Custom Bootcamp- 29/08/2023, Tuesday

Siddhi Saxena, Batch 05, Room 008, Associate Data Engineer (IDA)

Database management system

-Database Management System is a software or technology used to manage data from a database. Storing in the database, updating an existing database, delete from the database. DBMS is a system that enables you to store, modify and retrieve data in an organized way. It also provides security to the database.

Features of DBMS-

-Data modeling

-Data storage and retrieval

-Concurrency control

-Data integrity and security

-Backup and recovery

Types of Data Languages

-Data Definition Language (DDL)

-Data Manipulation Language(DML)

-Data Control Language(DCL)

-Transactional Control Language(TCL)

Types of database-

-Centralized database

-distributed database

-network database

-NoSQL database

-Cloud database

-relational database

-Object oriented database

-hierarchical database

Relational Data Model

1. It is the theoretical basis of relational databases, which is a technique or way of structuring data using relations, which are grid like mathematical structures consisting of columns and rows.
2. In the relationship model, all data is logically structured within relations, i.e., tavles, etc.
3. Each relation has a name and is formed from named attributes or columns of data.
4. Each tuple or row holds one value per attribute.

Structured Query Language

1. Standardized language used to interact eith relational databases.
2. Allows users to create modify, retrieve and

CODD’S rule-

Rule 0: The Foundation Rule

The database must be in relational form. So that the system can handle the database through its relational capabilities.

Rule 1: Information Rule

A database contains various information, and this information must be stored in each cell of a table in the form of rows and columns.

### Rule 2: Guaranteed Access Rule

Every single or precise data (atomic value) may be accessed logically from a relational database using the combination of primary key value, table name, and column name.

### Rule 3: Systematic Treatment of Null Values

This rule defines the systematic treatment of Null values in database records. The null value has various meanings in the database, like missing the data, no value in a cell, inappropriate information, unknown data and the primary key should not be null.

### Rule 4: Active/Dynamic Online Catalog based on the relational model

It represents the entire logical structure of the descriptive database that must be stored online and is known as a database dictionary. It authorizes users to access the database and implement a similar query language to access the database.

### Rule 5: Comprehensive Data SubLanguage Rule

The relational database supports various languages, and if we want to access the database, the language must be the explicit, linear or well-defined syntax, character strings and supports the comprehensive: data definition, view definition, data manipulation, integrity constraints, and limit transaction management operations. If the database allows access to the data without any language, it is considered a violation of the database.

### Rule 6: View Updating Rule

All views table can be theoretically updated and must be practically updated by the database systems.

### Rule 7: Relational Level Operation (High-Level Insert, Update and delete) Rule

A database system should follow high-level relational operations such as insert, update, and delete in each level or a single row. It also supports union, intersection and minus operation in the database system.

### Rule 8: Physical Data Independence Rule

All stored data in a database or an application must be physically independent to access the database. Each data should not depend on other data or an application. If data is updated or the physical structure of the database is changed, it will not show any effect on external applications that are accessing the data from the database.

### Rule 9: Logical Data Independence Rule

It is similar to physical data independence. It means, if any changes occurred to the logical level (table structures), it should not affect the user's view (application). For example, suppose a table either split into two tables, or two table joins to create a single table, these changes should not be impacted on the user view application.

### Rule 10: Integrity Independence Rule

A database must maintain integrity independence when inserting data into table's cells using the SQL query language. All entered values should not be changed or rely on any external factor or application to maintain integrity. It is also helpful in making the database-independent for each front-end application.

### Rule 11: Distribution Independence Rule

The distribution independence rule represents a database that must work properly, even if it is stored in different locations and used by different end-users. Suppose a user accesses the database through an application; in that case, they should not be aware that another user uses particular data, and the data they always get is only located on one site. The end users can access the database, and these access data should be independent for every user to perform the SQL queries.

### Rule 12: Non Subversion Rule

The non-submersion rule defines RDBMS as a [SQL](https://www.javatpoint.com/sql-tutorial) language to store and manipulate the data in the database. If a system has a low-level or separate language other than SQL to access the database system, it should not subvert or bypass integrity to transform data.

Difference between DBMS and RDBMS-

| **DBMS** | **RDBMS** |
| --- | --- |
| [DBMS](https://www.geeksforgeeks.org/introduction-of-dbms-database-management-system-set-1/) stores data as file. | [RDBMS](https://www.geeksforgeeks.org/rdbms-architecture/) stores data in tabular form. |
| Data elements need to access individually. | Multiple data elements can be accessed at the same time. |
| No relationship between data. | Data is stored in the form of tables which are related to each other. |
| Normalization is not present. | Normalization is present. |
| DBMS does not support distributed database. | RDBMS supports distributed database. |
| It stores data in either a navigational or hierarchical form. | It uses a tabular structure where the headers are the column names, and the rows contain corresponding values. |
| It deals with small quantity of data. | It deals with large amount of data. |
| Data redundancy is common in this model. | Keys and indexes do not allow Data redundancy. |
| It is used for small organization and deal with small data. | It is used to handle large amount of data. |
| Not all Codd rules are satisfied. | All 12 Codd rules are satisfied. |
| Security is less | More security measures provided. |
| It supports single user. | It supports multiple users. |
| Data fetching is slower for the large amount of data. | Data fetching is fast because of relational approach. |
| The data in a DBMS is subject to low security levels with regards to data manipulation. | There exists multiple levels of data security in a RDBMS. |
| Low software and hardware necessities. | Higher software and hardware necessities. |
| Examples:[XML](https://www.geeksforgeeks.org/xml-basics/), Window Registry, Forxpro, dbaseIIIplus etc. | Examples: [MySQL](https://www.geeksforgeeks.org/architecture-of-mysql/), [PostgreSQL](https://www.geeksforgeeks.org/what-is-postgresql-introduction/), SQL Server, Oracle, Microsoft Access etc. |

Entities- An entity may be any object, class, person or place. In the ER diagram, an entity can be represented as rectangles.

