In [1]:
```
print("mukesh")
```

```
Starting Spark application
```

ID	YARN Application ID	Kind	State	Spark UI
0	application_1550060762900_0004	pyspark3	idle	Link (http://hn1-suyalc.m3rwtj3rvkaenpwmwu0hf2c3aa.cx.internal.cloudapp.net:8088/proxy/application_1

```
SparkSession available as 'spark'.
mukesh
```

In [33]:
```
from pyspark.sql import *
Employee=Row("Id","Name","Age")
employee1=Employee(101, "sachin",40)
employee2=Employee(102, "zahir",41)
employee3=Employee(103, "virat",29)
employee4=Employee(104, "saurav",41)
employee5=Employee(105,"rohit",30)
employee6=Employee(105,"rohit",30)
employee7=Employee(105,"rohit",30)

a = [employee1,employee2,employee3,employee4,employee5,employee6,employee7]
df1 = spark.createDataFrame(a)

Table2=Row("Player_Id","Skill")
player1=Table2(101, "batsman")
player2=Table2(102, "bowler")
player3=Table2(103, "batsman")
player4=Table2(104, "batsman")
df2=spark.createDataFrame([player1,player2,player3,player4])

Table3=Row("Id","Name","Centuries")
player1=Table3(101, "sachin",100)
player2=Table3(103, "virat",50)
player3=Table3(104,  "saurav",45)
player4=Table3(105,"rohit",35)

df3 =spark.createDataFrame([player1,player2,player3,player4])
```

In [32]:
```
df.show()
df2.show()
df3.show()
```

```
+---+------+---+
| Id|  Name|Age|
+---+------+---+
|101|sachin| 40|
|102| zahir| 41|
|103| virat| 29|
|104|saurav| 41|
|105| rohit| 30|
+---+------+---+

+---------+-------+
|Player_Id|  Skill|
+---------+-------+
|      101|batsman|
|      102| bowler|
|      103|batsman|
|      104|batsman|
+---------+-------+

+---+------+---------+
| Id|  Name|Centuries|
+---+------+---------+
|101|sachin|      100|
|103| virat|       50|
|104|saurav|       45|
|105| rohit|       35|
+---+------+---------+
```

In [8]:
```
df.printSchema()
df1.printSchema()
df2.printSchema()
```

```
root
 |-- Id: long (nullable = true)
 |-- Name: string (nullable = true)
 |-- Age: long (nullable = true)

root
 |-- Id: long (nullable = true)
 |-- Name: string (nullable = true)
 |-- Age: long (nullable = true)

root
 |-- Player_Id: long (nullable = true)
 |-- Skill: string (nullable = true)
```

In [10]:
```
# Cartesian Join (m*n) combinations
df12 = df1.join(df2)
df12.show()
df12.count()
```

```
+---+------+---+---------+-------+
| Id|  Name|Age|Player_Id|  Skill|
+---+------+---+---------+-------+
|101|sachin| 40|      101|batsman|
|101|sachin| 40|      102| bowler|
|101|sachin| 40|      103|batsman|
|101|sachin| 40|      104|batsman|
|102| zahir| 41|      101|batsman|
|102| zahir| 41|      102| bowler|
|102| zahir| 41|      103|batsman|
|102| zahir| 41|      104|batsman|
|103| virat| 29|      101|batsman|
|103| virat| 29|      102| bowler|
|103| virat| 29|      103|batsman|
|103| virat| 29|      104|batsman|
|104|saurav| 41|      101|batsman|
|104|saurav| 41|      102| bowler|
|104|saurav| 41|      103|batsman|
|104|saurav| 41|      104|batsman|
|105| rohit| 30|      101|batsman|
|105| rohit| 30|      102| bowler|
|105| rohit| 30|      103|batsman|
|105| rohit| 30|      104|batsman|
+---+------+---+---------+-------+

20
```

