**Project 2**

**Project 2.1 - Music Data Analysis**

Student Name: Abarajithan SA

Course: Big Data Hadoop & Spark Training

Contents

[Project Overview 2](#_Toc503992066)

[Fields present in the data files 2](#_Toc503992067)

[LookUp Tables 2](#_Toc503992068)

[DATASET: 3](#_Toc503992069)

[Data Enrichment 3](#_Toc503992070)

[Data Analysis (SHOULD BE IMPLEMETED IN SPARK) 3](#_Toc503992071)

[Challenges and Optimizations: 4](#_Toc503992072)

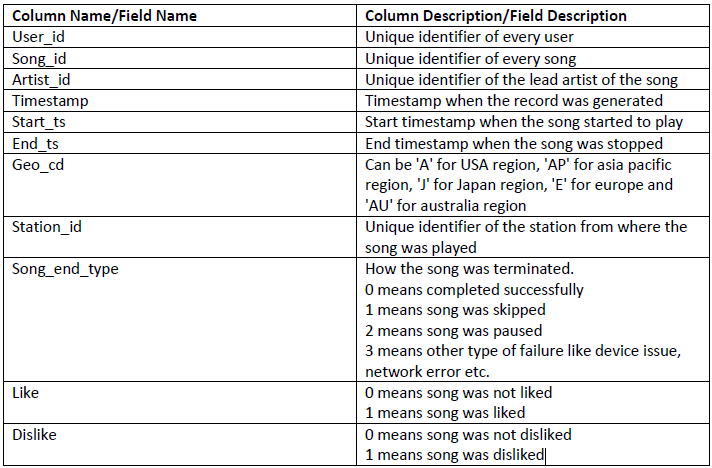
[Flow of operations 4](#_Toc503992073)

# Project Overview

A leading music-catering company is planning to analyze large amount of data received from varieties of sources, namely mobile app and website to track the behavior of users, classify users, calculate royalties associated with the song and make appropriate business strategies. The file server receives data files periodically after every 3 hours.

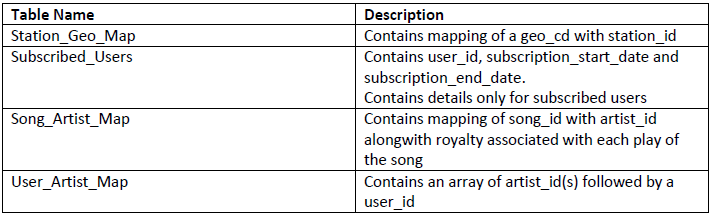
# Fields present in the data files

Data files contain below fields.



# LookUp Tables

There are some existing look up tables present in **NoSQL** databases. They play an important role in data enrichment and analysis.



# DATASET

1. Data coming from web applications reside in /data/web and has xml format.
2. Data coming from mobile applications reside in /data/mob and has csv format.
3. Data present in lookup directory should be used in HBase.

Below is the link for same.

<https://drive.google.com/drive/folders/0B_P3pWagdIrrMjJGVlNsSUEtbG8?usp=sharing>

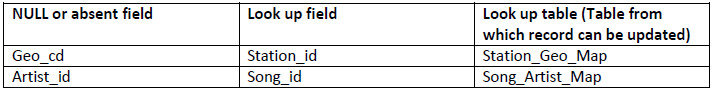
# Data Enrichment

Rules for data enrichment,

1. If any of like or dislike is NULL or absent, consider it as 0.
2. If fields like Geo\_cd and Artist\_id are NULL or absent, consult the lookup tables for fields

Station\_id and Song\_id respectively to get the values of Geo\_cd and Artist\_id.

1. If corresponding lookup entry is not found, consider that record to be invalid.



# Data Analysis (SHOULD BE IMPLEMETED IN SPARK)

It is not only the data which is important, rather it is the insight it can be used to generate important. Once we have made the data ready for analysis, we have to perform below analysis on a daily basis.

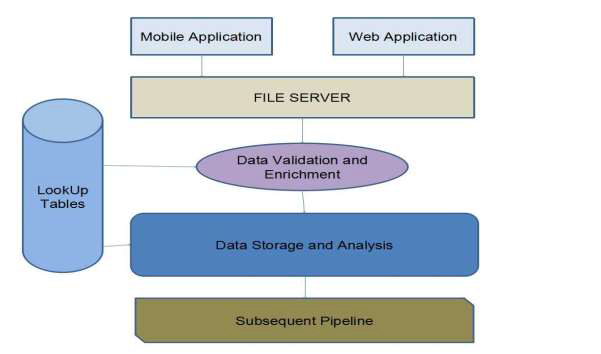
1. Determine top 10 station\_id(s) where maximum number of songs were played, which were liked by unique users.
2. Determine total duration of songs played by each type of user, where type of user can be 'subscribed' or 'unsubscribed'. An unsubscribed user is the one whose record is either not present in Subscribed\_users lookup table or has subscription\_end\_date earlier than the timestamp of the song played by him.
3. Determine top 10 connected artists. Connected artists are those whose songs are most listened by the unique users who follow them.
4. Determine top 10 songs who have generated the maximum revenue. Royalty applies to a song only if it was liked or was completed successfully or both.
5. Determine top 10 unsubscribed users who listened to the songs for the longest duration.

# Challenges and Optimizations:

1. LookUp tables are in NoSQL databases. Integrate them with the actual data flow.
2. Try to make joins as less expensive as possible.
3. Data Cleaning, Validation, Enrichment, Analysis and Post Analysis have to be automated. Try using schedulers.
4. Appropriate logs have to maintain to track the behavior and overcome failures in the pipeline.

# Flow of operations

A schematic flow of operations is shown below,



In the following sections, we are going to see the Music Data Analysis as per the above rules.