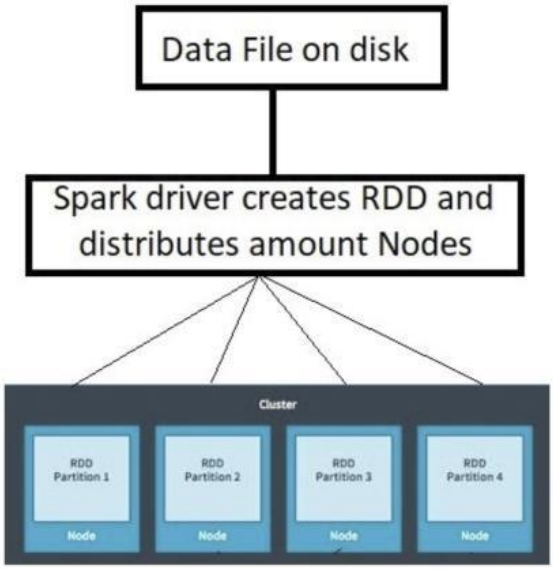
Task No 07

RDD

- RDD = Resilient Distributed Datasets



Creating RDDs. How to do it?

- Parallelizing an existing collection of objects
- Perform Action and Transformation
- From existing RDDs
- External datasets:
  Files in HDFS
  Objects in Amazon S3 bucket
  lines in a text ,file

There are limitations of RDDs. So, dataframes overcome that limitations of Rdds. In this manual yougo through how to use data frames in pyspark.
For detailed information, go through the documentation of Pyspark.

DataFrames

There are multiple ways to create DataFrames in Apache Spark:

- DataFrames can be created using an existing RDD
- You can create a DataFrame by loading a CSV file directly
- You can programmatically specify a schema to create a DataFrame