UNIVERSIDAD EAFIT

INGENIERÍA DE SISTEMAS

ST0263 TÓPICOS ESPECIALES EN TELEMÁTICA, 2022-1

UNIDAD 3 BIG DATA

DOCUMENTACIÓN y EVIDENCIAS DE LAB 5

ESTE DOCUMENTO DEBE SER ENTREGADO POR BUZON DE ENTREGAS DE LA MATERIA

Además entregar por buzón la URL del repo github donde tiene todo de los LABs 5 y 6

NOMBRE DEL ALUMNO: Santiago Ignacio Vanegas Pérez

EMAIL: svaneg11@eafit.edu.co

GITHUB DEL ALUMNO PARA LOS LABS: <https://github.com/svaneg11/svaneg11_ST0263/tree/main/U3-Lab1>

Realizar las actividades relacionadas con el lab5 en: <https://github.com/st0263eafit/st0263-2261/tree/main/bigdata>

|  |
| --- |
| LAB5-EVIDENCIAS DE INSTALACIÓN DE EMR EN AWS  párrafos descriptivos, Screenshots, códigos fuente, extractos de código, scripts, urls, etc.   * **Creando el cluster** |
| * **Agregar los puertos en EMR**      * **Luego se modifica el security group del nodo master del EMR para agregar los puertos.** * **Creacion del bucket S3 para los notebooks**      * **Crear el usuario ‘hadoop’ y entrar a Hue.**      * **Entrar a Jupyter y crear un notebook de PySpark.**      * **Entrar a Zeppelin y crear un notebook de spark** |
| LAB5-EVIDENCIAS DE COPIADO DE ARCHIVOS EN HDFS EN EL CLUSTER EMR  párrafos descriptivos, Screenshots, códigos fuente, extractos de código, scripts, urls, etc.   * **Entrar a la maquina EC2 del nodo master de EMR.**      * **Descargar los datasets del github de la materia, y moverlos al directorio del usuario en hdfs.**      * **Los archivos ahora se ven desde Hue.**      * **Se borra la carpeta de datasets para demostrar la subida de los datasets desde la interfaz de Hue.** * **En Hue, desde el directorio /user/hadoop se arrastra la carpeta de datasets para subirla.** |
|  |
| LAB5-EVIDENCIAS DE COPIADO DE ARCHIVOS EN AWS S3  párrafos descriptivos, Screenshots, códigos fuente, extractos de código, scripts, urls, etc.  OJO: Adjuntar la URL del bucket público donde tendrá todos los datos de entrada y salida de LAB5 y LAB6   * **URL del bucket publico** <https://svaneg11datalake.s3.amazonaws.com/> * **Archivos subidos en la zona raw a través de la consola de Hue.**      * **Borrar los archivos del bucket para demostrar la subida a través de la terminal (AWS CLI)**      * **Entrar a el nodo master del EMR desde EC2 usando ssh, descargar el dataset de githunb y subirlo a S3 con AWS CLI** |
|  |
| LAB5-EVIDENCIAS DEL DESARROLLO DE LOS PROGRAMAS EN MRJOB  párrafos descriptivos, Screenshots, códigos fuente, extractos de código, scripts, urls, etc. |
| * **datampleados.txt**  1. **El salario promedio por Sector Económico.**      1. **El salario promedio por Empleado.**      1. **Número de SE por Empleado que ha tenido a lo largo de la estadística**      * **dataempresas.txt**  1. **Por acción, día-menor-valor, día-mayor-valor.**      1. **Listado de acciones que siempre han subido o se mantienen estables.** |
|  |
|  |
| LAB5-evidencias de algo no considerado en las anteriores evidencias y requeridas en el LAB  párrafos descriptivos, Screenshots, códigos fuente, extractos de código, scripts, urls, etc.   * **Agregar las políticas para poder hacer los archivos del bucket S3 publicos y que puedan ser accedidos via URL.** |
|  |