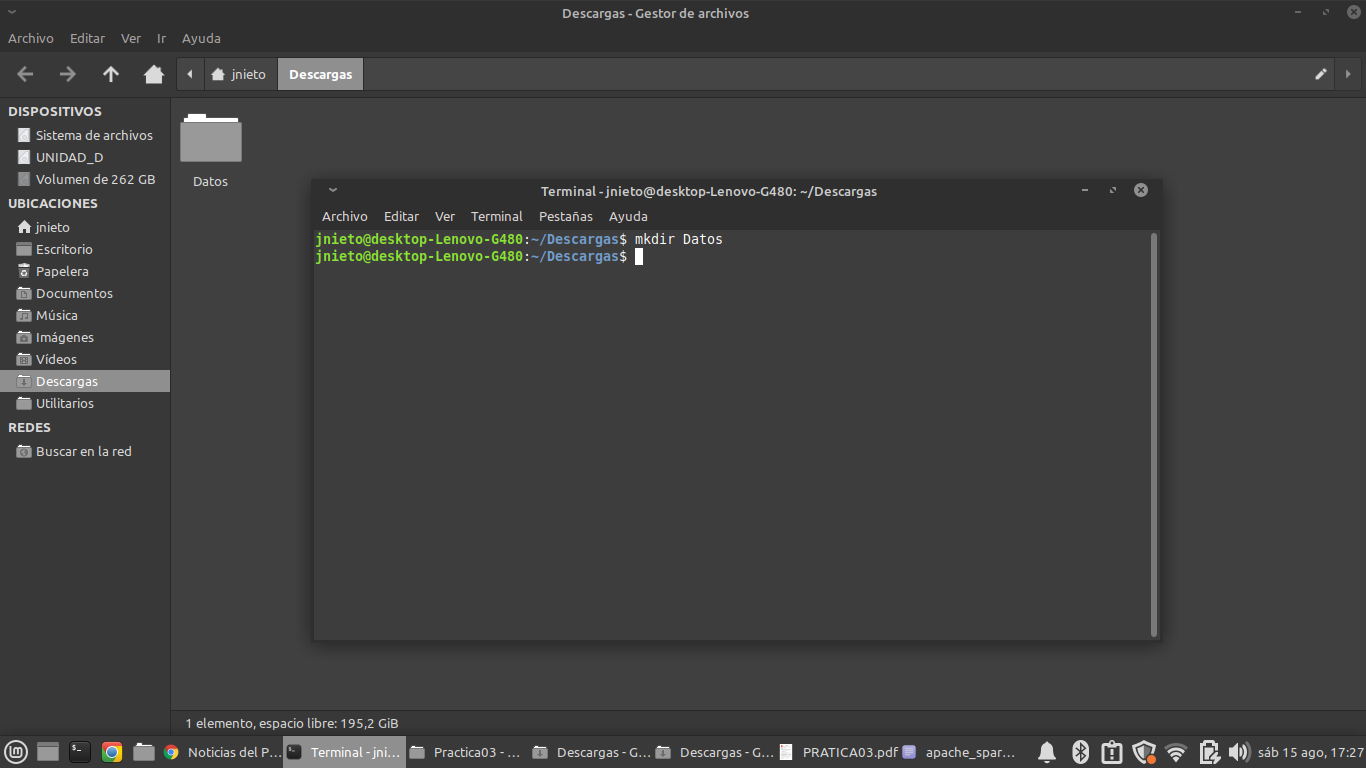
Python Developer – Práctica 03

Enlace a dataset: https://www.stats.govt.nz/assets/Uploads/Births-and-deaths/Births-and-deaths-Year-ended-December-2019/Download-data/births-and-deaths-year-ended-december-2019.zip

1) Crear una carpeta en Descargas llamada datos.



