

# Spotify Dataset ETL Project

Huijun (Sunnie) Yan

# Outline

- Objects
- ETL process
  - Extracting the data
  - Transforming the data
  - Loading the data
- Things to consider after ETL
- Limitation & Future work

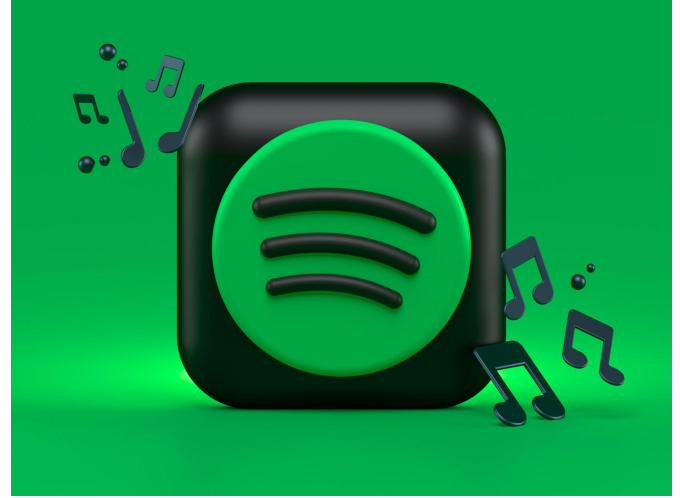
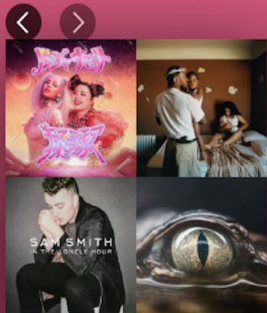


Photo by [Alexander Shatov](#) on [Unsplash](#)



PLAYLIST

# My Playlist









Huijun • 148 songs, about 9 hr



...

# TITLE

ALBUM

1	 <b>Lay Me Down</b> Sam Smith	In The Lonely Hour
2	 <b>Kiss Me More (feat. Naomi Watanabe)</b> Doja Cat, 渡辺直美	Kiss Me More (feat. Naomi Watanabe)
3	 <b>Glimpse of Us</b> Joji	Glimpse of Us
4	 <b>The Heart Part 5</b> Kendrick Lamar	Mr. Morale & The Big Steppers
5	 <b>關於小熊</b> Soft Lipa	收斂水
6	 <b>this is what falling in love feels like</b> JVKE	this is what falling in love feels like
7	 <b>10 Préludes, Op. 23: No. 5, Alla marcia in G Minor</b> Sergei Rachmaninoff, Ratko Delorko	Workday - Home Office - Work at Home: 8 h Classical Music
8	 <b>Caution</b> Cuco	Caution

## Objects:

- Presenting the process of ETL (extract, transform, load)
- Preparing the data relating to [My Playlist] for further analysis

