Data Intake Report

Name: **File ingestion and schema validation**
Report date: **1/11/2024**
Internship Batch: **LISUM28**
Version:**1.0**
Data intake by: **Rana Fakeeh**
Data intake reviewer:
Data storage location:
https://www.kaggle.com/datasets/ajohrn/bikeshare-usage-in-london-and-taipei-network?select=london.csv
https://www.kaggle.com/datasets/jfenley/citibike-combined-data-20132019?select=merge_pre_covid_reduced.csv

Tabular data details:

Table 1 - london.csv

Total number of observations	38,215,560
Total number of files	
Total number of features	9 (the primary key is 'rental_id')
Base format of the file	.csv
Size of the data	4.73 GB

Table 2 - merge_pre_covid_reduced.csv

Total number of observations	91,944,421
Total number of files	
Total number of features	9 (the primary key is composite)
Base format of the file	.csv
Size of the data	6.17 GB

Proposed Approach:

- Reading large CSV dataset:
    1. Converting the ***csv*** file into a ***parquet*** format using ***dask***.
    2. Reading and processing the ***parquet*** file using ***pyspark***.
- Dedup validation (identification):
    1. The ***dask*** creates an increasing numerical index for each row.
- Mention your assumptions (if you assume any other thing for data quality analysis).
    1. The ***london.csv*** file was renamed to ***bikes-1.csv***.
    2. The ***merge_pre_covid_reduced.csv*** was renamed to ***bikes-2.csv***.
    3. The ***bikes-1.csv*** is a valid while ***bikes-2.csv*** is invalid, see Figures 1, 2.
    4. The columns of ingested compressed pipe separated file were renamed as per the ***config*** file in addition to creating ***index*** and ***inbound_file*** columns, see Figure 3.

```
config = read_config_file('config.yml')
pprint(config)
```

```
{'column_name_splitter': '_',
 'columns': [{'bike_id': 'int'},
             {'start_time': 'timestamp'},
             {'end_time': 'timestamp'},
             {'start_station_name': 'string'},
             {'end_station_name': 'string'}],
 'inbound': {'delimiter': ',',
             'filename': 'bikes-1',
             'filetype': 'csv',
             'folder': 'data/input/',
             'header': True,
             'skip_rows': 0},
 'outbound': {'delimiter': '|',
              'filename': 'bikes-1',
              'filetype': 'gzip',
              'folder': 'data/output/',
              'header': True}}
```

```
root
 |-- rental_id: double (nullable = true)
 |-- duration: double (nullable = true)
 |-- bike_id: double (nullable = true)
 |-- end_rental_date_time: string (nullable = true)
 |-- end_station_id: double (nullable = true)
 |-- end_station_name: string (nullable = true)
 |-- start_rental_date_time: string (nullable = true)
 |-- start_station_id: double (nullable = true)
 |-- start_station_name: string (nullable = true)
 |-- __null_dask_index__: long (nullable = true)
```

```
Schema validation passed and file accepted.
Writing outbound compressed file ... Done
Reading ingested file ... Done
{'file_size_bytes': 819602386,
 'filename': 'data\\output\\bikes-1\\part-00000-58d06810-125a-45f2-9ded-1014d43a7eda-c000.csv.gz',
 'num_cols': 7,
 'num_rows': 38215560}
Following INBOUND columns are not in CONFIG ['duration', '__null_dask_index__', 'start_station_id', 'end_station_id', 'rental_i
d']
Wall time: 4min 1s
```

*Figure 1: config file, inbound schema, and validation results of **bikes-1.csv** (valid)*

```
config = read_config_file('config.yml')
pprint(config)
```

```
{'column_name_splitter': '_',
 'columns': [{'bike_id': 'int'},
             {'start_time': 'timestamp'},
             {'end_time': 'timestamp'},
             {'start_station_name': 'string'},
             {'end_station_name': 'string'}],
 'inbound': {'delimiter': ',',
             'filename': 'bikes-2',
             'filetype': 'csv',
             'folder': 'data/input/',
             'header': True,
             'skip_rows': 0},
 'outbound': {'delimiter': '|',
              'filename': 'bikes-2',
              'filetype': 'gzip',
              'folder': 'data/output/',
              'header': True}}
```

```
root
 |-- Trip Duration: double (nullable = true)
 |-- Start Time: string (nullable = true)
 |-- Stop Time: string (nullable = true)
 |-- Start Station ID: double (nullable = true)
 |-- End Station ID: double (nullable = true)
 |-- Bike ID: double (nullable = true)
 |-- User Type: double (nullable = true)
 |-- Birth Year: double (nullable = true)
 |-- Gender: double (nullable = true)
 |-- __null_dask_index__: long (nullable = true)
```

```
Schema validation failed and file rejected.
Following CONFIG columns are not in INBOUND ['end_station_name', 'end_time', 'start_station_name']
Following INBOUND columns are not in CONFIG ['Start Station ID', 'Gender', 'Stop Time', 'Birth Year', 'User Type', 'End Station
ID', 'Trip Duration', '__null_dask_index__']
Wall time: 0 ns
```

*Figure 2: config file, inbound schema, and validation results of **bikes-2.csv** (invalid)*

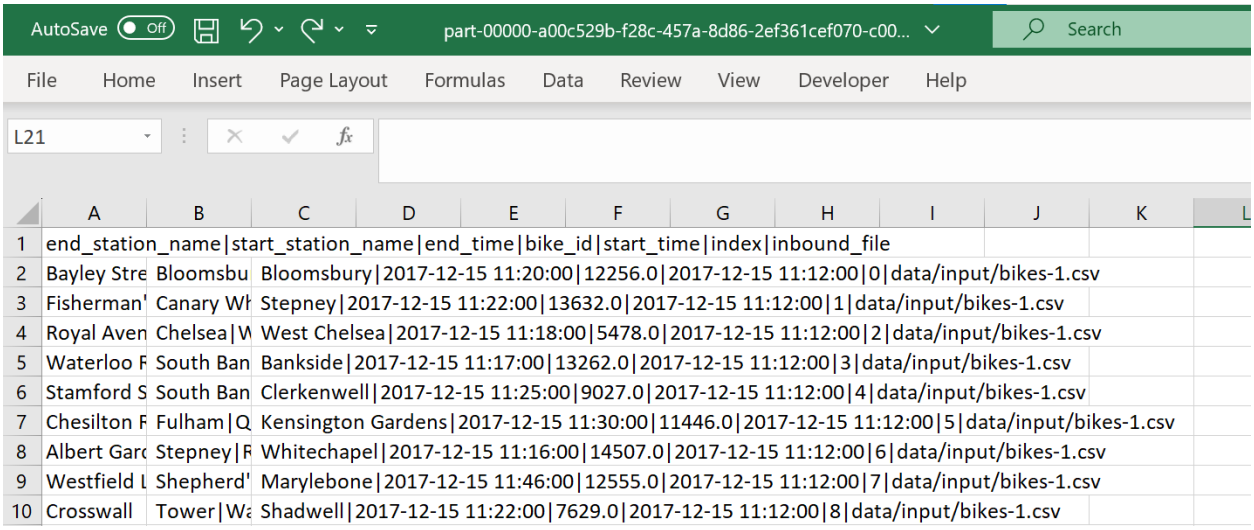

Figure 3: ingested compressed file