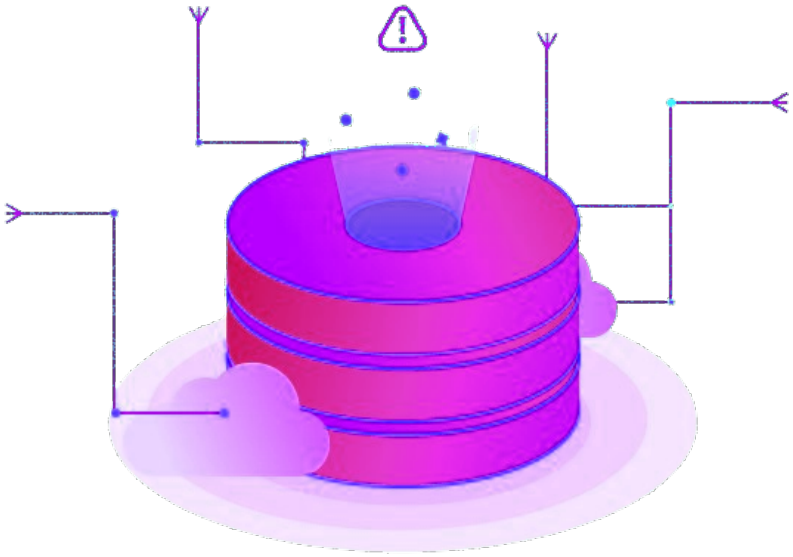
Universidad Nacional Autónoma De México
Facultad De Ciencias, 2024-I
Almacenes & Mineria de Datos



Práctica 03:

Integración de Datos.

Profesor:
Gerardo Avilés Rosas

Ayudante de teoría:
Gerardo Uriel Soto Miranda

Ayudante de laboratorio:
Ricardo Badillo Macías

Introducción.

Los macrodatos, el internet de las cosas, la actividad en la nube y muchas más herramientas están generando un auge en la cantidad tanto de fuentes de datos como de datos existentes en el mundo. La mayoria de estos datos ya se recopilaron y almacenaron en entornos aislados o almacenes de datos independientes. La integración de datos es el proceso que reúne esos datos para generar un mayor valor de datos y estadísticas.

Integración de Datos:

La **integración de datos es un proceso que consiste en reunir datos de diferentes fuentes para obtener una vista unificada y más valiosa de ellos**, de modo que tu empresa puede tomar mejores decisiones y con mayor rapidez.

La integración de datos puede consolidar todo tipo de datos (estructurados, no estructurados, por lotes y de transmisión) para realizar cualquier tipo de tareas, desde consultas básicas a bases de datos de inventarios hasta estadísticas predictivas complejas.

La integración de datos se implementa generalmente en un almacenamiento de datos mediante software especializado que aloja grandes almacenes de datos de recursos internos y externos. Los datos se extraen, se mezclan y se presentan de forma unificada.

Pasos de la integración:

i. **Acceso a los datos:** Dedes todas las fuentes y localizaciones se extraen los datos, tanto si se trata de lcoales, en la nube o de una combinación de ambos.
ii. **Integración de datos:** Los registros de una fuente de datos mapean registros en otra. Por ejemplo juntar 2 conjuntos de datos uno con los atributos (nombre,apellidos) y otro con (nom,ape), asegurando de que en ambos casos los datos van al lugar correcto.
iii. **Entrega de datos integrados**: Justo en el momento en que la empresa los necesita, por lotes, casi en tiempo real o en tiempo real

Figura 1: *Integración.*

Formas de llevar a cabo la integración de datos:

i. **Integración manual:** La persona que se va a encargar de la integración de los datos tendrá que recopilar y limpiar los datos de las distintas fuentes y después combinarlos dentro de un mismo almacén.

ii. **Integración con uso de middleware:** El middleware ayuda normalizar los datos de acuerdo a la aplicación de destino para que puedan ser usados. Este tipo de integración se suele usar cuando hay un sistema heredado, debido a la antigüedad del mismo.

iii. **Integración a partir de aplicaciones:** Este tipo de integración sólo es posible cuando la integración se va a realizar entre una cantidad no muy numerosa de aplicaciones. La herramienta localizara, extraera e integrara

los datos desde las distintas fuentes.

iv. **Integración de acceso uniforme:** Los datos se mantienen en la fuente original de los datos. Se crea una interfaz para que los datos parezcan coherentes cuando se accede alos datos desde otras fuentes.

v. **Integración de almacenamiento común:** Consiste en hacer una copia de los datos de las diferentes fuentes en un almacén de datos o data service. De esta forma, se consigue una visión unificada.

