Overview

The purpose of the project is to build a predictive model to recommend songs based on some characteristics of the user. The datasets include users file, artist file and the ‘listens’ file containing the song_id, genre, date.

Batch processing

The first step is to extract the data, load the data into separate tables, then it comes to the cleaning data, like constraint validation and cleansing pattern, regex validation, corrupt data validation etc.

The second step: Join the user table with the listens, and then join the artist table . The merged table would contain the user info, the songs have been played as well as the artist , genre info.

Above is the ETL process.

Streaming processing

When users are enjoying the songs, we update the merged table in the batch processing with the real-time data including the user_id, age, gender, songs_id, artist, date and so on.

In order to build the predictive model, following attributes have been considered: age, gender, songs played, artist, genre, listened_date.

Build a model

To build the predictive model, we use alternating least squares (ALS) algorithm. This approach analyses the implicit data like the user’s behaviour history, a list of songs they like and so on.

We evaluate the recommendation by measuring the Mean Squared Error of tracks prediction.

If the recommended songs have been played a lot of times, that means our algorithm works fine.

Summary:

The project could be implemented using Spark platform, which is easy to process the streaming data and the real-time machine algorithm.

Disadvantages:

--These systems often require a large amount of existing data on a user in order to make accurate recommendations

--In many of the environments in which these systems make recommendations, there are millions of users and products. Thus, a large amount of computation power is often necessary to calculate recommendations.

Ways to improve:

To get more accurate result, we need to collect more data include the following:

---Asking if a user is satisfied with the recommendations and get their feedback.

---Collecting the songs ,artists or genre a user searched in browser, Youtube.

---Presenting two items to a user and asking him/her to choose the better one of them.

--- analyze the list of favorite songs of the users