**RDBMS NORMALIZER**

GROUP – MALLIKARJUNA (12608389), ADHISHWAR (12608145)

This code normalizes a database by generating SQL queries to restructure data according to specified normal forms (1NF to 5NF). Each normal form has increasingly strict rules for data organization, intended to reduce redundancy and dependency.

Let's go through each part of the code in detail:

**1. Class Definitions: FunctionalDependency and MultivaluedDependency**

The code defines two classes to represent different types of dependencies:

* **Functional Dependency (FunctionalDependency)**:
  + This class takes two parameters: determinants (the columns that determine the value of other columns) and dependents (the columns dependent on the determinants).
  + Example: If column A determines column B, then (A → B) is a functional dependency, meaning the values in B depend on the values in A.
  + The attribute is\_multivalued is set to False, indicating it is not a multivalued dependency.
* **Multivalued Dependency (MultivaluedDependency)**:
  + Similar to a functional dependency, but with is\_multivalued set to True.
  + Represents cases where one attribute determines multiple values of another attribute independently.
  + Example: If A determines multiple independent values of B, it is represented as A -->> B.

**2. Parsing Functional and Multivalued Dependencies: parse\_fd\_file(filepath)**

This function reads functional and multivalued dependencies from a file and returns two lists: fds (functional dependencies) and mvds (multivalued dependencies).

* **File Format**:
  + Lines containing a functional dependency are in the format A -> B.
  + Lines with a multivalued dependency are formatted as A -->> B.
* **Parsing Logic**:
  + The function splits each line based on the dependency operator (-> or -->>).
  + For functional dependencies (->), it adds entries to the fds list.
  + For multivalued dependencies (-->>), it adds entries to the mvds list.

**3. Normalization Checks and Data Cleanup**

* **is\_in\_first\_normal\_form(df)**:
  + This function checks if the data in each column is in atomic form (1NF).
  + In 1NF, each cell must contain a single value (no lists, arrays, or nested dictionaries).
  + If any column contains lists or dictionaries, it returns False, indicating the data is not in 1NF.
* **remove\_duplicate\_rows(df)**:
  + This function removes duplicate rows in the dataset using pandas.drop\_duplicates().
  + Ensuring uniqueness is essential for normalization since duplicate rows could violate higher normal forms.

**4. SQL Query Generation Functions**

Each function here generates SQL code to restructure data according to a target normal form.

* **1NF**:
  + generate\_1nf(df): This function assumes that all data already fits 1NF requirements, so it generates a SQL statement to create the MainData table.
* **2NF**:
  + generate\_2nf(fds, primary\_keys): Applies only to tables with a composite primary key, where no non-key column should be partially dependent on part of the primary key.
  + For each functional dependency where determinants is a subset of primary\_keys, it creates a new table.
* **3NF**:
  + generate\_3nf(fds): 3NF requires that all non-key columns depend only on the primary key. This function generates tables for functional dependencies that meet this condition, with SQL statements to create new tables based on these dependencies.
* **BCNF (Boyce-Codd Normal Form)**:
  + generate\_bcnf(fds): BCNF is stricter than 3NF, requiring that all determinants are superkeys.
  + This function checks if each functional dependency's determinant is a superkey by calling is\_superkey.
  + If not, it generates a new table to remove partial dependencies.
* **4NF**:
  + generate\_4nf(mvds): In 4NF, tables should not contain multivalued dependencies that are not superkeys.
  + This function iterates over the list of multivalued dependencies and creates a new table for each one to eliminate these dependencies.
* **5NF**:
  + generate\_5nf(fds): This form requires decomposition into smaller tables without introducing redundancy. It generates separate tables to satisfy dependencies in 5NF.

**5. Utility Functions**

* **is\_superkey(determinants)**:
  + This placeholder function checks if a set of columns forms a superkey.
  + For simplicity, it always returns True in this code. A more complex implementation would analyze the schema to check if determinants uniquely identifies rows.
* **save\_queries\_to\_file(queries, filename)**:
  + This function saves SQL queries to a specified output file (filename).
  + Each query is written with a comment indicating the table it generates.

**6. Main Workflow: main()**

The main() function ties everything together, guiding the user through the normalization process:

1. **User Input**:
   * Prompts the user for primary keys (needed for determining which dependencies apply to specific normal forms) and the target normal form level (1 to 5).
2. **File Parsing and Data Loading**:
   * Reads dependencies from a text file (fd.txt) and data from an Excel file (data1.xlsx).
3. **Normalization Process**:
   * Depending on the chosen normal form:
     + Generates SQL queries for each level up to the target normal form.
     + Prints each generated query and adds it to all\_queries, a list of all queries to be saved.
4. **Saving SQL Queries**:
   * Calls save\_queries\_to\_file to write all generated queries to an output file, Output.sql.

**Summary**

This code is designed for a database normalization task, taking in functional and multivalued dependencies to analyze and restructure data according to the user-specified target normal form. Each normal form (1NF to 5NF) progressively reduces data redundancy and dependency, helping to maintain data integrity and streamline queries.