Algunas de estas formas, como la integración a partir de aplicaciones o la de almacenamiento común, se apoyan en herramientas de procesos de ETL para hacer la integración de datos, el cual extrae los datos del sistema de origen, los transforma para que sean compatibles en el sistema y, finalmente, se cargan los datos en el sistema destino.

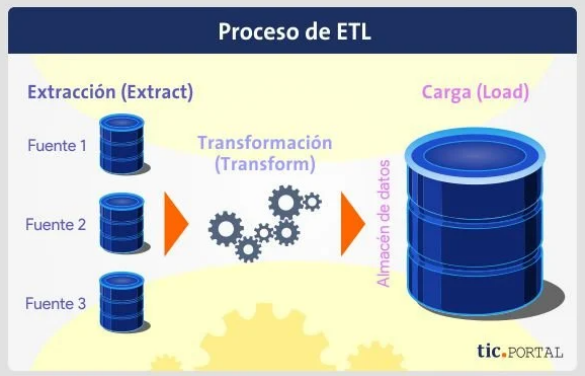


Figura 2: *Proceso ETL.*

Integración de esquemas:

La **integración de esquemas se utiliza para combinar dos o más esquemas de base de datos en un único esquema que puede almacenar datos de ambas bases de datos originales**.

Las tareas a realizar al hacer una integración de esquema son:

i. **Identificar correspondencias y conflictos entre los esquemas:** Como los esquemas se diseñan individualmente, es necesario especificar construcciones en los esquemas que representan el mismo concepto del mundo real. Algunos conflictos son :

   i. **Conflicto de nombres:** Los conflictos pueden ser de dos tipo sinónimos y homónimos. Un sinónimo ocurre cuando dos esquemas usan nombres diferentes para describir el mismo concepto. Un homónimo ocurre cuando dos esquemas usan el mismo nombre para describir conceptos diferentes.

   i. **Conflictos de tipos:** Un concepto similar puede representarse en dos esquemas mediante diferentes construcciones de modelado.

   i. **Conflictos de dominio:** Un solo atributo puede tener diferentes dominios en diferentes esquemas.

   i. **Conflictos entre restricciones:** Dos esquemas pueden imponer restricciones diferentes.

ii. **Modificar vistas para que se asemejen entre sí:** Algunos esquemas se modifican para que se ajusten más a otros esquemas.

iii. **Fusión de vistas y reestructuración :** Los esquemas globales se crean fusionando los esquemas individuales. Los conceptos correspondientes se representan solo una vez en el esquema global y se especifica la asignación entre las vistas y los esquemas globales. (Mucha intervención humana y negociación para resolver conflcitos).

SQL Server Integration Services (SSIS):

SQL Server Integration Service (SSIS) es un componente que permite generar procesos de migración de grandes cantidades de adtos de diferentes orígenes llamados ETL.

SSIS puede ser utilizados para **la extración, carga y transformación de datos mediante la extracción de datos de varios orignenes**, como la base de datos de SQL Server, la base de datos de Oracle y los archivos de Excel.

Dispone de un entorno de desarrollo gráfico integrado en **Visual Studio**. **SSIS** dispone de procesos que realizan "cosas" como ejecutar un Script SQL, leer datos de un fichero, leer datos de una tabla... y **cada proceso se une con otro mediante flujos de datos pudiendo comunicar ambos procesos**.

Los proyectos ETL de SSIS tienen sentido cuando se pueden ejecutar y automatizar sin la necesidad de abrir el proyecto con Visual Studio, para ello, SSIS puede generar un paquete que se podrá ejecutar desde el agente de programación SQL Server Agent o crear un paquete ejecutable.

Un paquete de SSIS es la colección de tareas ejecutadas de forma ordneada necesarias para combinar datos en un único conjunto de datos y cargar la tabla de destino en un solo paso en lugar de seguir un proceso paso a paso para guardar los archivos en un servidor SQL Server. Un paquete puede usar flujo de control, administrador, tareas, variables, controladores de eventos, parámetros y más para lograr esto.

