# Data Engineer technical test Overview

This test is for the Data Engineer position at DueDil. There is no time limit for this test.  Go to the following URL, and download the dataset on sampled Last.fm usage:

http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/lastfm-1K.html

Please answer the questions below using either Python, Scala or Java, preferably in Spark. For each question, describe the approach you used, and provide all output and scripts, supporting files, queries, commands, etc. that you wrote to solve the problem. Your solution should be completely runnable from the command line (assume a Unix based OS) and include clear documentation and testing.

### *Part A*

Create a list of user IDs, along with the number of distinct songs each user has played.

### *Part B*

Create a list of the 100 most popular songs (artist and title) in the dataset, with the number of times each was played.

### *Part C*

Say we define a user's "session" of Last.fm usage to be comprised of one or more songs played by that user, where each song is started within 20 minutes of the previous song's start time. Create a list of the top 10 longest sessions, with the following information about each session: userid, timestamp of first and last songs in the session, and the list of songs played in the session (in order of play).