PRÁCTICA 2 SISTEMAS DE INFORMACIÓN.

CURSO 2023-2024 TRATAMIENTO PARALELO DE DATOS UTILIZANDO MAP- REDUCE.

Índice

1. [Introducción 2](#_TOC_250005)
2. [Dataset 2](#_TOC_250004)
3. [Entorno de trabajo. 3](#_TOC_250003)
4. [Procedimiento 3](#_TOC_250002)
5. [Detalles generales de la aplicación 3](#_TOC_250001)
6. [Entrega de prácticas. 4](#_TOC_250000)

# Introducción.

A lo largo de esta práctica se estudiarán los conceptos básicos en la gestión de grandes cantidades de datos utilizando el paradigma Map-Reduce. Gracias a este paradigma podremos gestionar una gran cantidad de datos de forma paralela, en un cluster adecuado para ello.

# Dataset.

Contexto

El análisis de ventas es la práctica de generar información a partir de datos de ventas, tendencias y métricas para establecer objetivos y prever el rendimiento futuro de las ventas. El análisis de ventas consiste en extraer datos para evaluar el rendimiento del equipo de ventas en relación con sus objetivos. Proporciona información sobre los productos/servicios de mayor y menor rendimiento, los problemas en la venta y las oportunidades de mercado, la previsión de ventas y las actividades de ventas que generan ingresos.

Contenido

Order ID - Un ID de pedido es el sistema numérico que Amazon utiliza exclusivamente para realizar un seguimiento de los pedidos. Cada pedido recibe su propio ID de pedido que no se duplicará. Este número puede ser útil para el vendedor cuando intente averiguar ciertos detalles sobre un pedido, como la fecha o el estado del envío.

Product- El producto que se ha vendido.

Quantity Ordered- La cantidad pedida es la cantidad total de artículos pedidos en el pedido inicial (sin cambios).

Price Each- El precio de cada producto.

Order Date - Es la fecha en la que el cliente solicita el envío del pedido.

Purchase Address - La orden de compra es preparada por el comprador, a menudo a través de un departamento de compras. La orden de compra, o PO, suele incluir un número de PO, que es útil para cotejar los envíos con las compras; una fecha de envío; la dirección de facturación; la dirección de envío; y los artículos solicitados, las cantidades y el precio.

El dataset se encuentra almacenado **al completo** en el HDFS de Hadoop en el CPD del Centro Universitario de Mérida. Está formado por varios ficheros planos de datos, donde la información se encuentra separada por comas (,)

## [directorio donde está almacenado el dataset] /input/practica2

# Entorno de trabajo.

Una vez conocido el dataset, el siguiente paso es conocer el entorno de trabajo donde ejecutaremos las consultas descritas en la siguiente sección. Este entorno de trabajo es un cluster donde se ha instalado Hadoop. Esta distribución ha sido explicada en clases y es de todos conocido sus componentes y funcionalidades.

Al ecosistema Hadoop podemos acceder mediante su dirección IP:

158.49.96.124

El usuario de acceso a los diferentes nodos es **hadoop** el pass se dirá en clases.

Los pasos para la ejecución de un proyecto Map-Reduce han sido explicados en clase junto con la documentación correspondiente al Tema 5 de la asignatura.

# Procedimiento.

Se pide realizar una o varias aplicaciones Java utilizando el paradigma Map-Reduce, que una vez ejecutadas en el cluster de Hadoop nos permita contestar a las siguientes preguntas:

* 1. ¿Cuánto se ganó en 2019?
  2. ¿Cuál fue el mejor mes para las ventas? ¿Cuánto se ganó ese mes?
  3. ¿Qué ciudad tuvo el mayor número de ventas?
  4. ¿A qué hora debemos mostrar publicidad para maximizar la probabilidad de que el cliente compre el producto?
  5. ¿Qué producto vendió más? ¿Por qué crees que vendió más?

# Detalles generales de la aplicación.

La aplicación que se pide desarrollar debe ser implementada utilizando el lenguaje de programación JAVA bajo el paradigma Map-Reduce. El dataset a utilizar se encuentra alojado en el cluster de Hadoop del CPD del CUM, en el directorio de HDFS (Sistema de ficheros distribuido) llamado /input/practica2. Todos los alumnos harán uso de este dataset, sin necesidad de subir más datos extras al cluster.

Las salidas de las diferentes ejecuciones o de la ejecución general debe alojarse en un directorio específico de HDFS. Este directorio debe denominarse utilizando la siguiente nomenclatura:

/output\_nnnaaa\_2023: donde

nnn: tres letras del nombre del alumno. aaa: tres letras del apellido del alumno.

Se debe tener en cuenta que un conjunto de alumnos trabajará sobre el mismo directorio de trabajo en los diferentes nodos, en concreto sobre el directorio

/home/hadop por lo que los ficheros .jar que se coloquen deben ser con nomenclatura única para no machacar el de los compañeros. Se puede seguir la misma nomenclatura que la utilizada para el directorio de salida.

# Entrega de prácticas.

La práctica debe ir perfectamente documentada, utilizando para ello la documentación interna de JAVA y como salida el Javadoc generado.

La fecha de entrega de la práctica será el día **21 de diciembre de 2023** a través del campus virtual de la Universidad de Extremadura.

Se propondrá una fecha de consulta de prácticas por el profesor, para que cada alumno defienda la realización de la práctica en las horas de clase. Si el alumno no defiende la práctica, pero realiza la entrega, el profesor podrá proponer una prueba práctica de la misma, tal y como se indica en la ficha 12a de la asignatura.