Akilesh K

k.akilesh123@gmail.com

Data engineering - Batch 1

Date: 12-02-24

CODING ASSESSMENT - PYSPARK

Execute Manipulating, Droping, Sorting, Aggregations, Joining, GroupeBy DataFrames

Execute Pyspark -sparksql joins & Applying Functions in a Pandas DataFrame

Create Dataframe

```python
from pyspark.sql import SparkSession
spark =SparkSession.builder.appName("Practice").getOrCreate()


emp = [(1,"Smith","finance",30000),(2, "Rose","Marketing", 40000),(3,"Williams","HR",10000),(4, "Jones","finance",20000),(5,"Eliot","Marketing",80000),(6, "Jack","manager",45000)]
empColumns = ["emp_id","name","Department","salary"]

empDF = spark.createDataFrame(data=emp, schema = empColumns)
empDF.printSchema()
empDF.show()

```

from pyspark.sql import SparkSession

spark =SparkSession.builder.appName("Practice").getOrCreate()

emp = [(1,"Smith","finance",30000),(2, "Rose","Marketing", 40000),(3,"Williams","HR",10000),(4, "Jones","finance",20000),(5,"Eliot","Marketing",80000),(6, "Jack","manager",45000)]

empColumns = ["emp_id","name","Department","salary"]

empDF = spark.createDataFrame(data=emp, schema = empColumns)

empDF.printSchema()

empDF.show()

Schema and table values

▸ (3) Spark Jobs

▸ empDF: pyspark.sql.dataframe.DataFrame = [emp_id: long, name: string ... 2 more fields]

```
root
 |-- emp_id: long (nullable = true)
 |-- name: string (nullable = true)
 |-- Department: string (nullable = true)
 |-- salary: long (nullable = true)

+------+--------+----------+------+
|emp_id|    name|Department|salary|
+------+--------+----------+------+
|     1|   Smith|   finance| 30000|
|     2|    Rose| Marketing| 40000|
|     3|Williams|        HR| 10000|
|     4|   Jones|   finance| 20000|
|     5|   Eliot| Marketing| 80000|
|     6|    Jack|   manager| 45000|
+------+--------+----------+------+
```

Command took 15.31 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:17:15 AM on My Cluster

GROUP BY

Group By Sum of salary

```
1   empDF.groupBy("Department").sum("salary").show()
```

▸ (2) Spark Jobs

```
+----------+-----------+
|Department|sum(salary)|
+----------+-----------+
|   finance|      50000|
| Marketing|     120000|
|        HR|      10000|
|   manager|      45000|
+----------+-----------+
```

Command took 4.33 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:20:28 AM on My Cluster

Group By min of salary

Cmd 3

```
1   empDF.groupBy("Department").min("salary").show()
```

▸ (2) Spark Jobs

```
+----------+-----------+
|Department|min(salary)|
+----------+-----------+
|   finance|      20000|
| Marketing|      40000|
|        HR|      10000|
|   manager|      45000|
+----------+-----------+
```

Command took 1.95 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:21:42 AM on My Cluster

Group By max of salary

```
1   empDF.groupBy("Department").max("salary").show()
```

▸ (2) Spark Jobs

```
+----------+-----------+
|Department|max(salary)|
+----------+-----------+
|   finance|      30000|
| Marketing|      80000|
|        HR|      10000|
|   manager|      45000|
+----------+-----------+
```

Command took 1.39 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:21:57 AM on My Cluster

Group By average of salary

```
1    empDF.groupBy("Department").avg("salary").show()
```

▸ (2) Spark Jobs

```
+----------+-----------+
|Department|avg(salary)|
+----------+-----------+
|   finance|    25000.0|
| Marketing|    60000.0|
|        HR|    10000.0|
|   manager|    45000.0|
+----------+-----------+
```

Command took 1.56 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:22:38 AM on My Cluster

Group By mean of salary

```
1    empDF.groupBy("Department").mean("salary").show()
```

