Jose Pablo Castillo Rodas

**Prueba de Conocimiento**

**Herramientas:**

Enlace al GITHUB de esta prueba: https://github.com/jpcastillo2391/Bam\_test

Descripción de Base de datos local:

Relacional postgresql-x64-14 - PostgreSQL Server 14 Local

* Se uso esta DB Relacional dado a que es open source y gratuita.
* Dado que los datos entregados, son estructurados esta DB esta instancia es mas que suficiente.
* Ya se tenia previamente instalada y ejecutando, por lo que no conllevo a mayor tiempo/costo de instalación.

Anaconda – Spyder

Se uso la versión anaconda v4 , Python 3.8.5

* Su ejecución es fácil, al igual que VS Code.
* Visualización de variables en monitor
* Gratuita (la versión community), la versión para DSC on cloud tiene costo.
* Tiene varias librerías de DS listas para ejecución.
* Previamente lo use cuando fui profesor de Métodos Numéricos.

Parte I

Inciso 1

El programa “ETL\_cvs\_to\_postgres.py” se ejecuta y carga las tablas en la base de datos. Dado a que al principio se entendió que el inciso 1 tenia que ser en Python, esta se encuentra en el .py.

Las Demas soluciones serán entregadas en .sql

Se hizo uso para solventar de un grupo by y condiciones de suma en ciertos casos de status.

Inciso 2

Hice uso de pivotes para lograr establecer la tabla de tal manera que la ordinalidad del producto y cliente se colocase siempre en la columna que correspondía. Dado a que utilice el Nombre para establecer el mayor comprador en el caso del cliente, este no tenia ni nombre ni apellido en su mayoría, por lo que aparecen nulls.

Inciso 3

Este es una diferencia de agrupamientos, sin embargo los agrupamientos son diferentes por lo que se opto por un join de dos selects de agrupamientos distintos.

Inciso 4

Este estaba intrigante, ya que dada la restricción de cursores y otras funciones se analizo la alternativa de joins aninados por un desfase. La agrupación y sumatoria es para establecer quien cumple con la continuidad de código en estos 3 desfases, implicando continuidad y poder llevar el conteo de sucesiones no repetidas.

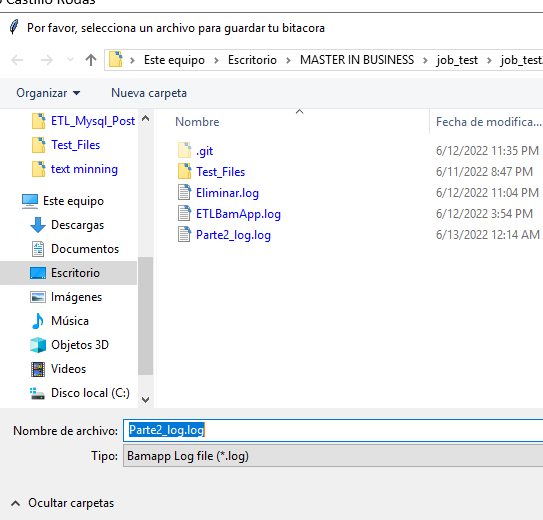
Inciso 5

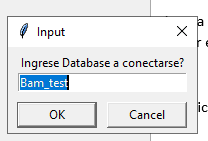
Es el cumpleaños de mi novia, voy a tomar un descanso para ir a atender a las visitas.

Parte II

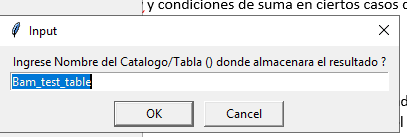
Ya había implementado varias de los requisitos desde que hice el ETL. Sin embargo, esta como el 2\_ParteII.py

Favor, instalar las librerías tkinter y pandastable para poder visualizar la tabla previo a su almacenamiento

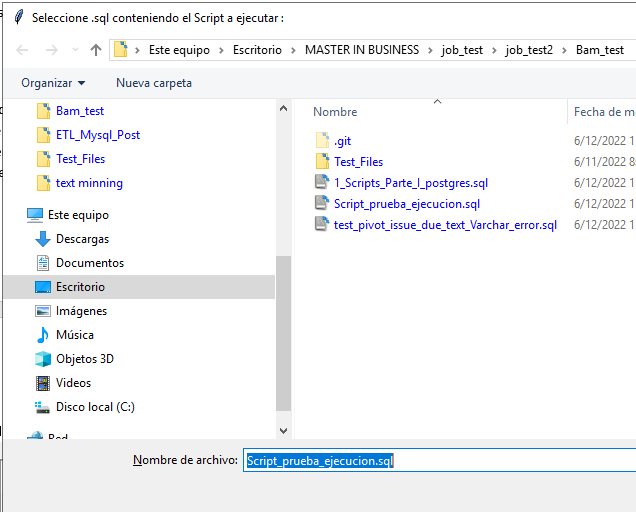
1. Lugar y nombre de almacenamiento de log
2. Nombre de database a conectarse



