Preguntas

1. ¿Qué diferencia existe entre una arquitectura Data warehouse y un Data lakehouse? ¿Qué tipo de arquitectura le recomendaría a WWI para complementar lo desarrollado en este laboratorio incluyendo fuentes de datos no estructuradas, análisis en tiempo real que incluyan simultáneamente tanto datos estructurados como no estructurados?

Rta: La diferencia es que el Data Warehouse es un sistema centralizado para almacenar datos estructurados, para ello, los datos pasan por una fase de limpieza y tranformación antes de ser procesados. Mientras que, el Lakehouse combina las ventajas del Warehouse con las ventajas de los Lakes, pues puede analizar e integrar tanto datos estructurados como no estructurados, para poder realizar analítica en tiempo real.

En el caso de WWI, le recomendaríamos usar un Data Lakehouse, para poder incluir datos estructurados y no estructurados. Además, será de gran utilidad para realizar análisis en tiempo real, como lo permite BigQuery.

2. ¿Qué ventajas y desventajas observa al momento de implementar un ETL utilizando este tipo de herramientas respecto a desarrollarlo utilizando Python, Pandas y demás herramientas vistas durante la primera parte del curso?

Rta: Las ventajas del uso de ETL son la escalabilidad, la facilidad de integración, la automatización de procesos y la sencillez de su interfaz gráficas. Mientras que, algunas desventajas son que, en algunos escenarios, puede generar un costo mayor y que es rígido en la medida en que se limita a las opciones disponibles de la interfaz.

Por otra parte, usar Python, Pandas u otras, permiten tener flexibilidad, un costo bajo. Sin embargo, se requieren mayores conocimientos técnicos y estar realizando constante mantenimiento para validar el funcionamiento de todo.

3. Investigue cómo se puede conectar a través de Tableau o Power BI a su Big Query, donde están las tablas finales creadas y con datos. Documente los hallazgos e intente conectar los datos creados del laboratorio con una de estas herramientas para crear tableros de control.

Rta: En el caso de Tableau, se puede seleccionar BigQuery directamente como la fuente de datos, seleccionar en el proyecto el dataset y seguidamente importar los datos de forma

directa desde BigQuery. Para el caso de PowerBI, solo se debe seleccionar la opción “obtener datos” y seleccionar la de BigQuery, luego pedirá las credenciales de inicio de sesión y se tendrá acceso a las tablas.

4. ¿Qué tipo de tablas de hechos y de medidas se identifican en el modelo multidimensional dado? justifique la respuests. Para la tabla de hechos indique si es factless, transaccional, periódica o snapshot acumulativo. Para las medidas indique su tipo (No aditiva, Semi-aditiva o Aditiva).

Rta: La tabla de hechos es transaccional, conteniendo información de la línea del pedido, cantidad, producto, etc. Las medidas son principalmente aditivas, puesto que se pueden sumar para llegar a obtener un total. Por otra parte, la tabla de hechos también es acumulativa, dado que con el paso del tiempo almacena mayor cantidad de valores asociados. Otra característica es la periodicidad de los hechos, pues pueden analizarse bajo este umbral. Finalmente, las tablas también tienen características del tipo factless, ya que algunas solo relacionan dimensiones de forma directa, es decir, registran la ocurrencia de eventos.

5. Suponga que a una de las dimensiones del modelo creado llega información nueva y que tiene sentido llevar historia de alguno de sus atributos. ¿Qué ajustes a la dimensión relacionada al proceso ETL se deben realizar para que al cargar la información se incluya un manejo de historia atributos de dimensión tipo 4? Describa la dimensión, los atributos afectados y el proceso a seguir.

Rta: La dimensión tipo 4 es la histórica, donde se registran los cambios asociados a los atributos de la dimensión. De modo que una tabla de versiones se encarga de “llevar la historia” de los atributos.

Para este caso, primero se debe crear una tabla adicional para llevar los datos de la historia de cualquier cambio de los atributos. Luego, a cualquier cambio realizado en el ETL o en la tabla de datos, debe registrarse en la nueva tabla de historia y versiones, registrando la versión actual y el resumen de los cambios elaborados.

6. ¿Qué ajustes al proceso ETL construido en este laboratorio hay que realizar para cargar nueva información que sea reportada por WWI en la tabla de hechos OrderLines?. Esto se considera en la literatura una carga incremental.

Rta: Para el proceso, es necesario agregar un nuevo dataset y asociar el archivo csv correspondiente (que tuvo que ser cargado de forma previa). Luego, por medio de una acción aditiva de Join, se debe añadir a al recipe del filter que tenía la demás información. Por lo tanto, se configura el join. Además, para evitar errores, hay que validar que el esquema del Output coincida con las columnas del filter alterado, pues podría conducir a errores.

7. ¿Qué errores se le presentaron en el desarrollo del laboratorio y qué solución plantearon? Haga énfasis en los que fueron más difíciles de solucionar.

Rta: Lo más difícil fue el tema de las sentencias SQL, pues no sabíamos por qué al hacer el llamado a los datasets no añadía la sentencia al job. La solución fue ubicar en BigQuery el ID directamente. Otro problema ocurrió al momento de hacer un Join con la nueva dimensión, pues no lo hacía de forma correcta, pues el job no aceptaba la cantidad de columnas, de modo que fue necesario editar el esquema del Job desde BigQuery.