**MÓDULO 8 - Optimización de procesos ETL**

**Objetivos de aprendizaje**

**Descubrir estrategias para crear un proceso ETL que funcione para satisfacer las necesidades organizativas y de las partes interesadas y cómo mantener un proceso ETL de forma eficiente**

**Introducir las herramientas utilizadas en ETL**

**Comprender los objetivos principales de las pruebas de calidad de ETL**

**Comprender los objetivos principales de la validación del esquema de datos**

**Desarrollar las mejores prácticas de pruebas de calidad ETL y validación de esquemas de datos**

**Identificar e implementar escenarios de prueba apropiados y puntos de control para QA en data pipelines**

**Explicar diferentes métodos para QA de datos en el pipeline**

**Crear escenarios de pruebas de rendimiento y medir el rendimiento en toda la canalización**

**Verificar las reglas de negocio**

**Realizar pruebas generales de rendimiento**

**Optimizando pipelines y procesos ETL**

**La importancia de las pruebas de calidad.**

****

**En BI, la prueba de calidad es un proceso de verificación de datos en busca de defectos para evitar fallas en el sistema.**

**Las pruebas de calidad implican seis elementos de validación: integridad, consistencia, conformidad, precisión, redundancia y puntualidad.**

**Comencemos con la verificación de la integridad. Esto implica confirmar que los datos contienen todos los componentes o medidas deseadas.**

***La integridad se refiere a la precisión, consistencia y confiabilidad de los datos a lo largo de su ciclo de vida***. En las pruebas de calidad, esto a menudo significa comprobar si faltan relaciones en los valores de los datos.

**A continuación, tenemos la consistencia, implica confirmar que los datos son compatibles y coinciden en todos los sistemas.**

**Lo siguiente es la conformidad. Este elemento tiene que ver con si los datos se ajustan al formato de destino requerido.** Si se está utilizando una base de datos de almacenamiento relacional, hay qué asegurarse de que no haya **redundancia** en los datos es otro elemento importante de las pruebas de calidad.

***La puntualidad implica confirmar que los datos están actualizados****.* Esta verificación se realiza específicamente para asegurarse de que los datos se hayan actualizado con la información más reciente que pueda proporcionar información relevante.

**Siete elementos de las pruebas de calidad.**

Considerar

* *Completitud: ¿Los datos contienen todos los componentes o medidas deseados?*
* *Coherencia: ¿Son los datos compatibles y están de acuerdo en todos los sistemas?*
* *Conformidad: ¿Los datos se ajustan al formato de destino requerido?*
* *Precisión: ¿Se ajustan los datos a la entidad real que se mide o describe?*
* *Redundancia: ¿Solo se mueven, transforman y almacenan los datos necesarios para su uso?*
* *Puntualidad: ¿Están actualizados los datos?*
* *Integridad: ¿Son los datos precisos, completos, consistentes y confiables? (La integridad está influenciada por las cualidades mencionadas anteriormente).*

