databricks Pyspark Practice 1 - 11 Nov, 2022

(https://databricks.com)

```
spark
```

SparkSession - hive

SparkContext

Spark UI

Version
v3.3.0
Master
local[8]
AppName
Databricks Shell

```
from pyspark.sql import SparkSession
```

```
spark=SparkSession.builder.appName('Practice').getOrCreate()
```

```
df=spark.read.option("header", True).option("inferSchema", True).csv("/FileStore/tables/test1-1.csv")
```

```
df.show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
|    Sunny| 29|         4| 20000|
|     Paul| 24|         3| 20000|
|   Harsha| 21|         1| 15000|
|  Shubham| 23|         2| 18000|
+---------+---+----------+------+
```

```
df.printSchema()
```

```
root
 |-- Name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- Experience: integer (nullable = true)
 |-- Salary: integer (nullable = true)
```

```
type(df)
```

```
Out[19]: pyspark.sql.dataframe.DataFrame
```

```
df.columns
```

```
Out[21]: ['Name', 'age', 'Experience', 'Salary']
```

```
df.head(3) #show top 3 items
```

```
Out[22]: [Row(Name='Krish', age=31, Experience=10, Salary=30000),
 Row(Name='Sudhanshu', age=30, Experience=8, Salary=25000),
 Row(Name='Sunny', age=29, Experience=4, Salary=20000)]
```

```
df.select("Name").show()
```

```
+---------+
|     Name|
+---------+
|    Krish|
|Sudhanshu|
|    Sunny|
|     Paul|
|   Harsha|
|  Shubham|
+---------+
```

```
df.select(["Name","age"]).show() #shows particular cols
```

```
+---------+---+
|     Name|age|
+---------+---+
|    Krish| 31|
|Sudhanshu| 30|
|    Sunny| 29|
|     Paul| 24|
|   Harsha| 21|
|  Shubham| 23|
+---------+---+
```

```
df["Name"]
```

```
Out[29]: Column<'Name'>
```

```
df.dtypes #gives data type of cols
```

```
Out[31]: [('Name', 'string'), ('age', 'int'), ('Experience', 'int'), ('Salary', 'int')]
```

```
df.describe().show()
```

```
+-------+------+------------------+-----------------+------------------+
|summary|  Name|               age|       Experience|            Salary|
+-------+------+------------------+-----------------+------------------+
|  count|     6|                 6|                6|                 6|
|   mean|  null|26.333333333333332|4.666666666666667|21333.333333333332|
| stddev|  null| 4.179314138308661|3.559026084010437| 5354.126134736337|
```

```
|    min|Harsha|                21|                1|             15000|
|    max| Sunny|                31|               10|             30000|
+-------+------+------------------+-----------------+------------------+
```

```
#Adding columns
df_addcol=df.withColumn('Experience after two years',df['Experience']+2)
```

```
df_addcol.show()
```

```
+---------+---+----------+------+--------------------------+
|     Name|age|Experience|Salary|Experience after two years|
+---------+---+----------+------+--------------------------+
|    Krish| 31|        10| 30000|                        12|
|Sudhanshu| 30|         8| 25000|                        10|
|    Sunny| 29|         4| 20000|                         6|
|     Paul| 24|         3| 20000|                         5|
|   Harsha| 21|         1| 15000|                         3|
|  Shubham| 23|         2| 18000|                         4|
+---------+---+----------+------+--------------------------+
```

```
#drop col
df_dropcol=df.drop("Experience after two years")
```

```
df_dropcol.show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
|    Sunny| 29|         4| 20000|
|     Paul| 24|         3| 20000|
|   Harsha| 21|         1| 15000|
|  Shubham| 23|         2| 18000|
+---------+---+----------+------+
```

```
#column rename
df_dropcol.withColumnRenamed("Name","New Name").show()
```

```
+---------+---+----------+------+
| New Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
|    Sunny| 29|         4| 20000|
|     Paul| 24|         3| 20000|
|   Harsha| 21|         1| 15000|
|  Shubham| 23|         2| 18000|
+---------+---+----------+------+
```

```
df1=spark.read.option("header", True).option("inferSchema",
True).csv("/FileStore/tables/test2.csv")


df1.show()
```

```
+---------+----+----------+------+
|     Name| age|Experience|Salary|
+---------+----+----------+------+
|    Krish|  31|        10| 30000|
|Sudhanshu|  30|         8| 25000|
|    Sunny|  29|         4| 20000|
|     Paul|  24|         3| 20000|
|   Harsha|  21|         1| 15000|
|  Shubham|  23|         2| 18000|
|   Mahesh|null|      null| 40000|
|     null|  34|        10| 38000|
|     null|  36|      null|  null|
+---------+----+----------+------+
```

```
#drop null values
df1.na.drop().show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
|    Sunny| 29|         4| 20000|
|     Paul| 24|         3| 20000|
|   Harsha| 21|         1| 15000|
|  Shubham| 23|         2| 18000|
+---------+---+----------+------+
```

```
df1.na.drop(how='any',thresh=2).show()
```

```
+---------+----+----------+------+
|     Name| age|Experience|Salary|
+---------+----+----------+------+
|    Krish|  31|        10| 30000|
|Sudhanshu|  30|         8| 25000|
|    Sunny|  29|         4| 20000|
|     Paul|  24|         3| 20000|
|   Harsha|  21|         1| 15000|
|  Shubham|  23|         2| 18000|
