Dataframe Queries

In [26]:

```
rdd1 = sc.textFile("wasb:///data/data.txt")
rdd1.collect()

Header=rdd1.first()
rdd1 = rdd1.filter(lambda x:x!=Header)
rdd1.take(5)
```

```
['I,2010,Hyd', 'I,2011,Bang', 'W,2012,Chennai', 'I,2013,Delhi', 'B,2014,Mu
mbay']
```

In [27]:

```
from pyspark.sql.types import StructType,StructField,StringType
from pyspark import *

schema = StructType([StructField('Match_id',StringType(),True),
                     StructField('Year',StringType(),True),
                     StructField('Location',StringType(),True) ])
df1 = spark.createDataFrame(rdd1.map(lambda x:x.split(',')),schema=schema)
df1
```

```
DataFrame[Match_id: string, Year: string, Location: string]
```

In [28]:

```
df1.show()
```

```
+--------+----+--------+
|Match_id|Year|Location|
+--------+----+--------+
|       I|2010|     Hyd|
|       I|2011|    Bang|
|       W|2012| Chennai|
|       I|2013|   Delhi|
|       B|2014|  Mumbay|
+--------+----+--------+
```

In [18]:

```
type(df1)
```

```
<class 'pyspark.sql.dataframe.DataFrame'>
```

In [19]:

```
type(schema)
```

```
<class 'pyspark.sql.types.StructType'>
```

In [20]:

```
df1.printSchema()
```

```
root
 |-- Match_id: string (nullable = true)
 |-- Year: string (nullable = true)
 |-- Location: string (nullable = true)
```

In [29]:

```
df1.head(2)
```

```
[Row(Match_id='I', Year='2010', Location='Hyd'), Row(Match_id='I', Year='2
011', Location='Bang')]
```

In [30]:

```
df1.count()
```

```
5
```

In [31]:

```
len(df1.columns)
```

```
3
```

In [32]:

```
df1.describe().show()
```

```
+-------+--------+------------------+--------+
|summary|Match_id|              Year|Location|
+-------+--------+------------------+--------+
|  count|       5|                 5|       5|
|   mean|    null|            2012.0|    null|
| stddev|    null|1.5811388300841898|    null|
|    min|       B|              2010|    Bang|
|    max|       W|              2014|  Mumbay|
+-------+--------+------------------+--------+
```

In [34]:

```
df1.describe('Year').show()
```

```
+-------+------------------+
|summary|              Year|
+-------+------------------+
|  count|                 5|
|   mean|            2012.0|
| stddev|1.5811388300841898|
|    min|              2010|
|    max|              2014|
+-------+------------------+
```

In [36]:

```
df1.describe('Year').show()
```

```
+-------+------------------+
|summary|              Year|
+-------+------------------+
|  count|                 5|
|   mean|            2012.0|
| stddev|1.5811388300841898|
|    min|              2010|
|    max|              2014|
+-------+------------------+
```

In [37]:

```
df1.select('year','location').show()
```

```
+----+--------+
|year|location|
+----+--------+
|2010|     Hyd|
|2011|    Bang|
|2012| Chennai|
|2013|   Delhi|
|2014|  Mumbay|
+----+--------+
```

In [40]:

```
df1.select('Match_id').distinct().show()
df1.select('Match_id').distinct().count()
```

```
+--------+
|Match_id|
+--------+
|       B|
|       W|
|       I|
+--------+

3
```

In [46]:

```
df1.crosstab('Match_id','Year').show()
```

```
+-------------+----+----+----+----+----+
|Match_id_Year|2010|2011|2012|2013|2014|
+-------------+----+----+----+----+----+
|            I|   1|   1|   0|   1|   0|
|            W|   0|   0|   1|   0|   0|
|            B|   0|   0|   0|   0|   1|
+-------------+----+----+----+----+----+
```

In [47]:

```
df1.crosstab('Match_id','Location').show()
```

```
+-----------------+----+-------+-----+---+------+
|Match_id_Location|Bang|Chennai|Delhi|Hyd|Mumbay|
+-----------------+----+-------+-----+---+------+
|                I|   1|      0|    1|  1|     0|
|                W|   0|      1|    0|  0|     0|
|                B|   0|      0|    0|  0|     1|
+-----------------+----+-------+-----+---+------+
```

In [49]:

```
df1.dropna().show()
```

```
+--------+----+--------+
|Match_id|Year|Location|
+--------+----+--------+
|       I|2010|     Hyd|
|       I|2011|    Bang|
|       W|2012| Chennai|
|       I|2013|   Delhi|
|       B|2014|  Mumbay|
+--------+----+--------+
```

In [59]:

```
df1.filter(df1.Match_id =='B').count()
df1.filter(df1.Match_id == 'I').count()

a = df1.filter(df1.Match_id =='B').count()
b = df1.filter(df1.Match_id == 'I').count()
print(a,b)
```

```
1 3
```

In [60]:

```
# train.groupby('Age').agg({'Purchase': 'mean'}).show()

df1.groupby('Match_id').agg({'Year':'max'}).show()
```

```
+--------+---------+
|Match_id|max(Year)|
+--------+---------+
|       B|     2014|
|       W|     2012|
|       I|     2013|
+--------+---------+
```

In [61]:

```
df1.groupby('Match_id').count().show()
```

```
+--------+-----+
|Match_id|count|
+--------+-----+
|       B|    1|
|       W|    1|
|       I|    3|
+--------+-----+
```

In [67]:

```
df1.orderBy(df1.Year.desc()).show()
```

```
+--------+----+--------+
|Match_id|Year|Location|
+--------+----+--------+
|       B|2014|  Mumbay|
|       I|2013|   Delhi|
|       W|2012| Chennai|
|       I|2011|    Bang|
|       I|2010|     Hyd|
+--------+----+--------+
```

In [74]:

```
df1.withColumn('Captain',df1.Match_id).select('Match_id','Captain').show()
```

```
+--------+-------+
|Match_id|Captain|
+--------+-------+
|       I|      I|
|       I|      I|
|       W|      W|
|       I|      I|
|       B|      B|
+--------+-------+
```

In [75]:

```
df1.drop('captain').columns
```

```
['Match_id', 'Year', 'Location']
```

In [ ]:

```

```