RDD

R- Resilient

D- Distributed

D- Dataset

Spark rdd have following features:

🡪Immutable distributed objects

🡪Fault tolerance

🡪Each dataset in rdd can be logically partitioned on different nodes of cluster

Creation of rdd by various ways:

🡪using parallelize()

data=[('Nikunj','IT'),('Ahad','CS'),('Vaibhav','AIDS')]  
rdd\_eg=spark.sparkContext.parallelize(data)  
print(rdd\_eg.collect())

🡪Using textfile

#TO CREATE A FILE

with open('tp.txt', 'w') as f:  
 f.write('Create a new text file!')

rdd=spark.sparkContext.textFile('tp.txt')  
print(rdd.collect())

RDD Operation

1. Transformation:

Transformation is an operation applied on a given rdd to create a new rdd.

It follows lazy Evaluation approach.

Lazy evaluation: the task will not get started unless action is being triggered.

Few transformation are:

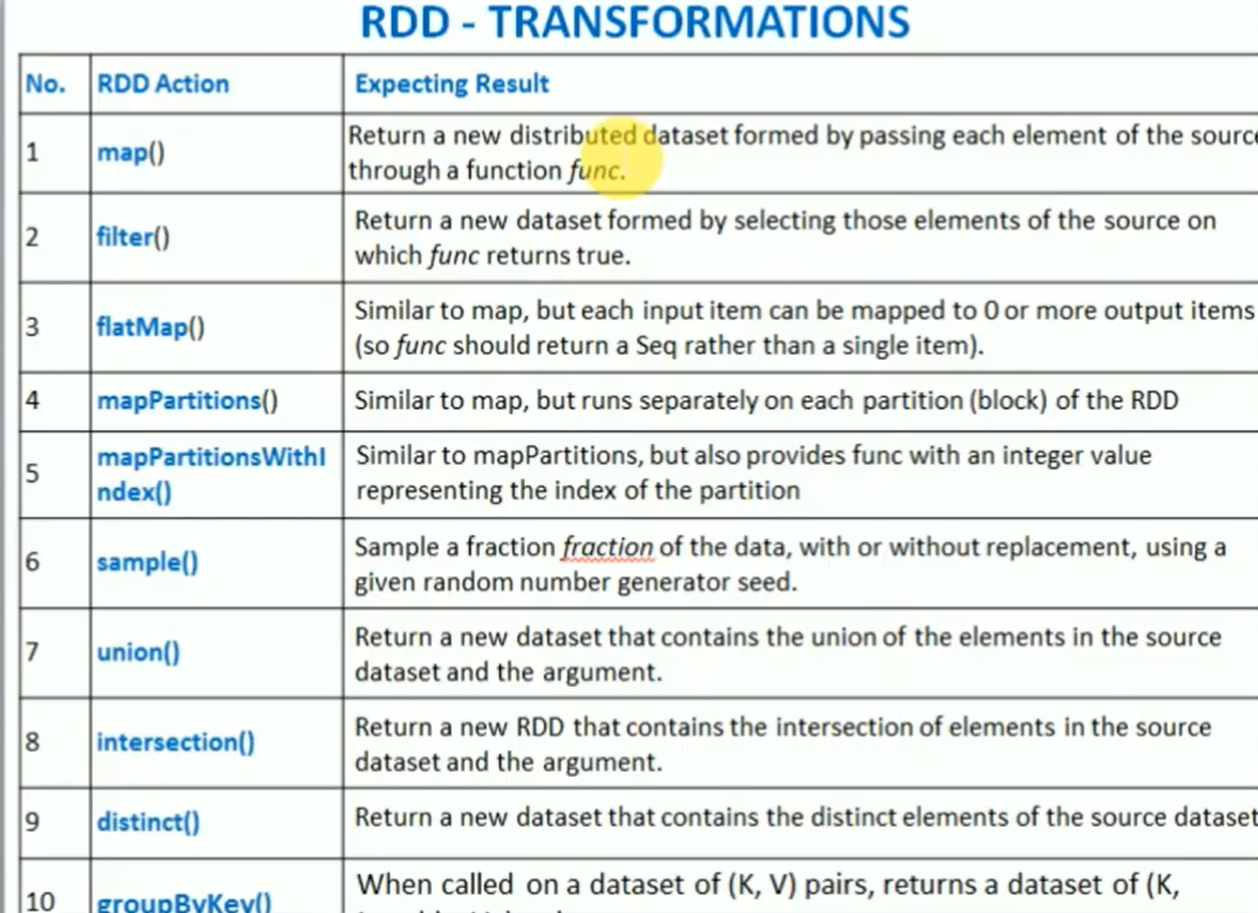
🡪map

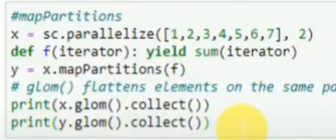
🡪flatmap

🡪filter

🡪reduceByKey

🡪sortBy

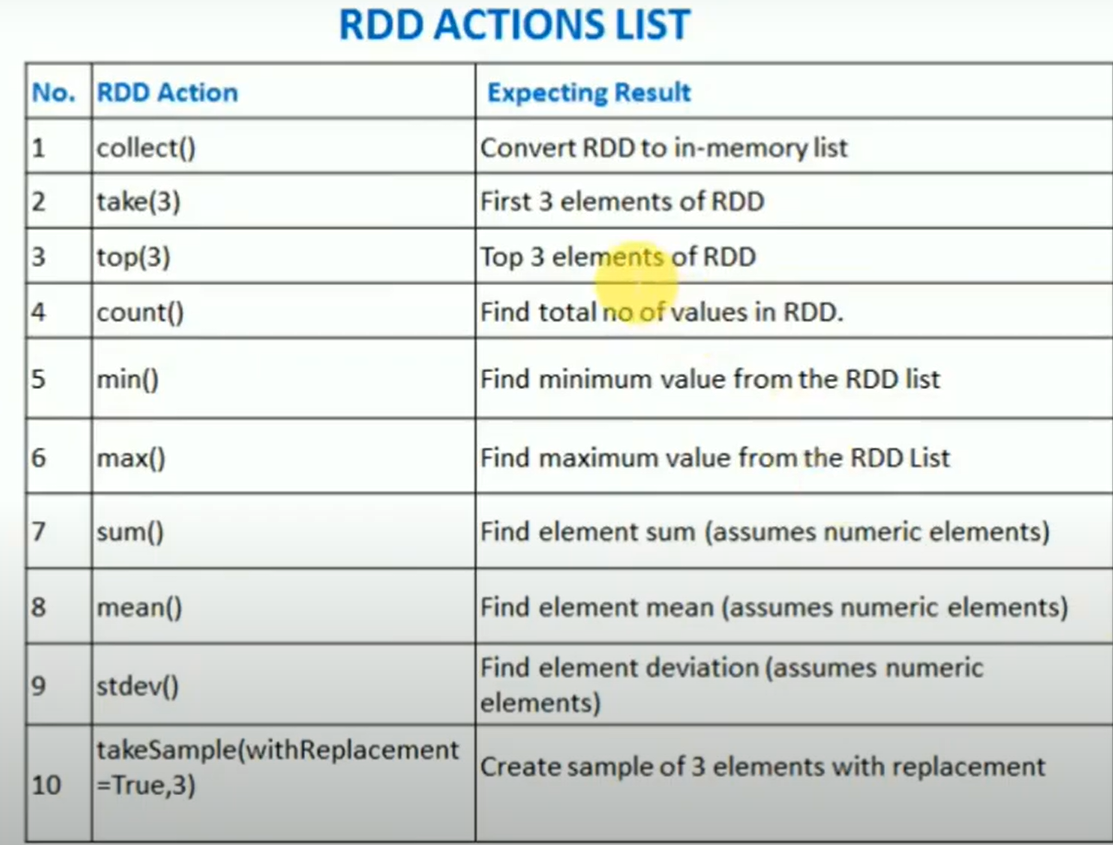




1. Actions:

Actions are the operations which are applied on an rdd, which returns a value to driver program after running a computation on dataset.

Some actions listed are:



To run actions first we need to set an environment on local , by initializing findspark.

rdd=spark.sparkContext.parallelize([1,5,9,4,6,7,2,3])  
print(rdd.collect())  
print(rdd.take(3))  
print(rdd.top(5))  
print(rdd.count())  
print(rdd.max())  
print(rdd.min())

Difference between RDD ,Dataset,dataframe