In [12]:
```
# Inner Join using a cloumn
df1_in_df2 = df1.join(df2,df1.Id==df2.Player_Id)
df1_in_df2.show()
```

```
+---+------+---+---------+-------+
| Id|  Name|Age|Player_Id|  Skill|
+---+------+---+---------+-------+
|103| virat| 29|      103|batsman|
|104|saurav| 41|      104|batsman|
|101|sachin| 40|      101|batsman|
|102| zahir| 41|      102| bowler|
+---+------+---+---------+-------+
```

In [22]:
```
# Inner Join using sequence of columns
df1_seq_df2 = df1.join(df2,df1.Id==df2.Player_Id,"left_outer")
df1_seq_df2.show()
```

```
+---+------+---+---------+-------+
| Id|  Name|Age|Player_Id|  Skill|
+---+------+---+---------+-------+
|103| virat| 29|      103|batsman|
|104|saurav| 41|      104|batsman|
|105| rohit| 30|     null|   null|
|101|sachin| 40|      101|batsman|
|102| zahir| 41|      102| bowler|
+---+------+---+---------+-------+
```

In [23]:
```
# Left Semi join
df1_semi_df2 = df1.join(df2,df1.Id==df2.Player_Id,"leftsemi")
df1_semi_df2.show()
```

```
+---+------+---+
| Id|  Name|Age|
+---+------+---+
|103| virat| 29|
|104|saurav| 41|
|101|sachin| 40|
|102| zahir| 41|
+---+------+---+
```

In [24]:
```
# Outer Join
df1_outer_df2 = df1.join(df2,df1.Id==df2.Player_Id,"outer")
df1_outer_df2.show()
```

```
+---+------+---+---------+-------+
| Id|  Name|Age|Player_Id|  Skill|
+---+------+---+---------+-------+
|103| virat| 29|      103|batsman|
|104|saurav| 41|      104|batsman|
|105| rohit| 30|     null|   null|
|101|sachin| 40|      101|batsman|
|102| zahir| 41|      102| bowler|
+---+------+---+---------+-------+
```

In [25]:
```
df1.show()
```

```
+---+------+---+
| Id|  Name|Age|
+---+------+---+
|101|sachin| 40|
|102| zahir| 41|
|103| virat| 29|
|104|saurav| 41|
|105| rohit| 30|
+---+------+---+
```

In [35]:
```
df1.crosstab("Id","Name").show()
df1.crosstab("Id","Age").show()
```

```
+-------+-----+------+------+-----+-----+
|Id_Name|rohit|sachin|saurav|virat|zahir|
+-------+-----+------+------+-----+-----+
|    101|    0|     1|     0|    0|    0|
|    102|    0|     0|     0|    0|    1|
|    105|    3|     0|     0|    0|    0|
|    103|    0|     0|     0|    1|    0|
|    104|    0|     0|     1|    0|    0|
+-------+-----+------+------+-----+-----+

+------+---+---+---+---+
|Id_Age| 29| 30| 40| 41|
+------+---+---+---+---+
|   101|  0|  0|  1|  0|
|   102|  0|  0|  0|  1|
|   105|  0|  3|  0|  0|
|   103|  1|  0|  0|  0|
|   104|  0|  0|  0|  1|
+------+---+---+---+---+
```

In [34]:
```
df1.show()
```

```
+---+------+---+
| Id|  Name|Age|
+---+------+---+
|101|sachin| 40|
|102| zahir| 41|
|103| virat| 29|
|104|saurav| 41|
|105| rohit| 30|
|105| rohit| 30|
|105| rohit| 30|
+---+------+---+
```

In [36]:
```
df1.groupBy("Id").max().show()
```

```
+---+-------+--------+
| Id|max(Id)|max(Age)|
+---+-------+--------+
|103|    103|      29|
|104|    104|      41|
|105|    105|      30|
|101|    101|      40|
|102|    102|      41|
+---+-------+--------+
```

