

Parquet

April 10, 2021

0.1 3.1.c Parquet

```
[1]: import os
import sys
import gzip
import json
from pathlib import Path
import csv

import pandas as pd
import s3fs
import pyarrow as pa
from pyarrow.json import read_json
import pyarrow.parquet as pq
import fastavro
import pygeohash
import snappy
import jsonschema
from jsonschema.exceptions import ValidationError

endpoint_url='https://storage.budsc.midwest-datascience.com'

current_dir = Path(os.getcwd()).absolute()
schema_dir = current_dir.joinpath('schemas')
schema_dir.mkdir(parents=True, exist_ok=True)
results_dir = current_dir.joinpath('results')
results_dir.mkdir(parents=True, exist_ok=True)

def read_jsonl_data():
    s3 = s3fs.S3FileSystem(
        anon=True,
        client_kwargs={
            'endpoint_url': endpoint_url
        }
    )
    src_data_path = 'data/processed/openflights/routes.jsonl.gz'
    with s3.open(src_data_path, 'rb') as f_gz:
```

```

        with gzip.open(f_gz, 'rb') as f:
            records = [json.loads(line) for line in f.readlines()]

    return records

records = read_jsonl_data()

```

```

[2]: def create_parquet_dataset():
    src_data_path = 'data/processed/openflights/routes.jsonl.gz'
    parquet_output_path = results_dir.joinpath('routes.parquet')
    s3 = s3fs.S3FileSystem(
        anon=True,
        client_kwargs={
            'endpoint_url': endpoint_url
        }
    )

    with s3.open(src_data_path, 'rb') as f_gz:
        with gzip.open(f_gz, 'rb') as f:
            records = [json.loads(line) for line in f.readlines()]
            df = pd.DataFrame(records)
            table = pa.Table.from_pandas(df)
            pq.write_table(table, parquet_output_path)
            return parquet_output_path

    #table2 = pq.read_table(parquet_output_path)
    #table2.to_pandas()
    create_parquet_dataset()

```

```

[2]: PosixPath('/home/jovyan/dsc650/schemas/results/routes.parquet')

```