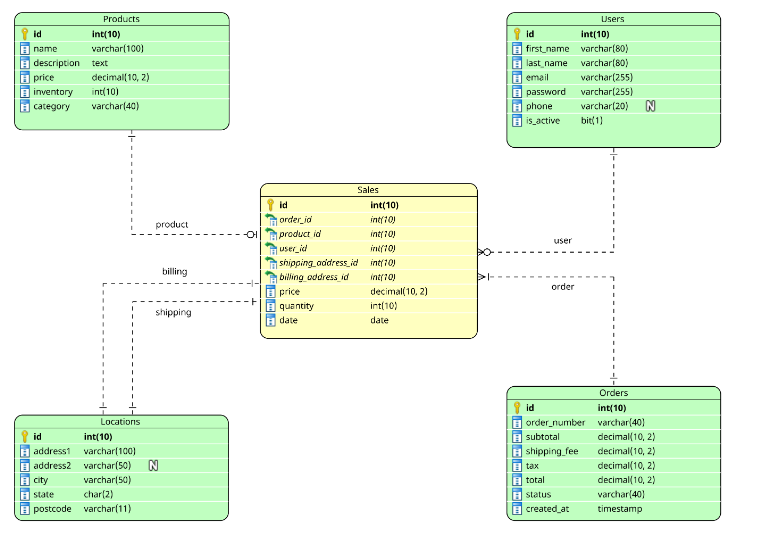
Problemas comunes

* *Verificar el mapeo de datos: ¿los datos del origen coinciden con los datos de la base de datos de destino?*
* *Comprobar si hay incoherencias: ¿Hay incoherencias entre el sistema de origen y el sistema de destino?*
* *Comprobar si hay datos inexactos: ¿Son correctos los datos y reflejan la entidad real que se está midiendo?*
* *Comprobar si hay datos duplicados: ¿Estos datos ya existen en el sistema de destino?*

**Monitorear la calidad de los datos con SQL**

El escenario

En Electrógena, una empresa de fabricación de productos electrónicos, un ingeniero de BI llamado Luis diseñó un almacén de datos para análisis e informes. Después del diseño del proceso ETL, Luis creó un diagrama del esquema.



El esquema de la base de datos sales\_warehouse. La tabla Ventas se conecta a las tablas Productos, Usuarios, Ubicaciones y Pedidos.

El diagrama del esquema de la base de datos sales\_warehouse contiene diferentes símbolos y conectores que representan dos piezas importantes de información: las tablas principales dentro del sistema y las relaciones entre estas tablas.

El esquema de la base de datos sales\_warehouse contiene cinco tablas:

Ventas

productos

Usuarios

Ubicaciones

Pedidos

Estas tablas están conectadas a través de claves. Las tablas contienen de cinco a ocho columnas (o atributos) que varían en tipo de datos. Los tipos de datos incluyen varchar o char (o carácter), entero, decimal, fecha, texto (o cadena), marca de tiempo y bit.

Las claves foráneas en la tabla Ventas enlazan con cada una de las otras tablas:

La clave externa "product\_id" enlaza con la tabla Productos

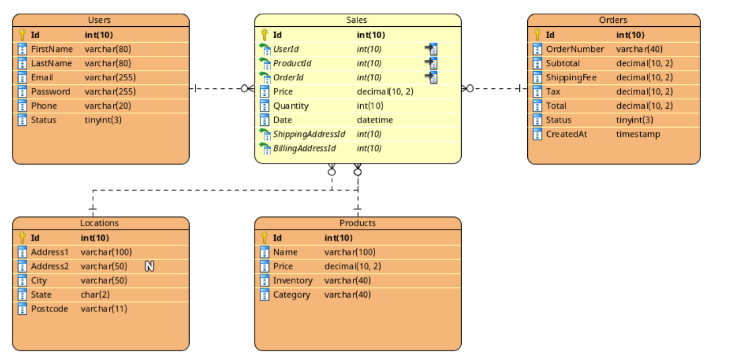
La clave externa "user\_id" enlaza con la tabla Usuarios

La clave externa "order\_id" enlaza con la tabla Pedidos

Las claves externas "shipping\_address\_id" y "billing\_address\_id" se vinculan a la tabla Ubicaciones

Después de que Luis creó la base de datos sales\_warehouse, el equipo de desarrollo realizó cambios en el sitio de ventas. Como resultado, la base de datos OLTP original cambió. Ahora, Luis debe asegurarse de que la canalización de ETL funcione correctamente y que los datos del almacén coincidan con la base de datos OLTP original.

Luis usó el esquema OLTP original de la base de datos de la tienda para diseñar el almacén.



El esquema de la base de datos de la tienda original. La tabla Ventas se conecta a las tablas Productos, Usuarios, Ubicaciones y Pedidos.