In [43]:
```
from pyspark.sql import *
Datarow = Row("Id","Name","Salary","Department","Age","Bonus","State")
row1 = Datarow(1,"Mukesh",10000,"CS",34,2000,"TA")
row2 = Datarow(2,"MFDFDFD",20000,"EEE",33,201000,"UP")
row3 = Datarow(3,"Mffukesh",10000,"CS",23,20020,"UK")
row4 = Datarow(4,"Mukesh",30000,"CS",40,20050,"UK")
row5 = Datarow(5,"Mukesh",10000,"EEE",54,2000,"DEL")
row6 = Datarow(6,"Mukffesh",40000,"CS",30,2000,"TA")
row7 = Datarow(7,"Mukesh",40000,"EEE",33,204300,"UP")
row8 = Datarow(8,"Mukffesh",40000,"CS",33,22000,"DEL")
row9 = Datarow(8,"ffMukesh",50000,"CS",30,2000,"TA")
row10 = Datarow(1,"ffMukesh",10000,"HR",33,233000,"TA")
row11 = Datarow(1,"ggMukesh",60000,"CS",33,2000,"DEL")
row12 = Datarow(9,"hMukesh",800,"CS",30,232000,"TA")
row13 = Datarow(10,"jMukesh",80000,"HR",33,122000,"TA")
row14 = Datarow(11,"Mukesh",10000,"CS",33,2000,"UP")
row15 = Datarow(12,"Mukesh",10000,"CS",30,22000,"UK")
row16 = Datarow(13,"Mukesh",90000,"HR",29,342000,"UK")
row17 = Datarow(14,"Mukesh",110000,"CS",33,672000,"UK")
row18 = Datarow(15,"Mukesh",10000,"HR",28,662000,"UK")
row19 = Datarow(116,"Mukesh",1000000,"CS",27,772000,"TA")

a = [row1,row2,row3,row4,row5,row6,row7,row8,row9,row10,row11,row12,row13,row14,row15,row16,row17,row18,row19,row10]

df = spark.createDataFrame(a)
```

In [44]:
```
df.show()
```

```
+---+--------+-------+----------+---+------+-----+
| Id|    Name| Salary|Department|Age| Bonus|State|
+---+--------+-------+----------+---+------+-----+
|  1|  Mukesh|  10000|        CS| 34|  2000|   TA|
|  2| MFDFDFD|  20000|       EEE| 33|201000|   UP|
|  3|Mffukesh|  10000|        CS| 23| 20020|   UK|
|  4|  Mukesh|  30000|        CS| 40| 20050|   UK|
|  5|  Mukesh|  10000|       EEE| 54|  2000|  DEL|
|  6|Mukffesh|  40000|        CS| 30|  2000|   TA|
|  7|  Mukesh|  40000|       EEE| 33|204300|   UP|
|  8|Mukffesh|  40000|        CS| 33| 22000|  DEL|
|  8|ffMukesh|  50000|        CS| 30|  2000|   TA|
|  1|ffMukesh|  10000|        HR| 33|233000|   TA|
|  1|ggMukesh|  60000|        CS| 33|  2000|  DEL|
|  9| hMukesh|    800|        CS| 30|232000|   TA|
| 10| jMukesh|  80000|        HR| 33|122000|   TA|
| 11|  Mukesh|  10000|        CS| 33|  2000|   UP|
| 12|  Mukesh|  10000|        CS| 30| 22000|   UK|
| 13|  Mukesh|  90000|        HR| 29|342000|   UK|
| 14|  Mukesh| 110000|        CS| 33|672000|   UK|
| 15|  Mukesh|  10000|        HR| 28|662000|   UK|
|116|  Mukesh|1000000|        CS| 27|772000|   TA|
|  1|ffMukesh|  10000|        HR| 33|233000|   TA|
+---+--------+-------+----------+---+------+-----+
```