Weak Entity- An entity that depends on another entity called a weak entity. The weak entity doesn't contain any key attribute of its own. The weak entity is represented by a double rectangle.

Attributes- The attribute is used to describe the property of an entity. Eclipse is used to represent an attribute.

Key Attribute- The key attribute is used to represent the main characteristics of an entity. It represents a primary key. The key attribute is represented by an ellipse with the text underlined.

Composite Attribute- An attribute that composed of many other attributes is known as a composite attribute. The composite attribute is represented by an ellipse, and those ellipses are connected with an ellipse.

Multivalued Attribute- An attribute can have more than one value. These attributes are known as a multivalued attribute. The double oval is used to represent multivalued attribute.For example, a student can have more than one phone number.

Derived Attribute-An attribute that can be derived from other attribute is known as a derived attribute. It can be represented by a dashed ellipse

Relationships- A relationship is used to describe the relation between entities. Diamond or rhombus is used to represent the relationship.

One-to-One Relationship- When only one instance of an entity is associated with the relationship,

One-to-many relationship-When only one instance of the entity on the left, and more than one instance of an entity on the right associates with the relationship

Many-to-one relationship- When more than one instance of the entity on the left, and only one instance of an entity on the right associates with the relationship

Many-to-many relationship- When more than one instance of the entity on the left, and more than one instance of an entity on the right associates with the relationship

ER Modelling-

Usage of ER Diagrams

1. Helps to conceptualizes the databases.
2. Gives you better understanding of the information to be stored in the database.
3. Reduces complexity.
4. Helps you to describe elements using ER models.
5. Allows users to get a preview of the logical structure of the database.

Symbols in ER Diagram

1. Rectangle: Entity
2. Double Rectangle: Weak Entity
3. Diamond: Relationship
4. Ellipse: Attribute
5. Dotted Ellipse: Derived Attribute
6. Double Ellipse: Multivalued Attribute
7. Lines: Links attributes to entity types and entity types with other relationship types
8. Underlines the attributes: Primary Key

Normalization-

-Normalization is the process of organizing the data in the database.

-Normalization is used to minimize the redundancy from a relation or set of relations. It is also used to eliminate undesirable characteristics like Insertion, Update, and Deletion Anomalies.

-Normalization divides the larger table into smaller and links them using relationships.

-The normal form is used to reduce redundancy from the database table.

1st normal form- The most basis form of data normalization is 1NF. Which ensures there are no 2 same entries in a row .The most impt rule is that each cell must contain a single value and each record should be unique

In 2NF table, all the subsets of data that can be created in multiple rows are kept in separate tables. The primary key should not be dependent on any candidate key. Should follow 1NF rules.

In 3NF, should not have any transitive functional dependencies. Should follow 1st and 2nd NF.

BCNF(BOYCE CODD NORMAL FORM)

Boyce-Codd Normal Form or BCNF is an extension to the third normal form, and is also known as 3.5 Normal Form.

Rules-

It should be in the Third Normal Form.

And, for any dependency A → B, A should be a super key.

The second point sounds a bit tricky, right? In simple words, it means, that for a dependency A → B, A cannot be a non-prime attribute, if B is a prime attribute.

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Database** | **Data Warehouse** |
| Purpose | Is designed to record | Is designed to analyze |
| Processing Method | The database uses the Online Transactional Processing (OLTP) | Data warehouse uses Online Analytical Processing (OLAP). |
| Usage | The database helps to perform fundamental operations for your business | Data warehouse allows you to analyze your business. |
| Tables and Joins | Tables and joins of a database are complex as they are normalized. | Table and joins are simple in a data warehouse because they are denormalized. |
| Orientation | Is an application-oriented collection of data | It is a subject-oriented collection of data |
| Storage limit | Generally limited to a single application | Stores data from any number of applications |
| Availability | Data is available real-time | Data is refreshed from source systems as and when needed |
| Usage | ER modeling techniques are used for designing. | Data modeling techniques are used for designing. |
| Technique | Capture data | Analyze data |
| Data Type | Data stored in the Database is up to date. | Current and Historical Data is stored in Data Warehouse. May not be up to date. |
| Storage of data | Flat Relational Approach method is used for data storage. | Data Ware House uses dimensional and normalized approach for the data structure. Example: Star and snowflake schema. |
| Query Type | Simple transaction queries are used. | Complex queries are used for analysis purpose. |
| Data Summary | Detailed Data is stored in a database. | It stores highly summarized data. |

Difference between Database and Data Warehouse-

Star schema-

A star schema is the elementary form of a dimensional model, in which data are organized into facts and dimensions. A fact is an event that is counted or measured, such as a sale or log in. A dimension includes reference data about the fact, such as date, item, or customer.