# Curso Hackea Tu Futuro – Big Data Zaragoza (Módulo 6)

Nombre: Nerea

Apellidos: Jimeno Noriega

Fecha: 13/04/2023

Responde a las siguientes preguntas. Justifica la respuesta.

1. ¿Qué es Apache Spark?

Es una plataforma para procesar datos a gran escala de forma rápida y estable basado en la distribución cluster.

Apache Spark es de código abierto y cuenta con bibliotecas integradas para procesamiento de datos, incluida la manipulación de datos estructurados, el procesamiento de gráficos…, además es compatible con bastantes lenguajes de programación.

OK

Para ello, se trabaja con un grupo de ordenadores denominado Cluster. El ordenador más potente (Master), a través de la variable compartida de tipo broadcast, divide la tarea (RDD) entre el resto de ordenadores (Workers) según sus características (sus capacidades de memoria, calidad de conexión a internet, etc) y les indica la transformación (Closure) a aplicar. Los trabajadores guardan dichas instrucciones en la memoria y las ejecutan cuando reciben órdenes de hacerlo (acción).

Esta forma de trabajar permite aumentar el rendimiento cuando se trabaja con grandes cantidades de datos, pero no es aconsejable utilizarlo en caso de disponer de pocos datos, puesto que ralentizará el proceso.

1. ¿Qué son las funciones lambda?

Las funciones lambda son funciones anónimas en Python, estas se pueden usar cuando haya que pasar una función como parámetro y cuentan con una única instrucción cuyo valor corresponde al valor devuelto.

OK

Gastan menos memoria, pero no pueden utilizarse para resolver todo tipo de problemas de código.

1. ¿Para qué sirven los map()? ¿Cómo se usan con RDDs?

La función map() en Python se utiliza para crear un nuevo RDD a partir de otro aplicando una transformación a cada elemento original.

La función que se pasa a map debe recibir un único parámetro, que serán elementos individuales del RDD de partida y devolver el elemento transformado.

OK

1. ¿Para qué sirven los filter()? ¿Cómo se usan con RDDs?

La función filter() en Python se utiliza para crear un nuevo RDD a partir de otro manteniendo solo los elementos de la lista original que cumplan una condición.

La función que se pasa a filter debe recibir un único parámetro, que serán los elementos individuales del RDD de partida y devolver True o False para indicar si el elemento pasa o no el filtro.

OK

1. ¿Cuál es la diferencia entre transformaciones y acciones?

Las transformaciones en Spark son operaciones que transforman un RDD en otro RDD, mientras que las acciones son operaciones que realizan algún tipo de acción en un RDD y devuelven un valor o escriben los datos en una variable compartida de tipo broadcast.

Las transformaciones son de evaluación perezosa (no se ejecutan hasta que se realiza una acción en el RDD).

OK

1. ¿Cuáles son los tipos primitivos en Python y qué valores pueden contener cada uno de ellos? Pista: Son 3 tipos.

Los tipos primitivos en Python son:

* Integer(Int): números enteros (por ejemplo, 1, -2, 0).
* Float: números decimales (por ejemplo, 1.44, -0.8).
* Boolean (Bool): valores booleanos (True o False).
* String (cadena): cadena de texto (‘Hola mundo’ o “Hola mundo”)

OK

1. Escribe la sintaxis para crear variables.

La sintaxis para crear variables es:

nombre\_variable = valor\_variable

1. Escribe la sintaxis para crear funciones.

La sintaxis para crear funciones es:

def nombre\_funcion(parámetros):

return valor\_retorno

OK

1. Escribe la sintaxis para llamar a variables.

La sintaxis para llamar a las variables es:

print(nombre\_variable)

OK

1. Escribe la sintaxis para llamar a funciones.

La sintaxis para llamar a funciones es:

nombre\_funcion(argumentos)

OK