**RDD (Resilient Distributed Dataset)**

* <https://www.tutorialspoint.com/apache_spark/apache_spark_rdd.htm>

**Spark Context**

* Created by the driver program
* Environment RDD’s run in
* Creates RDD’s
* Responsible for making RDD’s resilient

**RDD creation example**

* sc.textFile(‘<file-system>//somefile.txt’)
* hiveCtx = HiveContext(sc)
  + run SQL queries like so
    - rows = hiveCtx.sql(“SELECT name, age FROM users”)
* Also create from
  + Cassandra
  + HBase
  + Elasticsearch
  + JSON, CSV, or other file

**Transforming RDD’s**

* Map
  + 1:1 relationship
  + Example
    - Rdd = sc.parallelize([1,2,3,4])
    - squareRDD = rdd.map(lambda x: x\*x)
* Flatmap
  + Any type of relationship (1:m)
* Filter
  + Filter content from the RDD
* Distinct
  + Distinct values
* Sample
  + Random sampling
* Union, intersection, subtract, Cartesian
  + Merging RDD’s

**RDD actions**

* Lazy Evaluation
  + An action is where anything happens
  + This is where a job is spawn
* API that allows you to obtain values from an RDD