Data Eng Technical Challenge Scenario

Imagine we are a conglomerate with many brands, retail stores & e-commerce assets.

We have decided (for recruitment/hiring purposes!) that we want to implement some basic statistical calculations on customer purchase data via a CLI-based application.

The CLI application must:

- Allow input of purchases via a file, containing data in the JSON format specified below.
- Parse the purchase data and calculate the following statistics:
  - total volume of spend
  - average purchase value
  - maximum purchase value
  - median purchase value
  - number of unique products purchased
- Hint: purchase value will need to be computed
- Print results to STDOUT in a JSON format
- Run on either Linux or Mac OS X

purchases_v1.json:


```json
[
    {
        "brand": "newventure.co",
        "customer_id": "a45f2398-3f57-4d83-84bf-87afc31b483a",
        "items": [
            {
                "department": "Tools",
                "price": "249.00",
                "product_category": "Sausages",
                "product_name": "Intelligent Fresh Pizza",
                "quantity": 1
            },
            {
                "department": "Health",
                "price": "366.00",
                "product_category": "Mouse",
                "product_name": "Refined Wooden Hat",
                "quantity": 2
            }
        ],
        "purchase_id": "3655582c-4b0c-4db4-9b53-b2e0d06bba8d"
    },
    {
        "brand": "Hammerbarn",
        "customer_id": "df23cfd4-d200-4f02-962d-78a9e6031f24",
        "items": [
            {
                "department": "Outdoors",
```

```
                "price": "549.00",
                "product_category": "Computer",
                "product_name": "Licensed Soft Table",
                "quantity": 2
            },
            {
                "department": "Electronics",
                "price": "330.00",
                "product_category": "Cheese",
                "product_name": "Rustic Cotton Pizza",
                "quantity": 1
            }
        ],
        "purchase_id": "3731f03f-f7ac-4089-b43d-13d3845b67e0"
    },
```


You can assume the following:

- Implementation should be preferably be in Python - however you may use another language or technology if you are much more comfortable with that choice
- This is a CLI application to run on a single machine - i.e. we are primarily assessing programming skills rather than big data framework knowledge or data infrastructure configuration
- Expected time spent on this exercise is 2 hours
- As per the assessment criteria below, performance is considered but not the most important factor

We are looking for:

- Production grade code, documentation and tests
- All assumptions to be documented
- Nice to have: Dockerisation & providing a command to run

Extra notes:

- You are able to use any open source modules and frameworks you see fit.
- If you would like to implement your application in a stream-oriented fashion, or using a distributed data processing engine, instead of processing a batch file, go for it!
- Feel free to extend the sample purchases.json as you see fit!

How we will evaluate (ordered by importance/weighting):

- Did your CLI application meet the requirements?
- Was there appropriate testing and documentation?
- Is it easy to extend?
- What is the performance of your application?

File	Modified
purchases_v1.json	Jan 21, 2022 by Jonathan Gomez

Drag and drop to upload or browse for files