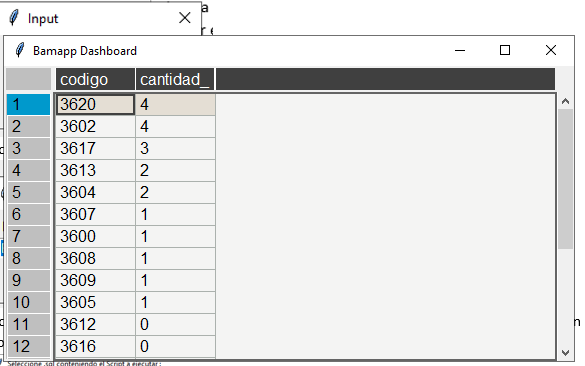
1. Nombre de la tabla donde se almacenara el resultado



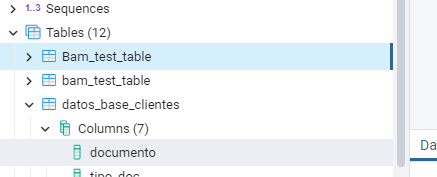
1. Nombre del script a ejecutar. Para efectos de este ejercicio, ya deje en el archivo un .sql con el nombre default para correrlo.



1. Ver el resultado



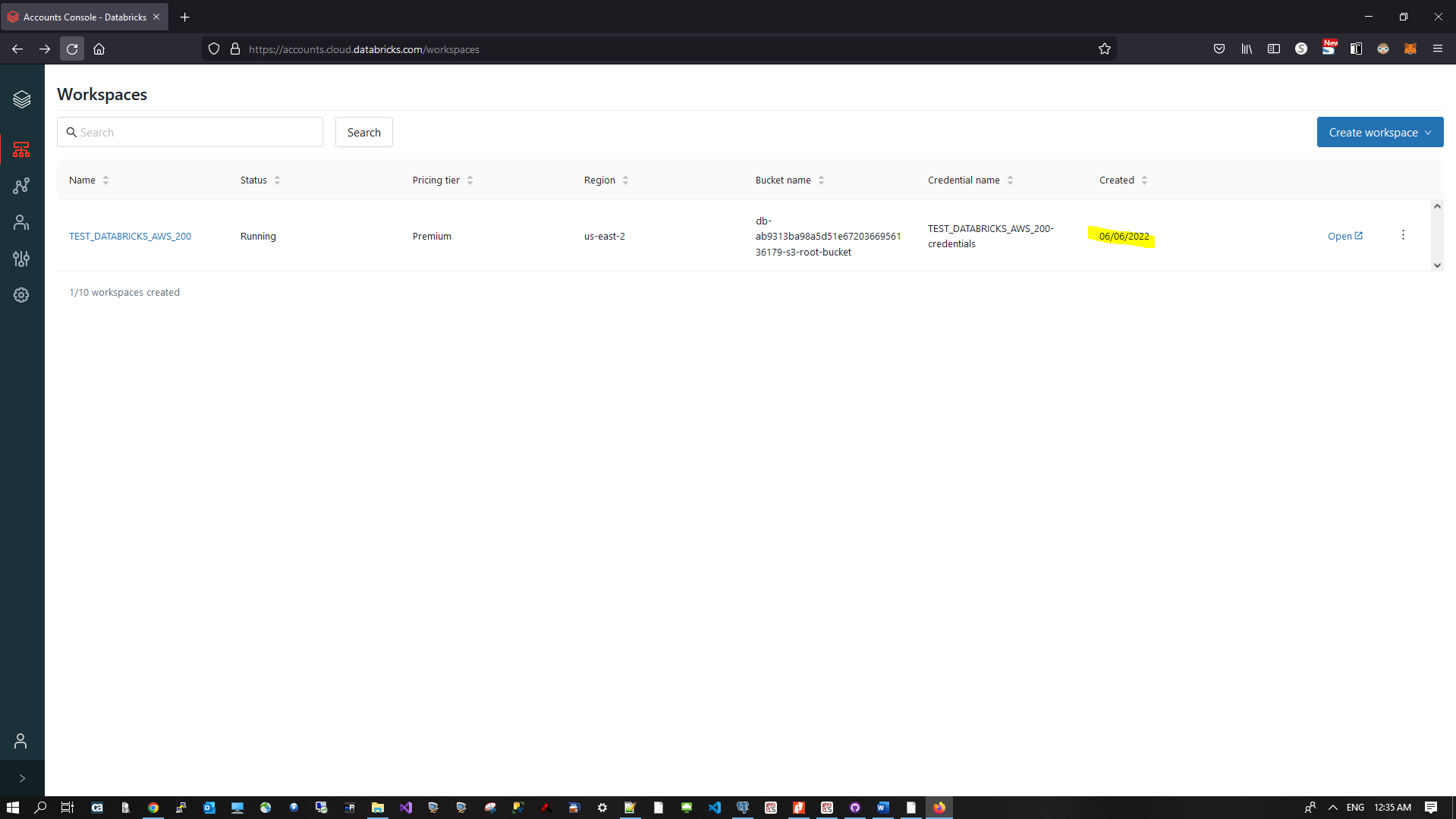
1. Verificar en base de datos



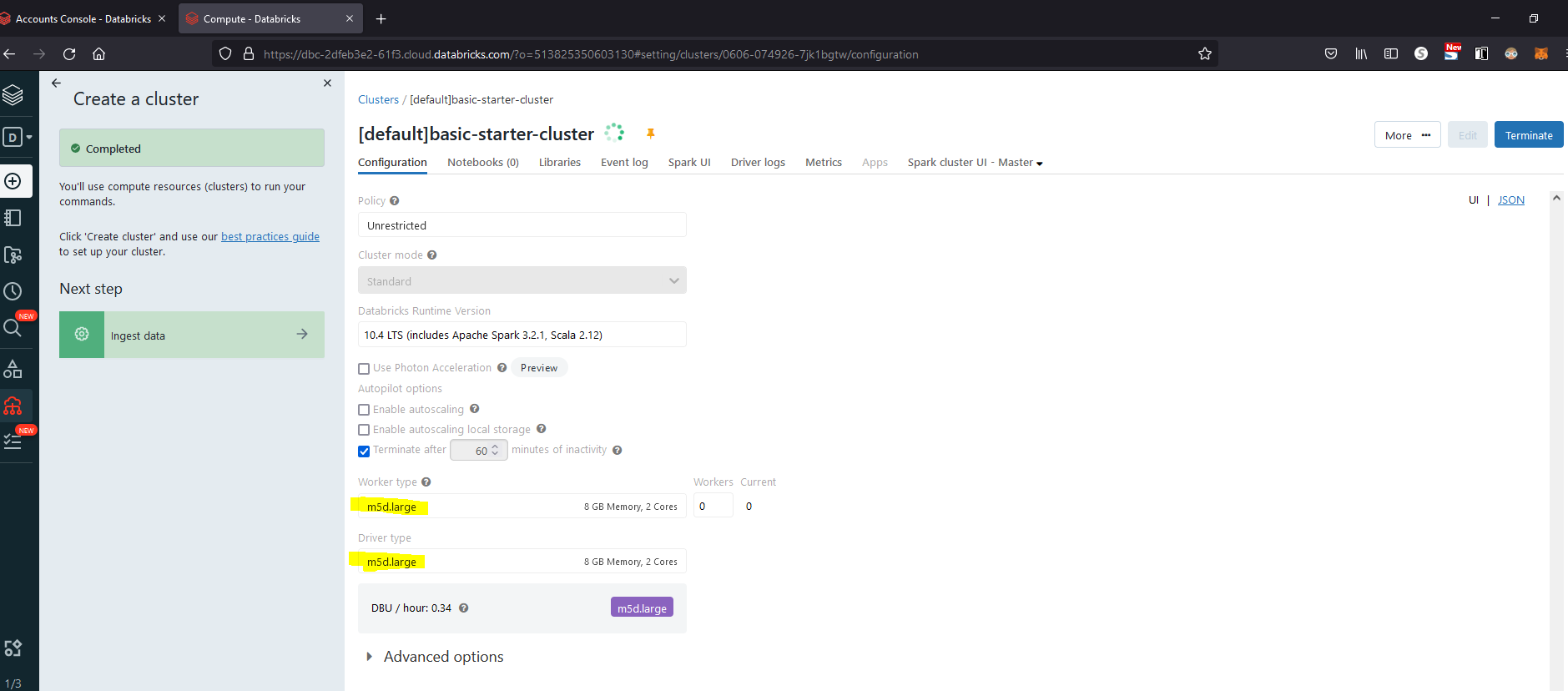
Parte III

Ya me acabe mi free trial de azure, como para levantar ambari y spark en una vm.

