Implementing data contracts: schema validation

Bruno Gonzalez
dataqualityguru.substack.com

Data pipelines failures can be a nightmare 😰

And it's annoying when you realize it's because of breaking changes in upstream data sources.

Bruno Gonzalez
dataqualityguru.substack.com

Data contracts can help 🚨

They are *agreements* between data producers and consumers that capture the schema, semantics, distributions, and enforcement policies of the data.

Chad Sanderson

→

Some data providers don't care 😔

You cannot always have an agreement because you consume **third-party data**, and you don't have control over the changes.

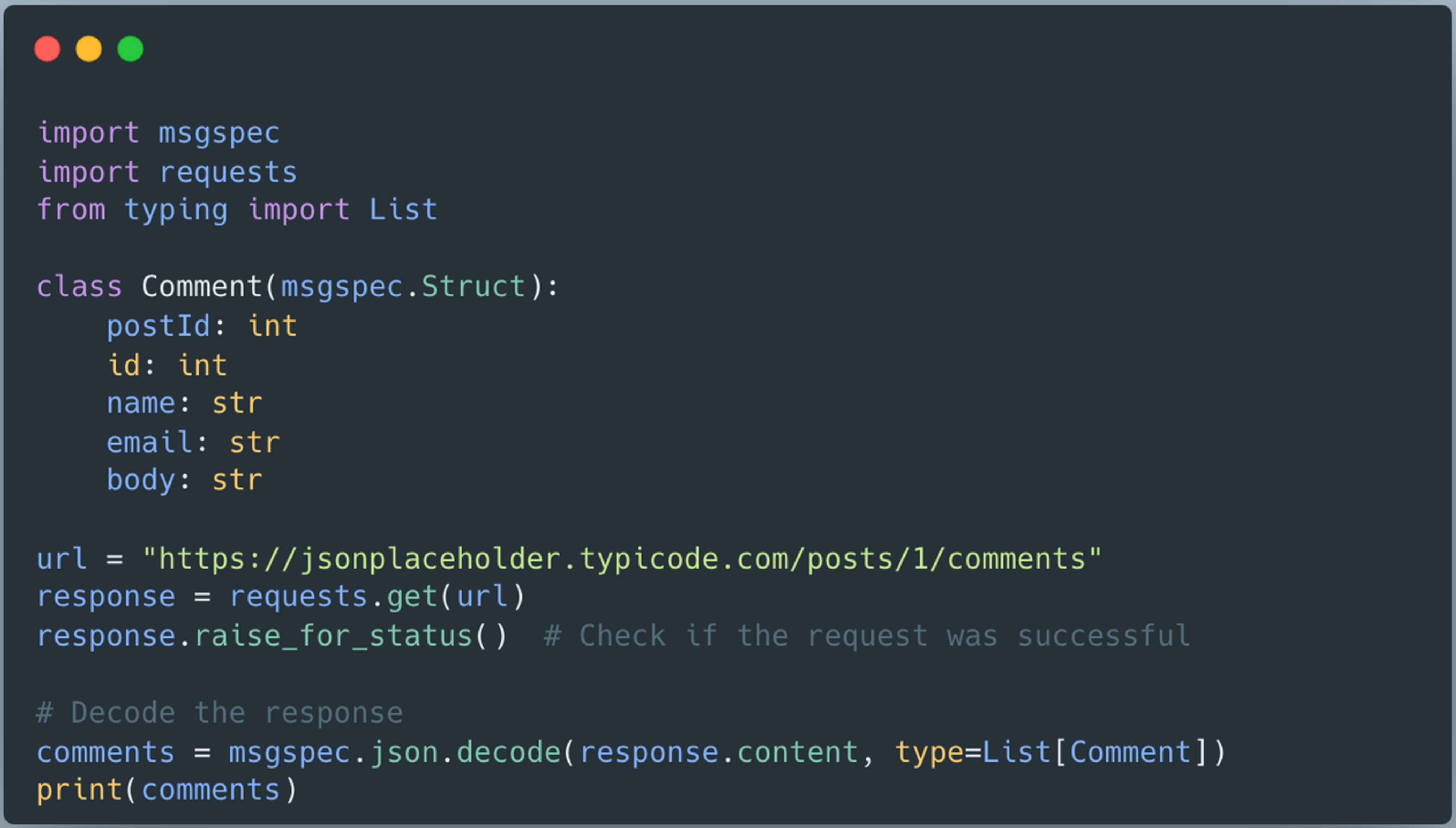
Bruno Gonzalez
dataqualityguru.substack.com

Schema validation ✅

If your data source is an API, you can **efficiently** validate the schema while decoding the API response.

Bruno Gonzalez
dataqualityguru.substack.com

Simple 🍰



```python
import msgspec
import requests
from typing import List

class Comment(msgspec.Struct):
    postId: int
    id: int
    name: str
    email: str
    body: str

url = "https://jsonplaceholder.typicode.com/posts/1/comments"
response = requests.get(url)
response.raise_for_status()  # Check if the request was successful

# Decode the response
comments = msgspec.json.decode(response.content, type=List[Comment])
print(comments)
```

Bruno Gonzalez

dataqualityguru.substack.com

Bonus 🎁

I left a ChatGPT prompt to simplify your validation definition 🤫

Bruno Gonzalez
dataqualityguru.substack.com

Want to read more?

Link to my article in the comments below

Bruno Gonzalez

dataqualityguru.substack.com