Data Engineering use case

Munich, November 2019

Pipeline Task

Please design and develop a process to perform the following task:

- Search for tweets mentioning #FlixBus, during the last week.
- Store the raw data in a format of your choice.
- From the raw data, extract the following data and store in a format of your choice:
    - Anonymized User id
    - Location of the user
    - Number of followers
    - Date time of the tweet
    - Hashtag (including all hashtags included in the tweet)
    - Number of tweets
    - Is it a retweet?
- The task must run automatically and handle possible situations such as (including, but not limited to):
    - Not availability of data source.
    - Empty results.
    - Changes in the API.
    - Not availability of destination location.
    - Logging.
- Please use Python for setting up the API call and perform the data transformations.
- You can choose the data storage specs.
- The task should be designed in such a way that it is possible to reprocess past dates (single day, date range).
- Show us some sample data collected and transformed with your code.

Pipeline Task

Please prepare a presentation for your case solution. We will discuss it further, but please include all the relevant information. It should include:

- How did you do it.
- How did you structure your ETL process.
- How would you improve your solution .
- Indicate how would you automate the task to run periodically, including (if possible) code snippets or samples.
- Tell us what would be your dream infrastructure to setup this pipeline if you have the time and resources.
- Please give us your feedback on this test ☺

Looking forward to your answer!