Week_6_pf

You are a backend developer for an e-commerce website. As part of your job, you should ensure the following while working with data/data requests:

- Make certain that the database does not include any duplicate data.

Let's start with defining a couple of things. For our e-commerce site we use Relational Database, postgresql or AuroraDB if we are going to use AWS. Our data is tabular, and exists in the form of tables. Each row in the table has a unique key, UUID. Next step would be to understand what is duplicate data. We can insert data into the database schema and afterwards we can select such data. SQL is a query language which makes our life easier. The first approach could be: INSERT INTO SELECT DISCTINCT. From the postgres tutorial we can see what DISCTINCT does: "the `DISTINCT` clause will evaluate the duplicate based on the combination of values of these columns." Another way of delineate duplicates is to use COUNT(*) = 0. So far, so good. However when we think our database contains duplicates, we need to stop and rethink the design of the schema. Normalization will help to handle it. Sometimes, we have to create additional tables which contain data which will be used in other tables, so we can use id keys and reference to these tables instead of duplicating data in multiple tables.

- When a client makes a request, he or she must receive valid and consistent data.

Ankur Gupta shows the 6 dimensions of data in the following image

When a client makes a request, a server asks the database layer for data. Database gets back response to the customer's query, because our data has no duplicates, integrated and unique, as result a client gets valid and complete data.

Your manager has asked you to prepare a report answering the following questions below.

1. Describe how you would set up backend support for your website. Your website must include the following: a database, a server setup, and an application interface (API) for the front end to interact with the back end.

Our front-end will be supported by the backend side. A database is postgresql, relational database, open sourced. There are two set-ups: for local development, testing and production. For production we use AuroraDB, amazon's database offering, completely management database. For local development we will use dockerized postgresql, docker-compose file.

Our compose file looks like this one. Only need to install docker, and run docker-compose up. Backend will be based on Spring Boot + JPA + Hibernate. For keeping track of database changes we will use flyaway.

APIs. In order to make our integration with front-end we defines the following API:

API	Description
GET /products	Gets list of available products
GET /products/product_id	Gets product details
POST /products	Insert a new product
PUT /products/product_id	Update existing product
DELETE /product/product_id	Delete existing product
GET /order/customer_id	Gets information related to order of specific customer
GET /payments	Get a list of available payment methods
GET /orders/order_id/status	Get a status of specific order

2. Describe the procedures to ensure the database doesn't include duplicate data. Describe two methods for accomplishing this.

We can use the UNIQUE key. For example (taken from stackoverflow):

```
CREATE TABLE tab(id SERIAL PRIMARY KEY,
            customer_id INT,
            call_time TIMESTAMP,
            employee_id INT,
            note VARCHAR(100),
            CONSTRAINT uc_tab UNIQUE (customer_id, call_time, employee_id)
        );
```

Another solution using EXISTS key

Example taken from another stackoverflow

```
DELETE FROM dupes d
WHERE  EXISTS (
  SELECT FROM dupes
  WHERE  key = d.key
  AND    ctid < d.ctid
  );
```

3. How do you ensure valid and consistent results data when a client requests it? Explain.

The linkedin team defines a number of steps:

- “Define your data quality criteria”. Basically it means standards and expectations that are set for the data. As we defined earlier, the common approach contains the following: ACTVC, accuracy, completeness, timelines, validity and consistency.
- “Perform data profiling.” It is a process which examines the data such as data types, date formats, metrics, ranges, distributions, frequencies and distributions. It is also helps to identify potential issues.
- “Apply data validation plan.” It could be checks and constraints which ensure that data meets the criteria. For example, a validation checkpoint which validates a phone number or email.
- “Conduct data verification tests”. Such tests ensure that our data is accurate, reliable and consistent.

4. Explain how back-end testing would be used to evaluate the solutions/methods you suggested in questions 2 and 3.

Our solution was two folded. One - is a REST API, and the second - database.
To test REST API we can use either: Postman or CURL. If we are going to automate our tests, a better solution would be curl or even python requests library. We also can mock a server, which will listen to specific ports and answer to our requests with defined response bodies. Next step would be integration tests, where a server will be real, and requests are coming from our code. Next, load testing. We would like to know how our backend server handles high load. Forthemore, we came to the database testing, i.e. data testing.

