Technical Challenge - Data Engineer

Challenge:

Endpoint: Please check the link https://ourworldindata.org/coronavirus-source-data for the download of the csv files.

The technical challenge consists of two tasks:

Task 1

Create a production-grade data pipeline using python that fetches data from the endpoint and inserts it to a database of your choice. Aside from the columns listed in the *Complete COVID-19 dataset*, extra columns should be provided containing values for all indicators fetched from the *hospitalizations dataset*. The pipeline should be able to update the existing rows if updates are present. Please explain the reasoning behind your decisions in your project *README.md*.

Task 2

Based on the data, critical decisions are made. Therefore we want to make sure the data is up to date and reliable (e.g. numbers from the weekend are sometimes delayed), hence we need to have data quality/availability measuring of some kind. Please implement the solution which would try to resolve this problem and explain reasoning behind it as well in *README.md*.

Timeline:

You are expected to submit your solutions within 7 days from the day you receive the challenge.

How do I submit?

Please upload all your solutions and supporting scripts in a **Private** github repository and make sure you add the contact persons to it:

- Ganesh
- Nacho
- Marko

Once you have completed the challenge please send an email to HR with your github repository name and project link.

All the best !