En aws tengo actualmente una instancia de trial de databricks la cual configure con cloudformation. Voy a ver si me puedo conectar a su spark clusterizado en ec2 spot.



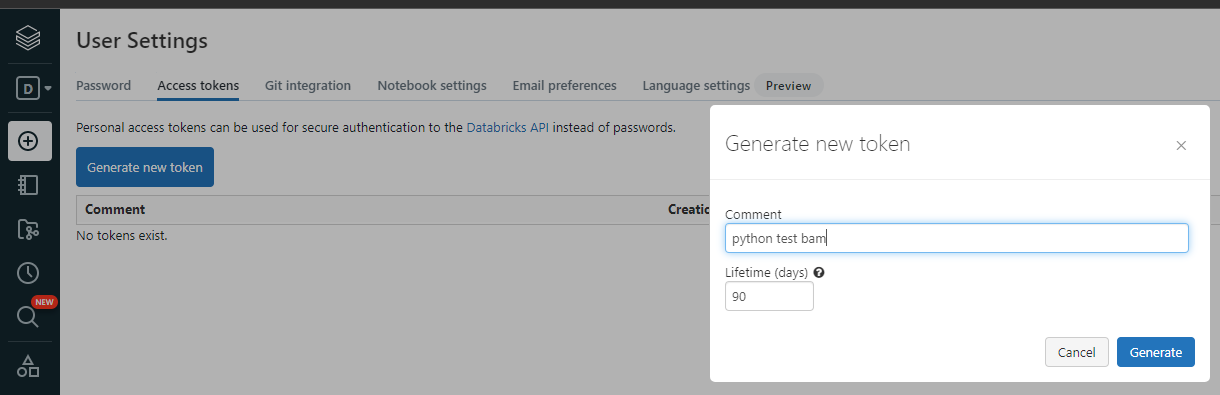
Ya por experiencia se que el spot de ec2 que levanta databricks como cluster como md5n.large es de los mas baratos. $0.02 la hora.



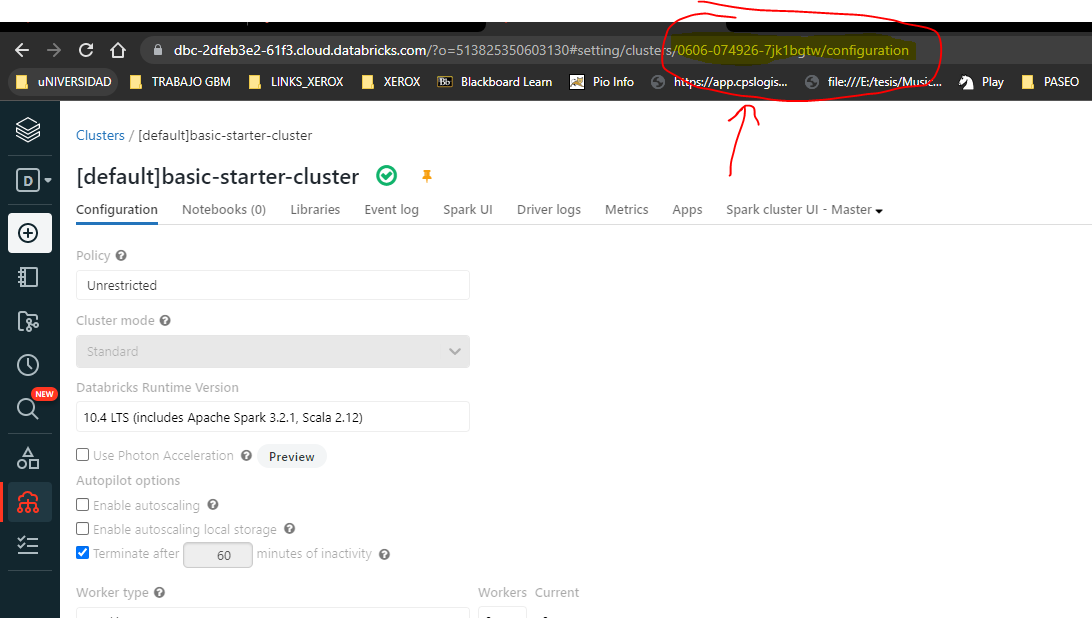
Encontré este ejemplo de como conectarme a mi databricks desde mi ide. Yo uso spyder, sin embargo creo que ejecutare esto desde visual code.