# ETL Process

# 1. Extracting data using Spotify API

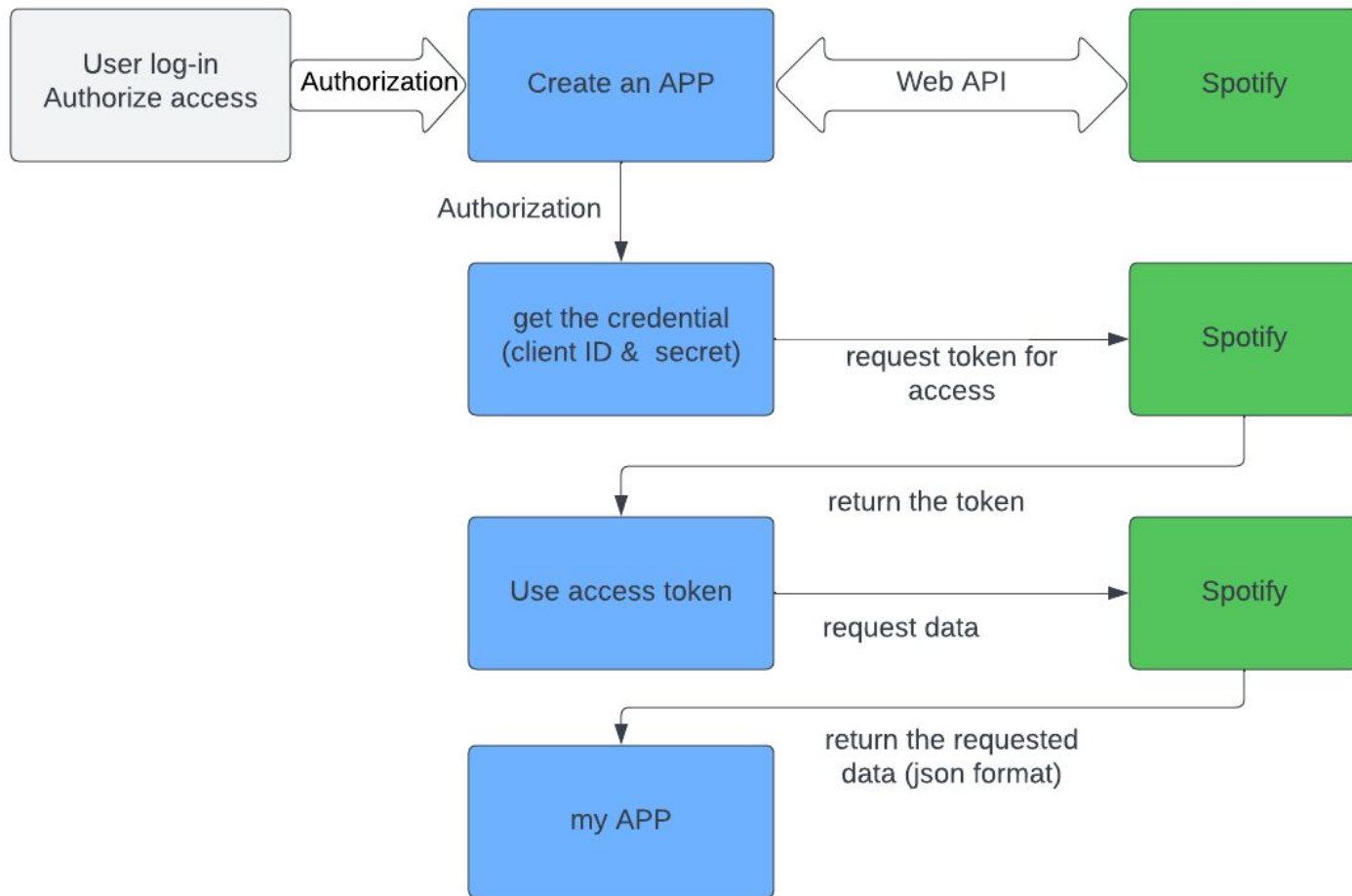
How does Spotify API work?

## Authorization

Authorization refers to the process of granting a user or application access permissions to Spotify data and features. Spotify implements the [OAuth 2.0](#) authorization framework:



Source: <https://developer.spotify.com/documentation/general/guides/authorization/>



# Dashboard



## DataEngineeringProject

CLIENT ID 8ee26094be3742b3a7870f9bb7174c0f

CapstoneProject

← BACK TO DASHBOARD > OVERVIEW

## DataEngineeringProject

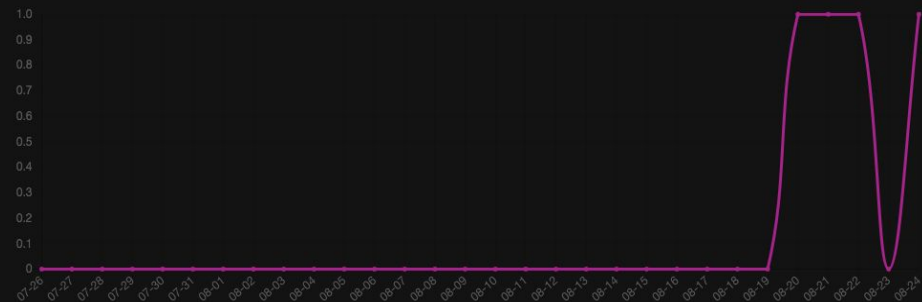
CapstoneProject

App Status Development mode ([what does this mean?](#))