▸ (2) Spark Jobs

```
+----------+-----------+
|Department|avg(salary)|
+----------+-----------+
|   finance|    25000.0|
| Marketing|    60000.0|
|        HR|    10000.0|
|   manager|    45000.0|
+----------+-----------+
```

Command took 1.07 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:22:40 AM on My Cluster

Group By count in each department

```
1    empDF.groupBy("Department").count().show()
```

▸ (2) Spark Jobs

```
+----------+-----+
|Department|count|
+----------+-----+
|   finance|    2|
| Marketing|    2|
|        HR|    1|
|   manager|    1|
+----------+-----+
```

Command took 1.72 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:23:27 AM on My Cluster

Group by multiple columns

```
1    empDF.groupBy("name","Department").sum("salary").show()
```

▸ (2) Spark Jobs

```
+--------+----------+-----------+
|    name|Department|sum(salary)|
+--------+----------+-----------+
|   Smith|   finance|      30000|
|    Rose| Marketing|      40000|
|Williams|        HR|      10000|
|   Jones|   finance|      20000|
|   Eliot| Marketing|      80000|
|    Jack|   manager|      45000|
+--------+----------+-----------+
```

Command took 1.67 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:25:56 AM on My Cluster

AGGREGATIONS

Aggregated sum of salary

```
1    empDF.agg(({"salary":"sum"})).show()
```

▸ (2) Spark Jobs

```
+-----------+
|sum(salary)|
+-----------+
|     225000|
+-----------+
```

Command took 1.19 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:27:06 AM on My Cluster

Aggregated min of salary

```
1    empDF.agg(({"salary":"min"})).show()
```

▸ (2) Spark Jobs

```
+-----------+
|min(salary)|
+-----------+
|      10000|
+-----------+
```

Command took 0.89 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:28:06 AM on My Cluster

Aggregated max of salary

```
1    empDF.agg(({"salary":"max"})).show()
```

▸ (2) Spark Jobs

```
+-----------+
|max(salary)|
+-----------+
|      80000|
+-----------+
```

Command took 0.75 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:28:21 AM on My Cluster

Aggregated mean of salary

```
1   empDF.agg(({"salary":"mean"})).show()
```

▸ (2) Spark Jobs

```
+-----------+
|avg(salary)|
+-----------+
|    37500.0|
+-----------+
```

Command took 0.74 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:28:48 AM on My Cluster

SORTING

Sorting in ascending order

```
1   empDF.sort("salary").show()
```

▸ (1) Spark Jobs

```
+------+--------+----------+------+
|emp_id|    name|Department|salary|
+------+--------+----------+------+
|     3|Williams|        HR| 10000|
|     4|   Jones|   finance| 20000|
|     1|   Smith|   finance| 30000|
|     2|    Rose| Marketing| 40000|
|     6|    Jack|   manager| 45000|
|     5|   Eliot| Marketing| 80000|
+------+--------+----------+------+
```

Command took 1.51 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:33:53 AM on My Cluster

Sorting in descending order

```
1   empDF.sort(empDF["salary"].desc()).show()
```

▸ (1) Spark Jobs

```
+------+--------+----------+------+
|emp_id|    name|Department|salary|
+------+--------+----------+------+
|     5|   Eliot| Marketing| 80000|
|     6|    Jack|   manager| 45000|
|     2|    Rose| Marketing| 40000|
|     1|   Smith|   finance| 30000|
|     4|   Jones|   finance| 20000|
|     3|Williams|        HR| 10000|
+------+--------+----------+------+
```

Command took 0.60 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:34:08 AM on My Cluster

Sorting using multiple column

```
1   empDF.sort("salary","name").show()
```

▸ (1) Spark Jobs

```
+------+--------+----------+------+
|emp_id|    name|Department|salary|
+------+--------+----------+------+
|     3|Williams|        HR| 10000|
|     4|   Jones|   finance| 20000|
|     1|   Smith|   finance| 30000|
|     2|    Rose| Marketing| 40000|
|     6|    Jack|   manager| 45000|
|     5|   Eliot| Marketing| 80000|
+------+--------+----------+------+
```

