Data Engineering on AWS (UDACITY)

Data Modelling

ACID Transactions

Properties of database transactions intended to guarantee validity even in the event of errors or power failures.

Atomicity: The whole transaction is processed or nothing is processed. A commonly cited example of an atomic transaction is money transactions between two bank accounts. The transaction of transferring money from one account to the other is made up of two operations. First, you have to withdraw money in one account, and second you have to save the withdrawn money to the second account. An atomic transaction, i.e., when either all operations occur or nothing occurs, keeps the database in a consistent state. This ensures that if either of those two operations (withdrawing money from the 1st account or saving the money to the 2nd account) fail, the money is neither lost nor created. Source Wikipedia for a detailed description of this example.

Consistency: Only transactions that abide by constraints and rules are written into the database, otherwise the database keeps the previous state. The data should be correct across all rows and tables. Check out additional information about consistency on Wikipedia.

Isolation: Transactions are processed independently and securely, order does not matter. A low level of isolation enables many users to access the data simultaneously, however this also increases the possibilities of concurrency effects (e.g., dirty reads or lost updates). On the other hand, a high level of isolation reduces these chances of concurrency effects, but also uses more system resources and transactions blocking each other. Source: Wikipedia

Durability: Completed transactions are saved to database even in cases of system failure. A commonly cited example includes tracking flight seat bookings. So once the flight booking records a confirmed seat booking, the seat remains booked even if a system failure occurs. Source: Wikipedia.

**When Not to Use a Relational Database**

Have large amounts of data: Relational Databases are not distributed databases and because of this they can only scale vertically by adding more storage in the machine itself. You are limited by how much you can scale and how much data you can store on one machine. You cannot add more machines like you can in NoSQL databases.

Need to be able to store different data type formats: Relational databases are not designed to handle unstructured data.

Need high throughput -- fast reads: While ACID transactions bring benefits, they also slow down the process of reading and writing data. If you need very fast reads and writes, using a relational database may not suit your needs.

Need a flexible schema: Flexible schema can allow for columns to be added that do not have to be used by every row, saving disk space.

Need high availability: The fact that relational databases are not distributed (and even when they are, they have a coordinator/worker architecture), they have a single point of failure. When that database goes down, a fail-over to a backup system occurs and takes time.

Need horizontal scalability: Horizontal scalability is the ability to add more machines or nodes to a system to increase performance and space for data.

Normalization

1. First Normal Form (1NF):

For a table to be in the First Normal Form, it should follow the following 4 rules:

It should only have single (atomic) valued attributes/columns.

Values stored in a column should be of the same domain

All the columns in a table should have unique names.

And the order in which data is stored, does not matter.

Example:

Consider a table:

|  |  |
| --- | --- |
| Customer ID | Purchased Items |
| 1 | Notebook, Pen, Eraser |
| 2 | Notebook, Pencil |
| 3 | Pen |

This table is not in 1NF because the 'Purchased Items' column has multiple values.

A 1NF version of this table would be:

| **Customer ID** | **Purchased Items** |
| --- | --- |
| 1 | Notebook |
| 1 | Pen |
| 1 | Eraser |
| 2 | Notebook |
| 2 | Pencil |
| 3 | Pen |

2. Second Normal Form (2NF):

A table is in the Second Normal Form if:

It is in the First Normal form.

And, it has no Partial Dependency. (A table has a partial dependency if a column depends not only on the primary key but also on another column of the table.)

Example:

Consider a table:

| **StudentID** | **Course** | **Professor** |
| --- | --- | --- |
| 1 | Math | Prof. A |
| 1 | English | Prof. B |
| 2 | Math | Prof. A |
| 3 | English | Prof. B |

This table is in 1NF but not in 2NF because 'Professor' depends on 'Course', not just 'StudentID' (the primary key).

The 2NF version of this table would be:

StudentCourses Table:

| **StudentID** | **Course** |
| --- | --- |
| 1 | Math |
| 1 | English |
| 2 | Math |
| 3 | English |

Courses Table:

| **Course** | **Professor** |
| --- | --- |
| Math | Prof. A |
| English | Prof. B |

3. Third Normal Form (3NF):

A table is in the Third Normal Form if:

It is in the Second Normal form.

