Avro

April 10, 2021

1 3.1

1.1 3.1.b Avro

[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]:
```
import fastavro
from fastavro.schema import load_schema
from fastavro import writer

def create_avro_dataset(records):
    schema_path = schema_dir.joinpath('routes.avsc')
    #global data_path
    data_path = results_dir.joinpath('routes.avro')
    parsed_schema = load_schema(schema_path)
    with open(data_path,'wb') as out:
        writer(out,parsed_schema,records)


create_avro_dataset(records)
```

[3]:
```
from fastavro import reader
with open('routes.avro', 'rb') as fo:
    avro_reader = reader(fo)
    for record in avro_reader:
        pass
print(record)
```

```
{'airline': {'airline_id': 19016, 'name': 'Apache Air', 'alias': 'Apache',
'iata': 'ZM', 'icao': 'IWA', 'callsign': 'APACHE', 'country': 'United States',
'active': True}, 'src_airport': {'airport_id': 2913, 'name': 'Osh Airport',
'city': 'Osh', 'iata': 'OSS', 'icao': 'UAFO', 'latitude': 40.6090011597,
'longitude': 72.793296814, 'timezone': 6.0, 'dst': 'U', 'tz_id': 'Asia/Bishkek',
'type': 'airport', 'source': 'OurAirports'}, 'dst_airport': {'airport_id': 2912,
'name': 'Manas International Airport', 'city': 'Bishkek', 'iata': 'FRU', 'icao':
'UAFM', 'latitude': 43.0612983704, 'longitude': 74.4776000977, 'timezone': 6.0,
'dst': 'U', 'tz_id': 'Asia/Bishkek', 'type': 'airport', 'source':
'OurAirports'}, 'codeshare': False, 'stops': 0, 'equipment': ['734']}
```