Client ID 8ee26094be3742b3a7870f9bb7174c0f

SHOW CLIENT SECRET

Daily Active Users



ENDPOINTS

Albums

Artists

Get Artist

Get Several Artists

Get Artist's Albums

Get Artist's Top Tracks

Get Artist's Related Artists

Shows

Episodes

Tracks

Get Track

Get Several Tracks

Get User's Saved Tracks

Save Tracks for Current User

Remove Tracks for Current User

Check User's Saved Tracks

Get Tracks' Audio Features

Get Track's Audio Features

Get Track's Audio Analysis

Get Recommendations

Search

Users

Playlists

Get Playlist

Change Playlist Details

Get Playlist Items

Add Items to Playlist

Update Playlist Items

Remove Playlist Items

Get Current User's Playlists

Get Artist's Top Tracks

OAuth 2.0

Get Spotify catalog information about an artist's top tracks by country.

Request

GET /artists/{id}/top-tracks

id string

The Spotify ID of the artist.

Example value: "0TnQYISbd1XYRBk9myaseg"

Query

market string

An ISO 3166-1 alpha-2 country code. If a country code is specified, only content that is available in that market will be returned. If a valid user access token is specified in the request header, the country associated with the user account will take priority over this parameter.

Note: If neither market or user country are provided, the content is considered unavailable for the client.

Users can view the country that is associated with their account in the account settings.

Example value: "ES"

Responses 200 401 403 429

A set of tracks

Body

tracks array of objects

album allOf

The album on which the track appears. The album object includes a link in href to full information about the album.

artists array of objects

The artists who performed the track. Each artist object includes a link in href to more detailed information about the artist.

available\_markets array of strings

A list of the countries in which the track can be played, identified by their ISO 3166-1 alpha-2 code.

## What data did I access?

a. The tracks of one of my playlist

[ My Playlist ]

b. The features of the tracks  
(danceability, beat,  
loudness,etc.)

