Data_Questionnaire

Question 1 (answers)

Data engineer is like a librarian. It is their responsibility to make sure the library runs smoothly. They manage data library to make sure data is collected, organized cleansed and available for people who needs the data to make decisions and learned from it. We could break down the roles of data engineers in a simple terminology which as follows

a) **Book Collection:** They collect books (data) from different places, like bookstores, publishers, and even handwritten notes.
b) **Book Organization:** They arrange the books on the shelves in a way that makes sense, like putting cookbooks together and history books in their section.
c) **Book Cleaning:** Sometimes, the books are dusty or damaged. The Data Engineer cleans them up and fixes torn pages (cleaning and transforming data).
d) **Library Security:** They make sure that only authorized people can access certain sections of the library (data security and privacy).
e) **Checking for Mistakes:** The Data Engineer keeps an eye out for books with missing pages or wrong information and fixes them (data quality checks).
f) **Updating the Library:** When new books arrive, they add them to the library and ensure that everything is in the right place (updating databases and data storage).
g) **Library's Performance:** They make sure the library is organized in a way that anyone can quickly find the book they need (optimizing data for speed and efficiency).
h) **Keeping Records:** Just like a librarian keeps a list of all the books, the Data Engineer maintains a record of all the data in the library (data documentation).
i) **Helping Others:** If someone comes looking for a specific book, the Data Engineer helps them find it (supporting data analysts and scientists).

From a technical point of view, data engineers are responsible for managing and optimizing data infrastructure, pipelines and systems. They make sure that the data is efficiently collected, processed, stored and made it accessible foe those who wants to gain an understanding and make important decisions based on historical data. Below are the key aspects of data engineers:

1. **Data Pipeline Development:**
    - Design and develop data pipelines that collect, process, and transform data from various sources (e.g., databases, APIs, logs, streams) into a usable format.
2. **Data Storage:**
    - Choose and implement appropriate data storage solutions, such as relational databases, NoSQL databases, data lakes, or cloud storage, based on data requirements.
3. **Data Transformation:**
    - Clean, preprocess, and transform raw data into formats suitable for analysis and reporting.

4. **Data Integration:**
    - Integrate data from different sources and systems to provide a unified view of the data landscape.
5. **ETL (Extract, Transform, Load) Processes:**
    - Develop ETL processes to extract data from source systems, transform it, and load it into the target data storage.
6. **Data Modeling:**
    - Create and maintain data models and schemas that facilitate efficient data querying and reporting.
7. **Data Quality and Validation:**
    - Implement data quality checks and validation processes to ensure data accuracy and consistency.
8. **Data Governance and Security:**
    - Implement data security measures and access controls to protect sensitive information. Ensure compliance with data privacy regulations.
9. **Performance Optimization:**
    - Optimize data pipelines and databases for performance, scalability, and reliability.
10. **Monitoring and Maintenance:**
    - Monitor data pipelines and infrastructure for issues and errors, and perform routine maintenance and troubleshooting.
11. **Documentation:**
    - Maintain comprehensive documentation of data pipelines, schemas, and processes for reference and future development.
12. **Collaboration:**
    - Collaborate with data analysts, data scientists, and business stakeholders to understand data requirements and provide the necessary data support.
13. **Continuous Learning:**
    - Stay updated with the latest data engineering technologies and best practices to ensure the organization's data infrastructure remains current and efficient.
14. **Cloud Technologies:**
    - Utilize cloud-based data platforms (e.g., AWS, Azure, GCP) to build scalable and cost-effective data solutions.
15. **Automation:**
    - Automate routine tasks and processes to improve efficiency and reduce manual intervention.

Question 2 (answers) MySQL

Section (a)

```
CREATE TABLE customers (
    customer_id INT PRIMARY KEY,
    name VARCHAR(100),
    email VARCHAR(100),
    tel VARCHAR(15),
    created_at TIMESTAMP,
    updated_at TIMESTAMP
);
```

```
CREATE TABLE invoice (
    invoice_id INT PRIMARY KEY,
    number VARCHAR(20),
    sub_total DECIMAL(10, 2),
    tax_total DECIMAL(10, 2),
    total DECIMAL(10, 2),
    customer_id INT,
    created_at TIMESTAMP,
    updated_at TIMESTAMP,
    FOREIGN KEY (customer_id)
REFERENCES customers
(customer_id)
);
```

```
CREATE TABLE invoice (
    invoice_id INT PRIMARY KEY,
    number VARCHAR(20),
    sub_total DECIMAL(10, 2),
    tax_total DECIMAL(10, 2),
    total DECIMAL(10, 2),
    customer_id INT,
    created_at TIMESTAMP,
    updated_at TIMESTAMP,
    FOREIGN KEY (customer_id)
REFERENCES customers
(customer_id)
);
```

Section (b)

```
SELECT c.customer_id, c.name, COUNT(il.invoice_id) AS books_purchased
FROM customers c
LEFT JOIN invoice i ON c.customer_id = i.customer_id
LEFT JOIN invoice_lines il ON i.invoice_id = il.invoice_id
GROUP BY c.customer_id, c.name
HAVING COUNT(il.invoice_id) > 5;
```

Section (c)

```
SELECT c.customer_id, c.name
FROM customers AS c
LEFT JOIN invoice i ON c.customer_id = i.customer_id
WHERE i.customer_id IS NULL;
```

Section (d)

```
SELECT c.name AS customer_name, il.description AS book_description
FROM customers AS c
JOIN invoice i ON c.customer_id = i.customer_id
JOIN invoice_lines il ON i.invoice_id = il.invoice_id;
```

Question 3