataproc Hadoop.

Uploading the data to Hadoop

Dataproc documentation provides an easy way to upload datasets to GCP ingest bucket and then directly copy them into the Dataproc Hadoop Distributed File System (HDFS). After uploading the dataset to the GCS bucket, I used Hadoop's distcp tool to transfer the files into my HDFS directory.



console.cloud.google.com/storage/browser/nyu-dataproc-hdfs-ingest/nyc_inspection_group21?pageState=("StorageObjectListTable":("f":"%255B%255D"))

Start your Free Trial with $300 in credit. Don't worry—you won't be charged if you run out of credits. Learn more

nyu-dataproc-hdfs-ingest

Buckets > nyu-dataproc-hdfs-ingest > nyc_inspection_group21

Show Live objects only

☐	Name	Size	Type	Created	Storage class	Last modified	Public access	Version h
☐	NYC_Restaurent_Inscp_Data.csv	156.6 MB	text/csv	Nov 27, 2025, 3:44:01 PM	Standard	Nov 27, 2025, 3:44:01 PM	Not public	–

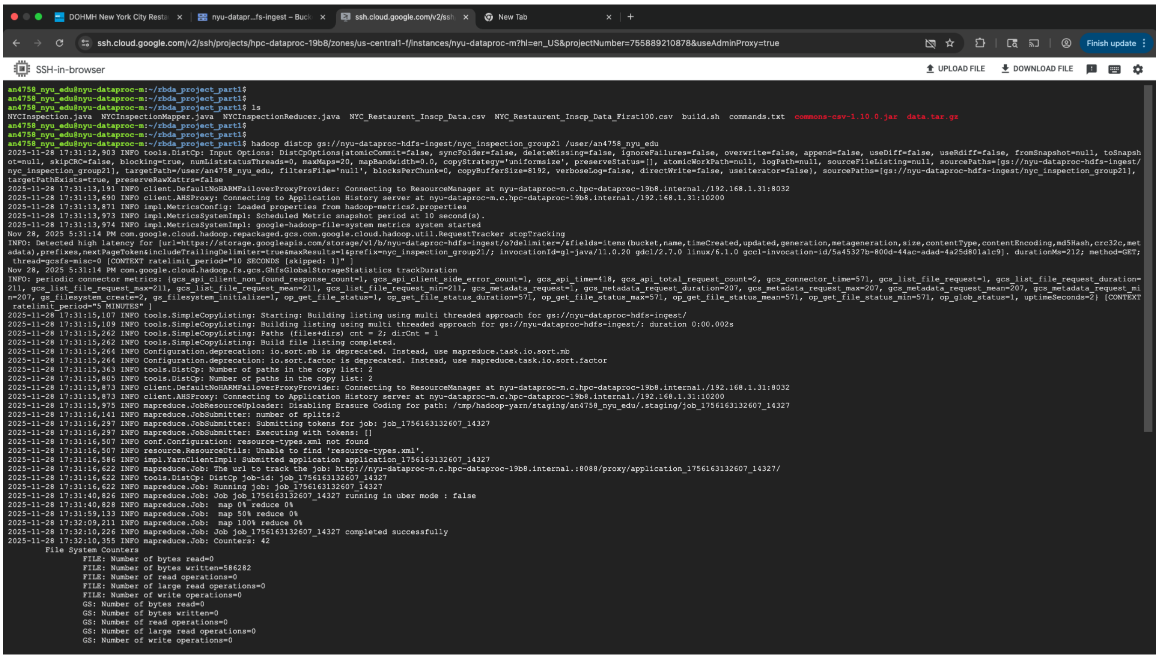
Uploads and operations

NYC_Restaurent_Inscp_Data.csv Complete

Figure 4: Step to upload dataset on GCP

To move the dataset from GCS into HDFS, I used the following command:

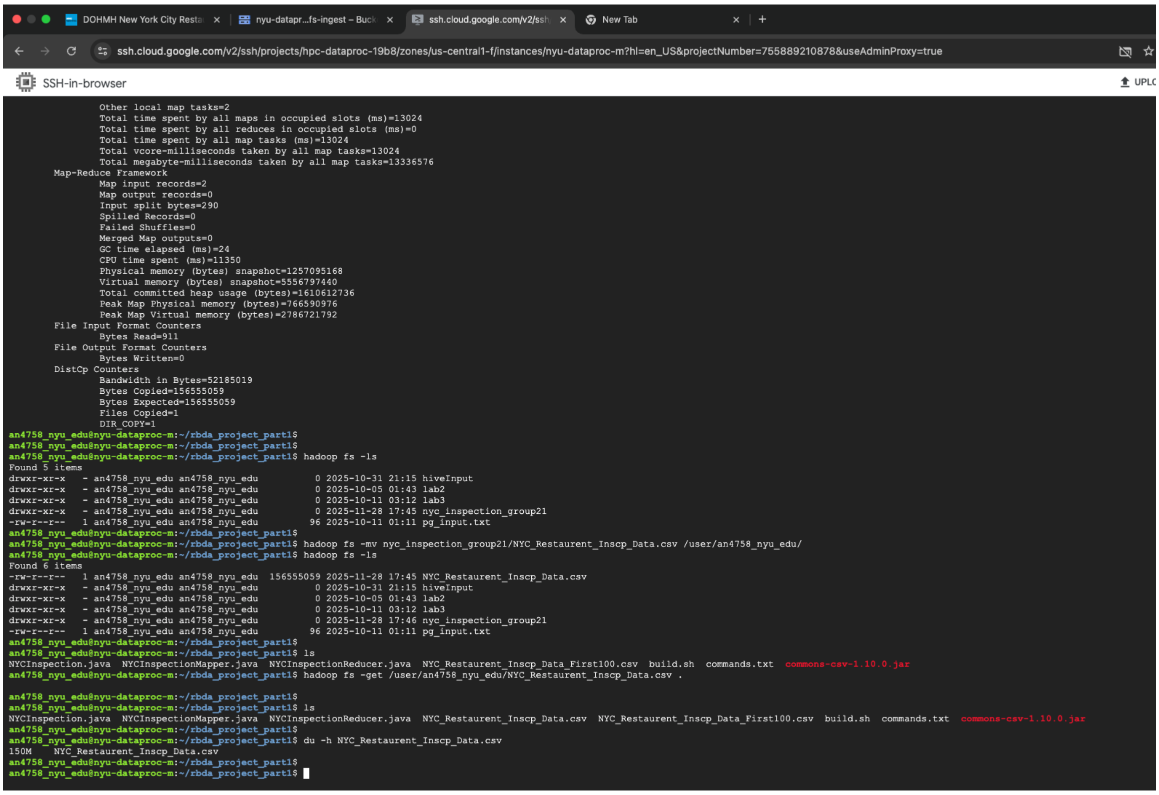
```
hadoop distcp gs://nyu-dataproc-hdfs-ingest/nyc_inspection_group21 /user/an4758_nyu_edu
```