c. The artists of those tracks

d. The top tracks of the artists



```
{
  "collaborative": false,
  "description": "",
  "external_urls": {
    "spotify": "https://open.spotify.com/playlist/1mRlX7vF1I4CKKavuMRczM"
  },
  "followers": {
    "href": null,
    "total": 0
  },
  "href": "https://api.spotify.com/v1/playlists/1mRlX7vF1I4CKKavuMRczM",
  "id": "1mRlX7vF1I4CKKavuMRczM",
  "images": [
    {
      "height": 640,
      "url": "https://mosaic.scdn.co/640/ab6716d00000b27302e38aa08451ffc986f76247ab6716d00000b2732e02117d76426a08ac7c174fab6716d000b273b11bdc91cb9ac6b14f5c1daaab6716d00000b273f798d46201c266747be5db2e",
      "width": 640
    },
    {
      "height": 300,
      "url": "https://mosaic.scdn.co/300/ab6716d00000b27302e38aa08451ffc986f76247ab6716d00000b2732e02117d76426a08ac7c174fab6716d000b273b11bdc91cb9ac6b14f5c1daaab6716d00000b273f798d46201c266747be5db2e",
      "width": 300
    },
    {
      "height": 60,
      "url": "https://mosaic.scdn.co/60/ab6716d00000b27302e38aa08451ffc986f76247ab6716d00000b2732e02117d76426a08ac7c174fab6716d000b273b11bdc91cb9ac6b14f5c1daaab6716d00000b273f798d46201c266747be5db2e",
      "width": 60
    }
  ],
  "name": "My Playlist",
  "owner": {
    "display_name": "Huijun",
    "external_urls": {
      "spotify": "https://open.spotify.com/user/31m5vsl7nvlpjstu5msbhfyftg4"
    },
    "href": "https://api.spotify.com/v1/users/31m5vsl7nvlpjstu5msbhfyftg4",
    "id": "31m5vsl7nvlpjstu5msbhfyftg4",
    "type": "user",
    "uri": "spotify:user:31m5vsl7nvlpjstu5msbhfyftg4",
    "primary_color": null,
    "public": false,
    "snapshot_id": "MTUwLdQ0YjKxMDRjNTM2YzE1MDNhYzQ2ZWNmNmJMDlKjMkMmExY2YjMTks",
    "tracks": {
      "href": "https://api.spotify.com/v1/playlists/1mRlX7vF1I4CKKavuMRczM/tracks?offset=0&limit=100",
      "items": [
        {
          "added_at": "2022-08-29T19:09:45Z",
          "added_by": {
            "external_urls": {
              "spotify": "https://open.spotify.com/user/31m5vsl7nvlpjstu5msbhfyftg4"
            },
            "href": "https://api.spotify.com/v1/users/31m5vsl7nvlpjstu5msbhfyftg4",
            "id": "31m5vsl7nvlpjstu5msbhfyftg4",
            "type": "user",
            "uri": "spotify:user:31m5vsl7nvlpjstu5msbhfyftg4",
            "is_local": false,
            "primary_color": null,
            "track": {
              "album": {
                "album_type": "album",
                "artists": [
                  {
                    "external_urls": {
                      "spotify": "https://open.spotify.com/artist/2wY79sveU1sp5g7SokK0iI"
                    },
                    "href": "https://api.spotify.com/v1/artists/2wY79sveU1sp5g7SokK0iI",
                    "id": "2wY79sveU1sp5g7SokK0iI",
                    "name": "Sam Smith",
                    "type": "artist",
                    "uri": "spotify:artist:2wY79sveU1sp5g7SokK0iI"
                  }
                ],
                "available_markets": [
                  "AD", "AE", "AG", "AL", "AM", "AO", "AR", "AT", "AU", "AZ", "BA", "BB", "BD", "BE", "BF", "BG", "BH", "BI", "BJ", "BN", "BO", "BR", "BS", "BT", "BW", "BY", "BZ", "CA", "CD", "CG", "CH", "CI", "CL", "CM", "CO", "CR", "CV", "CW", "CY", "CZ", "DE", "DJ", "DK", "DM", "DO", "DZ", "EC", "EE", "EG", "ES", "FI", "FJ", "FM", "FR", "GA", "GB", "GD", "GE", "GH", "GM", "GN", "GQ", "GR", "GT", "GW", "GY", "HK", "HN", "HR", "HT", "HU", "ID", "IE", "IL", "IN", "IO", "IS", "IT", "JM", "JO", "JP", "KE", "KG", "KH", "KI", "KM", "KN", "KR", "KW", "KZ", "LA", "LB", "LC", "LI", "LK", "LR", "LS", "LT", "LU", "LV", "LY", "MA", "MC", "MD", "ME", "MG", "MH", "MK", "ML", "MN", "MO", "MR", "MT", "MU", "MV", "MW", "MX", "MY", "MZ", "NA", "NE", "NG", "NI", "NL", "NO", "NP", "NR", "NZ", "OM", "PA", "PE", "PG", "PH", "PK", "PL", "PS", "PT", "PW", "PY", "QA", "RO", "RS", "RW", "SA", "SB", "SC", "SE", "SG", "SI", "SK", "SL", "SM", "SN", "SR", "ST", "SV", "SZ", "TD", "TG", "TH", "TJ", "TL", "TN", "TO", "TR", "TT", "TV", "TW", "TZ", "UA", "UG", "US", "UY", "UZ", "VC", "VE", "VN", "VU", "WS", "XK", "ZA", "ZM", "ZW"
                ],
                "external_urls": {
                  "spotify": "https://open.spotify.com/album/"
                }
              }
            }
          ]
        }
      ]
    }
  }
}
```

- a. Observe the data  
(check the keys,  
values)

