ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

FACULTAD DE INGENIERÍA EN ELECTRICIDAD Y COMPUTACIÓN
SISTEMAS DE BASES DE DATOS AVANZADOS
2020 – II PAO

PROYECTO DEL PRIMER PARCIAL

Fecha de Entrega: 7 diciembre 2020
Fecha de Presentación: 8 de diciembre 2020

DATASET

Para el proyecto tendrán que elegir un dataset (conjunto de datos), el que más les guste o les parezca interesante. El dataset elegido debe de contener al menos 10.000 (diez mil) registros. Existen varias páginas que publican datasets de diferentes áreas de conocimiento (ej. medicina, biología, economía). Pueden buscar en los siguientes sitios:

http://archive.ics.uci.edu/ml/index.php
https://sci2s.ugr.es/keel/datasets.php
https://www.kdnuggets.com/datasets/index.html
https://www.reddit.com/r/datasets/
https://www.kaggle.com

DESCRIPCION

El objetivo de este proyecto es de resolver alguna tarea utilizando un dataset seleccionado, guardar la información en una base de datos no relacional y aplicar map-reduce. Primero, deberán de guardar los datos en una base de datos no relacional. Como segundo paso, deberán definir una tarea a realizar, por ejemplo, calcular el total de ventas por producto, o el promedio de ventas por día de una empresa, etc). Luego, implementarán esta tarea en Python utilizando Hadoop. Deberán de registrar los tiempos de cómputo en Hadoop, tanto en la ejecución de la tarea completa, como la ejecución del código map-reduce.

ENTREGABLES

Responder la tarea (un estudiante por cada grupo) subiendo un .zip que contenga:

- El archivo .csv con el dataset
- Código de implementación de la base de datos no relacional (p.ej. estructura de documentos)
- Código de implementación de inserción de datos.
- Código de implementación mostrando al menos dos modificaciones.
- Código de implementación mostrando al menos dos eliminaciones.
- Código de implementación de al menos dos consultas.
- Código de implementación del map-reduce
- Documentación (PDF) con las siguientes secciones:
  1. Integrantes
  2. Título
  3. Descripción del dataset
  4. Descripción del código (con resultados) para operaciones CRUD.

5. Descripción de la tarea a realizar con map-reduce.
6. Descripción de la implementación realizada
7. Descripción de los resultados obtenidos.

AMBIENTE DE DESARROLLO

Para este proyecto se recomienda que tengan instalado en sus computadoras una máquina virtual en donde tengan instaladas Python Anaconda, mrjob, Hadoop, y MongoDB.

CALIFICACION

Los siguientes criterios de evaluación se aplicarán a este trabajo:

- 20% El dataset está bien descrito
- 10 % Se presenta evidencia de las operaciones CRUD.
- 10% La tarea map-reduce está definida claramente
- 20% El código es entendible y está documentado
- 20% El resultado de la implementación es correcto
- 20% Los tiempos de cómputo están bien calculados.