DOHMH New York City Resta
nyu-datapr...fs-ingest – Buck
ssh.cloud.google.com/v2/ssh
New Tab
ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889210878&useAdminProxy=true
Finish update
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

```
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls
NYCInspection.java  NYCInspectionMapper.java  NYCInspectionReducer.java  NYC_Restaurent_Inscp_Data.csv  NYC_Restaurent_Inscp_Data_First100.csv  build.sh  commands.txt  commons-csv-1.10.0.jar  data.tar.gz
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ hadoop distcp gs://nyu-dataproc-hdfs-ingest/nyc_inspection_group21 /user/an4758_nyu_edu
2025-11-28 17:31:12,903 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, append=false, useDiff=false, useRdiff=false, fromSnapshot=null, toSnapsh
ot=null, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=0.0, copyStrategy='uniformsize', preserveStatus=[], atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[gs://nyu-dataproc-hdfs-ingest/
nyc_inspection_group21], targetPath=/user/an4758_nyu_edu, filtersFile='null', blocksPerChunk=0, copyBufferSize=8192, verboseLog=false, directWrite=false, useiterator=false}, sourcePaths=[gs://nyu-dataproc-hdfs-ingest/nyc_inspection_group21],
targetPathExists=true, preserveRawXattrs=false
2025-11-28 17:31:13,191 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.31:8032
2025-11-28 17:31:13,690 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.31:10200
2025-11-28 17:31:13,871 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-11-28 17:31:13,973 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-11-28 17:31:13,974 INFO impl.MetricsSystemImpl: google-hadoop-file-system metrics system started
Nov 28, 2025 5:31:14 PM com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.util.RequestTracker stopTracking
INFO: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/nyu-dataproc-hdfs-ingest/o?delimiter=/&fields=items(bucket,name,timeCreated,updated,generation,metageneration,size,contentType,contentEncoding,md5Hash,crc32c,met
adata),prefixes,nextPageToken&includeTrailingDelimiter=true&maxResults=1&prefix=nyc_inspection_group21/; invocationId=gl-java/11.0.20 gdcl/2.7.0 linux/6.1.0 gccl-invocation-id/5a45327b-800d-44ac-adad-4a25d801a1c9]. durationMs=212; method=GET;
 thread=gcsfs-misc-0 [CONTEXT ratelimit_period="10 SECONDS [skipped: 1]" ]
Nov 28, 2025 5:31:14 PM com.google.cloud.hadoop.fs.gcs.GhfsGlobalStorageStatistics trackDuration
INFO: periodic connector metrics: {gcs_api_client_non_found_response_count=1, gcs_api_client_side_error_count=1, gcs_api_time=418, gcs_api_total_request_count=2, gcs_connector_time=571, gcs_list_file_request=1, gcs_list_file_request_duration=
211, gcs_list_file_request_max=211, gcs_list_file_request_mean=211, gcs_list_file_request_min=211, gcs_metadata_request=1, gcs_metadata_request_duration=207, gcs_metadata_request_max=207, gcs_metadata_request_mean=207, gcs_metadata_request_mi
n=207, gs_filesystem_create=2, gs_filesystem_initialize=1, op_get_file_status=1, op_get_file_status_duration=571, op_get_file_status_max=571, op_get_file_status_mean=571, op_get_file_status_min=571, op_glob_status=1, uptimeSeconds=2} [CONTEXT
 ratelimit_period="5 MINUTES" ]
2025-11-28 17:31:15,107 INFO tools.SimpleCopyListing: Starting: Building listing using multi threaded approach for gs://nyu-dataproc-hdfs-ingest/
2025-11-28 17:31:15,109 INFO tools.SimpleCopyListing: Building listing using multi threaded approach for gs://nyu-dataproc-hdfs-ingest/: duration 0:00.002s
2025-11-28 17:31:15,262 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 2; dirCnt = 1
2025-11-28 17:31:15,262 INFO tools.SimpleCopyListing: Build file listing completed.
2025-11-28 17:31:15,264 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
2025-11-28 17:31:15,264 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
2025-11-28 17:31:15,363 INFO tools.DistCp: Number of paths in the copy list: 2
2025-11-28 17:31:15,805 INFO tools.DistCp: Number of paths in the copy list: 2
2025-11-28 17:31:15,873 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.31:8032
2025-11-28 17:31:15,873 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.31:10200
2025-11-28 17:31:15,975 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/an4758_nyu_edu/.staging/job_1756163132607_14327
2025-11-28 17:31:16,141 INFO mapreduce.JobSubmitter: number of splits:2
2025-11-28 17:31:16,297 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1756163132607_14327
2025-11-28 17:31:16,297 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-11-28 17:31:16,507 INFO conf.Configuration: resource-types.xml not found
2025-11-28 17:31:16,507 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-11-28 17:31:16,586 INFO impl.YarnClientImpl: Submitted application application_1756163132607_14327
2025-11-28 17:31:16,622 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m.c.hpc-dataproc-19b8.internal.:8088/proxy/application_1756163132607_14327/
2025-11-28 17:31:16,622 INFO tools.DistCp: DistCp job-id: job_1756163132607_14327
2025-11-28 17:31:16,622 INFO mapreduce.Job: Running job: job_1756163132607_14327
2025-11-28 17:31:40,826 INFO mapreduce.Job: Job job_1756163132607_14327 running in uber mode : false
2025-11-28 17:31:40,828 INFO mapreduce.Job:  map 0% reduce 0%
2025-11-28 17:31:59,133 INFO mapreduce.Job:  map 50% reduce 0%
2025-11-28 17:32:09,211 INFO mapreduce.Job:  map 100% reduce 0%
2025-11-28 17:32:10,226 INFO mapreduce.Job: Job job_1756163132607_14327 completed successfully
2025-11-28 17:32:10,355 INFO mapreduce.Job: Counters: 42
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=586282
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                GS: Number of bytes read=0
                GS: Number of bytes written=0
                GS: Number of read operations=0
                GS: Number of large read operations=0
                GS: Number of write operations=0
```

Figure 5: Downloading data from GCP to Hadoop

To organize and download the dataset, I listed my HDFS contents, moved the CSV file from its folder into my HDFS root directory, and then copied it to my local filesystem as a backup copy.


DOHMH New York City Resta
nyu-datapr...fs-ingest – Buck
ssh.cloud.google.com/v2/ssh
New Tab
ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889210878&useAdminProxy=true
SSH-in-browser
UPLC

```
                Other local map tasks=2
                Total time spent by all maps in occupied slots (ms)=13024
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=13024
                Total vcore-milliseconds taken by all map tasks=13024
                Total megabyte-milliseconds taken by all map tasks=13336576
        Map-Reduce Framework
                Map input records=2
                Map output records=0
                Input split bytes=290
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=24
                CPU time spent (ms)=11350
                Physical memory (bytes) snapshot=1257095168
                Virtual memory (bytes) snapshot=5556797440
                Total committed heap usage (bytes)=1610612736
                Peak Map Physical memory (bytes)=766590976
                Peak Map Virtual memory (bytes)=2786721792
        File Input Format Counters
                Bytes Read=911
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bandwidth in Bytes=52185019
                Bytes Copied=156555059
                Bytes Expected=156555059
                Files Copied=1
                DIR_COPY=1
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ hadoop fs -ls
Found 5 items
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-10-31 21:15 hiveInput
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-10-05 01:43 lab2
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-10-11 03:12 lab3
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-11-28 17:45 nyc_inspection_group21
-rw-r--r--   1 an4758_nyu_edu an4758_nyu_edu         96 2025-10-11 01:11 pg_input.txt
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ hadoop fs -mv nyc_inspection_group21/NYC_Restaurent_Inscp_Data.csv /user/an4758_nyu_edu/
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ hadoop fs -ls
Found 6 items
-rw-r--r--   1 an4758_nyu_edu an4758_nyu_edu  156555059 2025-11-28 17:45 NYC_Restaurent_Inscp_Data.csv
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-10-31 21:15 hiveInput
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-10-05 01:43 lab2
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-10-11 03:12 lab3
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-11-28 17:46 nyc_inspection_group21
-rw-r--r--   1 an4758_nyu_edu an4758_nyu_edu         96 2025-10-11 01:11 pg_input.txt
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls
NYCInspection.java  NYCInspectionMapper.java  NYCInspectionReducer.java  NYC_Restaurent_Inscp_Data_First100.csv  build.sh  commands.txt  commons-csv-1.10.0.jar
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ hadoop fs -get /user/an4758_nyu_edu/NYC_Restaurent_Inscp_Data.csv .

an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls
NYCInspection.java  NYCInspectionMapper.java  NYCInspectionReducer.java  NYC_Restaurent_Inscp_Data.csv  NYC_Restaurent_Inscp_Data_First100.csv  build.sh  commands.txt  commons-csv-1.10.0.jar
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ du -h NYC_Restaurent_Inscp_Data.csv
150M    NYC_Restaurent_Inscp_Data.csv
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
```

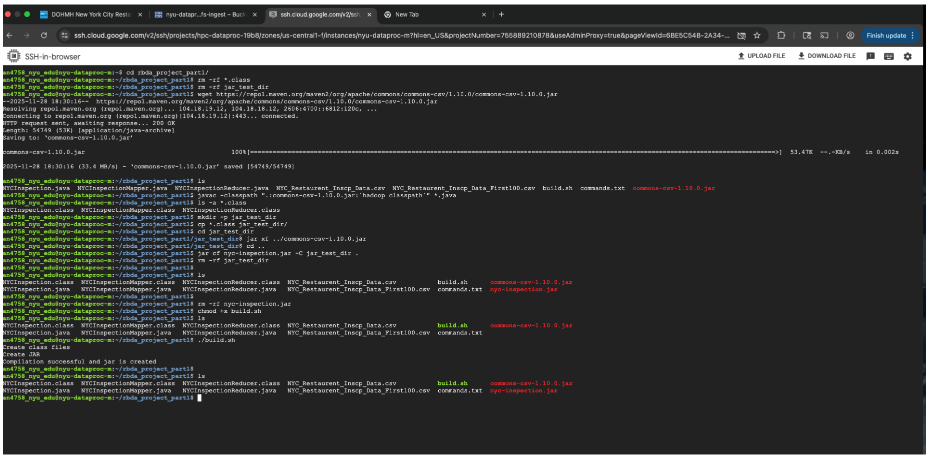
Figure 6: Renaming file and storing a copy on my local Dataproc FS

Commands:

```
hadoop fs -ls
hadoop fs -mv nyc_inspection_group21/NYC_Restaurent_Inscp_Data.csv /user/an4758_nyu_edu/
hadoop fs -ls
hadoop fs -get /user/an4758_nyu_edu/NYC_Restaurent_Inscp_Data.csv .
du -h NYC_Restaurent_Inscp_Data.csv
```

Compiling the code

There are two ways to compile code, I have included a build.sh file and also directly used javac to execute the commands and create a JAR file.



