

Apache Spark RDD Basics

What is RDD?

1. RDD is Spark's core abstraction, which is Resilient Distributed Dataset

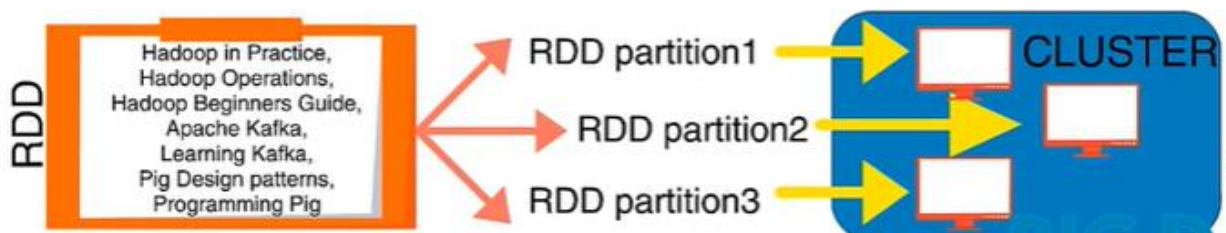
Resilient – Fault Tolerant, means ability to be re-computed from history

2. It is immutable distributed collection of objects

Immutable – can't be modified.

Distributed – loaded across various nodes of the cluster

3. Internally spark distributes the data in RDD, to different nodes across the cluster to achieve parallelization



RDD Creation

RDDs can be created

By loading an external dataset	By distributing collection of objects
for example, loading an external dataset books.txt can be done as below <code>val booksRDD = sc.textFile("/path/to/books.txt")</code>	for example, let's create a list collection and pass it to parallelize method of spark context <code>val colorsRDD = sc.parallelize(List["red", "blue"])</code>

Example 1:

RDD creation by loading an external dataset

```
$ val booksRDD = sc.textFile("/home/hduser/example.txt")
```

```
$ booksRDD.collect()
```

Example 2:

RDD creation by distributing collection

```
$ val colorsRDD = sc.parallelize(List("Red", "Green"))
```

```
$ colorsRDD.collect()
```