1. Project Overview

A fun attempt to analyse the trends in artists, genres and audio qualities of English and Hindi songs that have reigned the charts over the last forty years

1.1 Objective

The objective of this project is to build a customized base of various English and Hindi hits fetched from curated playlists on the Spotify website and use the same to analyse the trends in the songs of each language through various visualizations.

We will also try to find subtle differences in the Hindi and English songs the ruled the chart in each decade to get an idea of the listening tastes of the masses.

**1.2 Business Case**

NA.

**1.3 Risks**

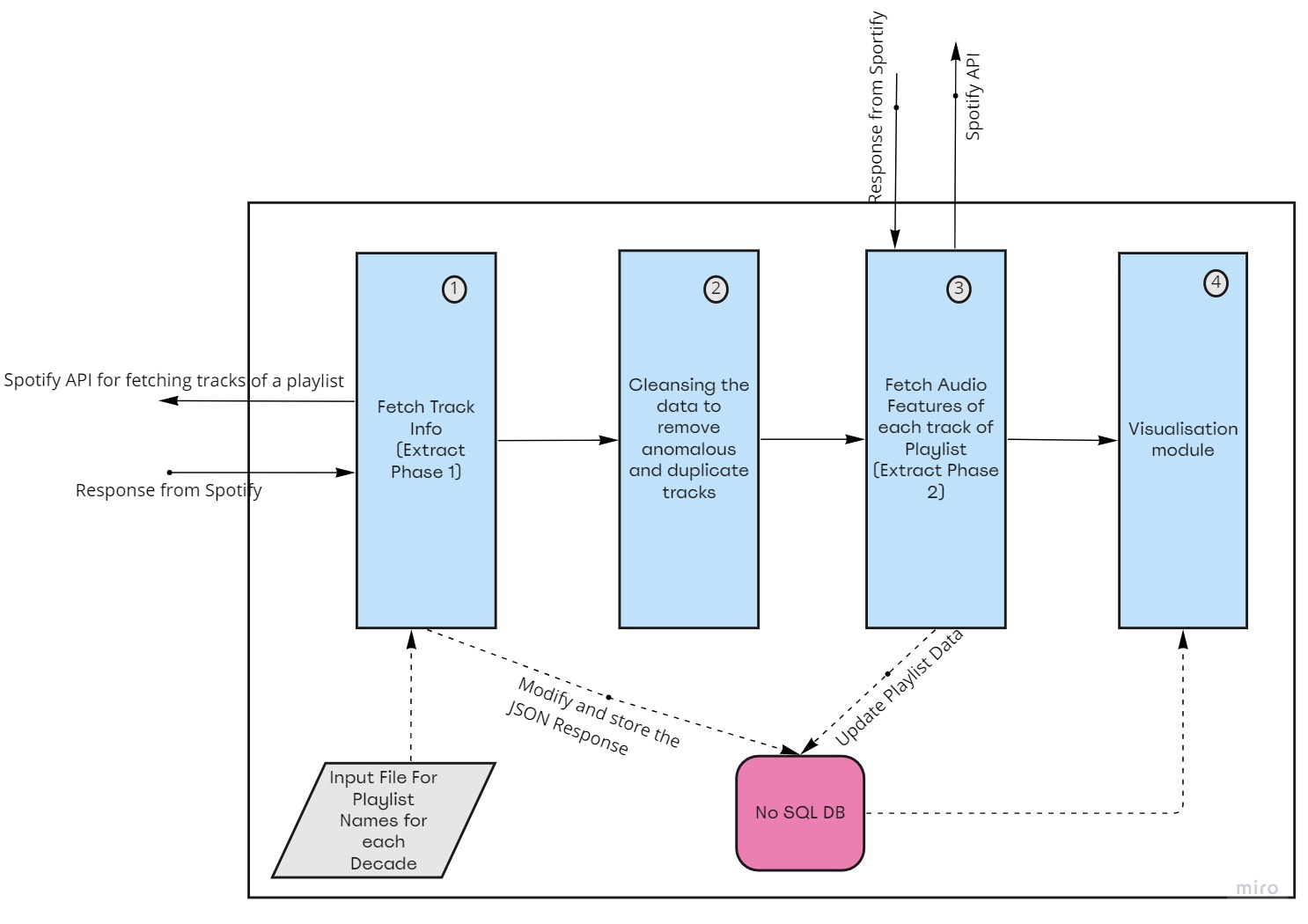
NA.

**1.4 Out of Scope**

This project does not attempt to be a recommendation system of any sort - it will only analyse the data of songs extracted by decade.