DOHMH New York City Resta
nyu-datapr...fs-ingest – Buck
ssh.cloud.google.com/v2/ssh
New Tab
ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889210878&useAdminProxy=true&pageViewId=6BE5C54B-2A34-...
Finish update
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

```
an4758_nyu_edu@nyu-dataproc-m:~$ cd rbda_project_part1/
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ rm -rf *.class
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ rm -rf jar_test_dir
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ wget https://repo1.maven.org/maven2/org/apache/commons/commons-csv/1.10.0/commons-csv-1.10.0.jar
--2025-11-28 18:30:16--  https://repo1.maven.org/maven2/org/apache/commons/commons-csv/1.10.0/commons-csv-1.10.0.jar
Resolving repo1.maven.org (repo1.maven.org)... 104.18.19.12, 104.18.18.12, 2606:4700::6812:120c, ...
Connecting to repo1.maven.org (repo1.maven.org)|104.18.19.12|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 54749 (53K) [application/java-archive]
Saving to: ‘commons-csv-1.10.0.jar’

commons-csv-1.10.0.jar                   100%[==================================================================================>]  53.47K  --.-KB/s    in 0.002s

2025-11-28 18:30:16 (33.4 MB/s) - ‘commons-csv-1.10.0.jar’ saved [54749/54749]

an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls
NYCInspection.java  NYCInspectionMapper.java  NYCInspectionReducer.java  NYC_Restaurent_Inscp_Data.csv  NYC_Restaurent_Inscp_Data_First100.csv  build.sh  commands.txt  commons-csv-1.10.0.jar
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ javac -classpath ".:commons-csv-1.10.0.jar:`hadoop classpath`" *.java
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls -a *.class
NYCInspection.class  NYCInspectionMapper.class  NYCInspectionReducer.class
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ mkdir -p jar_test_dir
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ cp *.class jar_test_dir/
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ cd jar_test_dir
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1/jar_test_dir$ jar xf ../commons-csv-1.10.0.jar
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1/jar_test_dir$ cd ..
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ jar cf nyc-inspection.jar -C jar_test_dir .
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ rm -rf jar_test_dir
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls
NYCInspection.class  NYCInspectionMapper.class  NYCInspectionReducer.class  NYC_Restaurent_Inscp_Data.csv           build.sh      commons-csv-1.10.0.jar
NYCInspection.java   NYCInspectionMapper.java   NYCInspectionReducer.java   NYC_Restaurent_Inscp_Data_First100.csv  commands.txt  nyc-inspection.jar
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ rm -rf nyc-inspection.jar
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ chmod +x build.sh
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls
NYCInspection.class  NYCInspectionMapper.class  NYCInspectionReducer.class  NYC_Restaurent_Inscp_Data.csv           build.sh      commons-csv-1.10.0.jar
NYCInspection.java   NYCInspectionMapper.java   NYCInspectionReducer.java   NYC_Restaurent_Inscp_Data_First100.csv  commands.txt
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ./build.sh
Create class files
Create JAR
Compilation successful and jar is created
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls
NYCInspection.class  NYCInspectionMapper.class  NYCInspectionReducer.class  NYC_Restaurent_Inscp_Data.csv           build.sh      commons-csv-1.10.0.jar
NYCInspection.java   NYCInspectionMapper.java   NYCInspectionReducer.java   NYC_Restaurent_Inscp_Data_First100.csv  commands.txt  nyc-inspection.jar
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
```

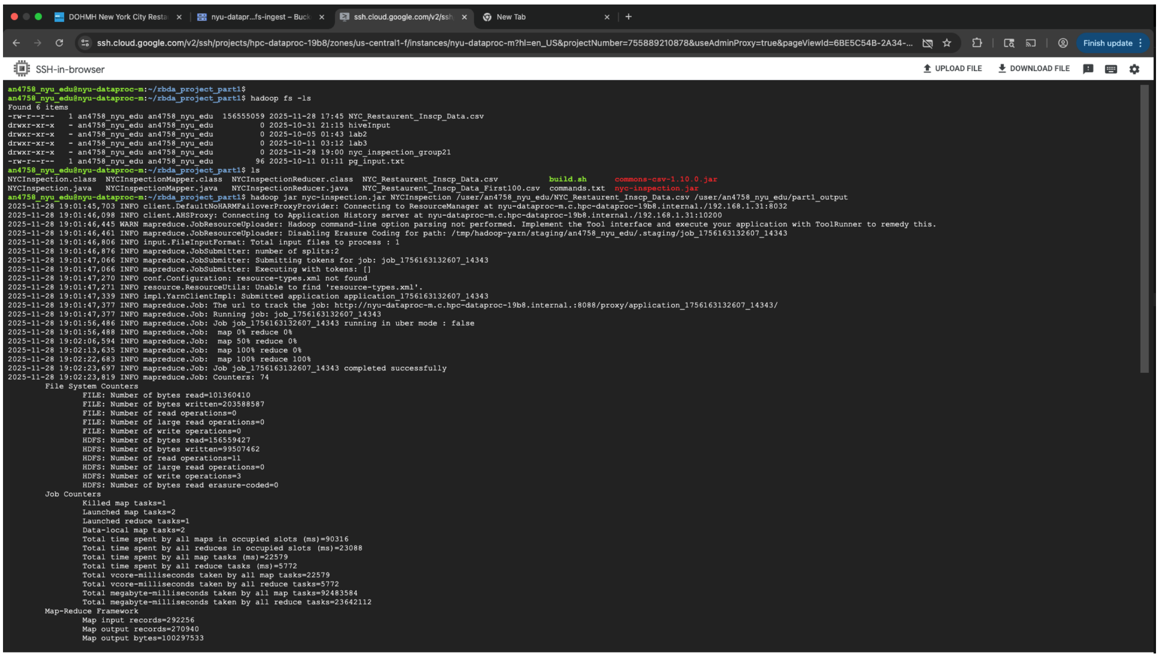
Figure 7: Compiling MapReduce code

```
# Download the required dependency
wget https://repo1.maven.org/maven2/org/apache/commons/commons-csv/1.10.0/commons-csv-1.10.0.jar

# Compile files with Hadoop classpath and CSV JAR
javac -classpath ".:commons-csv-1.10.0.jar:‘hadoop classpath‘" *.java

# Copy compiled class and unpack third party dependency.
mkdir -p jar_test_dir
cp *.class jar_test_dir/
cd jar_test_dir
jar xf ../commons-csv-1.10.0.jar

cd ..
# Create a complete JAR.
jar cf nyc-inspection.jar -C jar_test_dir .

rm -rf jar_test_dir
rm -rf nyc-inspection.jar

# To BUILD using build.sh
chmod +x build.sh
./build.sh
```

The above commands will build the code by compiling Java source files and will pack the dependencies by generating a final `nyc-inspection.jar`, that I will use to run the MapReduce job.