Often our data travels back and forth from UI to the backend and vice versa. We have to map data, in such a way that every required field in UI form, http response is correctly mapped into an appropriate table in the database. If CRUD (create/read/update/delete) occurred, the backend should provide support in the database, like insert, select, update or delete. Next, ACID properties validation.

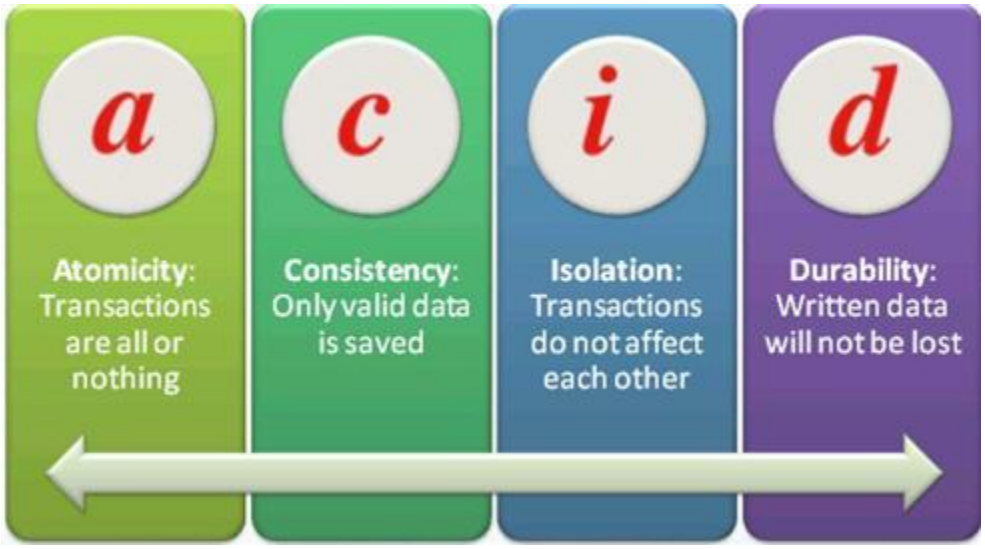

Image was taken from this source

The author shows us how to test such properties:
“**Atomicity test** will ensure any transaction performed on this table is all or none i.e. no records are updated if any step of the transaction is failed.

Consistency test will ensure that whenever the value in column A or B is updated, the sum always remains 100. It won’t allow insertion/deletion/update in A or B if the total sum is anything other than 100.

Isolation test will ensure that if two transactions are happening at the same time and trying to modify the data of the ACID test table, then these tractions are executing in isolation.

Durability test will ensure that once a transaction over this table has been committed, it will remain so, even in the event of power loss, crashes, or errors.”

Forthemore, check for data integrity. Mainly for CRUD operations everything where data is used should show the same and relevant data, on any screed, table etc.

Transactions - should follow: “winner takes at all”. I.e if during the transaction something fails, nothing happens. Only after the transaction succeeded, we updated our database.

Check database schemas, how defined tables, with its relations, primary and secondary keys. No duplicates allowed.
Another factor would be how we restore our database. If our server fails, hacked or something bad happened we have to restore our database. As a part of our testing also should be a step where we are restoring our database.

Self reflection
This week a challenge was to clearly understand how to test the backend, to develop a strategy, how we would test our database and also build a working solution by setting up a server, database and defining REST APIs.

References:

1. How to avoid data duplicates https://codingsight.com/sql-insert-into-select-5-easy-ways-to-handle-duplicates/
2. Examples of postgresql - https://www.postgresqltutorial.com/postgresql-tutorial/postgresql-select-distinct/
3. Ankur Gupta 6 dimensions of data quality https://www.collibra.com/us/en/blog/the-6-dimensions-of-data-quality
4. Principles of Database Management - https://assets.cambridge.org/97811071/86125/frontmatter/9781107186125_frontmatter.pdf
5. https://stackoverflow.com/questions/34132814/best-practices-to-prevent-duplicate-data-pandas-postgres
6. How to validate and verify data quality - https://www.linkedin.com/advice/0/how-do-you-check-data-accuracy-completeness-skills-data-analysis
7. How to test database - https://www.softwaretestinghelp.com/database-testing-process/