Link referencia: <https://docs.databricks.com/dev-tools/databricks-connect.html#language-python>

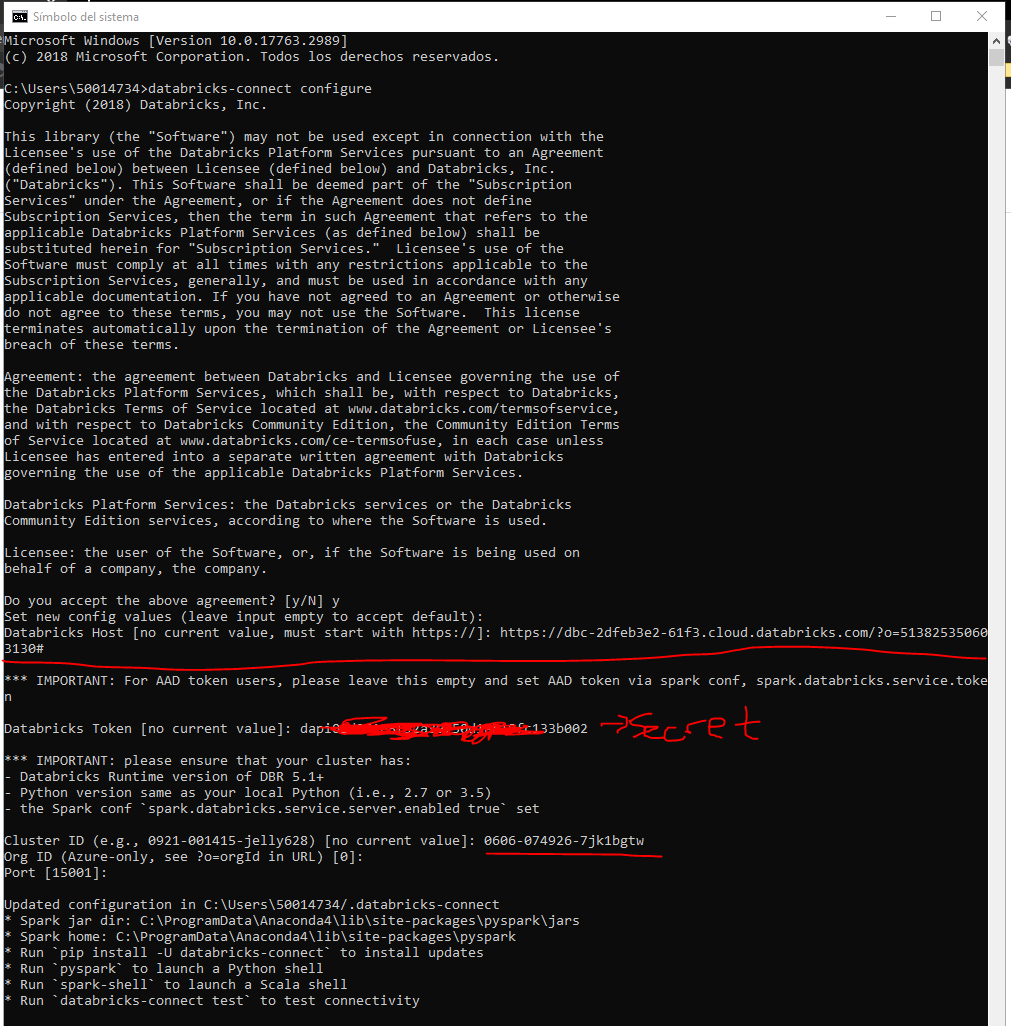
Creando mi secret token: https://docs.databricks.com/administration-guide/access-control/tokens.html