Componentes principales en un paquete SSIS:

- **Flujo de control:** El flujo de control le ayuda a organizar los componentes para facilitar la ejecución. Estos componentes incluyen tareas y contenedores.
  - **Tarea:** Una tarea se puede definir como la unidad de trabajo. Funciona exactamente como un lenguaje de programación. Sin embargo, no utiliza métodos de codificación para la ejecución. Debe arrastrar y soltar para configurar tareas.
  - **Contenedor:**
    - **Contenedor de secuencia:** Esto le permite organizar las tareas agrupándolas.
    - **Para el contenedor de bucle:** Esto le permite ejecutar una tarea varias veces en función de la evaluación.
    - **Contenedor Foreach loop:** Esto permite realizar bucles sobre un conjunto de objetos, como archivos en una carpeta.

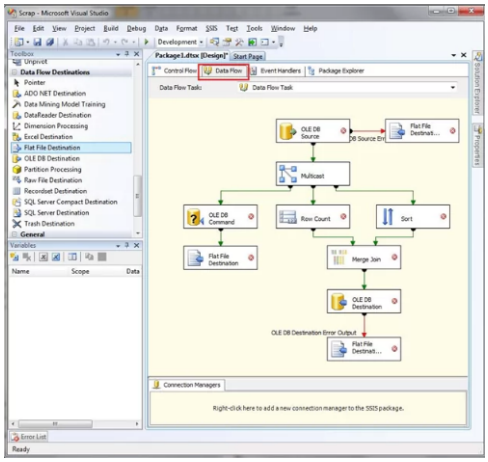


Figura 3: *SSIS*.

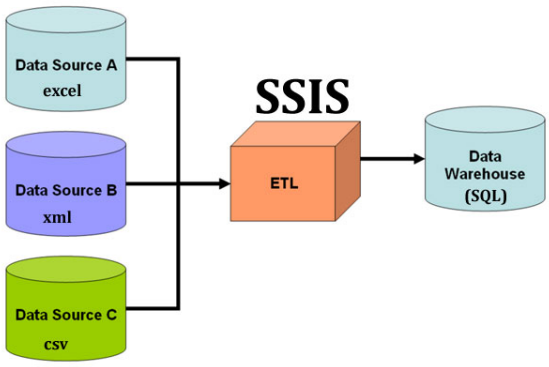


Figura 4: *ETL*.

Actividades.

i. Deberan hacer un resumen de el articulo **DataIntegration.pdf**, en donde plasmen los puntos mas importantes lo llamaran **Resumen.pdf**.

ii. Utilizando **SSIS**, deberan integrar los datos de las 4 tablas .csv, la idea es que primero integren los .csv con el mismo nombre, y de ahi mezclarlos. Se espera que despues del proyecto se tenga una integración con los datos completos de las 4 tablas en 1 solo archivo, ademas deberan agregarlo a una tabla dentro de una base de datos en SQL Server que creen. Deberan comprimir el proyecto entero y llamarlo **Integracion.zip**. Ademas deberan guardar el archivo resultante de la integración, y lo llamaran **integracion.csv**.

iii. Deberan escribir los pasos que realizaron para hacer la integración , asi como su analisis y lo pondran en **Práctica03.pdf**.

Figura 5: *Actividades.*

Entregables.

Deberán subir un archivo con formato *zip* a *Google Classroom*, de acuerdo a lo indicado en los lineamientos de entrega. Debe de estar organizado de la siguiente manera, (suponiendo que el nombre del equipo que está entregando es *Dream Team* y los integrantes del equipo son los profesores del curso).

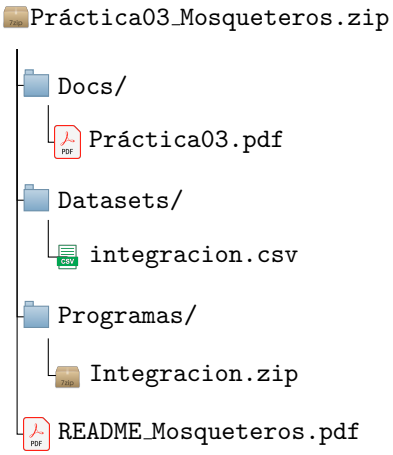


Figura 6: *Entregables.*

Nota.

Para cualquier duda o comentario que pudiera surgirles al hacer este trabajo, recuerden que cuentan con la asignación de este entregable en el grupo de *Classroom*, en donde seguramente encontrarás las respuestas que necesites.

Figura 7: *Nota.*