In [58]:
```
df.describe("Bonus").show()
df.groupBy("Bonus").max().show()
df.groupBy("Bonus").min().show()
df.groupBy("Bonus").agg({"Bonus":"max"}).show()
```

```
+-------+------------------+
|summary|             Bonus|
+-------+------------------+
|  count|                20|
|   mean|          188468.5|
| stddev|246404.54360290902|
|    min|              2000|
|    max|            772000|
+-------+------------------+
```

```
+------+-------+-----------+--------+----------+
| Bonus|max(Id)|max(Salary)|max(Age)|max(Bonus)|
+------+-------+-----------+--------+----------+
| 20050|      4|      30000|      40|     20050|
|232000|      9|        800|      30|    232000|
|342000|     13|      90000|      29|    342000|
| 20020|      3|      10000|      23|     20020|
|204300|      7|      40000|      33|    204300|
|233000|      1|      10000|      33|    233000|
| 22000|     12|      40000|      33|     22000|
|662000|     15|      10000|      28|    662000|
|122000|     10|      80000|      33|    122000|
|201000|      2|      20000|      33|    201000|
|672000|     14|     110000|      33|    672000|
|772000|    116|    1000000|      27|    772000|
|  2000|     11|      60000|      54|      2000|
+------+-------+-----------+--------+----------+
```

```
+------+-------+-----------+--------+----------+
| Bonus|min(Id)|min(Salary)|min(Age)|min(Bonus)|
+------+-------+-----------+--------+----------+
| 20050|      4|      30000|      40|     20050|
|232000|      9|        800|      30|    232000|
|342000|     13|      90000|      29|    342000|
| 20020|      3|      10000|      23|     20020|
|204300|      7|      40000|      33|    204300|
|233000|      1|      10000|      33|    233000|
| 22000|      8|      10000|      30|     22000|
|662000|     15|      10000|      28|    662000|
|122000|     10|      80000|      33|    122000|
|201000|      2|      20000|      33|    201000|
```

```
|672000|     14|     110000|      33|    672000|
|772000|    116|    1000000|      27|    772000|
|  2000|      1|      10000|      30|      2000|
+------+-------+-----------+--------+----------+
```

```
+------+----------+
| Bonus|max(Bonus)|
+------+----------+
| 20050|     20050|
|232000|    232000|
|342000|    342000|
| 20020|     20020|
|204300|    204300|
|233000|    233000|
| 22000|     22000|
|662000|    662000|
|122000|    122000|
|201000|    201000|
|672000|    672000|
|772000|    772000|
|  2000|      2000|
+------+----------+
```

In [72]:
```
df.groupBy("Bonus").agg({"Bonus":"max"}).show()
df.filter(df.Bonus > 10000).groupBy("Salary").agg({"Bonus":"sum"}).show()
```

```
+------+----------+
| Bonus|max(Bonus)|
+------+----------+
| 20050|     20050|
|232000|    232000|
|342000|    342000|
| 20020|     20020|
|204300|    204300|
|233000|    233000|
| 22000|     22000|
|662000|    662000|
|122000|    122000|
|201000|    201000|
|672000|    672000|
|772000|    772000|
|  2000|      2000|
+------+----------+

+-------+----------+
| Salary|sum(Bonus)|
+-------+----------+
|  10000|   1170020|
|  40000|    226300|
|  30000|     20050|
|1000000|    772000|
|    800|    232000|
| 110000|    672000|
|  90000|    342000|
|  20000|    201000|
|  80000|    122000|
+-------+----------+
```

In [53]:
```
df.describe().show()
```

```
+-------+------------------+-------+------------------+----------+------------------+------------------+-----+
|summary|                Id|   Name|            Salary|Department|               Age|             Bonus|State|
+-------+------------------+-------+------------------+----------+------------------+------------------+-----+
|  count|                20|     20|                20|        20|                20|                20|   20|
|   mean|             12.35|   null|           82040.0|      null|              32.6|          188468.5| null|
| stddev|24.828411145298844|   null|218245.32405915187|      null|6.0732372361299864|246404.54360290902| null|
|    min|                 1|MFDFDFD|               800|        CS|                23|              2000|  DEL|
|    max|               116|jMukesh|           1000000|        HR|                54|            772000|   UP|
+-------+------------------+-------+------------------+----------+------------------+------------------+-----+
```

In [52]:
```
df.printSchema()
```

```
root
 |-- Id: long (nullable = true)
 |-- Name: string (nullable = true)
 |-- Salary: long (nullable = true)
 |-- Department: string (nullable = true)
 |-- Age: long (nullable = true)
 |-- Bonus: long (nullable = true)
 |-- State: string (nullable = true)
```

In [ ]:
```

```