Technical Test

Business Intelligence Data Warehouse Developer

Overview

Before any love story begins, we need to make sure we know each other and meet each other's expectations. This is the part where we want to get to know you better.

Since our team builds data products on a daily basis, we want to make sure you do too. If you successfully pass this test, we are more assured that you are just like us and we can build awesome, impactful product together to drive growth, not only for our company but for us, individuals.

So, let's start building NOW!

What are your responsibilities?

- Handling End to End Data Pipelines from heterogeneous input and output to help Analytical purposes
- Become data architect by implementing Data Modelling concept based on data type and use case
- Responsible for managing reliable Single Version of Truth and turning Data into critical information and knowledge that can be used to make sound business decisions
- Define clear and well-scoped requirements documents, including specs and workflows, that is easily understood by technical and non-technical audiences
- Work with Product Managers, Engineers, BI Analytics and various other teams to ship data warehouse feature, updates, and data product improvements

Required qualifications

- Have at least 2 years experience in Data Environment as Data Warehouse Developer
- Deep understanding related to Data Warehouse concepts, architecture and data modelling
- Advanced level SQL querying
- Proficient in Batch and real-time data processing
- Experience using Git
- Python and shell scripting
- Analytical and data-driven, you love digging into the data to understand what's happening and define & measure success on every project
- You're passionate about your customers and always bring questions back to what will serve them best.

Technical Test

Business Intelligence Data Warehouse Developer

Instructions

A. There are 8 tasks in this technical test
B. For each task, we expect below outputs:
    1. SQL Files
    2. Source Code
    3. Python Script
    4. SQL Files
    5. ERD Diagram (in jpg or pdf format)
    6. Any notes application
    7. Any notes application
    8. Any notes application

Question 1

For analysis, create a SQL query to show how many customers use multiple GOJEK services on a daily basis, along with the combination of services used:

order_date	no_of_service	total_customer	detail		order_payment
			order_type	total_customer_per_order_type	
2018-04-03	1	20	CAR	5	CASH
			RIDE	4	
			FOOD	1	
			SEND	10	
2018-04-02	3	15	FOOD,CAR ,RIDE	10	CASH
			RIDE,SEND ,TIX	5	

Several rules for this task:

A. Take only order_status = “Completed”
B. No repetition on details of order_type
    a. E.g. same combinations like RIDE, CAR, SEND and CAR, RIDE, SEND are unacceptable, it needs to be unique
C. Group for each order_payment and its combination
    a. I.e. aggregations by CASH, GOPAY, CASH&GOPAY (ALL)
D. Used date in Jakarta timezone

You will use BigQuery for this task, please follow these steps to access the data source:

A. Join the group by clicking the link and click Ask to Join group
B. After your join request is approved, please go to this link
C. For your script connectivity, you can use this link to download the service account

Question 2

Based on your SQL Query from question 1, please create a job to run the SQL on a daily basis with any programming language and fulfill these conditions:

A. By default, it will run a query with D-1 date as start date and today as the end date
B. It can receive variables sent when executing the script which contains start date and end date for backfill purposes

Question 3

Given a CSV file on this link as the data source, please create a python script to reformat the data to JSON files with output like this

Question 4

You're given this query to look for how many transactions and amounts for this gopay_id = '1234' at Oct 1st,2020 along with its creation date

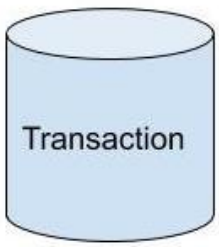
```
WITH transaction AS (
   SELECT * FROM ride
   UNION DISTINCT
   SELECT * FROM bills
   UNION DISTINCT
   SELECT * FROM food
   UNION DISTINCT
   SELECT * FROM pulsa
),
payments AS(
 SELECT
   gopay_id,
   DATE(creation_time) AS creation_date,
   (
   SELECT
     COUNT( DISTINCT transaction_id)
   FROM
     transaction
   WHERE
     DATE(transaction_time) = '2020-10-01'
     AND gopay_id = '1234' ) AS transactions
   (
   SELECT
     SUM( DISTINCT amount)
```