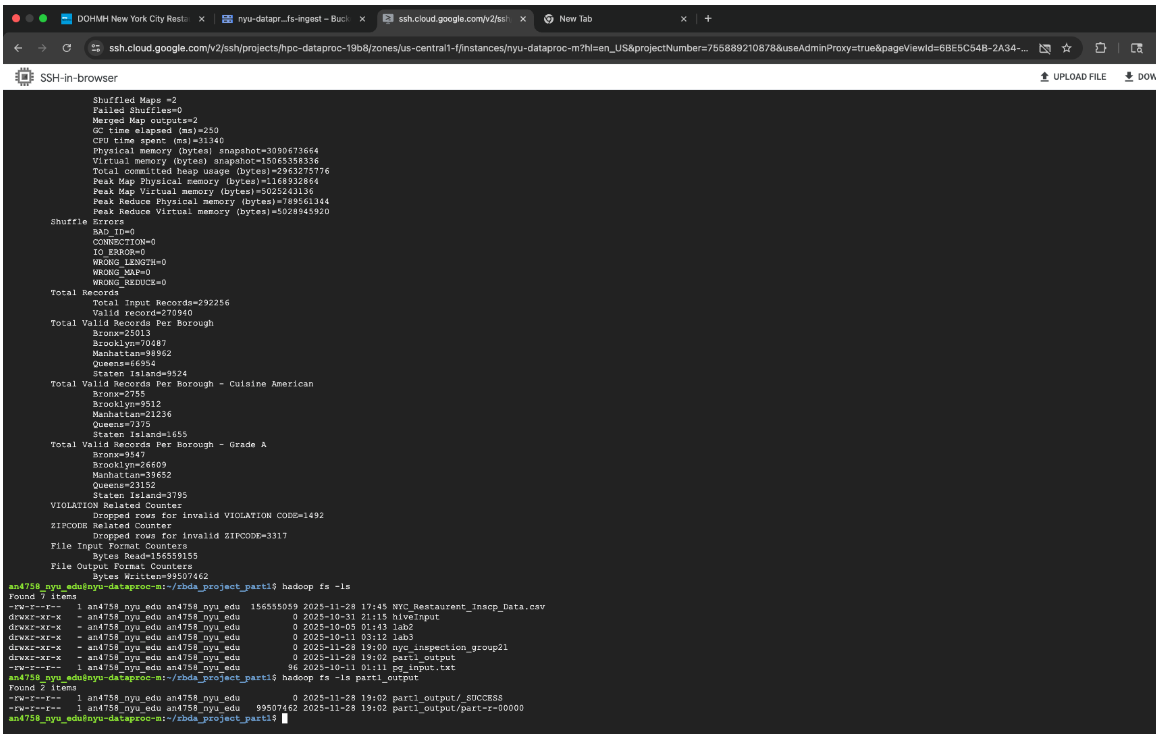
Running MapReduce job

Now that I have a Jar file, I will run the code on Dataproc and get the required results.


DOHMH New York City Resta
nyu-datapr...fs-ingest – Buck
ssh.cloud.google.com/v2/ssh
New Tab
ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889210878&useAdminProxy=true&pageViewId=6BE5C54B-2A34-...
Finish update
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

```
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ hadoop fs -ls
Found 6 items
-rw-r--r--   1 an4758_nyu_edu an4758_nyu_edu  156555059 2025-11-28 17:45 NYC_Restaurent_Inscp_Data.csv
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-10-31 21:15 hiveInput
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-10-05 01:43 lab2
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-10-11 03:12 lab3
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-11-28 19:00 nyc_inspection_group21
-rw-r--r--   1 an4758_nyu_edu an4758_nyu_edu         96 2025-10-11 01:11 pg_input.txt
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls
NYCInspection.class  NYCInspectionMapper.class  NYCInspectionReducer.class  NYC_Restaurent_Inscp_Data.csv           build.sh      commons-csv-1.10.0.jar
NYCInspection.java   NYCInspectionMapper.java   NYCInspectionReducer.java   NYC_Restaurent_Inscp_Data_First100.csv  commands.txt  nyc-inspection.jar
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ hadoop jar nyc-inspection.jar NYCInspection /user/an4758_nyu_edu/NYC_Restaurent_Inscp_Data.csv /user/an4758_nyu_edu/part1_output
2025-11-28 19:01:45,703 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.31:8032
2025-11-28 19:01:46,098 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.31:10200
2025-11-28 19:01:46,445 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-11-28 19:01:46,461 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/an4758_nyu_edu/.staging/job_1756163132607_14343
2025-11-28 19:01:46,806 INFO input.FileInputFormat: Total input files to process : 1
2025-11-28 19:01:46,876 INFO mapreduce.JobSubmitter: number of splits:2
2025-11-28 19:01:47,066 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1756163132607_14343
2025-11-28 19:01:47,066 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-11-28 19:01:47,270 INFO conf.Configuration: resource-types.xml not found
2025-11-28 19:01:47,271 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-11-28 19:01:47,339 INFO impl.YarnClientImpl: Submitted application application_1756163132607_14343
2025-11-28 19:01:47,377 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m.c.hpc-dataproc-19b8.internal.:8088/proxy/application_1756163132607_14343/
2025-11-28 19:01:47,377 INFO mapreduce.Job: Running job: job_1756163132607_14343
2025-11-28 19:01:56,486 INFO mapreduce.Job: Job job_1756163132607_14343 running in uber mode : false
2025-11-28 19:01:56,488 INFO mapreduce.Job:  map 0% reduce 0%
2025-11-28 19:02:06,594 INFO mapreduce.Job:  map 50% reduce 0%
2025-11-28 19:02:13,635 INFO mapreduce.Job:  map 100% reduce 0%
2025-11-28 19:02:22,683 INFO mapreduce.Job:  map 100% reduce 100%
2025-11-28 19:02:23,697 INFO mapreduce.Job: Job job_1756163132607_14343 completed successfully
2025-11-28 19:02:23,819 INFO mapreduce.Job: Counters: 74
        File System Counters
                FILE: Number of bytes read=101360410
                FILE: Number of bytes written=203588587
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=156559427
                HDFS: Number of bytes written=99507462
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=90316
                Total time spent by all reduces in occupied slots (ms)=23088
                Total time spent by all map tasks (ms)=22579
                Total time spent by all reduce tasks (ms)=5772
                Total vcore-milliseconds taken by all map tasks=22579
                Total vcore-milliseconds taken by all reduce tasks=5772
                Total megabyte-milliseconds taken by all map tasks=92483584
                Total megabyte-milliseconds taken by all reduce tasks=23642112
        Map-Reduce Framework
                Map input records=292256
                Map output records=270940
                Map output bytes=100297533
```

Figure 8: Running MapReduce job for NYC inspection data.


DOHMH New York City Resta
nyu-datapr...fs-ingest – Buck
ssh.cloud.google.com/v2/ssh
New Tab
ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889210878&useAdminProxy=true&pageViewId=6BE5C54B-2A34-...
SSH-in-browser
UPLOAD FILE
DOW

