converse now

Challenge

Context (hypothetical):

- We are building an ASR model for our platform. For training, we need a large and variable amount of transcripted audio, capturing the full extent of voices that we may encounter.
  There will be corner-cases where the model will have a sub-par performance. In those cases, we will have to search for new training data with the characteristics specific to those cases and train the model over them.
  Currently, the data is added to the training dataset by an analyst, which must download the necessary data, filtering it with an ad hoc script. The only tractability relies in a Jupyter Notebook.
  The transcriptions are used in a NLP pipeline, which classifies the intent of the audio transcriptions. Currently the NLP model also gets its data downloaded and processed by hand by an analyst. There is a separate process that tags the intent of the transcriptions.

Problem:

Given the context, we are building a platform to have a robust access to training data. We need you to:

- Download the Firefox's commonvoice dataset (50GB). Assume this is the data that our system generates and stores as a byproduct of production.
- Using that dataset as your example data, build a system that should:
  - Consume and store the data into your platform.
  - Make it queryable through an API.
  - Provide statistics on the stored dataset.
- Considerations
  - You can solve the problem by any means you see fit.
  - You can ask any questions you want to further clarify the problem.
  - The expected time to solve it is one week, please adjust the complexity of the solution accordingly.
  - You can plan possible extensions to your system (given more time and resources) and we can discuss them when reviewing your solution.
  - The system you have built will be presented and explained in another call.
  - Key Points we are interested in:
    - Design decisions and process
    - Assumptions made
    - Possible future extensions