Command took 0.63 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:34:14 AM on My Cluster

DROPPING

Showing the pivot description

```
empDF.groupBy("Department").pivot("name").sum("salary").show()
```

(7) Spark Jobs

```
+----------+-----+-----+-----+-----+-----+--------+
|Department|Eliot| Jack|Jones| Rose|Smith|Williams|
+----------+-----+-----+-----+-----+-----+--------+
|   finance| null| null|20000| null|30000|    null|
|        HR| null| null| null| null| null|   10000|
|   manager| null|45000| null| null| null|    null|
| Marketing|80000| null| null|40000| null|    null|
+----------+-----+-----+-----+-----+-----+--------+
```

Command took 4.65 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:31:13 AM on My Cluster

Adding a null value

(3) Spark Jobs

empDF1: pyspark.sql.dataframe.DataFrame = [emp_id: long, name: string ... 2 more fields]

```
+------+--------+----------+------+
|emp_id|    name|Department|salary|
+------+--------+----------+------+
|     1|   Smith|   finance| 30000|
|     2|    Rose| Marketing| 40000|
|     3|Williams|        HR| 10000|
|     4|   Jones|   finance| 20000|
|     5|   Eliot| Marketing| 80000|
|     6|    Jack|   manager| 45000|
|     7|  ronald|      null| 25000|
+------+--------+----------+------+
```

Command took 0.75 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:48:49 AM on My Cluster

Dropping the row with null value

```
empDF1.na.drop().show()
```

(3) Spark Jobs

```
+------+--------+----------+------+
|emp_id|    name|Department|salary|
+------+--------+----------+------+
|     1|   Smith|   finance| 30000|
|     2|    Rose| Marketing| 40000|
|     3|Williams|        HR| 10000|
|     4|   Jones|   finance| 20000|
|     5|   Eliot| Marketing| 80000|
|     6|    Jack|   manager| 45000|
+------+--------+----------+------+
```

Command took 0.67 seconds -- by kakilesh123@gmail.com at 2/12/2024, 11:48:55 AM on My Cluster

JOINING

Creating a new database for executing joins

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()

emp = [(1,"Smith",-1,"2018","10","M",3000),(2, "Rose",1 , "2010", "20","M", 4000),(3,"Williams",1,"2010","10","M",1000),(4, "Jones",2 ,"2005","10","F",2000),(5,"Brown",2,
"2010","40","",-1),(6, "Brown", 2, "2010","50","",-1)]
empColumns = ["emp_id","name","superior_emp_id","year_joined", "emp_dept_id","gender","salary"]

empDF = spark.createDataFrame(data=emp, schema = empColumns)
empDF.show()
```

(3) Spark Jobs

empDF: pyspark.sql.dataframe.DataFrame = [emp_id: long, name: string ... 5 more fields]

```
+------+--------+---------------+-----------+-----------+------+------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|
+------+--------+---------------+-----------+-----------+------+------+
|     1|   Smith|             -1|       2018|         10|     M|  3000|
|     2|    Rose|              1|       2010|         20|     M|  4000|
|     3|Williams|              1|       2010|         10|     M|  1000|
|     4|   Jones|              2|       2005|         10|     F|  2000|
|     5|   Brown|              2|       2010|         40|      |    -1|
|     6|   Brown|              2|       2010|         50|      |    -1|
+------+--------+---------------+-----------+-----------+------+------+
```

Command took 1.07 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:00:49 PM on My Cluster

```python
dept = [("Finance",10),("Marketing",20),("Sales",30),("IT",40)]
deptColumns = ["dept_name","dept_id"]
deptDF = spark.createDataFrame(data=dept, schema = deptColumns)
deptDF.show()
```

(3) Spark Jobs

deptDF: pyspark.sql.dataframe.DataFrame = [dept_name: string, dept_id: long]