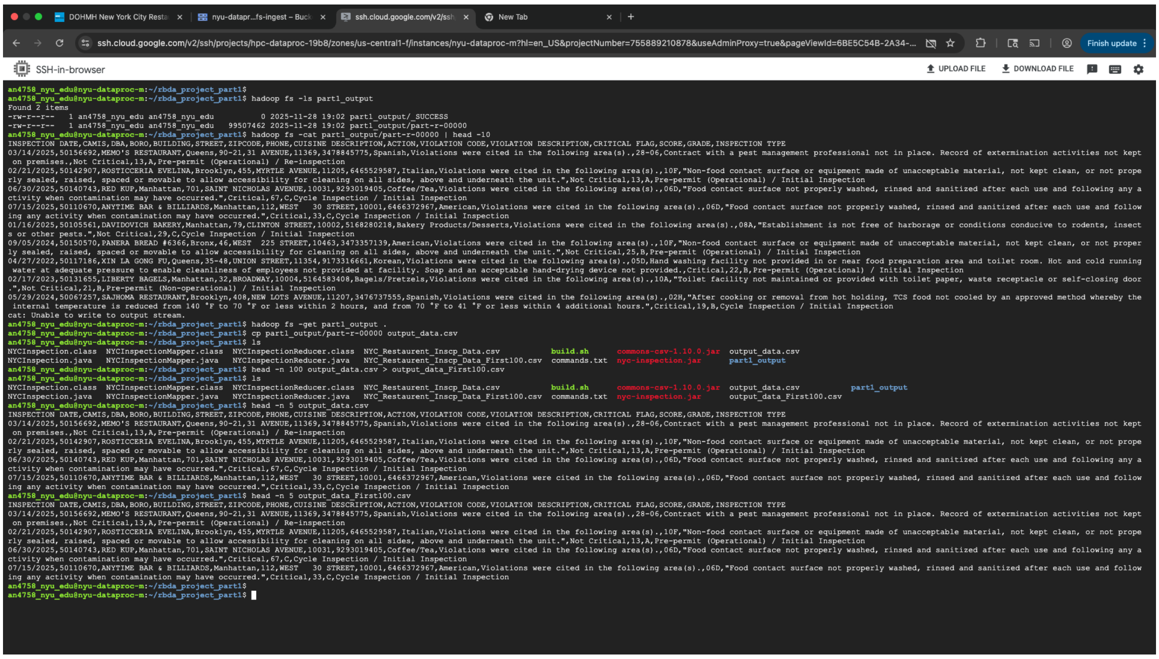
```
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=250
                CPU time spent (ms)=31340
                Physical memory (bytes) snapshot=3090673664
                Virtual memory (bytes) snapshot=15065358336
                Total committed heap usage (bytes)=2963275776
                Peak Map Physical memory (bytes)=1168932864
                Peak Map Virtual memory (bytes)=5025243136
                Peak Reduce Physical memory (bytes)=789561344
                Peak Reduce Virtual memory (bytes)=5028945920
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        Total Records
                Total Input Records=292256
                Valid record=270940
        Total Valid Records Per Borough
                Bronx=25013
                Brooklyn=70487
                Manhattan=98962
                Queens=66954
                Staten Island=9524
        Total Valid Records Per Borough - Cuisine American
                Bronx=2755
                Brooklyn=9512
                Manhattan=21236
                Queens=7375
                Staten Island=1655
        Total Valid Records Per Borough - Grade A
                Bronx=9547
                Brooklyn=26609
                Manhattan=39652
                Queens=23152
                Staten Island=3795
        VIOLATION Related Counter
                Dropped rows for invalid VIOLATION CODE=1492
        ZIPCODE Related Counter
                Dropped rows for invalid ZIPCODE=3317
        File Input Format Counters
                Bytes Read=156559155
        File Output Format Counters
                Bytes Written=99507462
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ hadoop fs -ls
Found 7 items
-rw-r--r--   1 an4758_nyu_edu an4758_nyu_edu  156555059 2025-11-28 17:45 NYC_Restaurent_Inscp_Data.csv
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-10-31 21:15 hiveInput
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-10-05 01:43 lab2
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-10-11 03:12 lab3
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-11-28 19:00 nyc_inspection_group21
drwxr-xr-x   - an4758_nyu_edu an4758_nyu_edu          0 2025-11-28 19:02 part1_output
-rw-r--r--   1 an4758_nyu_edu an4758_nyu_edu         96 2025-10-11 01:11 pg_input.txt
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ hadoop fs -ls part1_output
Found 2 items
-rw-r--r--   1 an4758_nyu_edu an4758_nyu_edu          0 2025-11-28 19:02 part1_output/_SUCCESS
-rw-r--r--   1 an4758_nyu_edu an4758_nyu_edu   99507462 2025-11-28 19:02 part1_output/part-r-00000
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
```

Figure 9: Job counters and output file.

My MapReduce job ran successfully, and I was able to see all my custom counters printed in the logs (Total Records, Borough-level information, Invalid ZIPCODES and much more). The reducer output is in the `part1_output` directory in HDFS, containing the result file.

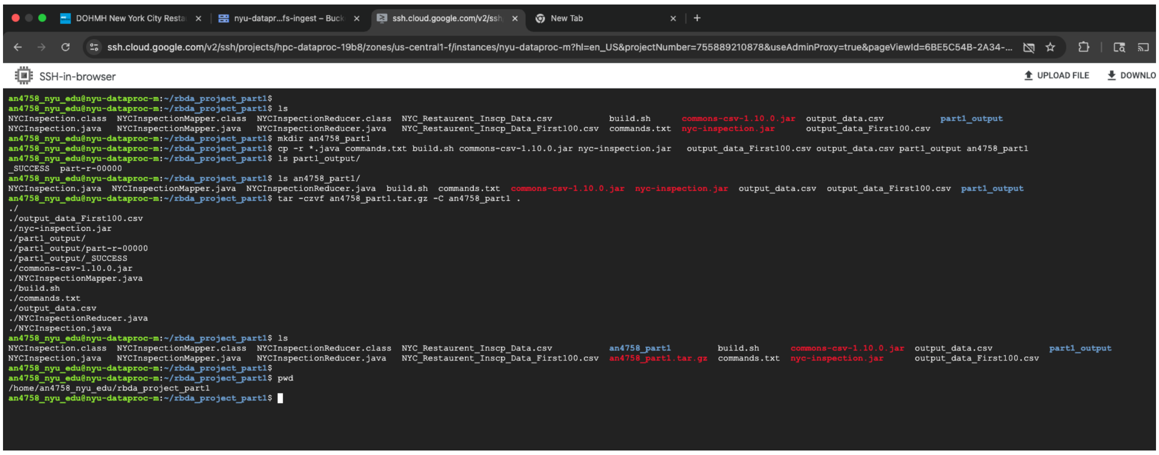
Output

In this step, I copied the processed output from HDFS into the local Dataproc FS and evaluated the rows to verify the MapReduce job output. As seen in the screenshot, I was able to clearly view the CSV header followed by properly formatted rows. This confirms that the reducer output is correct. After validating the dataset, I compressed all required source files, scripts, dependency JARs, and outputs into a `.tar.gz` format to prepare for my part1 submission.


