# Curso Actualízate – Cloud A Coruña (Módulo 6)

Nombre: Daniel

Apellidos: Silvarrey Fernandez

Fecha: 9/3/23

Responde a las siguientes preguntas. Justifica la respuesta.

1. ¿Qué es PySpark?

Es una librería de Python que permite realizar y procesar grandes volúmenes de datos de manera distribuida y se basa en el motor de procesado distribuido de Apache Spark y permite el procesamiento paralelo de datos mediante la creación y manipulación de RDDs

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

Resilient Distributed Datasets, un RDD es una colección inmutable y distribuida de objetos que se pueden procesar en paralelo

1. ¿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 las tareas a realizar en forma de bloques de datos, y les asigna el trabajo en función de la disponibilidad de recursos y la distribución de datos.

1. ¿Qué es un cluster?

Un cluster es un conjunto de ordenadores conectados que trabajan juntos para realizar tareas de manera colaborativa y compartida, con el objetivo de aumentar la capacidad de procesamiento y mejorar el rendimiento en la ejecución de aplicaciones y procesos, un cluster se refiere a un grupo de nodos que trabajan juntos para procesar grandes conjuntos de datos mediante la distribución de trabajo en múltiples máquinas.

1. Define transformación y acción con tus palabras.

Transformación es como un plan o una receta que le dices a Spark qué hacer con los datos en un RDD para transformarlos en otro RDD. Es como si estuvieras diciéndole a Spark: "Quiero que hagas esto con mis datos y me devuelvas el resultado en un RDD nuevo"

Acción es como una orden que le das a Spark para que haga algo con los datos en un RDD y te devuelva un resultado que puedes ver o guardar. Es como si le dijeras a Spark: "Quiero que hagas esto con mis datos y me muestres el resultado".

1. Menciona al menos dos transformaciones y dos acciones.

map y filter como Transformaciones.

count y collect como Acciones.

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

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

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

lista\_ejercicio = [2, 4, 6, 8, 10]

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

lista\_result = list(result)

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

from pyspark import SparkContext

sc = SparkContext("local", "ejercicio")

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

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.

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

1. Ahora desparaleliza el RDD rdd\_ejercicio.

lista\_desparalelizada = rdd\_ejercicio.collect()