

*Ejercicio de Laboratorio 4: Databricks*

Instituto Politécnico Nacional.

Escuela Superior de Cómputo.

Licenciatura en ciencia de datos.

Bases de Datos Avanzadas

Emiliano López Méndez.

**Introduccion**

En este eje

**Desarrollo de la Actividad:**

**Parte 1 Carga de archivo**

Como se le podría dar persistencia al servidor?

**Parte 2 Consultas con spark.sql**

Como se le podría dar persistencia al servidor?

**Parte 3 Utilizando el modo sql realizar 4 consultas y tabla temporal**

Como se le podría dar persistencia al servidor?

**Parte 4 Almacenar resultados en archivo parquet**

Como se le podría dar persistencia al servidor?

**Conclusión**

En este ejercicio de laboratorio, se utiliza Apache Hive para crear, poblar y consultar tablas a través del cliente Beeline, ejecutándose en un entorno Docker. Hive es una herramienta clave en el ecosistema de Big Data que permite gestionar grandes volúmenes de datos estructurados y semiestructurados, permitiendo realizar consultas SQL sobre estos datos almacenados en HDFS o en sistemas de archivos locales. A través de este ejercicio, se exploran las funcionalidades básicas de Hive, incluyendo la creación de tablas, la inserción de datos y la realización de consultas relacionales.