DOHMH New York City Resta
nyu-datapr...fs-ingest – Buck
ssh.cloud.google.com/v2/ssh
New Tab
ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889210878&useAdminProxy=true&pageViewId=6BE5C54B-2A34-...
Finish update
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

```
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ hadoop fs -ls part1_output
Found 2 items
-rw-r--r--   1 an4758_nyu_edu an4758_nyu_edu          0 2025-11-28 19:02 part1_output/_SUCCESS
-rw-r--r--   1 an4758_nyu_edu an4758_nyu_edu   99507462 2025-11-28 19:02 part1_output/part-r-00000
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ hadoop fs -cat part1_output/part-r-00000 | head -10
INSPECTION DATE,CAMIS,DBA,BORO,BUILDING,STREET,ZIPCODE,PHONE,CUISINE DESCRIPTION,ACTION,VIOLATION CODE,VIOLATION DESCRIPTION,CRITICAL FLAG,SCORE,GRADE,INSPECTION TYPE
03/14/2025,50156692,MEMO'S RESTAURANT,Queens,90-21,31 AVENUE,11369,3478845775,Spanish,Violations were cited in the following area(s).,28-06,Contract with a pest management professional not in place. Record of extermination activities not kept
 on premises.,Not Critical,13,A,Pre-permit (Operational) / Re-inspection
02/21/2025,50142907,ROSTICCERIA EVELINA,Brooklyn,455,MYRTLE AVENUE,11205,6465529587,Italian,Violations were cited in the following area(s).,10F,"Non-food contact surface or equipment made of unacceptable material, not kept clean, or not prope
rly sealed, raised, spaced or movable to allow accessibility for cleaning on all sides, above and underneath the unit.",Not Critical,13,A,Pre-permit (Operational) / Initial Inspection
06/30/2025,50140743,RED KUP,Manhattan,701,SAINT NICHOLAS AVENUE,10031,9293019405,Coffee/Tea,Violations were cited in the following area(s).,06D,"Food contact surface not properly washed, rinsed and sanitized after each use and following any a
ctivity when contamination may have occurred.",Critical,67,C,Cycle Inspection / Initial Inspection
07/15/2025,50110670,ANYTIME BAR & BILLIARDS,Manhattan,112,WEST   30 STREET,10001,6466372967,American,Violations were cited in the following area(s).,06D,"Food contact surface not properly washed, rinsed and sanitized after each use and follow
ing any activity when contamination may have occurred.",Critical,33,C,Cycle Inspection / Initial Inspection
01/16/2025,50105561,DAVIDOVICH BAKERY,Manhattan,79,CLINTON STREET,10002,5168280218,Bakery Products/Desserts,Violations were cited in the following area(s).,08A,"Establishment is not free of harborage or conditions conducive to rodents, insect
s or other pests.",Not Critical,29,C,Cycle Inspection / Initial Inspection
09/05/2024,50150570,PANERA BREAD #6366,Bronx,46,WEST  225 STREET,10463,3473357139,American,Violations were cited in the following area(s).,10F,"Non-food contact surface or equipment made of unacceptable material, not kept clean, or not proper
ly sealed, raised, spaced or movable to allow accessibility for cleaning on all sides, above and underneath the unit.",Not Critical,25,B,Pre-permit (Operational) / Initial Inspection
04/27/2022,50117186,XIN LA GONG FU,Queens,35-48,UNION STREET,11354,9173316661,Korean,Violations were cited in the following area(s).,05D,Hand washing facility not provided in or near food preparation area and toilet room. Hot and cold running
 water at adequate pressure to enable cleanliness of employees not provided at facility. Soap and an acceptable hand-drying device not provided.,Critical,22,B,Pre-permit (Operational) / Initial Inspection
02/17/2023,50131655,LIBERTY BAGELS,Manhattan,32,BROADWAY,10004,5164583408,Bagels/Pretzels,Violations were cited in the following area(s).,10A,"Toilet facility not maintained or provided with toilet paper, waste receptacle or self-closing door
.",Not Critical,21,B,Pre-permit (Non-operational) / Initial Inspection
05/29/2024,50067257,SAJHOMA RESTAURANT,Brooklyn,408,NEW LOTS AVENUE,11207,3476737555,Spanish,Violations were cited in the following area(s).,02H,"After cooking or removal from hot holding, TCS food not cooled by an approved method whereby the
 internal temperature is reduced from 140 °F to 70 °F or less within 2 hours, and from 70 °F to 41 °F or less within 4 additional hours.",Critical,19,B,Cycle Inspection / Initial Inspection
cat: Unable to write to output stream.
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ hadoop fs -get part1_output .
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ cp part1_output/part-r-00000 output_data.csv
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls
NYCInspection.class  NYCInspectionMapper.class  NYCInspectionReducer.class  NYC_Restaurent_Inscp_Data.csv           build.sh      commons-csv-1.10.0.jar  output_data.csv
NYCInspection.java   NYCInspectionMapper.java   NYCInspectionReducer.java   NYC_Restaurent_Inscp_Data_First100.csv  commands.txt  nyc-inspection.jar      part1_output
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ head -n 100 output_data.csv > output_data_First100.csv
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls
NYCInspection.class  NYCInspectionMapper.class  NYCInspectionReducer.class  NYC_Restaurent_Inscp_Data.csv           build.sh      commons-csv-1.10.0.jar  output_data.csv           part1_output
NYCInspection.java   NYCInspectionMapper.java   NYCInspectionReducer.java   NYC_Restaurent_Inscp_Data_First100.csv  commands.txt  nyc-inspection.jar      output_data_First100.csv
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ head -n 5 output_data.csv
INSPECTION DATE,CAMIS,DBA,BORO,BUILDING,STREET,ZIPCODE,PHONE,CUISINE DESCRIPTION,ACTION,VIOLATION CODE,VIOLATION DESCRIPTION,CRITICAL FLAG,SCORE,GRADE,INSPECTION TYPE
03/14/2025,50156692,MEMO'S RESTAURANT,Queens,90-21,31 AVENUE,11369,3478845775,Spanish,Violations were cited in the following area(s).,28-06,Contract with a pest management professional not in place. Record of extermination activities not kept
 on premises.,Not Critical,13,A,Pre-permit (Operational) / Re-inspection
02/21/2025,50142907,ROSTICCERIA EVELINA,Brooklyn,455,MYRTLE AVENUE,11205,6465529587,Italian,Violations were cited in the following area(s).,10F,"Non-food contact surface or equipment made of unacceptable material, not kept clean, or not prope
rly sealed, raised, spaced or movable to allow accessibility for cleaning on all sides, above and underneath the unit.",Not Critical,13,A,Pre-permit (Operational) / Initial Inspection
06/30/2025,50140743,RED KUP,Manhattan,701,SAINT NICHOLAS AVENUE,10031,9293019405,Coffee/Tea,Violations were cited in the following area(s).,06D,"Food contact surface not properly washed, rinsed and sanitized after each use and following any a
ctivity when contamination may have occurred.",Critical,67,C,Cycle Inspection / Initial Inspection
07/15/2025,50110670,ANYTIME BAR & BILLIARDS,Manhattan,112,WEST   30 STREET,10001,6466372967,American,Violations were cited in the following area(s).,06D,"Food contact surface not properly washed, rinsed and sanitized after each use and follow
ing any activity when contamination may have occurred.",Critical,33,C,Cycle Inspection / Initial Inspection
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ head -n 5 output_data_First100.csv
INSPECTION DATE,CAMIS,DBA,BORO,BUILDING,STREET,ZIPCODE,PHONE,CUISINE DESCRIPTION,ACTION,VIOLATION CODE,VIOLATION DESCRIPTION,CRITICAL FLAG,SCORE,GRADE,INSPECTION TYPE
03/14/2025,50156692,MEMO'S RESTAURANT,Queens,90-21,31 AVENUE,11369,3478845775,Spanish,Violations were cited in the following area(s).,28-06,Contract with a pest management professional not in place. Record of extermination activities not kept
 on premises.,Not Critical,13,A,Pre-permit (Operational) / Re-inspection
02/21/2025,50142907,ROSTICCERIA EVELINA,Brooklyn,455,MYRTLE AVENUE,11205,6465529587,Italian,Violations were cited in the following area(s).,10F,"Non-food contact surface or equipment made of unacceptable material, not kept clean, or not prope
rly sealed, raised, spaced or movable to allow accessibility for cleaning on all sides, above and underneath the unit.",Not Critical,13,A,Pre-permit (Operational) / Initial Inspection
06/30/2025,50140743,RED KUP,Manhattan,701,SAINT NICHOLAS AVENUE,10031,9293019405,Coffee/Tea,Violations were cited in the following area(s).,06D,"Food contact surface not properly washed, rinsed and sanitized after each use and following any a
ctivity when contamination may have occurred.",Critical,67,C,Cycle Inspection / Initial Inspection
07/15/2025,50110670,ANYTIME BAR & BILLIARDS,Manhattan,112,WEST   30 STREET,10001,6466372967,American,Violations were cited in the following area(s).,06D,"Food contact surface not properly washed, rinsed and sanitized after each use and follow
ing any activity when contamination may have occurred.",Critical,33,C,Cycle Inspection / Initial Inspection
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
```

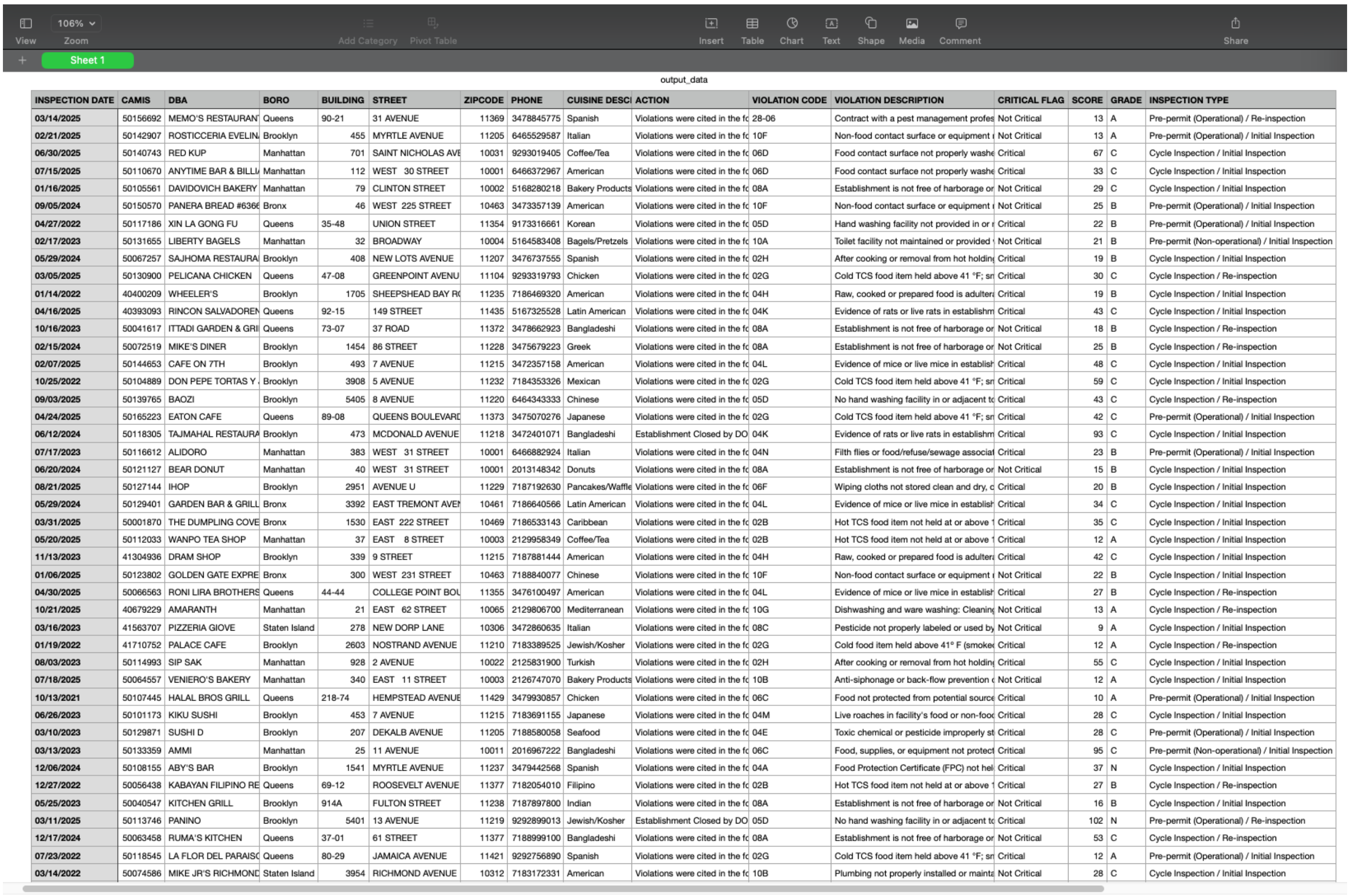
Figure 10: Checking output, copying file, analysing file and storing it as csv.