El esquema de la base de datos de la tienda también contiene cinco tablas: Ventas, Productos, Usuarios, Ubicaciones y Pedidos, que están conectadas mediante claves. Las tablas contienen de cuatro a ocho columnas que varían en tipo de datos. Los tipos de datos incluyen varchar o char, integer, decimal, date, text, timestamp, bit, tinyint y datetime.

Cada tabla en la base de datos de la tienda tiene un campo de identificación como clave principal. La base de datos contiene las siguientes tablas:

La tabla Ventas tiene columnas de precio, cantidad y fecha. Hace referencia a un usuario que realizó una venta (UserId), compró un producto (ProductId) y un pedido relacionado (OrderId). Además, hace referencia a la tabla Ubicaciones para las direcciones de envío y facturación (ShippingAddressId y BillingAddressId, respectivamente).

La tabla Usuarios tiene Nombre, Apellido, Correo electrónico, Contraseña y otras columnas relacionadas con el usuario.

La tabla Ubicaciones contiene información de dirección (Dirección1, Dirección2, Ciudad, Estado y Código postal).

La tabla Productos tiene Nombre, Precio, Número de inventario y Categoría de productos.

La tabla Pedidos tiene Número de pedido e información de compra (Subtotal, Tarifa de envío, Impuesto, Total y Estado).

Uso de SQL para encontrar problemas

Luis comparó la base de datos de sales\_warehouse con la base de datos original de la tienda para verificar la integridad, la consistencia, la conformidad, la precisión, la redundancia, la integridad y la puntualidad. Luis ejecutó consultas SQL para examinar los datos e identificar problemas de calidad. Luego, Luis preparó la siguiente tabla de listas, que incluye los tipos de problemas de calidad encontrados, las estrategias de calidad que se violaron, los códigos SQL utilizados para encontrar los problemas y las descripciones específicas de los problemas.

Pruebas de calidad sales\_warehouse

|  |  |  |  |
| --- | --- | --- | --- |
| Calidad probada | Estrategia de calidad | **SQL query** | **Observación de Luis** |
| Integridad | ¿Son los datos exactos, completos, consistentes y confiables? | **SELECT \***  **FROM Orders** | En la base de datos sales\_warehouse, el pedido con ID 7 tiene un valor total incorrecto. |
| Completitud | ¿Los datos contienen todos los componentes o medidas deseados? | **SELECT COUNT(\*)**  **FROM Locations** | La tabla Ubicaciones de la base de datos sales\_warehouse tiene una dirección adicional. En la base de datos de la tienda hay 60 registros, mientras que la tabla de la base de datos sales\_warehouse tiene 61. |
| Consistencia | ¿Los datos son compatibles y coinciden en todos los sistemas? | **SELECT Phone**  **FROM Users** | Varios usuarios dentro de la base de datos sales\_warehouse tienen teléfonos sin el prefijo "+". |
| Conformidad | ¿Los datos se ajustan al formato de destino requerido? | **SELECT id, postcode**  **FROM sales\_warehouse.Locations** | El código postal de ubicación para el registro con ID 6 en la base de datos sales\_warehouse es 722434213, lo cual es incorrecto. El código postal de Argentina contiene ocho caracteres alfanuméricos. |

