CASE STUDY – ETL Engineer

Hi and welcome to the trivago ETL Engineer challenge. Your application made us curious and therefore we would like to take you to the next step. Here is the chance for you to convince us that you are the right person for the job!
We wish you good luck!

PREPARATION TIME: approximately 120-150 minutes

SUBMISSION DEADLINE: Please send back your results by email within one week of receiving this study

HOW TO SUBMIT: Please provide us with your solution in an archive as an email attachement or link to a cloud upload

THE CHALLENGE

Situation: Attached you will find a csv file. Using this file, please work on the following tasks:

Task 1: Write a query that returns the ten most searched destinations for each 10 minute interval since 00:00:00

Task 2: Write a query that for each user returns the second distinct hotel that they clicked on, or null if the user did not click on two distinct hotels.

Task 3: Assuming that input files identical to the provided one are transferred to your file system at /uploads/clicklog_yyyy-mm-dd/, create a job that inserts the data in those files into a table in your database, and then writes the result from task 1 and 2 to separate tables.

Task 4: schedule this job to execute as soon as the input file is available on the filesystem and describe your approach
You can come up with a solution for the systems and tools you are most comfortable with, e.g. a mySQL server with stored procedures, cronjobs or hadoop and Hive.

Please provide us with:
• the queries for tasks 1-3
• the table and job definitions from task 3
• a description of your approach for task 4

Bonus points for (but not mandatory to proceed!):

- guidelines on how to reproduce your solution
- detailed instructions on requirements and dependencies
- insights into your line of thinking
- version history provided through a VCS
- solutions that would scale well with an increasing data set (up to billions of records)