And, it has no Transitive Dependency. (A transitive dependency in a table is a condition where a non-prime attribute depends on another non-prime attribute.)

Example:

Consider a table:

| **StudentID** | **Course** | **Professor** | **Professor Office** |
| --- | --- | --- | --- |
| 1 | Math | Prof. A | Office 101 |
| 1 | English | Prof. B | Office 102 |
| 2 | Math | Prof. A | Office 101 |
| 3 | English | Prof. B | Office 102 |

This table is in 2NF but not in 3NF because 'Professor Office' depends on 'Professor', not directly on 'StudentID' (the primary key).

The 3NF version of this table would be:

StudentCourses Table:

| **StudentID** | **Course** |
| --- | --- |
| 1 | Math |
| 1 | English |
| 2 | Math |
| 3 | English |

Courses Table:

| **Course** | **Professor** |
| --- | --- |
| Math | Prof. A |
| English | Prof. B |

Professors Table:

| **Professor** | **Professor Office** |
| --- | --- |
| Prof. A | Office 101 |
| Prof. B | Office 102 |

Remember that normalization, while helpful in most situations, is not always the best approach. It can lead to increased complexity and poorer performance, especially in read-heavy workloads. Therefore, it is important to consider the specific requirements and context of your database before deciding to normalize.

SQL Statement:

Now let's assume that the customer moved and we need to update the customer's address. However we do not want to add a new customer id. In other words, if there is any conflict on the customer\_id, we do not want that to change.

This would be a good candidate for using the **ON CONFLICT DO NOTHING** clause.

INSERT INTO customer\_address (customer\_id, customer\_street, customer\_city, customer\_state)

VALUES

(

432, '923 Knox Street', 'Albany', 'NY'

)

ON CONFLICT (customer\_id)

DO NOTHING;

Now, let's imagine we want to add more details in the existing address for an existing customer. This would be a good candidate for using the **ON CONFLICT DO UPDATE** clause.

INSERT INTO customer\_address (customer\_id, customer\_street)

VALUES

(

432, '923 Knox Street, Suite 1'

)

ON CONFLICT (customer\_id)

DO UPDATE

SET customer\_street = EXCLUDED.customer\_street;

**Data Modeling**

Data modeling is the process of creating a structured plan for data, organizing and defining data elements, and their interrelationships. It's like creating a blueprint before constructing a building.

**When to use data modeling**

Data modeling is used when designing a database. It helps ensure that data is accurate, consistent, and high-quality, and is used across various fields, such as software engineering, business intelligence, system analysis, and more.

**Data Modeling Process**

The data modeling process usually involves three stages:

1. Conceptual: This high-level model defines the main entities, attributes, and relationships.
2. Logical: This detailed model describes the data in as much detail as possible, without regard to how it will be physically implemented.
3. Physical: This model includes all the details needed for implementation, like the specific database system.

**Properties of Relational Data Models**

Relational data models use tables (relations) with rows and columns to represent data. They also include primary keys (unique identifiers for each record in a table) and foreign keys (links between tables).

**ACID Transactions**

ACID stands for Atomicity, Consistency, Isolation, and Durability. It's a set of properties that guarantee reliable processing of database transactions. For example, when you withdraw money from an ATM, ACID ensures the transaction is processed reliably.

**Normalization**

Normalization is the process of structuring a relational database to reduce redundancy and improve data integrity. For example, in a student database, rather than having one table with all the details, you'd separate it into multiple tables (like 'Personal Details', 'Course Details') and link them via keys.

**Fact and Dimension Table Modeling**

This is used in data warehousing. Fact tables contain measurable, quantitative data (like sales), and dimension tables contain descriptive data (like dates, product information). For example, in a sales data warehouse, a fact table might store units sold, and dimension tables could store data on the salesperson, product, and date.