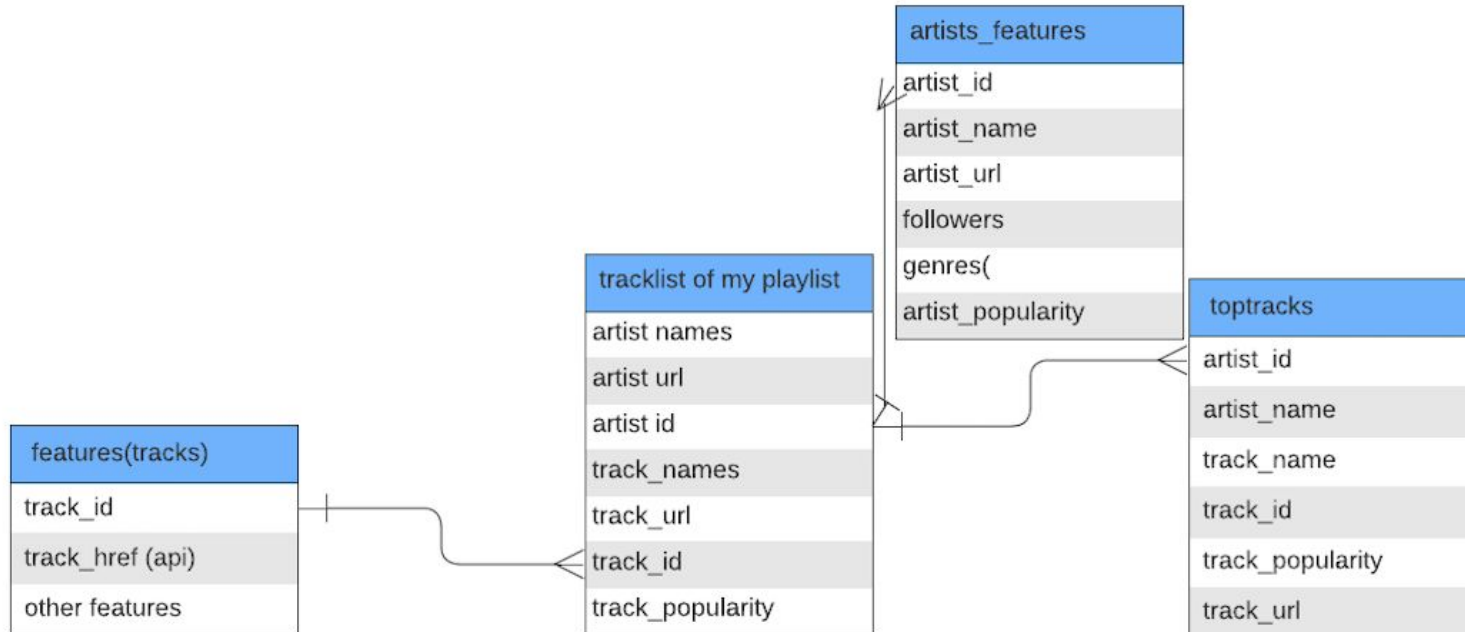
b. Choose the data

### c. Create dataframes

## 2. Transforming the data

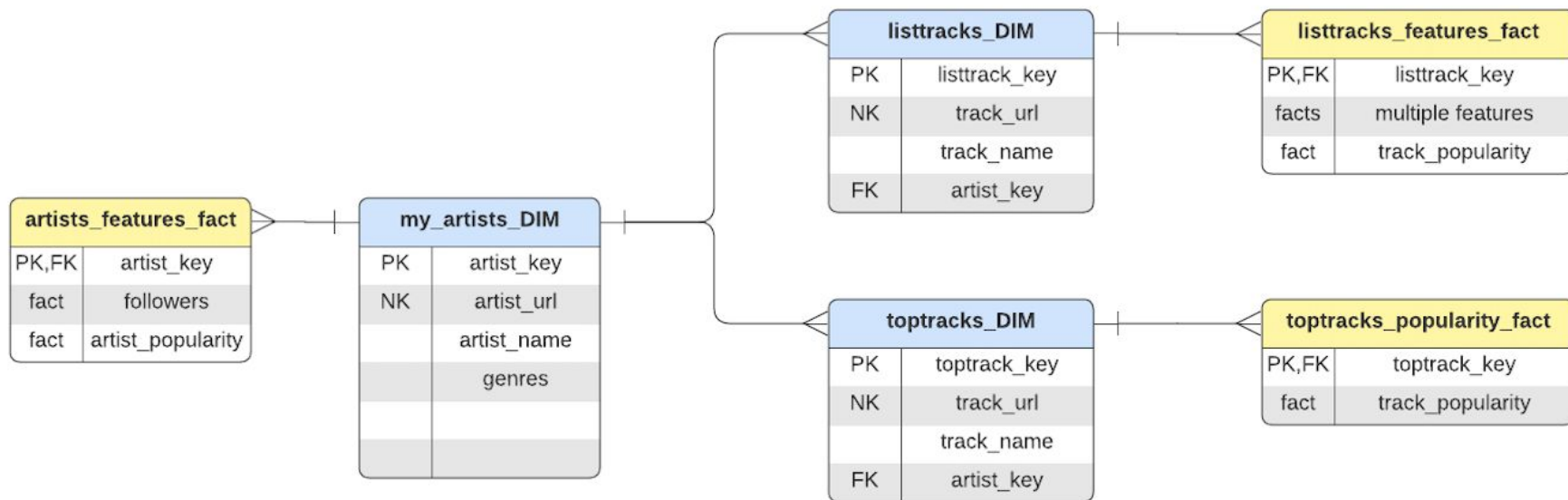
- a. Convert the json format to UTF-8 csv (with the 4 dataframe created)
- b. Clean the data
  - i. Drop duplicate rows
  - ii. Drop meaningless columns
  - iii. Check null values
  - iv. Format the strings
- c. Design the schema to structure the data
- d. Recreate tables based on the schema

## Original unstructured data:



## Spotify dataset schema design

174748594 | August 25, 2022



### 3. Loading the data to Sandbox SQL

a. Create the tables in Sandbox

b. Upload the tables (.CSV)

Challenges:

- i. some toptracks have different artist\_keys (IDs) which are not in my artist list (when there are two or more artists)

**Solution:** only keep the toptracks which are the work of the artists in the artist list.

- ii. Empty strings in the column of 'genres', which are not shown as NULL

**Solution:** replace them into 'not known'

Things to consider after ETL

# After ETL

## a. Quality check

The number of columns,  
rows, nulls, data types etc.

## b. Grant access permission

## c. Security:

Hide the client ID and client  
secret for my APP in Spotify

```
-----GRANT ACCESS-----  
GRANT ALL on TABLE  
