

Project:
Data Modeling with Postgres

SEARCH

RESOURCES

CONCEPTS

1. Introduction

2. Project Datasets

3. Project Instructions

4. Project Workspace

5. Project Cheat Sheet

6. Project: Data Modeling with Postg...

Project Datasets

Song Dataset

The first dataset is a subset of real data from the [Million Song Dataset](#). Each file contains metadata about a song and the artist of that song. The files are partitioned by letters of each song's track ID. For example, here are filepaths to two files in this dataset:

song_data/A/B/C/TRABCEI128F424C983.json
song_data/A/A/B/TRAABJL12903CDCF1A.json

And below is an example of what a single song file, TRAABJL12903CDCF1A.json, contains:

{"num_songs": 1, "artist_id": "ARJIE2Y1187B994AB7", "artist_latitude": null, "artist_longitude": null, "title": "Chasing Cars", "year": 2002, "album": "No Fixed Address", "track": 1}

Log Dataset

The second dataset consists of log files in JSON format generated by this [event simulator](#). These simulate activity logs from a music streaming service. The log files in the dataset you'll be working with are partitioned by year and month. Here are filepaths to two files in this dataset:

log_data/2018/11/2018-11-12-events.json
log_data/2018/11/2018-11-13-events.json

And below is an example of what the data in a log file, 2018-11-12-events.json, looks like:

	artist	auth	firstName	gender	itemInSession	lastName	length	level	location	method	page	registration	sessionId	songId
0	None	Logged In	Celeste	F	0	Williams	NaN	free	Klamath Falls, OR	GET	Home	1.541078e+12	438	Nothing But The Rain
1	Pavement	Logged In	Sylvie	F	0	Cruz	99.16036	free	Washington-Arlington-Alexandria, DC-VA-MD-WV	PUT	NextSong	1.540266e+12	345	Mercy: The Last Laugh
2	Barry Tuckwell/Academy of St Martin-in-the-Fields	Logged In	Celeste	F	1	Williams	277.15873	free	Klamath Falls, OR	PUT	NextSong	1.541078e+12	438	Hallelujah
3	Gary Allan	Logged In	Celeste	F	2	Williams	211.22567	free	Klamath Falls, OR	PUT	NextSong	1.541078e+12	438	Nothing But The Rain
4	None	Logged In	Jacqueline	F	0	Lynch	NaN	paid	Atlanta-Sandy Springs-Roswell, GA	GET	Home	1.540224e+12	389	Nothing But The Rain

Mentor Help

Ask a mentor on our Q&A platform

1 Peer Chat

Chat with peers and alumni

If you would like to look at the JSON data within log_data files, you will need to use a dataframe to read the data. Remember to first import JSON and pandas libraries:

```
df = pd.read_json(filepath, lines=True)
```

For example,

```
df = pd.read_json('data/log_data/2018/11/2018-11-01-events.json', lines=True)
```

the data file 2018-11-01-events.json.

In case you need a refresher on JSON file formats, [here is a helpful video](#).