DOHMH New York City Resta
nyu-datapr...fs-ingest – Buck
ssh.cloud.google.com/v2/ssh
New Tab
ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889210878&useAdminProxy=true&pageViewId=6BE5C54B-2A34-...
SSH-in-browser
UPLOAD FILE
DOWNLO

```
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls
NYCInspection.class  NYCInspectionMapper.class  NYCInspectionReducer.class  NYC_Restaurent_Inscp_Data.csv           build.sh      commons-csv-1.10.0.jar  output_data.csv           part1_output
NYCInspection.java   NYCInspectionMapper.java   NYCInspectionReducer.java   NYC_Restaurent_Inscp_Data_First100.csv  commands.txt  nyc-inspection.jar      output_data_First100.csv
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ mkdir an4758_part1
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ cp -r *.java commands.txt build.sh commons-csv-1.10.0.jar nyc-inspection.jar   output_data_First100.csv output_data.csv part1_output an4758_part1
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls part1_output/
_SUCCESS  part-r-00000
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls an4758_part1/
NYCInspection.java  NYCInspectionMapper.java  NYCInspectionReducer.java  build.sh  commands.txt  commons-csv-1.10.0.jar  nyc-inspection.jar  output_data.csv  output_data_First100.csv  part1_output
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ tar -czvf an4758_part1.tar.gz -C an4758_part1 .
./
./output_data_First100.csv
./nyc-inspection.jar
./part1_output/
./part1_output/part-r-00000
./part1_output/_SUCCESS
./commons-csv-1.10.0.jar
./NYCInspectionMapper.java
./build.sh
./commands.txt
./output_data.csv
./NYCInspectionReducer.java
./NYCInspection.java
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ ls
NYCInspection.class  NYCInspectionMapper.class  NYCInspectionReducer.class  NYC_Restaurent_Inscp_Data.csv           an4758_part1         build.sh      commons-csv-1.10.0.jar  output_data.csv           part1_output
NYCInspection.java   NYCInspectionMapper.java   NYCInspectionReducer.java   NYC_Restaurent_Inscp_Data_First100.csv  an4758_part1.tar.gz  commands.txt  nyc-inspection.jar      output_data_First100.csv
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$ pwd
/home/an4758_nyu_edu/rbda_project_part1
an4758_nyu_edu@nyu-dataproc-m:~/rbda_project_part1$
```

Figure 11: Preparing the files to download

Finally, I downloaded the file opened the processed output CSV to verify if it is correctly processed. As shown below, the file contains the correct header followed by the correct row entries generated by the reducer. I also can see that my Grade and Score values are properly processed.