    de_challenge.my_artists_DIM,  
    de_challenge.artists_features_fact,  
    de_challenge.listtracks_DIM,  
    de_challenge.listtracks_features_fact,  
    de_challenge.toptracks_DIM,  
    de_challenge.toptracks_popularity_fact  
to "DF_Student";  
-----
```

```
121 SELECT  
122 listtracks.track_name,  
123 artists.artist_name,  
124 artists.artist_key  
125  
126 FROM de_challenge.my_artists_DIM artists  
127 INNER JOIN de_challenge.listtracks_DIM listtracks  
128 on artists.artist_key = listtracks.artist_key;  
129
```

Data output Messages Notifications

	track_name character varying (255) 🔒	artist_name character varying (50) 🔒	artist_key character varying (100) 🔒
1	Lay Me Down	Sam Smith	2wY79sveU1sp5g7Sok...
2	Kiss Me More (feat. Na...	Doja Cat	5cj0LLjcoR7YOSnhnX0...
3	Glimpse of Us	Joji	3MZsBdqDrRTJihTHQr...
4	The Heart Part 5	Kendrick Lamar	2YZyLoL8N0Wb9xBt1...
5	關於小熊	Soft Lipa	3Xp3DA50zRP4TYOtN...
6	this is what falling in lo...	JVKE	164Uj4eKjI6zTBKfJLFK...
7	10 Préludes, Op. 23: N...	Sergei Rachmaninoff	0Kekt6CKSo0m5mivKc...
8	Caution	Cuco	2Tglaf8nvDzwSQnpSrj...

# Limitation & Future work



# Limitation & Future work

- a. Chinese/Korean characters
- b. Limited data size (100 tracks per day of the playlist)
- c. Automation
- d. Interaction (Streamlit)

Thank you!

Questions?