**Star and Snowflake Schemas**

These are methods of organizing tables in a relational database. The star schema is simple and has one fact table with multiple dimension tables. The snowflake schema is a more complex version, with dimension tables also having related tables. For example, a star schema for a bookstore might have a fact table for sales and dimension tables for books and customers. The snowflake schema could further break down the books table into tables for authors and publishers.

**Data Definitions and Constraints**

Data definitions are rules that dictate what data can be entered into a database, like the type of data (integer, text). Constraints are restrictions on the data, like a primary key constraint that ensures the key is unique.

**Properties of NoSQL Data Models**

NoSQL databases are non-relational and can handle a variety of data models, including document, key-value, columnar, and graph. They can handle unstructured data and are often used when scalability and speed are needed over transaction consistency.

**When to use NoSQL Databases**

NoSQL databases are used when dealing with large volumes of data, need to quickly process data, or when data is not structured and can't easily fit into a table format.

**Distributed Database Design**

Distributed databases store data across multiple locations but operate as a single system. They can improve performance, reliability, and scalability. However, their design needs to balance data consistency, availability, and partition tolerance (CAP Theorem).

**CAP Theorem**

CAP Theorem states that in a distributed system, it is impossible to simultaneously provide more than two out of three guarantees: Consistency (all nodes have the same data), Availability (every request receives a response), and Partition tolerance (the system works despite network failures).

**How to create Relational Data Models & Relational Data Modeling with Postgres**

Creating relational data models involves identifying entities, attributes, and relationships. Tools like ER diagrams are often used. With Postgres, you create tables to represent entities, define columns for attributes, and use primary and foreign keys to define relationships.

**How to create NoSQL Data Models & NoSQL Data Modeling with Apache Cassandra**

Creating NoSQL data models involves identifying how data will be accessed, as NoSQL models are often designed around the queries that will be run. In Apache Cassandra, you create tables based on how you want to query data, with each unique query typically requiring its own table.

**Main characteristics of the Kimball Bus Architecture:**

The Kimball Bus Architecture is a popular data warehouse architecture developed by Ralph Kimball, one of the original architects of data warehousing. It's an approach focused on the usability and performance of the data warehouse from a business analysis perspective. The main characteristics of the Kimball Bus Architecture are as follows:

1. **Dimensional Modeling:** The Kimball methodology uses a bottom-up approach where data marts are first built and then combined together to form a complete data warehouse. The data is organized into fact and dimension tables, which is known as a star schema or a snowflake schema. This makes it easy for business users to understand the data warehouse structure and to generate queries.
2. **Integrated Bus Structure:** In the Kimball methodology, all data marts are connected using a consistent and unified bus structure. The bus is essentially a collection of conformed dimensions, which are dimensions that are used in the same way across different fact tables and data marts. This provides an integrated view of the entire business.
3. **Data Mart Driven:** The Kimball architecture is data-mart oriented. A data mart is a subset of the data warehouse that pertains to a specific business line, like sales or finance. Data marts are built focusing on delivering business objectives for departments in the organization.
4. **ETL System:** In the Kimball approach, the ETL (Extract, Transform, Load) system is the cornerstone of the data warehouse. Data from different source systems is cleaned, transformed, and loaded into the data warehouse, ensuring data quality and consistency.
5. **Incremental Development:** The Kimball architecture allows for an iterative and incremental development approach. Data marts can be built one at a time, each time adding new context and value to the data warehouse. This allows for early wins and the ability to show value from the data warehouse project sooner.
6. **Performance Optimized:** The Kimball methodology optimizes for query performance. The use of dimensional models (star schemas or snowflake schemas) helps in achieving faster query results, as they require fewer joins, and are simpler and more intuitive for users to understand.
7. **Business-Friendly:** The dimensional approach used in Kimball is very business-friendly and is readily understandable by business users. The dimensions and facts of a business process are modeled in such a way that it is intuitive to business users.

Overall, the Kimball Bus Architecture is designed to be understandable and high-performing for business users, and flexible and incremental for IT to implement.

