Suponga que se presenta la necesidad de ingestar archivos planos de distintos aplicativos origen, realizarles pequeñas transformaciones (mayoritariamente de formato) y guardarlo en formato Apache Parquet en HDFS.

Dada esta necesidad se pide generar una solución utilizando Pyspark o Spark Scala que contemple las siguientes características:

1.- Debe tener la flexibilidad suficiente para procesar archivos de texto ancho fijo y delimitados.

2.- Como input debe tomar un archivo de configuración que contenga los siguientes parámetros:

a.- El path de origen del archivo a ser ingestado;

b.- El path destino donde se escribirá ese archivo;

c.- La metadata necesaria para procesar el archivo crudo (e.g. delimitador, header, columnas, anchos de columnas);

d.- La metadata necesaria para generar el archivo en parquet que crea necesarias (e.g. particiones, cantidad de archivos, compresión);

e.- Permitir el agregado de columnas al data frame con valores predefinidos.

Rta: Como lo comente en la entrevista mi experiencia se basa más en la suite de Microsoft, no obstante dejo debajo algunos códigos de este punto que he trabajado apuntando a un datalake ya que no sé como conectarlo al equipo local.

**Pregunta 1**

Como resolvería este tipo de petición? Explique detalladamente el proceso de limpieza y transformación del modelo inicial. Que tecnologías utilizaría y por que?

Rta: Propongo confeccionar un DataLake con tecnología Azure Cloud con los siguientes recursos:

Blob Storage con sus etapas bien marcadas, etapa bronce, silver y oro. Realizar una ingesta masiva en el raw de la etapa Silver en formato parquet. Las carpetas a crear serían: Event, Segment, Session y User, que contendrá todo el contenido de las dimensionales. Como la tabla principal es grande la segmentaria mediante databricks para aperturarlas mediante Spark que esta pensado para trabajar con grandes volúmenes de registros.

Una vez creadas las tres etapas en el blob storage (sin realizar modificaciones y transformaciones del dato en las dos primeras etapas) realizaría una copia espejo de la etapa Oro alojándola en Databricks.

Es decir que en el DataLake tendríamos el transaccional y luego colocaría un DW de SQL modelando el dato para disponibilizarlo en vistas y consumir la información ya procesada directamente del DW. En caso que se desee visualizar la información confeccionar los dashboard con PowerBI.

Todo este proceso estaría orquestado mediante DataFactory.

**Ejercicio 1**

Realice el DER que de soporte al modelo dimensional solicitado por la banca privada.

Realizado

**Ejercicio 2**

Escriba las queries necesarias partiendo de la tabla inicial y que de como resultado el modelo planteado en el ejercicio anterior.

Realizado

**Ejercicio 3**

Escriba la consulta necesaria para obtener el KPI de retención de clientes para los 10 clientes que mas veces se hayan logueado en el último mes.

Realizado

**Pregunta 2**

Suponga que la ingesta de estos datos se realiza utilizando Apache Spark debido a que la tabla cruda tiene billones de registros. Que parametros de spark tendría en cuenta a la hora de realizar dicha ingesta? Explique brevemente en que consta cada uno de ellos. En que formato de compresión escribiría los resultados? Por que?

Sin responder.

**Pregunta 3**

Existen varios problemas en cuanto a la calidad de datos de la tablas que consultan los usuarios de la banca privada, se esta investigando como mejorar y prevenir estos incidentes. Describa brevemente que implementaría para garantizar la confiabilidad de los datos.

Se podría implementar triggers que validen la información antes de insertar esos datos. Dentro de esa validación podemos colocar la lógica validando en exactitud lo que querramos que ingrese a la tabla. También podemos insertar en una tabla stage y luego aplicarle la calidad del dato a esa tabla, una vez que el dato se encuentra Ok se lo inserta a la tabla fisíca.

**Bonus Track!!!**

* En qué requerimiento implementarías una cola de mensajes en una solución orientada a datos? Que lenguaje utilizarías y porque?

Rta: No he trabajado con cola de mensajes, si estuve trabajando en recibir desde una Api los datos que dejaba el flujograma de un BOT para Banco Galicia, donde generaba un gran volumen. Lo que realice fue que inserte en una tabla stage y validar que cada uno de esos registros se encontraban en la tabla física y luego eliminarla de la stage. De esta manera garantizaba que no existan datos duplicados en la tabla física (además de su PK) y no afectar la performance ya que sobre ella corrían proceso de modelado de datos cada quince minutos.

* Que experiencia posees sobre py spark o spark scala? Contar breves experiencias, en caso de contar experiencia con otro framework de procesamiento distribuido, dar detalles también.

Rta: Mi gran experiencia la tengo mediante la suite completa de Microsoft, principalmente con SQL (Sql base de datos, Sql datwarehouse, Sql Synapse Analitycs) y servicio Cloud. No obstante, en el 2021 estuve asignados a dos grandes proyectos (Nestle e Ypf) que utilizaban Py Spark mediante Databricks. Con respecto a Spark Scala es baja la experiencia que tengo.

* Qué funcionalidad podrías expandir desde el area de ingeniería de datos con una API y arquitectónicamente como lo modelarías?

Rta: Es baja mi experiencia mediante Api’s, únicamente las he trabajado mediante LoggicApps y Azure Functions, dos recursos del servicio Cloud de Microsoft. No obstante, me encuentro a total disposición para capacitarme en las herramientas que necesite la empresa para desarrollar mi posición.