



**17\_Select in pyspark**

Write a pyspark code perform below function

* Write a pyspark code for combine FirstName and LastName and display it as "Name" (also include white space between first name & last name)
* Select employee detail whose name is "Vikas"
* Get all employee detail from EmployeeDetail table whose "FirstName" start with letter 'a'.

**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)**

**])**

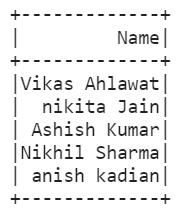
**emp\_df=spark.createDataFrame(data,schema)**

**#1. Write a pyspark code for combine FirstName and LastName and display it as "Name" (also include white space between first name & last name)**

**from pyspark.sql.functions import concat\_ws**

**emp\_df.select(concat\_ws(" ","First\_Name","Last\_Name")\**

**.alias("Name")).show()**

****

**# 2. Select employee detail whose name is "Vikas"**

**# Methos 1**

**from pyspark.sql.functions import col**

**emp\_df.filter(col("First\_Name") == 'Vikas' ).show(truncate=False)**

**# Methos 2**

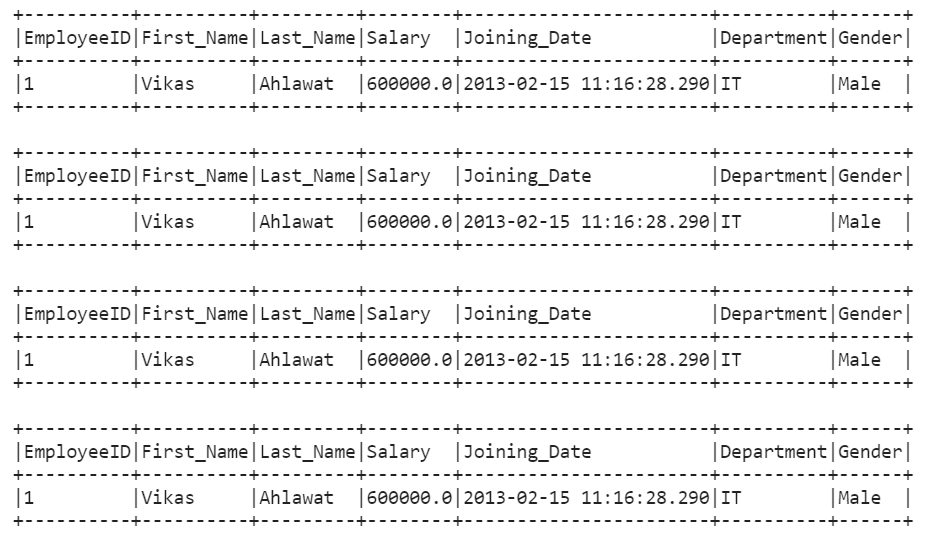
**emp\_df.filter(emp\_df.First\_Name == 'Vikas' ).show(truncate=False)**

**# Methos 3**

**emp\_df.filter(emp\_df['First\_Name'] == 'Vikas' ).show(truncate=False)**

**# Methos 4**

**emp\_df.where(emp\_df['First\_Name'] == 'Vikas' ).show(truncate=False)**

****

* Get all employee detail from EmployeeDetail table whose "FirstName" start with letter 'a'.

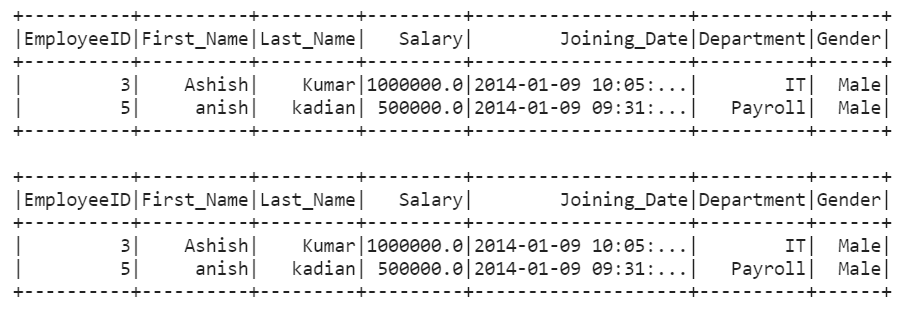
**# Method 1**

**from pyspark.sql.functions import lower**

**emp\_df.filter(lower(emp\_df['First\_Name']).like("a%")).show()**

**# Method 2**

**emp\_df.filter((emp\_df['First\_Name'].like("a%")) | (emp\_df['First\_Name'].like("A%")) ).show()**

****

