

In [1]:

```

import json
from pathlib import Path
import os

import pandas as pd
import s3fs

def read_cluster_csv(file_path, endpoint_url='https://storage.budsc.midwest-datasience
    s3 = s3fs.S3FileSystem(
        anon=True,
        client_kwargs={
            'endpoint_url': endpoint_url
        }
    )
    return pd.read_csv(s3.open(file_path, mode='rb'))


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

people_json = kv_data_dir.joinpath('people.json')
visited_json = kv_data_dir.joinpath('visited.json')
sites_json = kv_data_dir.joinpath('sites.json')
measurements_json = kv_data_dir.joinpath('measurements.json')

```

In [2]:

```

class KVDB(object):
    def __init__(self, db_path):
        self._db_path = Path(db_path)
        self._db = {}
        self._load_db()

    def _load_db(self):
        if self._db_path.exists():
            with open(self._db_path) as f:
                self._db = json.load(f)

    def get_value(self, key):
        return self._db.get(key)

    def set_value(self, key, value):
        self._db[key] = value

    def save(self):
        with open(self._db_path, 'w') as f:
            json.dump(self._db, f, indent=2)

```

In [3]:

```

def create_sites_kvdb():
    db = KVDB(sites_json)
    df = read_cluster_csv('data/external/tidynomicon/site.csv')
    for site_id, group_df in df.groupby('site_id'):
        db.set_value(site_id, group_df.to_dict(orient='records')[0])
    db.save()

```

```

def create_people_kvdb():
    db = KVDB(people_json)
    df_ppl = read_cluster_csv('data/external/tidynomicon/person.csv')
    for person_id, group_df in df_ppl.groupby('person_id'):
        db.set_value(person_id, group_df.to_dict(orient='records')[0])
    db.save()

def create_visits_kvdb():
    db = KVDB(visited_json)
    df_vis = read_cluster_csv('data/external/tidynomicon/visited.csv')
    for visit_id, group_df in df_vis.groupby('visit_id'):
        db.set_value(str(visit_id), group_df.to_dict(orient='records')[0])
    db.save()

def create_measurements_kvdb():
    db = KVDB(measurements_json)
    df_measr = read_cluster_csv('data/external/tidynomicon/measurements.csv')
    for visit_id, group_df in df_measr.groupby('visit_id'):
        db.set_value(visit_id, group_df.to_dict(orient='records')[0])
    db.save()

```

In [4]:

```

create_sites_kvdb()
create_people_kvdb()
create_visits_kvdb()
create_measurements_kvdb()

```

In [5]:

```

## Trying out writing and reading separately
new_json = kv_data_dir.joinpath('new.pickle')

print(new_json)
kvdb = KVDB(new_json)
key = "(619, 'DR-1')"
print(key)
value = dict(
    visit_id=619,
    site_id='DR-1',
    visit_date='1927-02-08'
)
kvdb.set_value(key, value)
retrieved_value = kvdb.get_value(key)
print("Retrieved Value:", retrieved_value)
kvdb.save()

```

```

/home/jovyan/dsc650/dsc650/assignments/assignment02/results/kvdb/new.pickle
(619, 'DR-1')
Retrieved Value: {'visit_id': 619, 'site_id': 'DR-1', 'visit_date': '1927-02-08'}

```

In []: