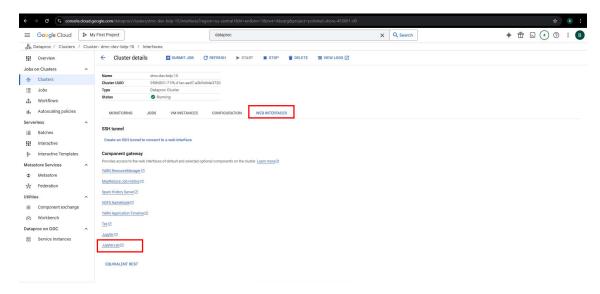
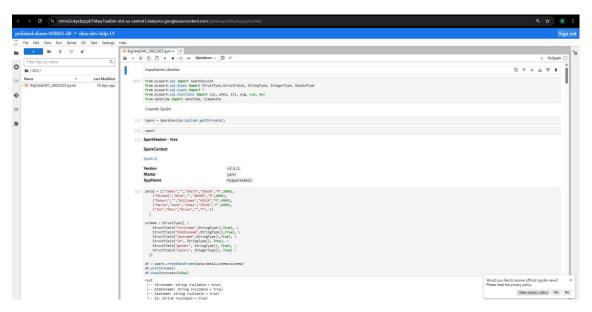
Laboratorio Desarrollo en Notebook con Apache Spark



Jupyter Notebooks sobre Apache Spark en Google Cloud Platform

Importación de Módulos

Crear Sesión de Spark



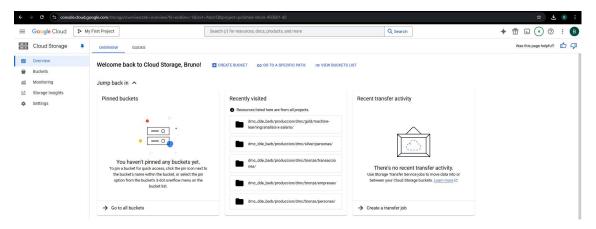
El Spark Console se crea la sesión por defecto.



Dataframes con Schema



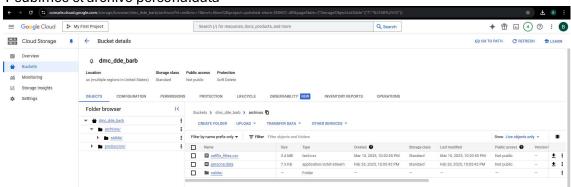
Creación de bucket para Cloud Storage en Google Cloud Plataform

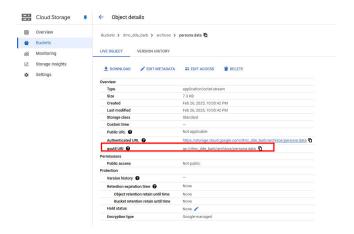


Creamos un bucket dmc_dde_barb

Una carpeta archivos dentro del bucket dmc_dde_barb/archivos

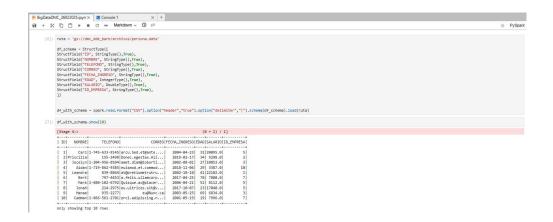
Y subimos el archivo persona.data



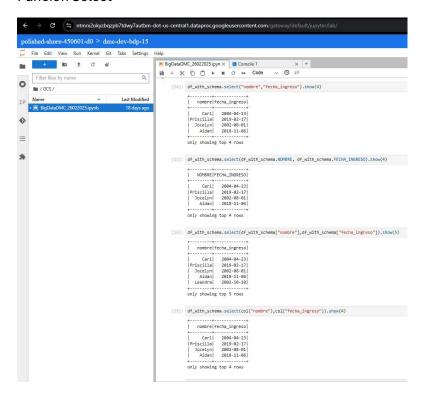


printSchema() para ver el esquema del dataframe.

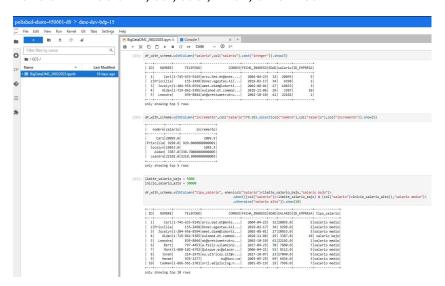
Almacenar en un dataframe la lectura de archivos externos con spark.read.format Función show para mostrar datos de un dataframe Tipos StrucType y StructField para definir esquemas.



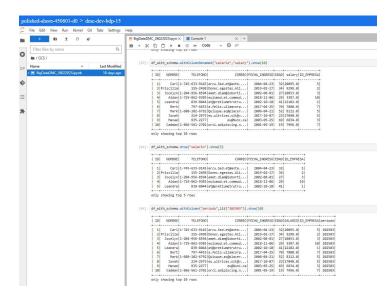
Función Select



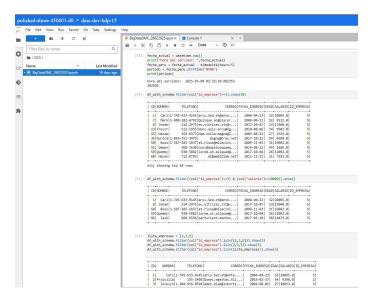
Función withColumn, col, cast, when, otherwise



