viu Universidad Internacional de Valencia

Actividad 2
Programación Apache Spark

Procesamiento de datos masivos

Y. Cardinale

Introducción:

El procesamiento masivo de datos no siempre se puede realizar con tecnologías tradicionales. En muchas ocasiones se tienen que utilizar tecnologías Big Data como Hadoop MapReduce y Spark.

Objetivo:

Conocer el modelo de procesamiento MapReduce y las principales herramientas Big Data.

Desarrollar programas Big Data utilizando el framework Apache Spark.

Trabajo previo:

Lectura del material docente de la parte específica que se encuentra disponible desde el comienzo del curso en la carpeta: Recursos y materiales>1. Materiales docentes:

Visualización de las videoconferencias “VC Procesamiento Big Data en memoria”, “AG Apache Spark”, y “VC Transformaciones y acciones Spark” que se encontrarán disponibles en: Videoconferencias> grabaciones

Metodología:

En las videoconferencias teóricas (VC) se expondrá al alumno conocimientos, material e indicaciones suficientes para que pueda elaborar una unidad didáctica basada en el aprendizaje y enseñanza por competencias en matemáticas e informática. En la videoconferencia de actividad guiada (AG) se establecerá las pautas concretas y la dinámica que el alumnado deberá seguir para realizar la actividad propuesta. Las actividades se centrarán en poner en práctica y asentar los conocimientos adquiridos en la videoconferencia teórica anterior.

Actividades a elaborar:

Desarrollo de dos programas Big Data utilizando el framework Spark. Se podrá utilizar el lenguaje Python o Java.

1. Dado un dataset que contenga entradas con la forma “persona;método_pago;dinero_gastado”, crea un programa llamado personaGastosConTarjetaCredito que para cada persona indique la suma del dinero gastado con tarjeta de crédito. Ejemplo:

Entrada	Salida
`Alice;Tarjeta de crédito;100`	`Alice;250`
`Alice;Tarjeta de crédito;150`	`Bob;201`
`Alice;Bizum;200`	

```
Bob;Tarjeta de crédito;201
```

Notar que Alice gasta en total 450 euros, pero sólo 250 son con tarjeta de crédito (100 + 150).

2. Dado un dataset que contenga entradas con la forma “persona;método_pago;dinero_gastado”, crea un programa llamado personaYMetodosDePago que:

a) Por cada persona indique en cuántas compras pagó más de 1500 euros con tarjeta de crédito. La solución se tiene que guardar en un archivo comprasCreditoMayorDe1500.
b) Por cada persona indique en cuántas compras pagó menos de 1500 euros con tarjeta de crédito. La solución se tiene que guardar en un archivo comprasCreditoMenorDe1500.

Se valorará positivamente la optimización del programa, por ejemplo a través de la funcionalidad Combiner.

Ejemplo:

Entrada	Salida (a)	Salida (b)
`Alice;Tarjeta de crédito;1000`	`Alice;2`	`Alice;1`
`Alice;Tarjeta de crédito;1800`	`Bob;0`	`Bob;0`
`Alice;Tarjeta de crédito;2100`		
`Bob;Bizum;2000`		

Notar que si bien Bob hace una compra superior a 1500 euros, no la hace con tarjeta de crédito.

Sobre la entrega:

- La tarea se entregará en algún formato comprimido (gzip, zip, etc.), con dos carpetas, uno por programa, con el código fuente de los ejercicios, documentos en PDF (con la explicación de la solución y los print screen de las ejecuciones), así como instrucciones para la compilación y ejecución de dichos códigos estándares y tendrá el siguiente formato:
  AG_2-02MBIG-P1(o 2)-Apellido-Nombre.gz
- Cada estudiante deberá explicar su solución y demostrarla al profesor (no necesariamente en horas de clase)
- Esta actividad tiene un peso de 20%