Ex. No. 5	**CREATING DATA FRAMES USING APACHE SPARK**
Date of Exercise	**08/09/2025**

Aim

To create and manipulate DataFrames using Spark SQL in order to efficiently handle and query structured data.

Description

In Apache Spark, a DataFrame is a distributed collection of data organized into named columns, similar to a relational database table.

- DataFrames provide higher-level abstractions for structured data processing.
- They support SQL queries, aggregation, filtering, and joins.
- DataFrames can be created from RDDs, CSV/JSON files, databases, or in-memory data.
- Using Spark SQL, we can interact with DataFrames using both SQL queries and DataFrame API.

Key features:

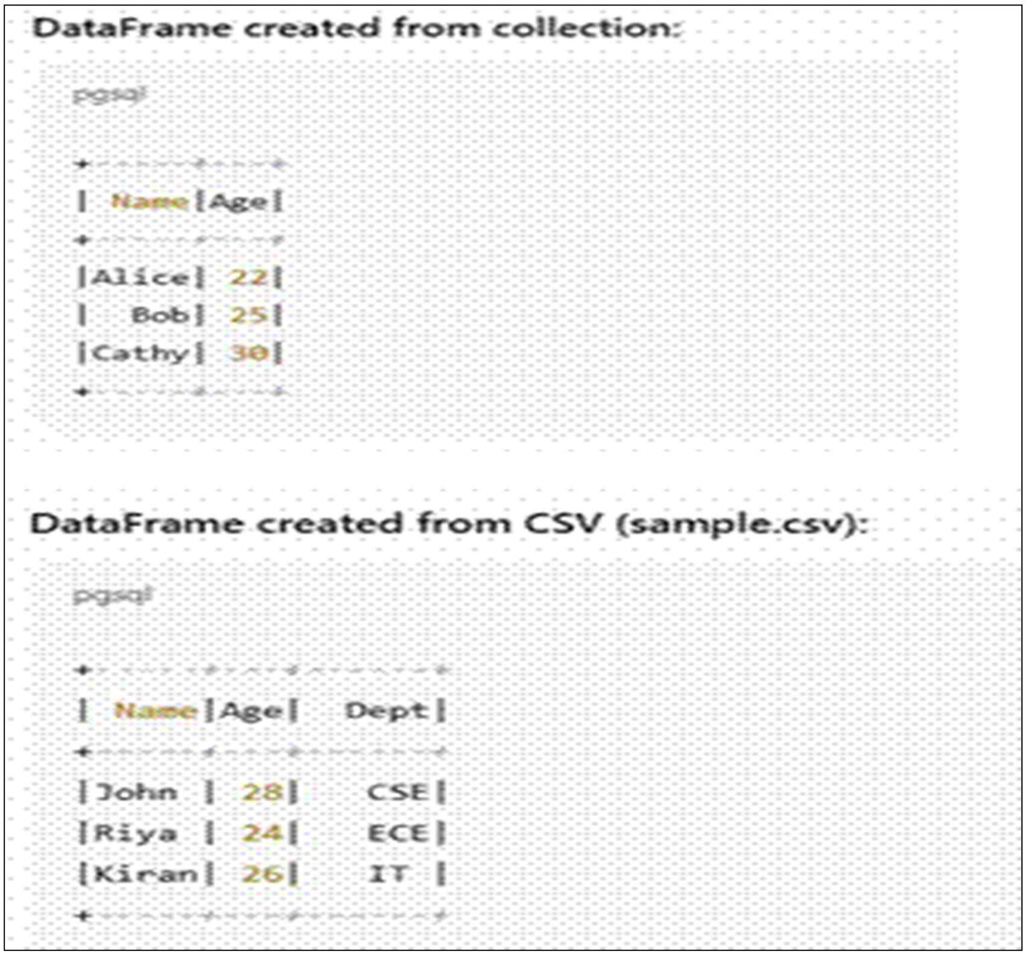
1. Supports structured and semi-structured data (JSON, CSV, Parquet, etc.).
2. Optimized execution using Catalyst optimizer.
3. Provides integration with SQL queries.

Program

```
# Import necessary libraries from pyspark.sql import SparkSession

# Create SparkSession spark =
SparkSession.builder.appName("DataFrameExample").getOrCreate()

# 1. Create DataFrame from collection data = [("Alice", 22), ("Bob", 25), ("Cathy", 30)]
columns = ["Name", "Age"] df1 = spark.createDataFrame(data, columns) print("DataFrame
created from collection:") df1.show()

# 2. Create DataFrame from external CSV file
```

# (Assume sample.csv contains: Name,Age,Dept) df2 = spark.read.csv("sample.csv", header=True, inferSchema=True) print("DataFrame created from CSV:") df2.show() # 3. Perform operations print("Selecting only Name column:") df1.select("Name").show() print("Filtering rows where Age > 23:") df1.filter(df1.Age > 23).show() # Stop Spark session spark.stop()

Output



DataFrame created from collection:

```
pgsql

+-----+---+
| Name|Age|
+-----+---+
|Alice| 22|
|  Bob| 25|
|Cathy| 30|
+-----+---+
```

DataFrame created from CSV (sample.csv):

```
pgsql

+-----+---+----+
| Name|Age|Dept|
+-----+---+----+
|John | 28| CSE|
|Riya | 24| ECE|
|Kiran| 26| IT |
+-----+---+----+
```

Result:

Thus, a DataFrame was successfully created in Spark SQL from a collection and an external CSV file, and basic operations such as selection and filtering were performed