2. Technical Design Diagram

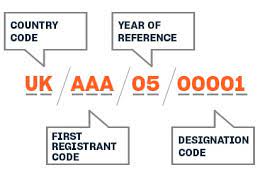
The following diagrams provides the high level flow of the system with inputs/outputs and storage systems.



Before the process begins, a csv file containing the playlist names and playlist ids( Spotify provided id) of all potential playlists for each decade needs to be prepared.

This step has been kept manual as there is no proper naming convention for playlists to determine the containing tracks, hence automating the same is difficult

1. **Fetch (Extract Phase 1)** : Based on the input file, the tracks belonging to each playlist id will be fetched from Spotify. All the relevant fields like track name, track id, duration, album details, artist details and available markets need to be extracted and stored in the database
2. **Cleanse** : Once all the tracks for each playlist are fetched and written into decade-wise collections (Refer Data Model below), the following cleansing operations need to be performed for ensuring data quality:
   1. For each collection, remove the tracks that do not belong to the titled decade ie. The collection titled English Playlists of the 90s should not contain tracks that may have been release before or after the 1990s.
   2. Duplicate tracks in each collection need to be identified and deleted. As tracks from multiple user-curated playlists are being procured, there are high chances of duplicates being present across playlists
   3. The last step in this phase is a transformation. Add a country field to each track for identifying the country of origin. This can be derived from the ISRC of the track that is generated at the time of recording. The first 2 characters of an ISRC code represent the country code of recording. Sample given below



1. **Fetch (Extract Phase 2)** : Once the cleansing is done, the audio features of all the playlists present in the database need to be fetched via Sportify provided APIs for the same.

The audio features include characteristics like acousticness, danceability, tempo, loudness, liveness and speechiness to describe the attributes of a song calculated by spotify.

***Note***: The reason why the cleansing and transformation (Step 2) is to be performed before this step is to restrict the number of tracks being queried in this step and reduce the amount of data being transferred and decrease the number of API calls.

1. **Visualize** : This phase will use graphs and other pictorial representations to bring out the characteristics of each decade. A minimum of three visualization screens are expected:
   1. Identify and represent the top 5 artists of each decade (Hindi/English)
   2. Compare danceability and acousticness quotient of English v/s Hindi Songs
   3. Identify the top few countries making a mark in the Billboard English hits.

3. Technical Specifications

Provide high-level technical specifications with enough detail to develop an infrastructure cost estimate.

3.1 Servers

3.1.1 Application Servers

OS: Windows 8 and above

RAM: Min 4 GB

CPU: Intel i3 (10 Gen) or above

Browser : Google Chrome Version 100.0.0.0 or higher, Edge Version 101.0.1210.47 or higher

3.1.2 File Servers

**NA**

3.1.3 Web Servers

OS: Windows 10 or above

RAM: Min 8 GB

CPU: Intel i3 (10 Gen) or above

Browser : Google Chrome Version 100.0.0.0 or higher, Edge Version 101.0.1210.47 or higher

Software:

Python 3.7

3.1.4 Database Servers

OS: Windows 10 or above

RAM: Min 8 GB

CPU: Intel i3 (10 Gen) or above

**3.2 Access Requirements**

NA

3.3 Databases

Software: MongoDB 5.0.6, Mongosh 1.2.2

Data Requirement: Approx 200 MB

4. Data Model

The following is the data model of a single collection in mongodb. There will be a total of 8 collections , all of which will be following the same data model as below.

Collection List:

4 collections for Hindi - One for each decade 1980s, 1990s, 2000s and 2010s.

Similar for English

Tentative collection names:

* Hindi\_playlist\_1980s
* Hindi\_playlist\_1990s … etc

As the analysis is to be done at the track level, all the details ie. Album, artists, audio features have been kept in the same collection to promote data locality and prevent unnecessary joins and lookups.

Hindi\_playlist\_1980s

\_id: ObjectId

Track\_id: String

Name: String

album: {

href: String,

name: String,

release\_date: String

},

Artists: [{

Href: String

Id: String

Name: String

Type: String

URI: String

}

],

Available\_markets: <Array of country codes>

Duration\_ms: Integer

Explicit: Boolean

External\_ids : { isrc: String}

Popularity: Integer

Country: String

acousticness: Double, danceability: Double,

energy: Double,

instrumentalness: Double,

key: Integer,

liveness: Double,

loudness: Double,

mode: Integer,

speechiness: Double,

tempo: Double,

time\_signature: Integer,

Uri : String,

valence: Double

Document Tracking

The following chart is used to log of all changes made to this document.

| **Version** | **Date of edit/change** | **Who made the edit/change** | **Description of edit/change** |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |