Crear un RDD en python con pyspark

"Un Dataset Distribuido Resiliente (RDD), la abstracción básica en Spark. representa una colección inmutable y dividida de elementos que pueden ser operados en paralelo.
Esta clase contiene las operaciones básicas disponibles en todos los RDD "

Existen 2 formas de crear un RDD.

1.A partir de una colección

2.A partir de otra fuente como un fichero

Para crear un RDD la principal forma es usar el método parallelize() de la clase SparkContenxt a partir de una colección de datos.

- Crear una colección de datos
- Usar el método parallelize() que nos permite crear un RDD a partir de una lista o tupla de elementos

lines=sc.parallelize([“pandas”,”pspark”])

Hay otras formas de crear un RDD como a través de un recurso externo(fichero,bd)

lines=sc.textFile(“ruta_fichero”)

Este comando utiliza el método textFile de la clase SparkContext para crear una instancia de conjunto de datos resiliente y distribuido(RDD)