# Curso Actualízate – Cloud A Coruña (Módulo 6)

Nombre: Víctor

Apellidos: Beneito López

Fecha: 09/03/2023

Responde a las siguientes preguntas. Justifica la respuesta.

1. **¿Qué es PySpark?**

Librería de Python para poder utilizar Apache Spark. Y Apache Spark es un motor de código abierto desarrollado específicamente para el procesamiento y análisis de datos a gran escala.

OK

1. **¿Qué es un RDD y de qué son las siglas?**

RDDs (Resilient Distributed Dataset). Colección inmutable y distribuida de elementos.

OK

1. **¿Qué cosas les envía el maestro a los trabajadores? ¿Y cómo le comparte la faena el maestro a los trabajadores?**

El Master divide la estructura de datos y designa los Workers a cada una de las divisiones mediante variables compartidas (tipo Broadcast)

OK

1. **¿Qué es un cluster?**

Conjunto de ordenadores trabajando entre sí, cooperando para lograr un objetivo.

OK

1. **Define transformación y acción con tus palabras.**

**Transformación**: cambios/tareas que hay que realizar dentro de cada closure. Tras aplicar una transformación, obtenemos un nuevo y modificado RDD basado en el original.

**Acción**: cuando el Master les proporciona las transformaciones a realizar a los Workers. Es aplicar una operación sobre un RDD y obtener un valor como resultado, que dependerá del tipo de operación.

OK

1. **Menciona al menos dos transformaciones y dos acciones.**

Transformaciones: map y filter

Acciones: count y collect

OK

1. **Crea una lista de Python y guárdalo en una variable que se llame lista\_ejercicio.**

lista\_ejercicio = [1,2,3,4,5]

OK

1. **Aplica un map() a la lista con una lambda que multiplique por tres a cada elemento.**

map(lambda x : x\*3, lista\_ejercicio)

OK

1. **Ahora paraleliza la lista lista\_ejercicio y guárdalo en la variable rdd\_ejercicio.**

sc = SparkContext(master = "local", appName="Ejercicio")

rdd\_ejercicio = sc.parallelize(lista\_ejercicio)

OK

1. **Aplica un map() al RDD rdd\_ejercicio que multiplique por tres a cada elemento. Fijate en las diferencias de sintaxis. El ejercicio 1 y 2 son secuenciales y el 3 y 4 paralelizados.**

~~map(lambda x : x\*3, rdd\_ejercicio))~~

nuevo\_rdd = rdd\_ejercicio.map(lambda x : x\*3)

1. **Ahora desparaleliza el RDD rdd\_ejercicio.**

nuevo\_rdd.collect()

sc.stop()