2) Obtener la cabecera de los datos

Después de inicar las instancias master y slave de spark, habilitar la shell de spark para python con:

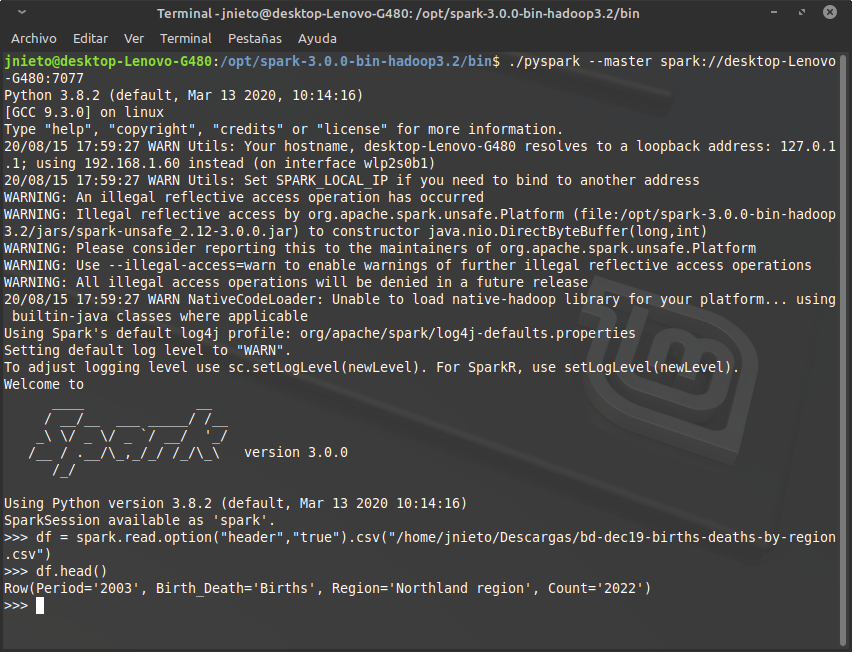
$ ./pyspark --master spark://<URL>

Iniciada la shell, ingresar

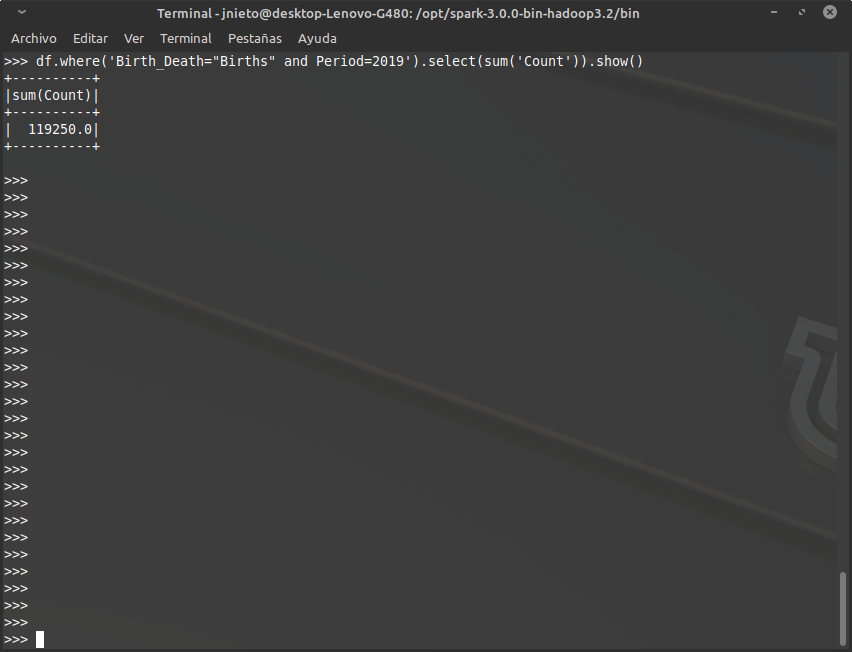
df = spark.read.option("header","true").csv("/home/jnieto/Descargas/bd-dec19-births-deaths-by-region.csv")

Para mostrar datos de la cabecera

df.head()

