Apache Spark 2 using SQL - Basic DDL and DML

1. Understanding Spark SQL DDL & DML

Spark SQL supports **DDL (Data Definition Language)** and **DML (Data Manipulation Language)** operations similar to traditional RDBMS, but with some differences:

Type	Purpose	Examples in Spark SQL
DDL	Define/modify database structures	CREATE DATABASE, CREATE TABLE, DROP TABLE, ALTER TABLE
DML	Manipulate data in tables	INSERT INTO, SELECT, LOAD DATA, INSERT OVERWRITE

Spark SQL executes these commands on **tables in Hive Metastore** or **temporary views** (in-memory) depending on the setup.

2. Basic Flow in Spark 2 with Scala

1. **Create SparkSession** (with Hive support if persistent tables are needed)
2. **Use DDL to create tables/databases**
3. **Use DML to insert, update, or query data**
4. **Execute SQL queries using spark.sql()**

3. Sample Scala Code

Here's a working example:

```scala
import org.apache.spark.sql.SparkSession
object SparkSQLDDL_DML extends App {
 // Create SparkSession with Hive support (needed for permanent tables)
 val spark = SparkSession.builder()
   .appName("SparkSQL DDL DML Example")
   .master("local[*]")
   .config("spark.sql.warehouse.dir", "spark-warehouse")
   .enableHiveSupport()
   .getOrCreate()

 spark.sparkContext.setLogLevel("ERROR")

 // DDL: Create Database
 spark.sql("CREATE DATABASE IF NOT EXISTS sales_db")
 // Use the database
 spark.sql("USE sales_db")
 // DDL: Create Table
```

```
spark.sql(
  """
    |CREATE TABLE IF NOT EXISTS customers (
    |  id INT,
    |  name STRING,
    |  city STRING
    |) USING parquet
  """.stripMargin)
// DML: Insert data into the table
spark.sql("INSERT INTO customers VALUES (1, 'Anil', 'Chennai')")
spark.sql("INSERT INTO customers VALUES (2, 'Priya', 'Bangalore')")
// DML: Select data
val df = spark.sql("SELECT * FROM customers")
df.show()
// DML: Overwrite table
spark.sql("INSERT OVERWRITE TABLE customers VALUES (3, 'Vijay', 'Delhi')")
// View final data
spark.sql("SELECT * FROM customers").show()
}
```

4. Example Output

```
+---+-----+---------+
|id |name |city     |
+---+-----+---------+
| 1 |Anil |Chennai  |
| 2 |Priya|Bangalore|
+---+-----+---------+
+---+-----+-----+
|id |name |city |
+---+-----+-----+
| 3 |Vijay|Delhi|
+---+-----+-----+
```

5. Common DDL Commands in Spark SQL

Command	Description	Example
CREATE DATABASE	Creates a new database	CREATE DATABASE sales_db
USE	Switch database	USE sales_db
CREATE TABLE	Creates a table	CREATE TABLE products (id INT, name STRING) USING parquet
DROP TABLE	Deletes a table	DROP TABLE customers
ALTER TABLE	Modifies table metadata	ALTER TABLE customers RENAME TO clients

6. Common DML Commands in Spark SQL

Command	Description	Example
INSERT INTO	Appends data	INSERT INTO customers VALUES (4, 'Sara', 'Pune')
INSERT OVERWRITE	Replaces existing data	INSERT OVERWRITE customers VALUES (5, 'Raj', 'Mumbai')
SELECT	Query data	SELECT * FROM customers

LOAD DATA	Load file into table	LOAD DATA LOCAL INPATH '/path/file.csv' INTO TABLE customers

7. Quizzes

Q1. In Spark SQL, INSERT OVERWRITE will:

 a) Append new rows to existing table

 b) Delete table schema and create a new table

 c) Replace existing table data

 d) Add data to a temporary view

Q2. Which statement is valid for creating a table in Spark SQL?

 a) CREATE TABLE customers (id INT, name STRING) USING parquet

 b) CREATE TABLE customers id INT, name STRING USING parquet

 c) CREATE TABLE USING parquet customers (id INT, name STRING)

 d) None of the above

Q3. Which is **NOT** a DML command in Spark SQL?

 a) INSERT

 b) SELECT

 c) ALTER TABLE

 d) LOAD DATA

8. Good GitHub Repositories

Here are some **active repositories** you can explore:

1. **Spark SQL Basics** – https://github.com/spark-examples/spark-sql-examples
2. **Spark with Scala & Hive** – https://github.com/awesome-spark/spark-sql
3. **Big Data Spark SQL Examples** – https://github.com/PacktPublishing/Apache-Spark-2x-for-Java-Developers