# /Curso Actualízate – Cloud A Coruña (Módulo 6)

Nombre: Esteban

Apellidos: Vásquez González

Fecha: 09/03/2023

Responde a las siguientes preguntas. Justifica la respuesta.

1. ¿Qué es PySpark? Es una librería API para Python que permite utilizar los clusters de spark y tener acceso a las funcionalidades de paralelizar las tareas de manejo de altos volúmenes de datos
2. ¿Qué es un RDD y de qué son las siglas? Resilent Distributed Database. Hace referencia a los datos paralelizados por el Master y la asignación de los mismos a los Workers.
3. ¿Qué cosas les envía el maestro a los trabajadores? ¿Y cómo le comparte la faena el maestro a los trabajadores? El maestro envía a los trabajadores los RDDs y las instrucciones de qué acciones aplicar sobre los mismos (clousures)

El maestro se encarga de dividir la data en secciones según la cantidad de nodos disponibles (RDD) y asignarlos a los nodos trabajadores utilizando un programa Driver

1. ¿Qué es un cluster? Es un grupo de ordenadores interconectados como nodos que distribuyen tareas de procesamiento de datos, en cada cluster hay un nodo maestro y múltiples nodos trabajadores. El maestro se encarga de asignar la tereas a los nodos y luego unir los resultados
2. Define transformación y acción con tus palabras. La transformación son los cambios que se van a realizar a los RDDs y las acciones las instrucciones de ejecución de dichos cambios
3. Menciona al menos dos transformaciones y dos acciones. Dos transformaciones map(), filter(). Dos acciones count(), take()
4. Crea una lista de Python y guárdalo en una variable que se llame lista\_ejercicio. Lista\_ejercicio = [1,2,3,4,5,6,7]
5. Aplica un map() a la lista con una lambda que multiplique por tres a cada elemento. list(map(lambda x : x\*3, lista\_ejercicio))
6. Ahora paraleliza la lista lista\_ejercicio y guárdalo en la variable rdd\_ejercicio.

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

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.

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

1. Ahora desparaleliza el RDD rdd\_ejercicio. rdd\_ejercicio.collect()