Curso Actualízate – Cloud A Coruña (Módulo 6)

Nombre: Tania

Apellidos: Sánchez Sánchez Fecha: 9/3/2023

Responde a las siguientes preguntas. Justifica la respuesta.

# ¿Qué es PySpark?

Es una librería de Python que pertenece a Apache Fundation para hacer código gratuito. Procesa de manera distribuida las grandes cantidades de información.

OK

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

Es una estructura de datos paralelizada, es decir dividida en partes. Los RDD cargan colecciones de datos.

Las siglas significan Resilient Distributed Dataset.

OK

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

El maestro les envía los datos en forma de RDD, es decir, su trozo.

El maestro indica a los trabajadores lo que tienen que realizar con su RDD mediante la Clausura (Clousure) a través del programa driver mediante variables compartidas de tipo broadcast.

# ¿Qué es un cluster?

Es una agrupación de ordenadores conectados entre sí.

OK

# Define transformación y acción con tus palabras.

**Transformación** es lo que le dice el maestro a los trabajadores que tienen que realizar. ( teneis que hacer esta tarea)

**Acción** es la orden que le da el maestro a los trabajadores para que realicen lo que les ha indicado en la transformación ( hacedlo!!!!)

OK

# Menciona al menos dos transformaciones y dos acciones.

TRANSFORMACIONES: .map() y .filter() ACCIONES: count() y collect()

OK

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

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

OK

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

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

l~~istamap = lista\_ejercicio.map(lambda numero : 3\*numero)~~

listamap = list(map(lambda numero : 3\*numero, lista\_ejercicio))

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

rdd\_ejercicio = sc.parallelize(lista\_ejercicio) (no sé si es la original o si tendría que poner la listamap ya con los números multiplicados \*3)

OK la original, como lo has hecho

# 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\_ejerciciomap= rdd\_ejercicio.map(lambda numerordd : 3\*numerordd)

¡Perfecto! Fijate como cambia la sintaxis del map utilizado con estructuras de datos secuenciales vs a RDDs paralelizados.

# Ahora desparaleliza el RDD rdd\_ejercicio.

rdd\_ejerciciomap.collect()