Master PySpark: From Zero to Big Data Hero!!

PySpark Column Selection & Manipulation: Key Techniques

1. Different Methods to Select Columns

In PySpark, you can select specific columns in multiple ways:

- Using col() function/ column() / string way:

```
#Using col() function
df.select(col("Name")).show()

#Using column() function
df.select(column("Age")).show()

#Directly using string name
df.select("Salary").show()
```

2. Selecting Multiple Columns Together

You can combine different methods to select multiple columns:

```
#multiple column

df2 = df.select("ID", "Name", col("Salary"), column("Department"),
df.Phone)
df2.show()
```

3. Listing All Columns in a DataFrame

To get a list of all the column names:

```
#get all column name
df.columns
```

4.Renaming Columns with alias()

You can rename columns using the alias() method:

```
df.select(
  col("Name").alias('EmployeeName'),  # Rename "Name" to "EmployeeName"
  col("Salary").alias('EmployeeSalary'),  # Rename "Salary" to
"EmployeeSalary"
  column("Department"),  # Select "Department"
  df.Joining_Date  # Select "Joining_Date"
).show()
```

5. Using selectExpr() for Concise Column Selection

selectExpr() allows you to use SQL expressions directly and rename columns concisely:

```
df.selectExpr("Name as EmployeeName", "Salary as EmployeeSalary",
"Department").show()
```

Summary

- Use col(), column(), or string names to select columns.
- Use expr() and selectExpr() for SQL-like expressions and renaming.
- Use alias() to rename columns.
- Get the list of columns using df.columns.