```
+---------+-------+
|dept_name|dept_id|
+---------+-------+
|  Finance|     10|
|Marketing|     20|
|    Sales|     30|
|       IT|     40|
+---------+-------+
```

Command took 0.82 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:01:03 PM on My Cluster

Inner join

```python
empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"inner") .show()
```

(3) Spark Jobs

```
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
|     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
|     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
|     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```

Command took 2.35 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:02:17 PM on My Cluster

Outer join

```
empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"outer").show()
```

(3) Spark Jobs

```
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
|     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
|     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
|     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
|  null|    null|           null|       null|       null|  null|  null|    Sales|     30|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
|     6|   Brown|              2|       2010|         50|      |    -1|     null|   null|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```

Command took 1.32 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:02:21 PM on My Cluster

Left join

```
empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"left").show()
```

(6) Spark Jobs

```
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
|     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
|     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
|     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
|     6|   Brown|              2|       2010|         50|      |    -1|     null|   null|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```

Command took 1.81 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:02:27 PM on My Cluster

Right Join

```
empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"right").show()
```

(6) Spark Jobs

```
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
|     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
|     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
|     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
|  null|    null|           null|       null|       null|  null|  null|    Sales|     30|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```

Command took 1.67 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:02:33 PM on My Cluster

Left Semi Join

```
empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"leftsemi").show()
```

(3) Spark Jobs

```
+------+--------+---------------+-----------+-----------+------+------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|
+------+--------+---------------+-----------+-----------+------+------+
|     1|   Smith|             -1|       2018|         10|     M|  3000|
|     3|Williams|              1|       2010|         10|     M|  1000|
|     4|   Jones|              2|       2005|         10|     F|  2000|
|     2|    Rose|              1|       2010|         20|     M|  4000|
|     5|   Brown|              2|       2010|         40|      |    -1|
+------+--------+---------------+-----------+-----------+------+------+
```

Command took 1.50 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:02:39 PM on My Cluster

Left Anti Join

```
empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"leftanti").show()
```

▸ (4) Spark Jobs

```
+------+-----+---------------+-----------+-----------+------+------+
|emp_id| name|superior_emp_id|year_joined|emp_dept_id|gender|salary|
+------+-----+---------------+-----------+-----------+------+------+
|     6|Brown|              2|       2010|         50|      |    -1|
+------+-----+---------------+-----------+-----------+------+------+
```

Command took 1.43 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:02:42 PM on My Cluster

JOINS USING SPARK SQL

Create a dataframe

```
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("Join").getOrCreate()

emp_data = [(1, "John", 1), (2, "Jane", 2), (3, "Doe", 2)]
dept_data = [(1, "Engineering"), (2, "Sales")]

emp_schema = ["emp_id", "emp_name", "dept_id"]
dept_schema = ["dept_id", "dept_name"]

emp_df = spark.createDataFrame(emp_data, schema=emp_schema)
dept_df = spark.createDataFrame(dept_data, schema=dept_schema)

emp_df.createOrReplaceTempView("employees")
dept_df.createOrReplaceTempView("departments")


```

▸ emp_df: pyspark.sql.dataframe.DataFrame = [emp_id: long, emp_name: string ... 1 more field]

▸ dept_df: pyspark.sql.dataframe.DataFrame = [dept_id: long, dept_name: string]

ommand took 0.30 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:45:58 PM on My Cluster

Join using sparkSQL

```
joined_df = spark.sql("""
    SELECT e.emp_id, e.emp_name, d.dept_name
    FROM employees e
    JOIN departments d ON e.dept_id = d.dept_id
""")
joined_df.show()
```

(3) Spark Jobs

joined_df: pyspark.sql.dataframe.DataFrame = [emp_id: long, emp_name: string ... 1 more field]

