UNIVERSIDAD MAYOR DE SAN SIMÓN
FACULTAD DE CIENCIAS Y TECNOLOGÍA
DIRECCIÓN DE POSGRADO

DIPLOMADO ESTADÍSTICA APLICADA A LA TOMA DE DECISIONES

SEGUNDA VERSIÓN

LABORATORIO DATABRICKS

POSTULANTE : HEBERT JUAN DE DIOS DELGADILLO FERNANDEZ
DOCENTE : DANNY LUIS HUANCA SEVILLA

Cochabamba – Bolivia
2023

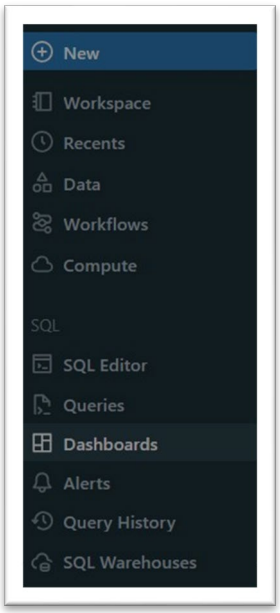
Laboratorio Databricks

1. ¿Qué es Databricks?

Databricks es una plataforma de análisis de datos que esta en la nube que se emplea para el proceso, análisis y transformación de datos, facilitando el diseño de modelos de aprendizaje automático complejos.

Fue desarrollada por el equipo que creó Apache Spark, que es un sistema de procesamiento de datos en memoria, y está diseñada para simplificar la implementación y gestión de entornos de análisis de datos en la nube.

¿Cuál el objetivo de los siguientes menus?



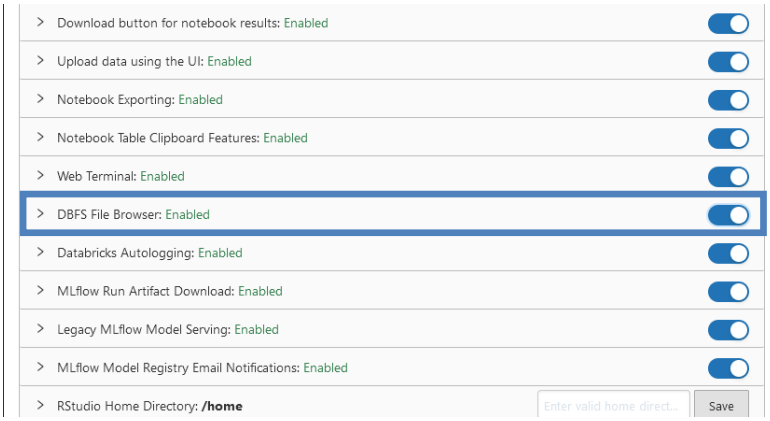
Workspace: Es nuestro entorno de trabajo, donde podemos ver los archivos que tenemos tanto los de Databricks como los de otros repositorios como GitHub, es principalmente un entorno de desarrollo colaborativo. Nos proporciona una interfaz basada en la web donde los usuarios pueden crear, editar y colaborar en cuadernos (notebooks) interactivos que contienen código y comentarios del código.

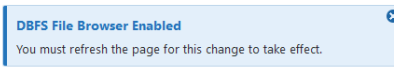
Data: La sección "Data" en Databricks se centra en la administración y acceso a los recursos de datos. Aquí podemos realizar la exploración, limpieza y transformación de datos. Pueden utilizar lenguajes como SQL para consultar y manipular datos de manera eficiente.

Compute: Es nuestro sistema de computo que usaremos en todo Databricks, aquí podemos brindarle muchas características a nuestro poder de procesamiento, pero mientras más características pongamos, mayor será el costo, el limite es nuestra billetera.

SQL Editor: Es como un DBeaver integrado en Databricks, nos permite ejecutar comandos SQL desde la interfaz de Databricks, ya sea en tablas gestionadas dentro de Databricks o en datos externos que se pueden acceder desde la plataforma.

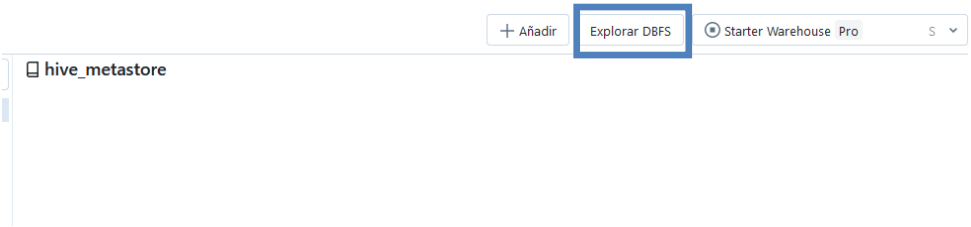
2. Habilitar el Navegador de DBFS (Browse DBFS). Para tal efecto ir al menú “Admin Settings”