### **Tienda Testeo de Calidad**

|  |  |  |  |
| --- | --- | --- | --- |
| **Característica** | **Estrategia de Calidad** | **SQL query** | **Observación de Luis** |
| Integridad | ¿Son los datos exactos, completos, consistentes y confiables? | **DESCRIBE Users** | Users.Status de la base de datos de la tienda y Users.is\_active de la base de datos sales\_warehouse parecen ser campos relacionados. Sin embargo, no es obvio cómo se transforma la columna Estado en la columna booleana is\_active. ¿Es posible que con un nuevo valor de estado, la canalización de ETL falle? |
| Consistencia | ¿Los datos son compatibles y coinciden en todos los sistemas? | **DESCRIBE Products** | Products.Inventory de la base de datos de la tienda tiene el tipo varchar en lugar del int(10) en el campo Products.inventory de la base de datos sales\_warehouse. Esto puede ser un problema si hay un valor con caracteres. |
| Exactitud | ¿Los datos se ajustan a la entidad real que se mide o describe? | **DESCRIBE Sales** | El tipo de datos de Sales.Date en la base de datos de la tienda es diferente de su tipo de datos en sales\_warehouse (fecha frente a fecha y hora). Puede que no sea un problema si el tiempo no es importante para la tabla de hechos de la base de datos sales\_warehouse. |
| Redundancia | ¿Solo se mueven, transforman y almacenan los datos necesarios para su uso? | **DESCRIBE Sales** | La tabla Sales de la base de datos sales\_warehouse tiene una restricción de índice única en las columnas OrderId, ProductId, UserId. Se puede agregar al esquema del almacén. |

Conclusiones clave

Probar la calidad de los datos es una habilidad esencial de un profesional de BI que garantiza buenos análisis e informes. Tal como lo hace Luis en este ejemplo, se pueden usar comandos SQL para examinar las bases de datos de BI y encontrar posibles problemas

**Validación de esquemas de datos**

**Conformidad desde el origen hasta el destino**

**Para garantizar la conformidad, desde el origen hasta el destino, los analistas de BI cuentan con tres herramientas muy eficaces; validación de esquemas, diccionarios de datos y linajes de datos.. En primer lugar, la validación del esquema es un proceso para garantizar que el esquema de datos del sistema de origen coincida con el esquema de datos de la base de datos de destino.**

**Las propiedades de validación del esquema deben garantizar tres cosas. Las claves siguen siendo válidas después de la transformación. Las relaciones de las tablas se han conservado y las convenciones son coherentes en toda la base de datos.**

**Un diccionario de datos es una colección de información que describe el contenido, el formato y la estructura de los objetos de datos dentro de una base de datos, así como sus relaciones**.

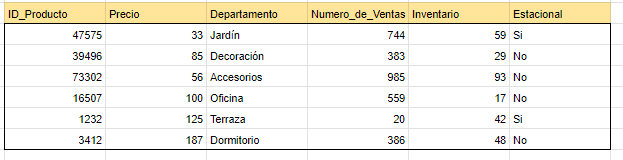
**El linaje de datos describe un proceso de identificación del origen de los datos, dónde se han movido a lo largo del sistema y cómo se han transformado con el tiempo.** Esto es útil porque si se obtiene un error, se puede rastrear el linaje de ese dato y comprender qué sucedió en el camino para causar el problema. Luego,se pueden establecer estándares para evitar el mismo problema en el futuro. El uso de validación de esquemas, diccionarios de datos y linajes de datos realmente ayuda a los analistas de BI a promover la coherencia a medida que los datos se mueven desde el origen hasta el destino. Esto significa que todos los usuarios pueden confiar en las soluciones de BI que se están creando.

**Ejemplo de diccionario de datos y linaje de datos**

Diccionarios de datos

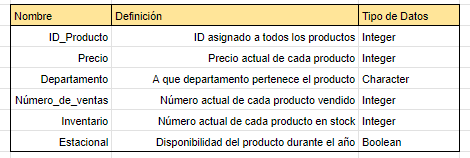
Aquí hay un ejemplo de una tabla de productos que existe dentro de una base de datos de ventas:

TABLA PRODUCTO



Esta tabla es en realidad la tabla de destino final para los datos recopilados de múltiples fuentes. Es importante garantizar la coherencia desde el origen hasta el destino porque estos datos provienen de diferentes lugares dentro del sistema. Aquí es donde entra el diccionario de datos:

DICCIONARIO DE DATOS



Linajes de datos

Un linaje de datos describe el proceso de identificar el origen de los datos, dónde se han movido a lo largo del sistema y cómo se han transformado con el tiempo. Esto puede ser realmente útil para los analistas de BI, porque cuando encuentran un error, pueden rastrearlo hasta el origen utilizando el linaje. Luego, pueden implementar comprobaciones para evitar que vuelva a ocurrir el mismo problema.