```
+------+--------+-----------+
|emp_id|emp_name|  dept_name|
+------+--------+-----------+
|     1|    John|Engineering|
|     2|    Jane|      Sales|
|     3|     Doe|      Sales|
+------+--------+-----------+
```

Command took 1.04 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:46:02 PM on My Cluster

Left Join using sparkSQL

```
joined_df =spark.sql("""
    SELECT e.emp_id, e.emp_name, d.dept_name
    FROM employees e
    LEFT JOIN departments d ON e.dept_id = d.dept_id
""")
joined_df.show()
```

(4) Spark Jobs

joined_df: pyspark.sql.dataframe.DataFrame = [emp_id: long, emp_name: string ... 1 more field]

```
+------+--------+-----------+
|emp_id|emp_name|  dept_name|
+------+--------+-----------+
|     1|    John|Engineering|
|     2|    Jane|      Sales|
|     3|     Doe|      Sales|
+------+--------+-----------+
```

Command took 1.37 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:49:37 PM on My Cluster

Right Join using sparkSQL

```
joined_df =spark.sql("""
    SELECT e.emp_id, e.emp_name, d.dept_name
    FROM employees e
    RIGHT JOIN departments d ON e.dept_id = d.dept_id
""")
joined_df.show()
```

(6) Spark Jobs

joined_df: pyspark.sql.dataframe.DataFrame = [emp_id: long, emp_name: string ... 1 more field]

```
+------+--------+-----------+
|emp_id|emp_name|  dept_name|
+------+--------+-----------+
|     1|    John|Engineering|
|     3|     Doe|      Sales|
|     2|    Jane|      Sales|
+------+--------+-----------+
```

Command took 1.80 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:50:17 PM on My Cluster

APPLY FUNCTIONS

Creating a dataframe

```python
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName('codingchallenge').getOrCreate()
columns = ["Seqno","Name"]
data = [("1", "john jones"),
    ("2", "tracey smith"),
    ("3", "amy sanders")]

df = spark.createDataFrame(data=data,schema=columns)
df.show()
```

▸ (3) Spark Jobs

▸ df: pyspark.sql.dataframe.DataFrame = [Seqno: string, Name: string]

```
+-----+------------+
|Seqno|        Name|
+-----+------------+
|    1|  john jones|
|    2|tracey smith|
|    3| amy sanders|
+-----+------------+
```

Command took 1.01 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:29:41 PM on My Cluster

Applying a function using column

```python
from pyspark.sql.functions import upper
df.withColumn("Upper_Name", upper(df.Name)) \
  .show()
```

▸ (3) Spark Jobs

```
+-----+------------+------------+
|Seqno|        Name|  Upper_Name|
+-----+------------+------------+
|    1|  john jones|  JOHN JONES|
|    2|tracey smith|TRACEY SMITH|
|    3| amy sanders| AMY SANDERS|
+-----+------------+------------+
```

Command took 0.57 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:29:58 PM on My Cluster

Applying a function using select statement

```
df.select("Seqno","Name", upper(df.Name)) \
    .show()
```

(3) Spark Jobs

```
+-----+------------+------------+
|Seqno|        Name| upper(Name)|
+-----+------------+------------+
|    1|  john jones|  JOHN JONES|
|    2|tracey smith|TRACEY SMITH|
|    3| amy sanders| AMY SANDERS|
+-----+------------+------------+
```

Command took 0.74 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:30:13 PM on My Cluster

Applying a function by creating a new view table

```
df.createOrReplaceTempView("TAB")
spark.sql("select Seqno, Name, UPPER(Name) from TAB") \
    .show()
```

(3) Spark Jobs

```
+-----+------------+------------+
|Seqno|        Name| upper(Name)|
+-----+------------+------------+
|    1|  john jones|  JOHN JONES|
|    2|tracey smith|TRACEY SMITH|
|    3| amy sanders| AMY SANDERS|
+-----+------------+------------+
```

Command took 0.96 seconds -- by kakilesh123@gmail.com at 2/12/2024, 12:30:54 PM on My Cluster