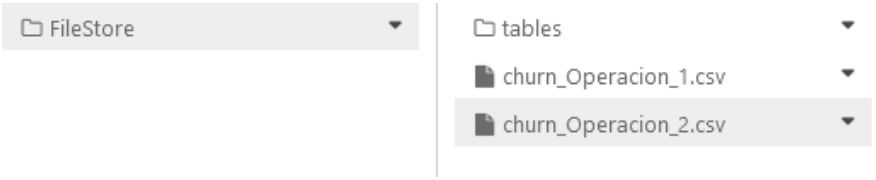


Con esta acción se habilita el browser del DBFS

Regresando al menú Data, se puede observar que se habilitó la opción “Browse DBFS”



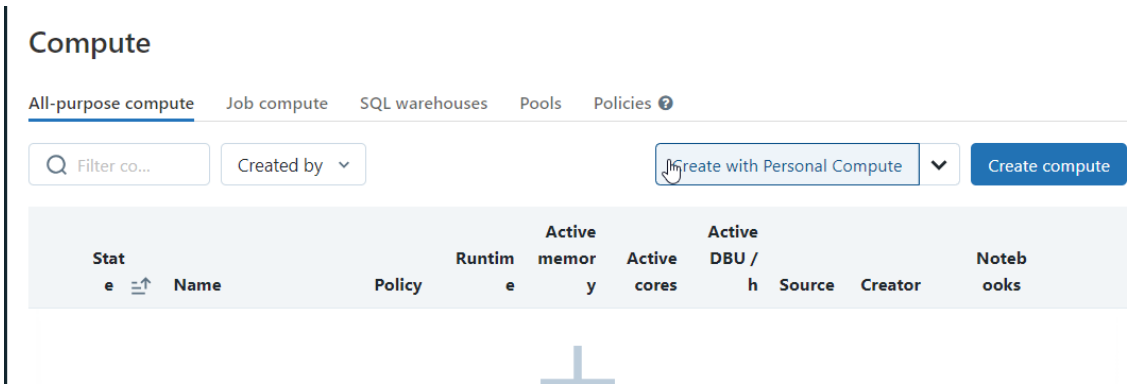
3. Ir al menú Data y suba los archivos csv del caso TELCO.



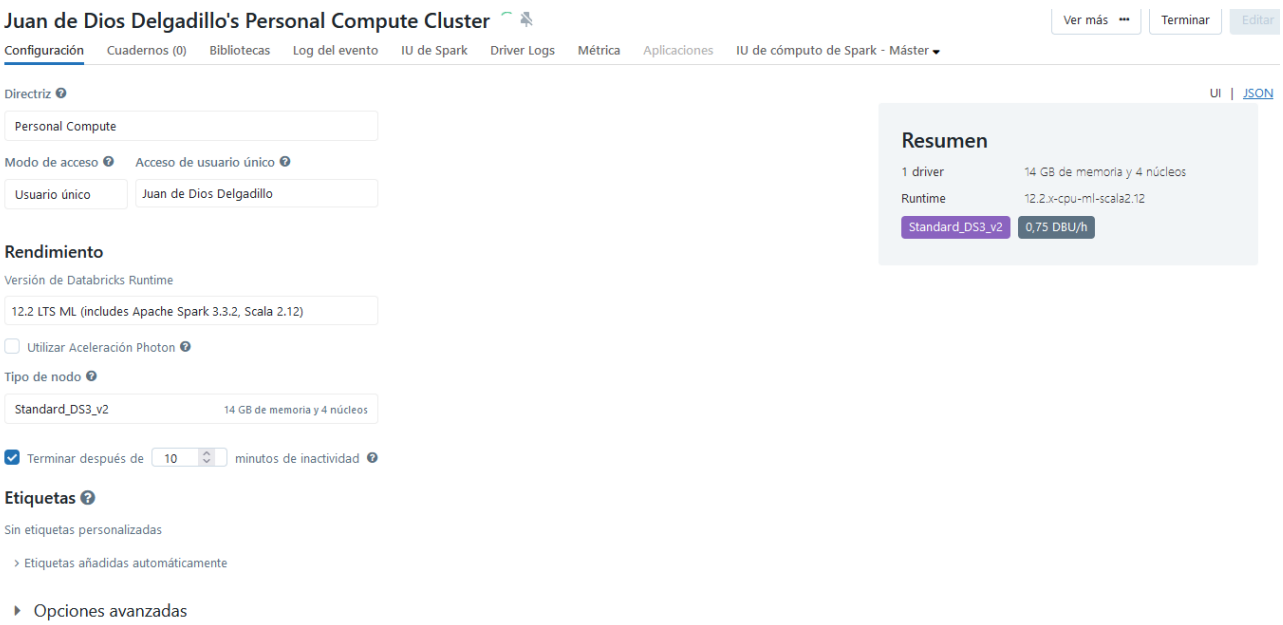
Reflexión: Hicimos esta operación sin necesidad de crear un clúster de cómputo, ya que el clúster esta en la nube, además que fue bastante sencillo en comparación con Hive que es con línea de comandos, esta es mas interactiva y no se necesita conocimientos básicos.