Por ejemplo, el sistema marcó un error con algunos datos entrantes sobre el número de ventas de un artículo en particular. Puede ser difícil encontrar dónde ocurrió este error si no se conoce el linaje de ese dato en particular, pero al seguir la ruta de esos datos a través del sistema, puede averiguar dónde crear una verificación.

**Reglas de negocio y testeo de rendimiento**

**Verificar reglas de negocio**

**En BI, una regla comercial es una declaración que crea una restricción en partes específicas de una base de datos.** Por ejemplo, una base de datos de envío podría imponer una regla comercial que establezca que las fechas de envío no pueden ser anteriores a las fechas de pedido. Esto evita que las fechas de pedido y las fechas de envío se mezclen y causen errores dentro de este sistema. Las reglas comerciales se crean de acuerdo con la forma en que una organización en particular usa sus datos.de base de datos.

**Las reglas comerciales qué se creen afectarán muchas de las bases de datos, diseñarán qué datos se recopilan y almacenan, cómo se definen las relaciones, qué tipo de información proporciona la base de datos y la seguridad de los datos.**

Esto ayuda a garantizar que la base de datos funcione según lo previsto. Las reglas comerciales son diferentes en cada organización porque la forma en que las organizaciones interactúan con sus datos siempre es diferente.

**Imposición de reglas de negocio**

Las reglas comerciales dependen en gran medida de la organización y sus necesidades de datos. Esto significa que las reglas comerciales son diferentes para cada organización. Esta es una de las razones por las que es tan importante verificar las reglas comerciales; estas comprobaciones ayudan a garantizar que la base de datos realmente esté haciendo el trabajo que necesita que haga.

Por ejemplo, una empresa de artículos del hogar, tiene una base de datos que administra las solicitudes de órdenes de compra ingresadas por los empleados. Las órdenes de compra de más de $500000 necesitan la aprobación del gerente. Para automatizar este proceso, se puede imponer un conjunto de reglas en la base de datos que entrega automáticamente las solicitudes de más de $500000 a una tabla de informes pendiente de la aprobación del gerente. Otras reglas comerciales que pueden aplicarse en este ejemplo son: los precios deben ser valores numéricos (el tipo de datos debe ser un número entero); o para que exista una solicitud, es obligatorio un motivo (el campo de la tabla no puede ser nulo).



El diagrama de flujo representa las tres reglas comerciales: solicitudes de pedidos de productos, solicitudes pendientes de aprobación y pedidos de productos aprobados

Para cumplir con este requisito comercial, hay tres reglas en juego en este sistema:

Las solicitudes de pedidos inferiores a $500000 se envían automáticamente a la tabla de solicitudes de pedidos de productos aprobados

Las solicitudes de más de $ 500000 se envían automáticamente a la tabla de solicitudes pendientes de aprobación

Las solicitudes aprobadas se envían automáticamente a la tabla de solicitudes de pedidos de productos aprobados

Estas reglas afectan inherentemente la forma de este sistema de base de datos para satisfacer las necesidades de esta organización en particular.

***Conclusiones clave***

***Las reglas comerciales determinan qué datos se recopilan y almacenan, cómo se definen las relaciones, qué tipo de información proporciona la base de datos y la seguridad de los datos. Estas reglas influyen en gran medida en cómo se diseña una base de datos y cómo funciona una vez que se ha configurado. Comprender las reglas comerciales y por qué son importantes es útil como analista de BI porque esto puede ayudar a comprender cómo funcionan los sistemas de bases de datos existentes, diseñar nuevos sistemas de acuerdo con las necesidades comerciales y mantenerlos para que sean útiles en el futuro.***

**Actividad:**

**¿Cómo diseñarías un pipeline eficiente?**