Este es el ID de mi cluster. Es temporal, asi que solo dura 1h



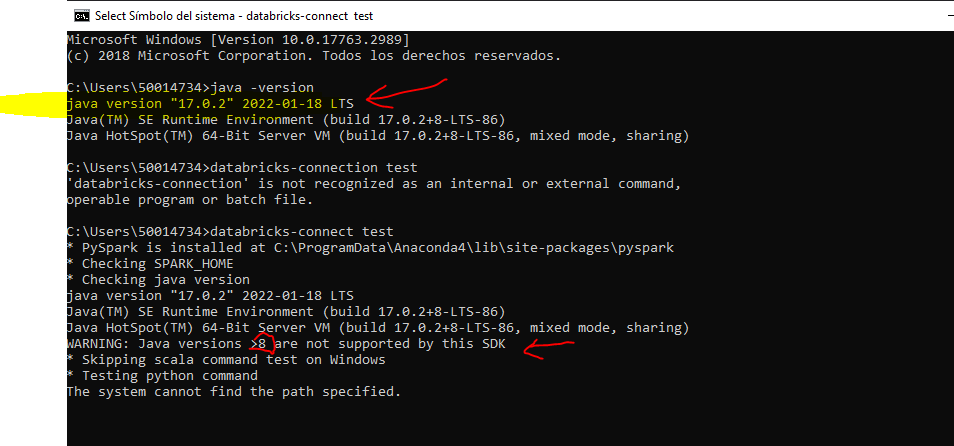
Deployando conexión



Para ejecutar el test: databricks-connect test

En caso de fallo, como este: Java gateway process exited before sending its port number , verificar versión de jala, y que este configurado JAVA\_HOME

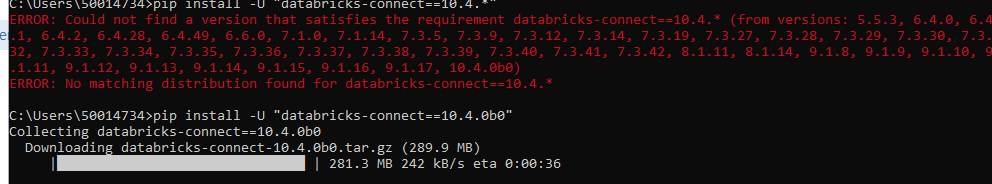
Abrir cuenta y descargar desde https://www.oracle.com/java/technologies/downloads/#java8-windows



Tomar encuenta la versoin de java. Para la versión de data brciks instalada, no acepta java mayore a la versión 8. Actualmente tengo la 17.

Reinstalando databricks connection

Pip install -U “databricks-connect==10.4.0b0” (es versión beta aparentemente.



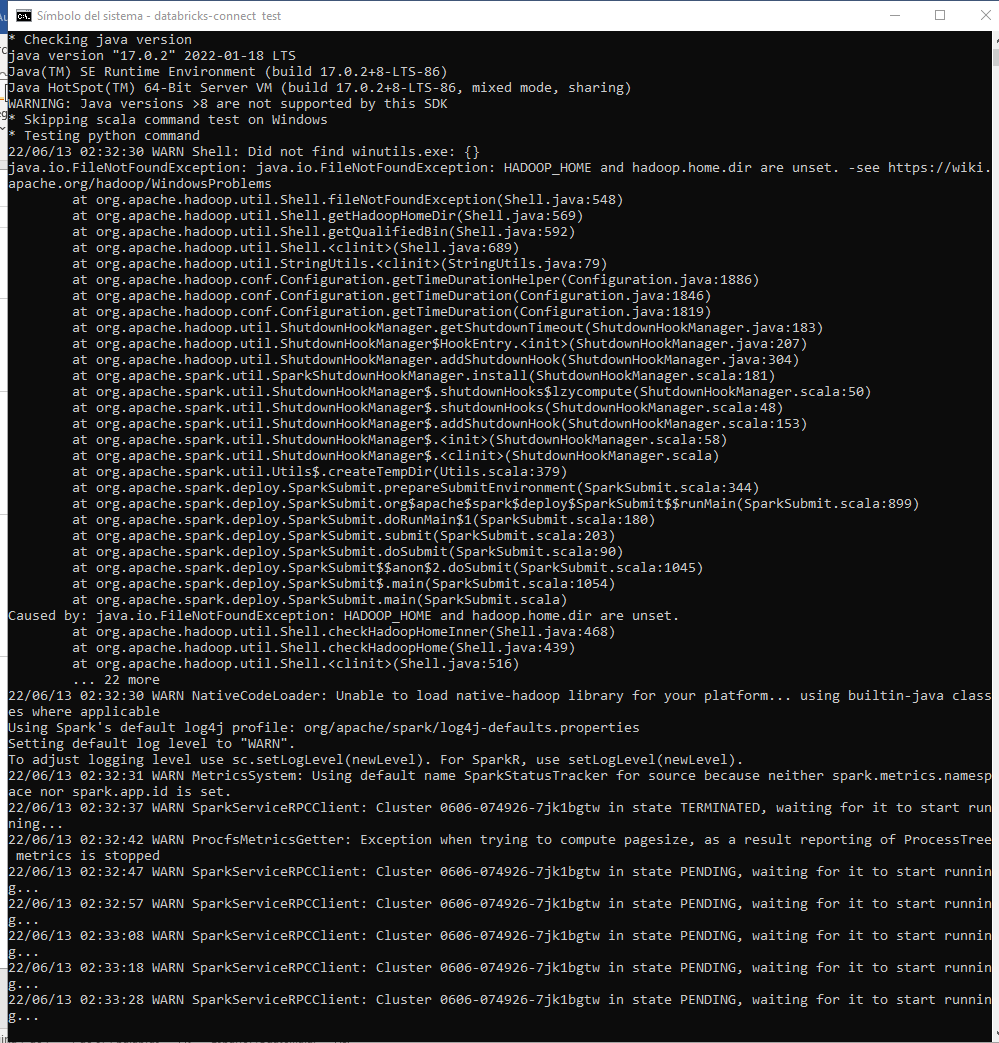
Funciono, ahora el error es.