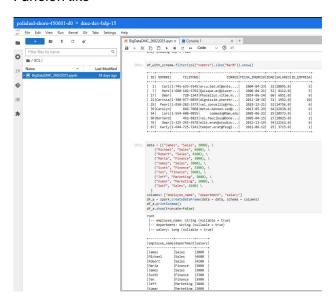
Función with Column Renamed, drop, lit



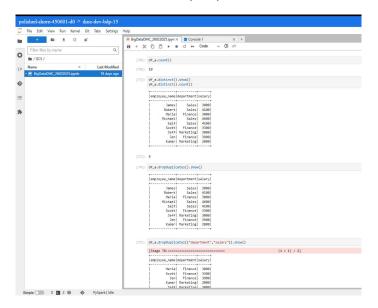
Función Filter e isin



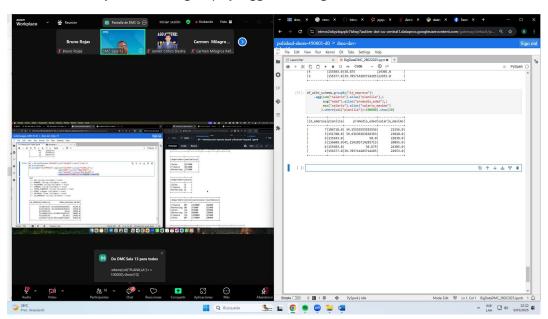
Función like



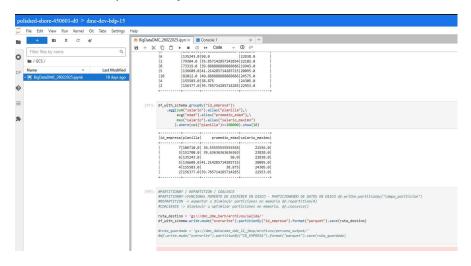
Función count, distinct, dropDuplicates



Función orderBy, asc, desc, groupBy, agg, sum, avg, max

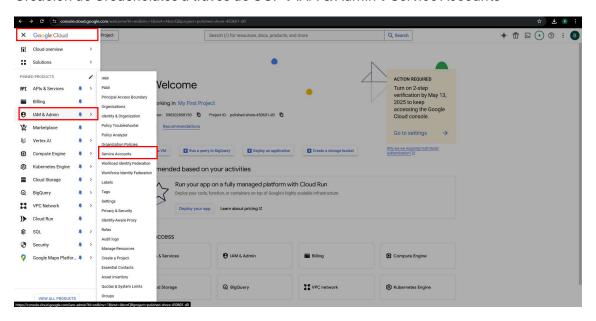


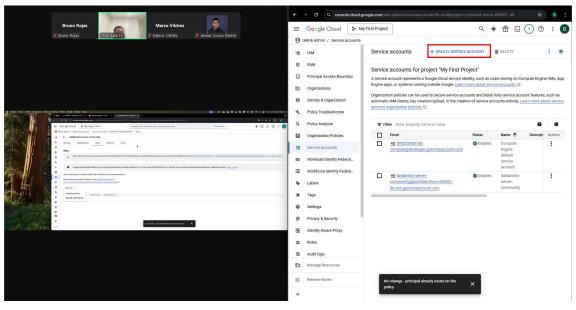
Función where, partitionby, write.mode, format, save

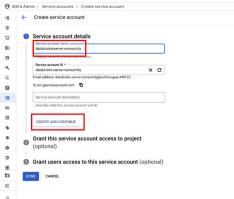


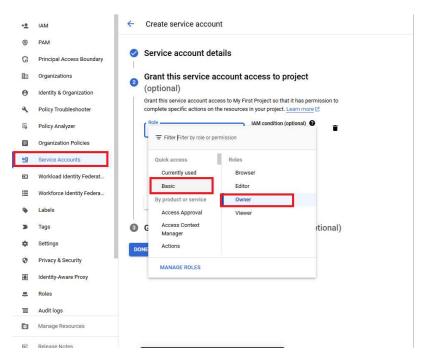
Apache Spark en Databrick

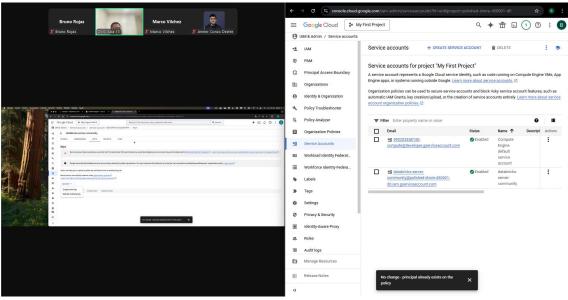
Creación de Credenciales a través de GCP→IAM & Admin→Service Accounts

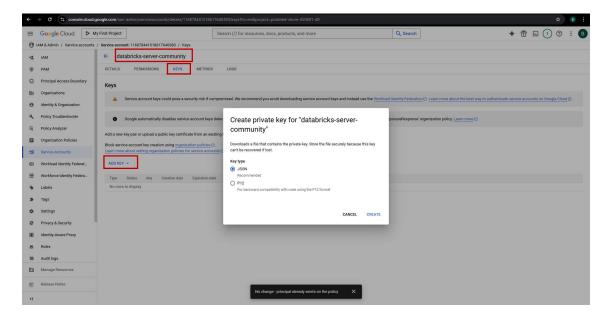




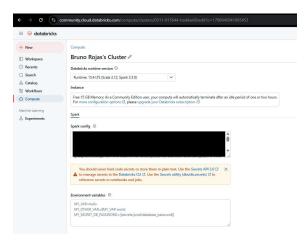




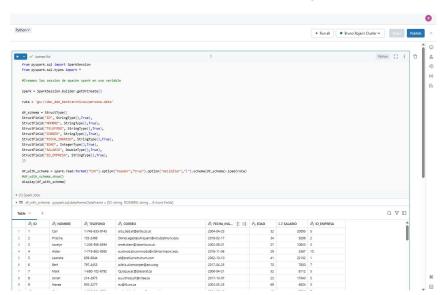




Las credenciales creadas las seteamos en la configuración del cluster en databricks



Conectamos databricks con nuestro bucket de gcp



Subimos un nuevo archivo a nuestro bucket que hemos descargado de kaggle y lo leemos desde databricks.