|   Mahesh|null|      null| 40000|
|     null|  34|        10| 38000|
+---------+----+----------+------+
```

```
df1.na.drop(how='any',subset=['Experience','Name']).show() #delete null in particular columns
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
|    Sunny| 29|         4| 20000|
|     Paul| 24|         3| 20000|
|   Harsha| 21|         1| 15000|
|  Shubham| 23|         2| 18000|
+---------+---+----------+------+
```

```
#filling missing value
df1.na.fill('Hari').show()
```

```
+---------+----+----------+------+
|     Name| age|Experience|Salary|
+---------+----+----------+------+
|    Krish|  31|        10| 30000|
|Sudhanshu|  30|         8| 25000|
|    Sunny|  29|         4| 20000|
|     Paul|  24|         3| 20000|
|   Harsha|  21|         1| 15000|
|  Shubham|  23|         2| 18000|
|   Mahesh|null|      null| 40000|
|     Hari|  34|        10| 38000|
|     Hari|  36|      null|  null|
+---------+----+----------+------+
```

```
df1.na.fill('Hari',['Name']).show() #or df1.na.fill('Hari',['Name','age']).show() for multiple columns
```

```
+---------+----+----------+------+
|     Name| age|Experience|Salary|
+---------+----+----------+------+
|    Krish|  31|        10| 30000|
|Sudhanshu|  30|         8| 25000|
|    Sunny|  29|         4| 20000|
|     Paul|  24|         3| 20000|
|   Harsha|  21|         1| 15000|
|  Shubham|  23|         2| 18000|
|   Mahesh|null|      null| 40000|
|     Hari|  34|        10| 38000|
|     Hari|  36|      null|  null|
+---------+----+----------+------+
```

```
# Filter operations
df2=spark.read.option("header", True).option("inferSchema", True).csv("/FileStore/tables/test1-1.csv")
```

```
df2.show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
|    Sunny| 29|         4| 20000|
|     Paul| 24|         3| 20000|
|   Harsha| 21|         1| 15000|
|  Shubham| 23|         2| 18000|
+---------+---+----------+------+
```

```
df2.filter("Salary<=20000").show()
```

```
+-------+---+----------+------+
|   Name|age|Experience|Salary|
+-------+---+----------+------+
|  Sunny| 29|         4| 20000|
|   Paul| 24|         3| 20000|
| Harsha| 21|         1| 15000|
|Shubham| 23|         2| 18000|
+-------+---+----------+------+
```

```
df2.filter("Salary<=20000").select(["Name","age"]).show()
```

```
+-------+---+
|   Name|age|
+-------+---+
|  Sunny| 29|
|   Paul| 24|
| Harsha| 21|
|Shubham| 23|
+-------+---+
```

```
df2.filter(df2['Salary']<=20000).show()
```

```
+-------+---+----------+------+
|   Name|age|Experience|Salary|
+-------+---+----------+------+
|  Sunny| 29|         4| 20000|
|   Paul| 24|         3| 20000|
| Harsha| 21|         1| 15000|
|Shubham| 23|         2| 18000|
+-------+---+----------+------+
```

```
df2.filter((df2['Salary']<=20000) & (df2['Salary']>=15000)).show()
```

```
+-------+---+----------+------+
|   Name|age|Experience|Salary|
+-------+---+----------+------+
|  Sunny| 29|         4| 20000|
|   Paul| 24|         3| 20000|
```

```
| Harsha| 21|         1| 15000|
|Shubham| 23|         2| 18000|
+-------+---+----------+------+
```

```
df2.filter(~(df2['Salary']<=20000)).show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
+---------+---+----------+------+
```

```
df3=spark.read.option("header", True).option("inferSchema",
True).csv("/FileStore/tables/test3.csv")
```

```
df3.show()
```

```
+---------+------------+------+
|     Name| Departments|salary|
+---------+------------+------+
|    Krish|Data Science| 10000|
|    Krish|         IOT|  5000|
|   Mahesh|    Big Data|  4000|
|    Krish|    Big Data|  4000|
|   Mahesh|Data Science|  3000|
|Sudhanshu|Data Science| 20000|
|Sudhanshu|         IOT| 10000|
|Sudhanshu|    Big Data|  5000|
|    Sunny|Data Science| 10000|
|    Sunny|    Big Data|  2000|
+---------+------------+------+
```

```
#groupBy ang Agg together
df3.groupBy("Name").sum().show() #Group by Name
```

```
+---------+-----------+
|     Name|sum(salary)|
+---------+-----------+
|Sudhanshu|      35000|
|    Sunny|      12000|
|    Krish|      19000|
|   Mahesh|       7000|
+---------+-----------+
```

```
df3.groupBy("Departments").sum().show() #GroupBy Departments
```

```
+------------+-----------+
| Departments|sum(salary)|
+------------+-----------+
|         IOT|      15000|
```

```
|    Big Data|      15000|
|Data Science|      43000|
+------------+-----------+
```

```
df3.groupBy("Departments").count().show()
```

```
+------------+-----+
| Departments|count|
+------------+-----+
|         IOT|    2|
|    Big Data|    4|
|Data Science|    4|
+------------+-----+
```

```
df3.agg({'salary':'sum'}).show()
```

```
+-----------+
|sum(salary)|
+-----------+
|      73000|
+-----------+
```

```
df3.groupBy("Name").max().show()
```

```
+---------+-----------+
|     Name|max(salary)|
+---------+-----------+
|Sudhanshu|      20000|
|    Sunny|      10000|
|    Krish|      10000|
|   Mahesh|       4000|
+---------+-----------+
```

```
df3.groupBy("Name").min().show()
```

```
+---------+-----------+
|     Name|min(salary)|
+---------+-----------+
|Sudhanshu|       5000|
|    Sunny|       2000|
|    Krish|       4000|
|   Mahesh|       3000|
+---------+-----------+
```

End