4. Crear un clúster de cómputo

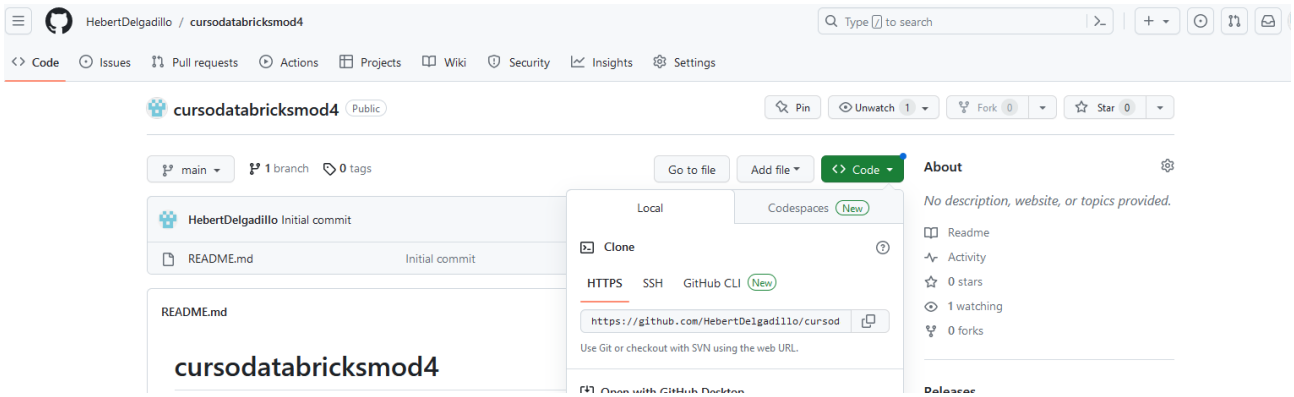
Hacer click en create compute



Crear el cluster con las opciones indicadas en las gráficas siguientes:

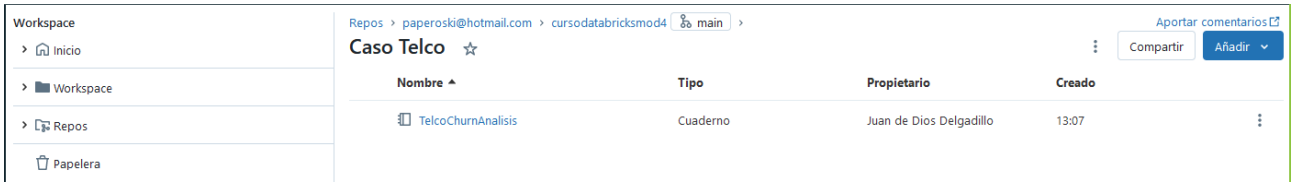


5. Crear una cuenta en GitHub



6. Importe el archivo dbc proporcionado TelcoChurnAnalisis.dbc

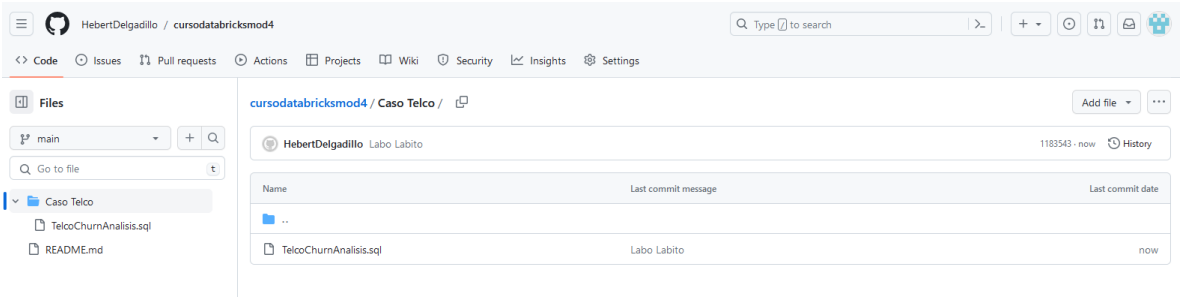
Para que se encuentre versionado crear una carpeta denominada “Casotelco” en el repositorio.



7. Ejecute los scripts que se encuentran en el notebook

8. Realice un commit and Push del código creado en la carpeta para tener actualizado el repositorio.

9. Verificar que este en GitHub



Conclusiones

Aprendi a usar el sistema de procesamiento de datos de DataBricks, cree una cuenta de prueba para ver como funciona, aprendi a crear un cómputo y ahora se que podemos obtener una capacidad de computo mayor solo configurándolo, claro que con un costo mayor, aprendi a usar el Workspace de Databricks, a subir archivos al mismo y a vincular Databricks con una cuenta de GitHub.