



**19\_Select in pyspark**

Write a pyspark code perform below function

* Get all employee detail from emp\_df whose "Gender" end with 'le' and contain 4 letters. The Underscore(\_) Wildcard Character represents any single character.
* Get all employee detail from EmployeeDetail table whose "FirstName" start with # 'A' and contain 5 letters.
* Get all unique "Department" from EmployeeDetail table.
* Get the highest "Salary" from EmployeeDetail table.

**Difficult Level :** EASY

**DataFrame:**

**data = [**

**[1, "Vikas", "Ahlawat", 600000.0, "2013-02-15 11:16:28.290", "IT", "Male"],**

**[2, "nikita", "Jain", 530000.0, "2014-01-09 17:31:07.793", "HR", "Female"],**

**[3, "Ashish", "Kumar", 1000000.0, "2014-01-09 10:05:07.793", "IT", "Male"],**

**[4, "Nikhil", "Sharma", 480000.0, "2014-01-09 09:00:07.793", "HR", "Male"],**

**[5, "anish", "kadian", 500000.0, "2014-01-09 09:31:07.793", "Payroll", "Male"],**

**]**

**# Create a schema for the DataFrame**

**schema = StructType([**

**StructField("EmployeeID", IntegerType(), True),**

**StructField("First\_Name", StringType(), True),**

**StructField("Last\_Name", StringType(), True),**

**StructField("Salary", DoubleType(), True),**

**StructField("Joining\_Date", StringType(), True),**

**StructField("Department", StringType(), True),**

**StructField("Gender", StringType(), True)**

**])**



**# Creating Spark Session**

**from pyspark.sql import SparkSession**

**from pyspark.sql.types import StructType,StructField,IntegerType,StringType**

**#creating spark session**

**spark = SparkSession. \**

**builder. \**

**config('spark.shuffle.useOldFetchProtocol', 'true'). \**

**config('spark.ui.port','0'). \**

**config("spark.sql.warehouse.dir", "/user/itv008042/warehouse"). \**

**enableHiveSupport(). \**

**master('yarn'). \**

**getOrCreate()**

**# Create a list of rows from the image**

**data = [**

**[1, "Vikas", "Ahlawat", 600000.0, "2013-02-15 11:16:28.290", "IT", "Male"],**

**[2, "nikita", "Jain", 530000.0, "2014-01-09 17:31:07.793", "HR", "Female"],**

**[3, "Ashish", "Kumar", 1000000.0, "2014-01-09 10:05:07.793", "IT", "Male"],**

**[4, "Nikhil", "Sharma", 480000.0, "2014-01-09 09:00:07.793", "HR", "Male"],**

**[5, "anish", "kadian", 500000.0, "2014-01-09 09:31:07.793", "Payroll", "Male"],**

**]**

**# Create a schema for the DataFrame**

**schema = StructType([**

**StructField("EmployeeID", IntegerType(), True),**

**StructField("First\_Name", StringType(), True),**

**StructField("Last\_Name", StringType(), True),**

**StructField("Salary", DoubleType(), True),**

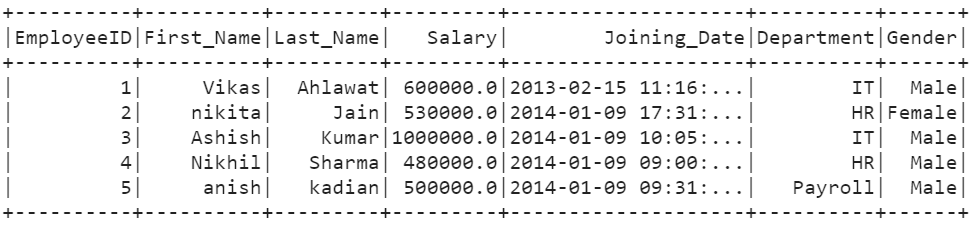
**StructField("Joining\_Date", StringType(), True),**

**StructField("Department", StringType(), True),**

**StructField("Gender", StringType(), True)**

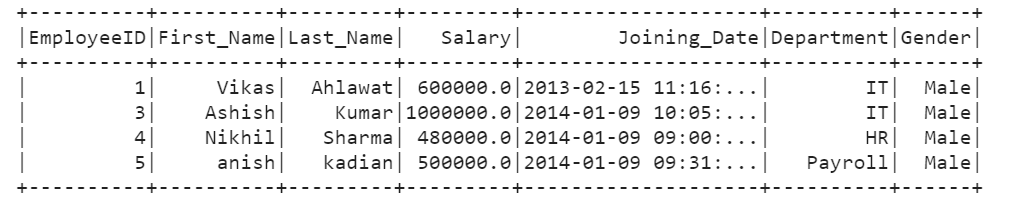
**])**

**emp\_df=spark.createDataFrame(data,schema)**

****

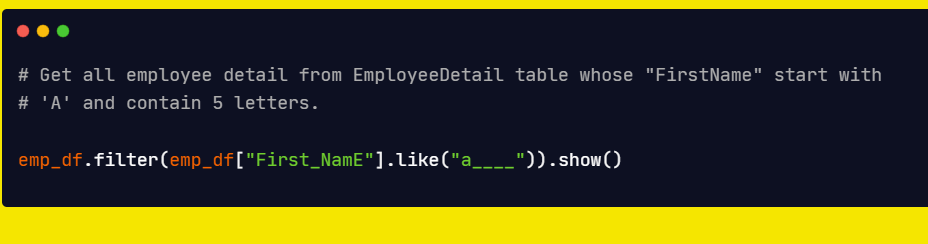
**Get all employee detail from emp\_df whose "Gender" end with 'le'and contain 4 letters. The Underscore(\_) Wildcard Character represents any single character.**

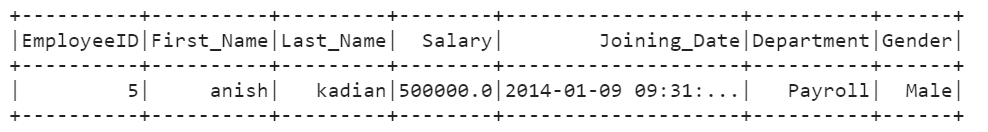
****

****

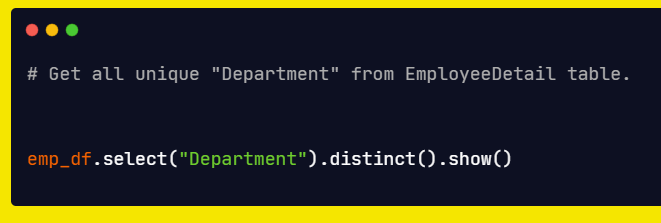
**# Get all employee detail from EmployeeDetail table whose "FirstName" start with**

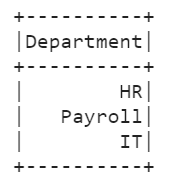
**# 'A' and contain 5 letters.**

****

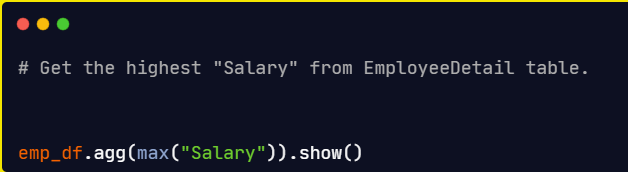
****

**# Get all unique "Department" from EmployeeDetail table.**

****

****

**# Get the highest "Salary" from EmployeeDetail table.**

****

****