Missing Python executable 'python3', defaulting to 'C:\ProgramData\Anaconda4\Lib\site-packages\pyspark\bin\..' for SPARK\_HOME environment variable. Please install Python or specify the correct Python executable in PYSPARK\_DRIVER\_PYTHON or PYSPARK\_PYTHON environment variable to detect SPARK\_HOME safely.

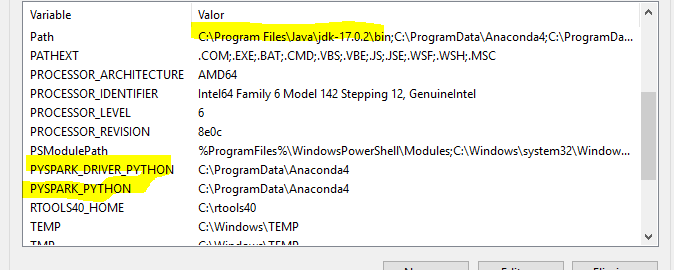
The system cannot find the path specified.

Se logro solventar la conneccion con Databricks y sparl en ec2 spot luego de ver esta solución en stackoverflow: https://stackoverflow.com/questions/70183411/databricks-connect-test-hangs-indefinitely-on-the-system-cannot-find-the-path-s/72599870#72599870

