**Ejercicio 3**

1. ¿Qué formas de hacer scheduling de una tarea en linux conoce?

Sinceramente nunca trabajé con scheduling en Linux pero se que estas actividades se hacen con CRON y CRONTAB. Por lo que sé de Airflow, que hace poco empece a aprenderlo, la sintaxis de CRON se puede utilizar también allí para scheduling. También por eso se dice que Airflow es CRON con esteroides :D. Desde Databricks una vez me tocó configurar un simple job y escribirlo con la sintaxis de CRON ya que con la interfaz de usuario no era posible configurar el scheduling que se necesitaba.

1. ¿Cómo y con qué comandos guardaría la mayor cantidad de detalle sobre las salidas de un script python que desea ejecutar de forma diaria a las 6AM?

Otra vez, nunca me tocó hacerlo, pero usaría la librería logging de Python que permite escribir en archivos los logs de la ejecución en general, entre ellos las salidas del proceso.

Luego de retornar una salida, podría con la librería logging con logging.info() dejaría resguardadas las salidas generadas por el proceso. Logging permite escribir en un archivo de log en donde también quedan registrados los timestamps.

1. ¿Qué comando o serie de comandos utilizaría para subir todos los contenidos de un directorio a un bucket de S3?

No tengo experiencia con s3 pero seguramente desde la CLI de aws instalada en la máquina donde se encuentra el directorio que se quiere copiar al bucket existe un comando para copiar estos archivos al bucket de s3.

Según investigué:

aws s3 cp myDir s3://mybucket/ --recursive

Con ese comando se podría copiar un directorio local a un bucket de s3 con la cli de aws.

Otra opción es hacerlo desde un orquestador como Airflow o Azure Data Factory. Desde Data Factory tengo más experiencia y es un entorno más codeless. Allí se configura la conexión al bucket de s3 y la conexión a un server, filesystem, en donde se encuentre el directorio. Luego con una actividad de Copy Data se puede copiar un directorio completo de un lugar a otro.

Con Airflow seguro existe un operador del provider de AWS de S3 que haga esto también.

d- Si una instancia de Redshift utilizada para reporting se está quedando sin espacio y se impone la necesidad de sacar algunos datos antiguos de la base, pero a pesar de que los datos de más de seis meses de antigüedad no se utilicen para reporting, se los requiere para entrenar y validar modelos predictivos, además de hacer algunos análisis ad-hoc en SQL **a un precio razonable considerando tanto infraestructura como costos de consultas** ¿Que tipo de solución propondría para poder consultar los datos usando servicios cloud en AWS? ***Intentar ser lo más descriptivo posible.***

Dependerá de la arquitectura, pero si se utiliza un datalake en s3 como instancia previa al datawarehouse, se podría construir la FACT\_TABLE con sólo la info de 6 meses anteriores y eso enviar a RedShift para que sea consumido por herramientas de reporting. La información anterior se puede almacenar en alguna fact histórica sin necesidad de enviarlo a Redshift. Si esa info se queda en el datalake, el almacenamiento en s3 es mucho más barato que en Redshift. Con herramientas como Databricks, se pueden entrenar y validar modelos predictivos y hacer análisis ad-hoc con SQL con los datos en s3.

En ese caso el datalake estaría en s3 y el procesamiento de los datos sobre el datalake se haría con databricks, servicio que se puede desplegar en AWS.

También (por lo menos en Azure Data Lake Storage o Azure Object Storage está la opción) se podría inclusive abaratar más los costos al elegir alguna opción de almacenamiento frío en datos muy viejos. En azure existe la opción de almacenamiento caliente (es más caro el almacenamiento pero más barato el acceso a ellos), almacenamiento frío (menos caro el almacenamiento y más caro el acceso a ellos) y archive (todavía más barato almacenarlos pero hay más costo de acceso). Es decir, si hay datos que sean más viejos que se los usaría ocasionalmente para algún entrenamiento o análisis ad-hoc pero no todos los días, sería más óptimo almacenarlos en algún bucket con características de almacenamiento frío o archive (suponiendo que en s3 exista lo mismo o algo parecido)