Pyspark Datafrmes

- Filter Operation
- &, |, ==
- ~

In [1]:
```
from pyspark.sql import SparkSession
```

In [2]:
```
spark = SparkSession.builder.appName('dataframe').getOrCreate()
```

In [5]:
```
df_pyspark = spark.read.csv('test3.csv', header=True, inferSchema=True)
```

In [6]:
```
df_pyspark.show()
```

```
+-------+---+----------+------+
|   name|age|experience|salary|
+-------+---+----------+------+
|Ajinkya| 32|        10| 30000|
|  Anish| 30|         8| 25000|
| Nikhil| 29|         4| 20000|
| Hitesh| 24|         3| 20000|
|  Onkar| 21|         1| 15000|
|  Ketan| 23|         2| 18000|
+-------+---+----------+------+
```

Filter Operations

In [8]:
```
### Salary of the people less than or equal to 20000
df_pyspark.filter('salary <= 20000').show()
```

```
+------+---+----------+------+
|  name|age|experience|salary|
+------+---+----------+------+
|Nikhil| 29|         4| 20000|
|Hitesh| 24|         3| 20000|
| Onkar| 21|         1| 15000|
| Ketan| 23|         2| 18000|
+------+---+----------+------+
```

In [10]:
```
df_pyspark.filter(df_pyspark.age > 25).collect()
```

Out[10]:
```
[Row(name='Ajinkya', age=32, experience=10, salary=30000),
 Row(name='Anish', age=30, experience=8, salary=25000),
 Row(name='Nikhil', age=29, experience=4, salary=20000)]
```

In [11]:
```
df_pyspark.where(df_pyspark.experience > 5).collect()
```

Out[11]:
```
[Row(name='Ajinkya', age=32, experience=10, salary=30000),
 Row(name='Anish', age=30, experience=8, salary=25000)]
```

In [12]:
```
df_pyspark.where('age <= 30').show()
```

```
+------+---+----------+------+
|  name|age|experience|salary|
+------+---+----------+------+
| Anish| 30|         8| 25000|
|Nikhil| 29|         4| 20000|
|Hitesh| 24|         3| 20000|
| Onkar| 21|         1| 15000|
| Ketan| 23|         2| 18000|
+------+---+----------+------+
```

In [13]:

```
df_pyspark.filter('salary > 20000').select(['name', 'age']).show()
```

```
+-------+---+
|   name|age|
+-------+---+
|Ajinkya| 32|
|  Anish| 30|
+-------+---+
```

In [14]:

```
df_pyspark.filter(df_pyspark['salary'] >= 20000).show()
```

```
+-------+---+----------+------+
|   name|age|experience|salary|
+-------+---+----------+------+
|Ajinkya| 32|        10| 30000|
|  Anish| 30|         8| 25000|
| Nikhil| 29|         4| 20000|
| Hitesh| 24|         3| 20000|
+-------+---+----------+------+
```

In [15]:

```
df_pyspark.filter((df_pyspark['salary'] >= 20000 ) & (df_pyspark.age <=30)).show()
```

```
+------+---+----------+------+
|  name|age|experience|salary|
+------+---+----------+------+
| Anish| 30|         8| 25000|
|Nikhil| 29|         4| 20000|
|Hitesh| 24|         3| 20000|
+------+---+----------+------+
```

In [16]:

```
df_pyspark.filter((df_pyspark['salary'] >= 20000 ) | (df_pyspark.age <=30)).show()
```

```
+-------+---+----------+------+
|   name|age|experience|salary|
+-------+---+----------+------+
|Ajinkya| 32|        10| 30000|
|  Anish| 30|         8| 25000|
| Nikhil| 29|         4| 20000|
| Hitesh| 24|         3| 20000|
|  Onkar| 21|         1| 15000|
|  Ketan| 23|         2| 18000|
+-------+---+----------+------+
```

In [28]:

```
df_pyspark.filter(f'age > 29 and salary >20000').show()
```

```
+-------+---+----------+------+
|   name|age|experience|salary|
+-------+---+----------+------+
|Ajinkya| 32|        10| 30000|
|  Anish| 30|         8| 25000|
+-------+---+----------+------+
```

In [30]:

```
df_pyspark.filter('salary < 20000 or age < 29').show()
```

```
+------+---+----------+------+
|  name|age|experience|salary|
+------+---+----------+------+
|Hitesh| 24|         3| 20000|
| Onkar| 21|         1| 15000|
| Ketan| 23|         2| 18000|
+------+---+----------+------+
```

In [39]:

```
df_pyspark.filter('''not (salary >= 20000 and age >= 27)''').show()
```

```
+------+---+----------+------+
|  name|age|experience|salary|
+------+---+----------+------+
|Hitesh| 24|         3| 20000|
| Onkar| 21|         1| 15000|
| Ketan| 23|         2| 18000|
+------+---+----------+------+
```

In [40]:

```
df_pyspark.filter('''(salary >= 20000 and age >= 27)''').show()
```

```
+-------+---+----------+------+
|   name|age|experience|salary|
+-------+---+----------+------+
|Ajinkya| 32|        10| 30000|
|  Anish| 30|         8| 25000|
| Nikhil| 29|         4| 20000|
+-------+---+----------+------+
```

In [ ]:

In [ ]: