Data Engineer - Take Home Assignment

Question 1 - Data Pipeline Design

Objectives

- We would like to know your thought process and how you understand & handle the high level thinking around data.

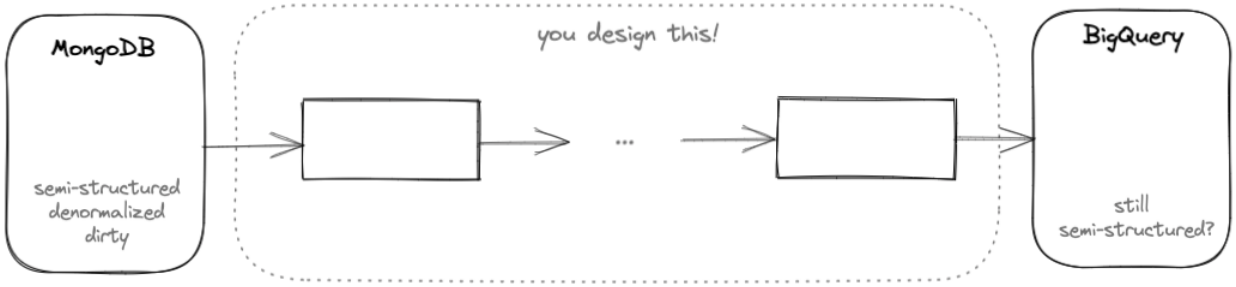
Problems

Assuming that we have

- MongoDB as our source of data (Backend of Micro-services)
    - Master (e.g. Store, Product) & Transactional data (e.g. Sales)
    - Semi structured
    - Denormalized
    - Dirty
- BigQuery as our serving layer
    - Our users have various level of Data Literacy & Competency
        - e.g., Business users with basic SQL knowledge & our friendly Data Scientists
- We mainly rely on GCP/Open-source.

Could you please

1. design the high-level daily-batch data ingestion pipeline to make both Master & Transactional data available at the serving layer (BigQuery)?
    a. hint: Scalable/quality/exception-handling
2. share your thought on how should the data on the serving layer look like? (We don't want to make our users run away, do we? hehe)
    a. hint: Do business people know about Semi-structured data or how to query Array data type?



Next Step

1. Submit your designs/diagrams/notes back to us as .zip, so we can review your answer in advance.
2. Prepare to present/walk us through what you've designed, your thought-process and how you've tackled the problems.

Notes

- If the assignment sound a bit unclear to you, feels free to make any assumptions.
- You don't need to write an essay for this assignment, just a basic diagram with key point noted would be fine for us, however it's completely your choice.
- **There is no right or wrong way to design this.**
    - **We are interested in the assumptions/concepts/thought-processes/decisions that you make and the reasons behind them or how you tackle the problem.**
- It's totally okay if you are not able to fully complete the assignment, let's talk about the ideas and what obstacles you've faced together.

Question 2 - Text Sanitizer

Objectives

- We would like to know how you design the application and how you implement the code based on the given requirements.

Problems

Must-have

Please write a "text sanitizer" application in any OOP languages (Python 3 preferred)

- receive CLI arguments "source" *& "target"*
- read a text file from "source" as an input data
- sanitize the input text (receive string and return string)
    - lowercase the input
    - replace "tab" with "____"
- generate simple statistic
    - count number of occurrence of each alphabet.
- print output(both sanitized text & statistic) to console.

Nice-to-have

Please write the extensible code to support the following requirements

1. the source of input & the form of output might be changed in the future. (e.g. it might read data from database or write to file at the specified arg "location" instead.)
2. we might have more steps to "sanitize" text in the future as well as new statistic calculation.
3. we might receive the "source" & "target" arguments from config file instead of relying on CLI args.

If you have time, we do appreciate if you can also show/tell us how you would make the project as PROD-ready.

Next Step

1. (Preferred) Push to remote git repository (e.g. Github) and send the link back to us or just directly send your project as .zip , so we can review your answer in advance.
2. Prepare to present/walk us through/demo your code and discuss on how you've designed and the reasons behind it.

Notes

- If the assignment sound a bit unclear to you, feels free to make any assumptions.
- You don't need to write state-of-art code, just give us demo/run-able code that cover all the requirements and comment your thought on it if you need to.
- **There is no right or wrong way to do this.**
    - **We are interested in the assumptions/concepts/thought-processes/decisions that you make and the reasons behind them or how you tackle the problem.**
- It's totally okay if you are not able to fully complete the assignment, let's talk about the ideas and what obstacles you've faced together.

Question 3 - SQL

Objectives

- We would like to know your thought process and how you implement SQL based on the given requirements.

Problems

Please write SQL to extract the product names and product classes for the top 2 sales for each product class in our product universe, ordered by class and then by sales. If there are any tie breakers, use the lower quantity to break the tie.

Sales Transaction

transaction_id	product_id	quantity
1	1	5
1	2	7
2	3	1
3	2	3
..	..	..

Product

product_id	product_name	retail_price	product_class_id
1	aa	10	1
2	bb	20	1
3	cc	30	2
..	..	..	..

Product Class

product_class_id	product_class_name
1	Class A
2	Class B
3	Class C
..	..

Expected Output

product_class_name	rank	product_name	sales_value
Class A	1	aa	12345
Class A	2	bb	9999
Class B	1	cc	2500
Class B	2	dd	2500
..	..	..	..

Next Step

1. (Preferred) Push to remote git repository (e.g. Github) and send the link back to us or just directly send your project as .zip , so we can review your answer in advance.
2. Prepare to present/walk us through/demo your code and discuss on your thought process.

Notes

- use BigQuery SQL syntax if possible.
- If the assignment sound a bit unclear to you, feels free to make any assumptions.
- **There is no right or wrong way to do this.**
    - **We are interested in the assumptions/concepts/thought-processes/decisions that you make and the reasons behind them or how you tackle the problem.**
- It's totally okay if you are not able to fully complete the assignment, let's talk about the ideas and what obstacles you've faced together.