Write a function that that will insert data type of double into a file and ensure that elements have evenly distributed the values. The function should take two arguments: N (Number of elements) and the full name of the file (ie fully qualified path).

1. Create a file with 50,000 elements of data type double using the function you have created as stated earlier.
2. Read the file created earlier, read into RDD and compute the average and convert the double into Integer find the sum of Integer. (use scala with spark without using data frame and dataset)
3. Rewrite the program for Question no 2 by using Dataframe and Dataset.
4. Using a DataFrame create a random sample of about 100 elements from the file created.

In [1]:
```
from pyspark.sql import *
from pyspark.sql import DataFrame
from pyspark.sql.functions import col, when, max, min
from pyspark.sql.functions import count, sum,avg
```

Function Creation

In [2]:
```
def create_file(N,Path):
    import numpy as np
    def f(x):
        return np.random.randint(1000)

    x = sc.parallelize([0] * N).map(f).map(lambda x: float(x))
    print(x.count())
    x.saveAsTextFile(Path)
    print("file created successfully")
```

Question.1

In [3]:
```
create_file(50000,"C:\Users\Ananya Chandraker\Documents\Assignment.csv")
```

```
50000
file created successfully
```

Question.2

In [5]:
```
rdd = sc.textFile("C:\Users\Ananya Chandraker\Documents\Assignment.csv")
rdd1 = rdd.map(lambda x: float(x))
rdd1.count()
```

Out[5]: 50000

Question 2 part 2 Datasets are bit new I have not worked with it, but definetly I can learn it

Question.3

In [21]:
```
from pyspark.sql import SQLContext
#sqlContext = SQLContext(sc)
df1 = spark.read.csv("C:\Users\Ananya Chandraker\Documents\Assignment.csv")
type(df1)
df1.printSchema()
df2 = df1.withColumn('sample', df1['_c0'].cast('int'))
df2.printSchema()
df_avg = df2.select(avg("sample"))
df_sum = df2.select(sum("sample"))
df_sum.show()
df_avg.show()
```

Out[21]:
```
pyspark.sql.readwriter.DataFrameReader
```

Currently due to some version issue I am not able to run the question.3 code in same jupyter notebook as it is giving some error, but when I am running the same in my cloud m/c getting the output as below:

root |-- _c0: string (nullable = true) root |-- _c0: string (nullable = true) |-- sample: integer (nullable = true) +----------+ |sum(sample)| +-----------+ | 25029726| +-----------+ +-----------+ |avg(sample)| +-----------+ | 500.59452| +----------+

Question.4

In [1]:
```
sample_data = sc.textFile("C:\Users\Ananya Chandraker\Documents\Assignment.csv").map(lambda x: float(x)).map(lambda x: (x,))
new_df = spark.createDataFrame(sample_data,['Value']).limit(100)
new_df.show(5)
```

```
+-----+
|Value|
+-----+
|107.0|
|888.0|
|390.0|
|513.0|
|627.0|
+-----+
only showing top 5 rows
```

In [ ]:
```

```