**Big Data Frameworks CSE3120**

Experiment: User Defined Function using Scala in spark framework

**Name:** Naveen Nidadavolu

**Roll No:** 22MIA1049

**Aim:**

To implement a **User Defined Function (UDF) using Scala** in the **Apache Spark framework**, which transforms the input data by capitalizing the first letter of each word in a given column.

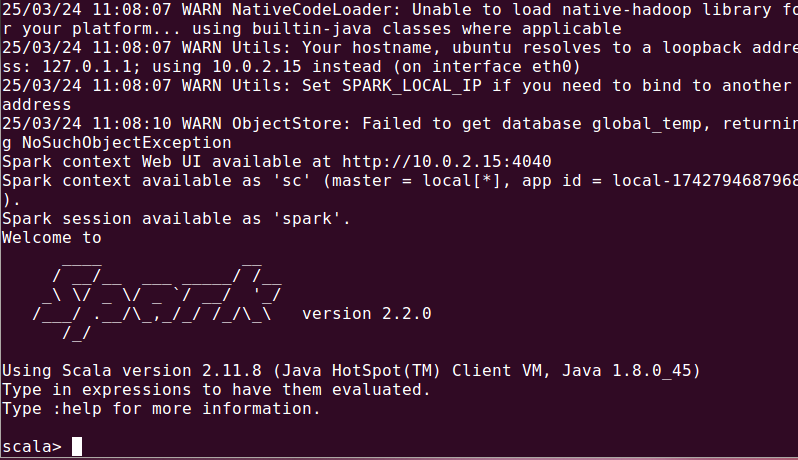
**Algorithm:**

1. **Start**
2. **Initialize Spark Session**:
   * Import necessary libraries.
   * Create a Spark session.
3. **Define Column Names**:
   * Create a sequence of column names.
4. **Create Sample Data**:
   * Define a sequence of tuples containing sample data.
5. **Convert Data into DataFrame**:
   * Use toDF() to convert the sequence into a Spark DataFrame.
6. **Define a User Defined Function (UDF)**:
   * Create a function that takes a string, splits it into words, capitalizes the first letter of each word, and then rejoins them.
7. **Register the UDF in Spark**:
   * Convert the function into a UDF using udf().
8. **Apply the UDF to the DataFrame**:
   * Use the select() method to apply the UDF to the specific column.
   * Rename the transformed column appropriately.
9. **Display the Output**:
   * Show the DataFrame with the transformed data.
10. **End**

**Procedure**

**Step 1: Start Spark Shell**

* Open the terminal and start the Spark shell.
* Ensure Spark is running successfully.



**Step 2: Define Column Names**

* Define a sequence of column names.

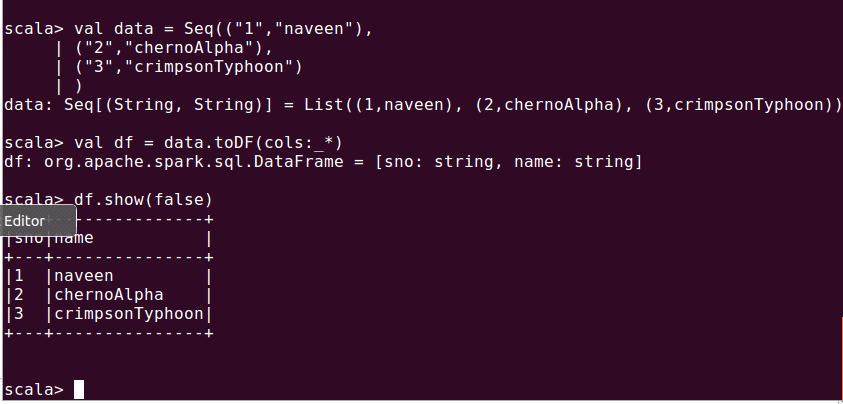
A computer screen with white text

AI-generated content may be incorrect.

* This sequence will be used for naming the DataFrame columns.

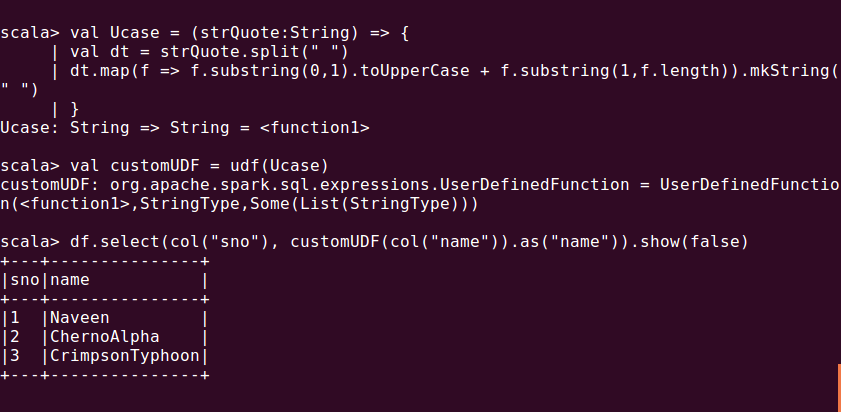
**Step 3: Create Sample Data**

* Define a sequence of tuples containing sample records.
* Convert it into a DataFrame with the specified column names.
* The DataFrame is displayed with sno and name columns.



**Step 4: Define User Defined Function (UDF)**

* Define a function to capitalize the first letter of each word in the "name" column.
* Convert the function into a Spark User Defined Function (UDF).



**Step 5: Apply UDF on DataFrame**

* Apply the UDF to modify the "name" column and display the updated DataFrame.

A computer screen shot of white text

AI-generated content may be incorrect.

**Result:**

The Spark User Defined Function (UDF) successfully transformed the "name" column by capitalizing the first letter of each word. The updated DataFrame displayed the modified names while retaining the "sno" column values.