output_data

INSPECTION DATE	CAMIS	DBA	BORO	BUILDING	STREET	ZIPCODE	PHONE	CUISINE DESCI	ACTION	VIOLATION CODE	VIOLATION DESCRIPTION	CRITICAL FLAG	SCORE	GRADE	INSPECTION TYPE
03/14/2025	50156692	MEMO'S RESTAURANT	Queens	90-21	31 AVENUE	11369	3478845775	Spanish	Violations were cited in the fo	28-06	Contract with a pest management profes	Not Critical	13	A	Pre-permit (Operational) / Re-inspection
02/21/2025	50142907	ROSTICCERIA EVELIN	Brooklyn	455	MYRTLE AVENUE	11205	6465529587	Italian	Violations were cited in the fo	10F	Non-food contact surface or equipment	Not Critical	13	A	Pre-permit (Operational) / Initial Inspection
06/30/2025	50140743	RED KUP	Manhattan	701	SAINT NICHOLAS AVE	10031	9293019405	Coffee/Tea	Violations were cited in the fo	06D	Food contact surface not properly washe	Critical	67	C	Cycle Inspection / Initial Inspection
07/15/2025	50110670	ANYTIME BAR & BILLI	Manhattan	112	WEST 30 STREET	10001	6466372967	American	Violations were cited in the fo	06D	Food contact surface not properly washe	Critical	33	C	Cycle Inspection / Initial Inspection
01/16/2025	50105561	DAVIDOVICH BAKERY	Manhattan	79	CLINTON STREET	10002	5168280218	Bakery Products	Violations were cited in the fo	08A	Establishment is not free of harborage or	Not Critical	29	C	Cycle Inspection / Initial Inspection
09/05/2024	50150570	PANERA BREAD #6366	Bronx	46	WEST 225 STREET	10463	3473357139	American	Violations were cited in the fo	10F	Non-food contact surface or equipment	Not Critical	25	B	Pre-permit (Operational) / Initial Inspection
04/27/2022	50117186	XIN LA GONG FU	Queens	35-48	UNION STREET	11354	9173316661	Korean	Violations were cited in the fo	05D	Hand washing facility not provided in or	Critical	22	B	Pre-permit (Operational) / Initial Inspection
02/17/2023	50131655	LIBERTY BAGELS	Manhattan	32	BROADWAY	10004	5164583408	Bagels/Pretzels	Violations were cited in the fo	10A	Toilet facility not maintained or provided	Not Critical	21	B	Pre-permit (Non-operational) / Initial Inspection
05/29/2024	50067257	SAJHOMA RESTAURA	Brooklyn	408	NEW LOTS AVENUE	11207	3476737555	Spanish	Violations were cited in the fo	02H	After cooking or removal from hot holdin	Critical	19	B	Cycle Inspection / Initial Inspection
03/05/2025	50130900	PELICANA CHICKEN	Queens	47-08	GREENPOINT AVENU	11104	9293319793	Chicken	Violations were cited in the fo	02G	Cold TCS food item held above 41 °F; sn	Critical	30	C	Cycle Inspection / Re-inspection
01/14/2022	40400209	WHEELER'S	Brooklyn	1705	SHEEPSHEAD BAY R	11235	7186469320	American	Violations were cited in the fo	04H	Raw, cooked or prepared food is adulter	Critical	19	B	Cycle Inspection / Initial Inspection
04/16/2025	40393093	RINCON SALVADOREN	Queens	92-15	149 STREET	11435	5167325528	Latin American	Violations were cited in the fo	04K	Evidence of rats or live rats in establishm	Critical	43	C	Cycle Inspection / Initial Inspection
10/16/2023	50041617	ITTADI GARDEN & GRI	Queens	73-07	37 ROAD	11372	3478662923	Bangladeshi	Violations were cited in the fo	08A	Establishment is not free of harborage or	Not Critical	18	B	Cycle Inspection / Re-inspection
02/15/2024	50072519	MIKE'S DINER	Brooklyn	1454	86 STREET	11228	3475679223	Greek	Violations were cited in the fo	08A	Establishment is not free of harborage or	Not Critical	25	B	Cycle Inspection / Re-inspection
02/07/2025	50144653	CAFE ON 7TH	Brooklyn	493	7 AVENUE	11215	3472357158	American	Violations were cited in the fo	04L	Evidence of mice or live mice in establish	Critical	48	C	Cycle Inspection / Initial Inspection
10/25/2022	50104889	DON PEPE TORTAS Y	Brooklyn	3908	5 AVENUE	11232	7184353326	Mexican	Violations were cited in the fo	02G	Cold TCS food item held above 41 °F; sn	Critical	59	C	Cycle Inspection / Initial Inspection
09/03/2025	50139765	BAOZI	Brooklyn	5405	8 AVENUE	11220	6464343333	Chinese	Violations were cited in the fo	05D	No hand washing facility in or adjacent t	Critical	43	C	Cycle Inspection / Re-inspection
04/24/2025	50165223	EATON CAFE	Queens	89-08	QUEENS BOULEVARD	11373	3475070276	Japanese	Violations were cited in the fo	02G	Cold TCS food item held above 41 °F; sn	Critical	42	C	Pre-permit (Operational) / Initial Inspection
06/12/2024	50118305	TAJMAHAL RESTAURA	Brooklyn	473	MCDONALD AVENUE	11218	3472401071	Bangladeshi	Establishment Closed by DO	04K	Evidence of rats or live rats in establishm	Critical	93	C	Cycle Inspection / Initial Inspection
07/17/2023	50116612	ALIDORO	Manhattan	383	WEST 31 STREET	10001	6466882924	Italian	Violations were cited in the fo	04N	Filth flies or food/refuse/sewage associat	Critical	23	B	Pre-permit (Operational) / Initial Inspection
06/20/2024	50121127	BEAR DONUT	Manhattan	40	WEST 31 STREET	10001	2013148342	Donuts	Violations were cited in the fo	08A	Establishment is not free of harborage or	Not Critical	15	B	Cycle Inspection / Initial Inspection
08/21/2025	50127144	IHOP	Brooklyn	2951	AVENUE U	11229	7187192630	Pancakes/Waffle	Violations were cited in the fo	06F	Wiping cloths not stored clean and dry, c	Critical	20	B	Cycle Inspection / Initial Inspection
05/29/2024	50129401	GARDEN BAR & GRILL	Bronx	3392	EAST TREMONT AVEN	10461	7186640566	Latin American	Violations were cited in the fo	04L	Evidence of mice or live mice in establish	Critical	34	C	Cycle Inspection / Initial Inspection
03/31/2025	50001870	THE DUMPLING COVE	Bronx	1530	EAST 222 STREET	10469	7186533143	Caribbean	Violations were cited in the fo	02B	Hot TCS food item not held at or above 1	Critical	35	C	Cycle Inspection / Initial Inspection
05/20/2025	50112033	WANPO TEA SHOP	Manhattan	37	EAST 8 STREET	10003	2129958349	Coffee/Tea	Violations were cited in the fo	02B	Hot TCS food item not held at or above 1	Critical	12	A	Cycle Inspection / Initial Inspection
11/13/2023	41304936	DRAM SHOP	Brooklyn	339	9 STREET	11215	7187881444	American	Violations were cited in the fo	04H	Raw, cooked or prepared food is adulter	Critical	42	C	Cycle Inspection / Initial Inspection
01/06/2025	50123802	GOLDEN GATE EXPRE	Bronx	300	WEST 231 STREET	10463	7188840077	Chinese	Violations were cited in the fo	10F	Non-food contact surface or equipment	Not Critical	22	B	Cycle Inspection / Initial Inspection
04/30/2025	50066563	RONI LIRA BROTHERS	Queens	44-44	COLLEGE POINT BOU	11355	3476100497	American	Violations were cited in the fo	04L	Evidence of mice or live mice in establish	Critical	27	B	Cycle Inspection / Re-inspection
10/21/2025	40679229	AMARANTH	Manhattan	21	EAST 62 STREET	10065	2129806700	Mediterranean	Violations were cited in the fo	10G	Dishwashing and ware washing: Cleanin	Not Critical	13	A	Cycle Inspection / Re-inspection
03/16/2023	41563707	PIZZERIA GIOVE	Staten Island	278	NEW DORP LANE	10306	3472860635	Italian	Violations were cited in the fo	08C	Pesticide not properly labeled or used by	Not Critical	9	A	Cycle Inspection / Initial Inspection
01/19/2022	41710752	PALACE CAFE	Brooklyn	2603	NOSTRAND AVENUE	11210	7183389525	Jewish/Kosher	Violations were cited in the fo	02G	Cold food item held above 41° F (smoke	Critical	12	A	Cycle Inspection / Re-inspection
08/03/2023	50114993	SIP SAK	Manhattan	928	2 AVENUE	10022	2125831900	Turkish	Violations were cited in the fo	02H	After cooking or removal from hot holdin	Critical	55	C	Cycle Inspection / Initial Inspection
07/18/2025	50064557	VENIERO'S BAKERY	Manhattan	340	EAST 11 STREET	10003	2126747070	Bakery Products	Violations were cited in the fo	10B	Anti-siphonage or back-flow prevention	Not Critical	12	A	Cycle Inspection / Initial Inspection
10/13/2021	50107445	HALAL BROS GRILL	Queens	218-74	HEMPSTEAD AVENUE	11429	3479930857	Chicken	Violations were cited in the fo	06C	Food not protected from potential source	Critical	10	A	Pre-permit (Operational) / Initial Inspection
06/26/2023	50101173	KIKU SUSHI	Brooklyn	453	7 AVENUE	11215	7183691155	Japanese	Violations were cited in the fo	04M	Live roaches in facility's food or non-foo	Critical	28	C	Cycle Inspection / Initial Inspection
03/10/2023	50129871	SUSHI D	Brooklyn	207	DEKALB AVENUE	11205	7188580058	Seafood	Violations were cited in the fo	04E	Toxic chemical or pesticide improperly st	Critical	28	C	Pre-permit (Operational) / Initial Inspection
03/13/2023	50133359	AMMI	Manhattan	25	11 AVENUE	10011	2016967222	Bangladeshi	Violations were cited in the fo	06C	Food, supplies, or equipment not protect	Critical	95	C	Pre-permit (Non-operational) / Initial Inspection
12/06/2024	50108155	ABY'S BAR	Brooklyn	1541	MYRTLE AVENUE	11237	3479442568	Spanish	Violations were cited in the fo	04A	Food Protection Certificate (FPC) not hel	Critical	37	N	Cycle Inspection / Initial Inspection
12/27/2022	50056438	KABAYAN FILIPINO RE	Queens	69-12	ROOSEVELT AVENUE	11377	7182054010	Filipino	Violations were cited in the fo	02B	Hot TCS food item not held at or above 1	Critical	27	B	Cycle Inspection / Re-inspection
05/25/2023	50040547	KITCHEN GRILL	Brooklyn	914A	FULTON STREET	11238	7187897800	Indian	Violations were cited in the fo	08A	Establishment is not free of harborage or	Not Critical	16	B	Cycle Inspection / Initial Inspection
03/11/2025	50113746	PANINO	Brooklyn	5401	13 AVENUE	11219	9292899013	Jewish/Kosher	Establishment Closed by DO	05D	No hand washing facility in or adjacent t	Critical	102	N	Pre-permit (Operational) / Re-inspection
12/17/2024	50063458	RUMA'S KITCHEN	Queens	37-01	61 STREET	11377	7188999100	Bangladeshi	Violations were cited in the fo	08A	Establishment is not free of harborage or	Not Critical	53	C	Cycle Inspection / Re-inspection
07/23/2022	50118545	LA FLOR DEL PARAISC	Queens	80-29	JAMAICA AVENUE	11421	9292756890	Spanish	Violations were cited in the fo	02G	Cold TCS food item held above 41 °F; sn	Critical	12	A	Pre-permit (Operational) / Initial Inspection
03/14/2022	50074586	MIKE JR'S RICHMOND	Staten Island	3954	RICHMOND AVENUE	10312	7183172331	American	Violations were cited in the fo	10B	Plumbing not properly installed or mainta	Not Critical	28	C	Cycle Inspection / Initial Inspection

Figure 12: Finished processed output_data.csv file

Conclusion

In this part of the project, I used MapReduce to process the NYC restaurant inspection dataset. I implemented validation, filtering, and column-level transformations in mapper and produced a processed and well-formatted CSV output using the reducer. I also collected specific information using specific counters. I also manually opened the processed csv file and checked it. This completes the data ingestion phase. The cleaned dataset is ready for further analysis using Hive, Trino and Tableau in the next part of the project.