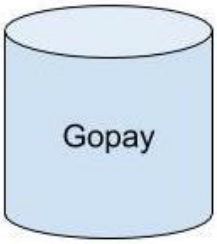
```
   FROM
     transaction
   WHERE
     DATE(transaction_time) = '2020-10-01'
     AND gopay_id = '1234' ) AS amount
 FROM
   gopay
  WHERE
   gopay_id = '1234'
)
SELECT * FROM payments
```

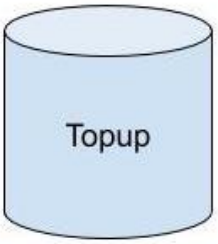
Is there anything that you want improve/optimize from this query? if so please write the optimized version of it in SQL file.

Question 5

Given you have these three tables in lowest/raw layer with datapoints as mentioned







Transaction	• transaction_id • gopay_id • name • amount • description • transaction_time	Gopay	• gopay_id • name • creation_time	Topup	• topup_id • amount • gopay_id • topup_time • transaction_id

With Transaction tables contains these transaction : Topup, Withdraw, Food, Ride, Pulsa, and Bills

Given above materials, draw the ERD diagram on how you will model these in the data warehouse upper layer
You're free to add any additional datapoints or assumptions to make your data model more robust and beneficial to the users.

Question 6

Company X has a large set of data, they still store the data in RDBMS and have difficulty doing reporting tasks. They need daily reporting and send it to management by email. For now, they query the database manually at 7am before operational hours start and save it to another table for reporting and send the report at 9am to management.

A. Do you have an idea for ideal architecture to solve this problem? Please create a diagram and explain your idea.
B. Please mention at least 3 tools or more that we can utilize to solve this problem? Please explain the function of each tool.
C. The company also has data inconsistency and redundancy issues, please explain how to solve this issue?
D. Please explain how to optimize query and increase query performance (at least 3 or more).

Question 7

Please spare 30 minutes to watch below videos about how GOJEK Journey to Next Generation Data Warehouse

- If you have a chance to build the architecture with us, what technology or process that you will replace or add to Next-Gen Data Warehouse, and share us your reasons

Question 8

Here at GOJEK we have 10 core values that we follow:

1) IT'S NOT ABOUT YOU
• Puts the company above themselves
• Obsesses about customer problems, not personal problems
• Has a purpose beyond personal success

2) STAND UP FOR WHAT YOU BELIEVE IN
• Says what they mean
• Has the courage to disagree
• Has a strong moral compass

3) COLLABORATE WITH COMPASSION
• Is a pleasure to work with
• Supports others in areas beyond their scope
• Considers how their actions affect others

4) BE FAST AND FEARLESS

INTERNAL

• Takes calculated risk
• Values failure as much as success
• Has a strong sense of urgency

5) EARN YOUR TITLE
• Walks the talk
• Gets down in the trenches
• Trusts their team

6) BECOME A SCIENTIST
• Follows the numbers intensively
• Finds solutions in unexpected places

• Learns independently and shares knowledge

7) ALWAYS BE PREPARED
• Does their homework
• Plans out every scenario
• Takes action to reduce risk

8) CRITICISM IS A GIFT
• Actively seeks feedback from others
• Gives helpful feedback to others unprompted
• Genuinely acts upon feedback given

9) COMMUNICATE WITH PURPOSE
• States objectives in every interaction
• Aligns early and consistently
• Focuses on what is actionable

10) SHOOT FOR GREATNESS
• Goes the extra mile
• Thinks big
• Loves to challenge themselves

A. Please choose three Core Values from the ten values above that you think is your strength, share us one specific example where you showed it
B. Please choose three Core Values from the ten values above that you think is your weakness, share us one specific example where you showed it
C. Please order the ten Core Values based on your priority if you work at GOJEK and why