

Viooh Data Engineer technical challenge



This challenge is meant to demonstrate your technical ability and your proficiency with Data technologies. You are welcome to use any software package and consult external reference materials (web, textbooks, articles, product manuals, etc.), but all work must be your own.

Please be aware that this test requires a sizeable download of an open dataset, so you may want to do the download in advance of starting the test.

Please describe the approach you used, and provide any scripts, queries, commands, etc. that you wrote to solve the problem, you will be asked to give detailed commentary of your work on the face to face interview, if we found your work satisfactory.

Please download the following dataset which contains the song played by [Last.fm](http://last.fm)'s users.

<http://ocelma.net/MusicRecommendationDataset/lastfm-1K.html>

We recommend reading the documentation page and familiarise with the dataset before starting. Using any functional programming language (preferably Clojure) and any Big Data tool/framework (preferably Kafka) on any infrastructure of your choice, please answer the following question.

"If we define a user 'session' 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 songs played in the top 50 longest sessions by tracks count. Please provide a tab separated text file with your answer, along with all scripts, tools and info needed to execute the test."

If you have any doubt about this test feel free to send us your questions via your recruitment agent.