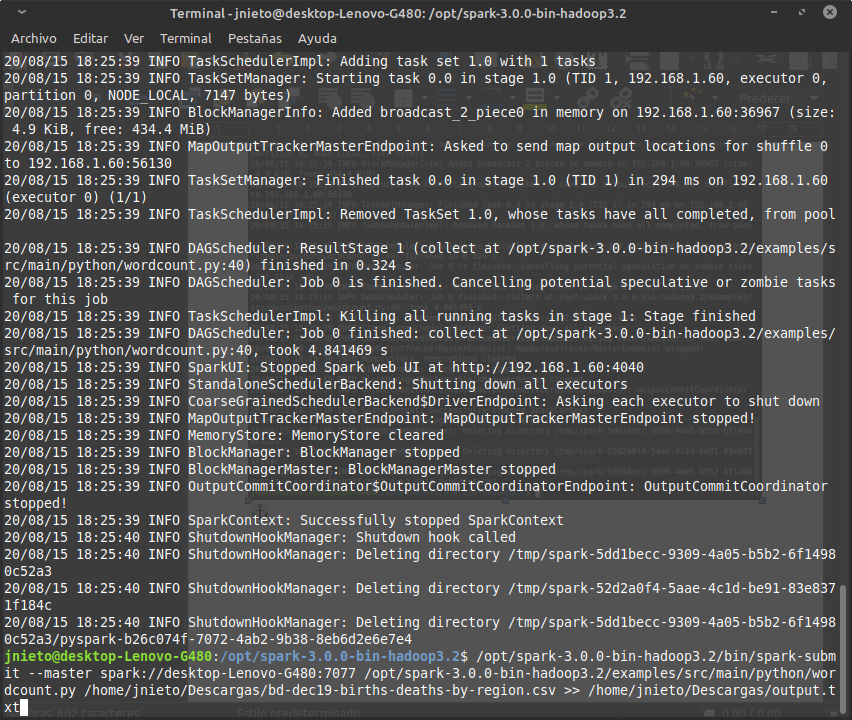


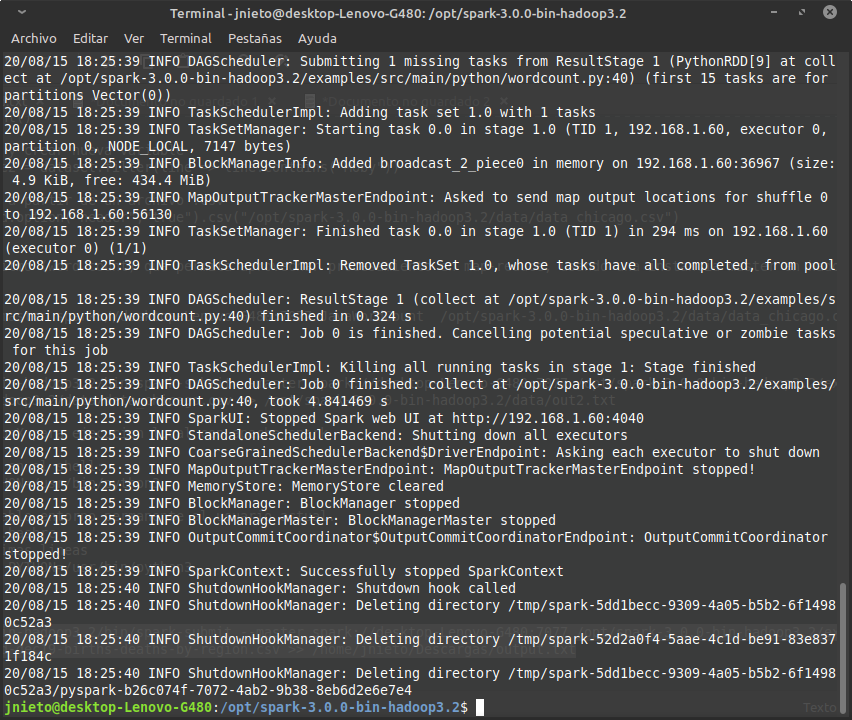
c. Usar SQL con el dataframe.

Sumar cantidad total de nacimientos en el año 2019, para todas las regiones.

d. Realizar el ejercicio de conteo de palabras empleando JavaWordCount.

El comando ingresado (esta ocasión se utiliza python-wordcount):



Del cual se obtiene una salida en la linea de comandos

y el correspondiente archivo .dump