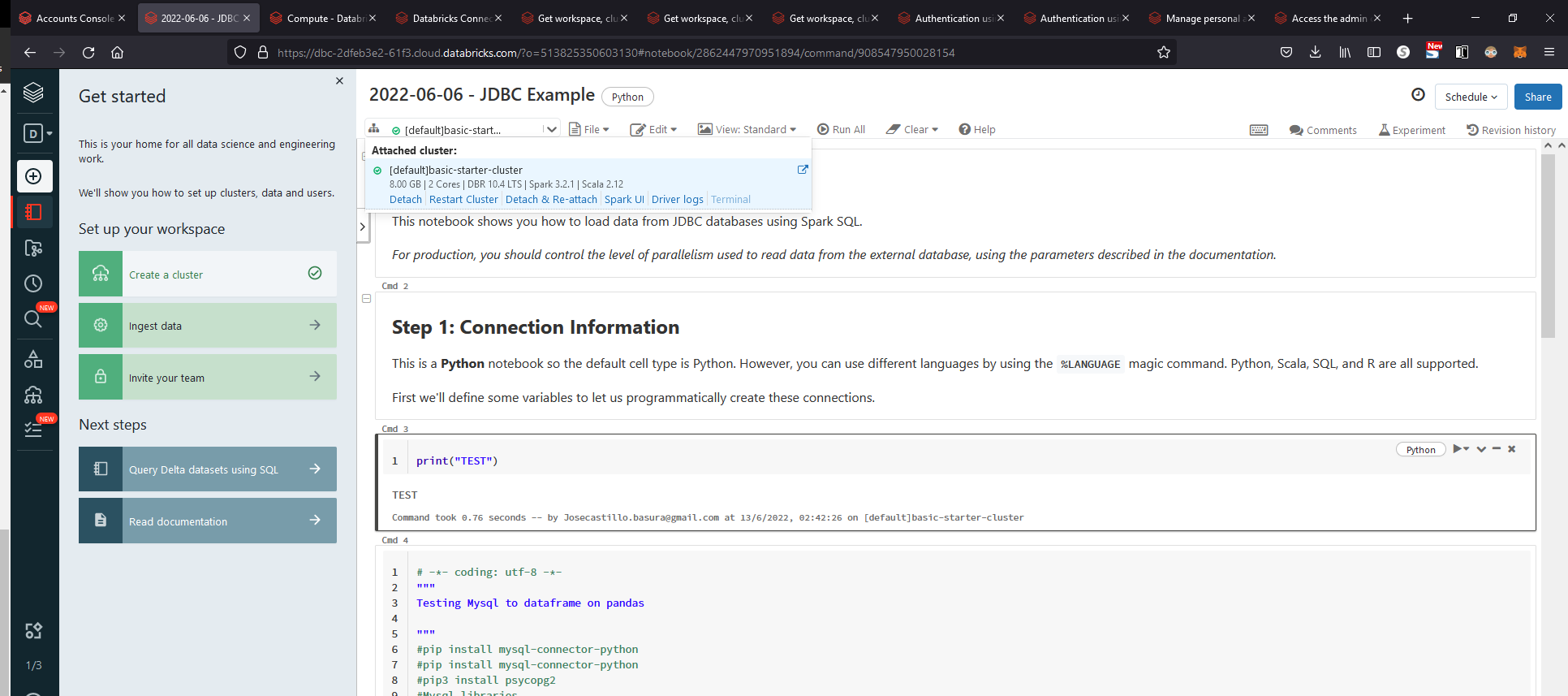
Ya sirviendo el databricks connect test



Ya Esta corriendo la instancia del cluster

Cambios importantes que fueron útiles: 

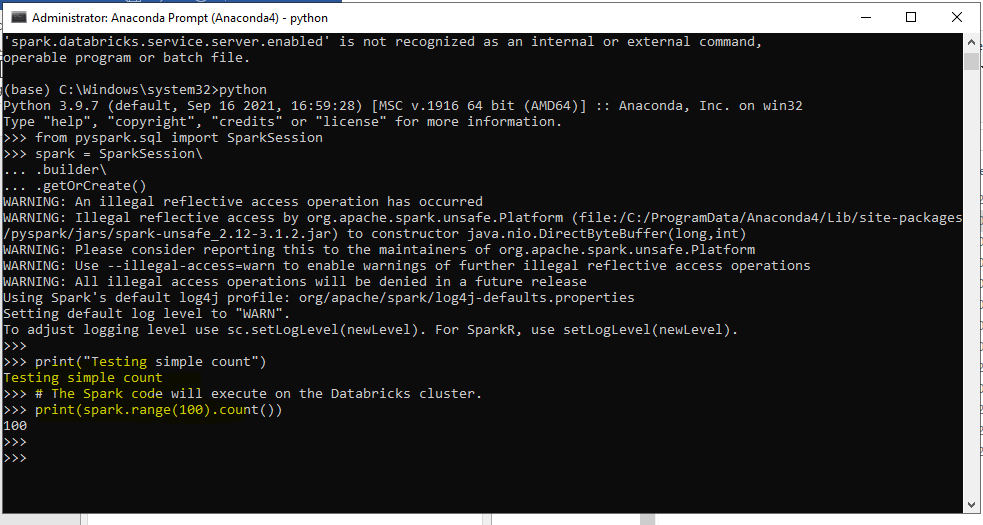
Verificacion en databricks.



Cometí el error de resetiar mi PC, y ya no me esta conectando.

Estoy siguiendo esta instalación de librerías: <https://sparkbyexamples.com/pyspark/install-pyspark-in-anaconda-jupyter-notebook/> . En algún momento mi ambiente se desconfiguro.

Corriendo desde conda, versión 4, directamente en Python si corre.



Quedo pendiendo levantar la instancia de hadoop

Muchas Gracias por la oportunidad BAM.

Referencias

<https://docs.databricks.com/dev-tools/databricks-connect.html#language-python>

<https://docs.databricks.com/workspace/workspace-details.html#workspace-url>

<https://docs.databricks.com/workspace/workspace-details.html#workspace-url>

<https://stackoverflow.com/questions/70183411/databricks-connect-test-hangs-indefinitely-on-the-system-cannot-find-the-path-s/72599870#72599870>

<https://runawayhorse001.github.io/LearningApacheSpark/pyspark.pdf>

<https://www.pythonfixing.com/2021/11/fixed-null-entry-in-command-string.html>

<https://stackoverflow.com/questions/31841509/pyspark-exception-java-gateway-process-exited-before-sending-the-driver-its-po>

<https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz>

<https://brain-mentors.com/hadoopinstallation/>