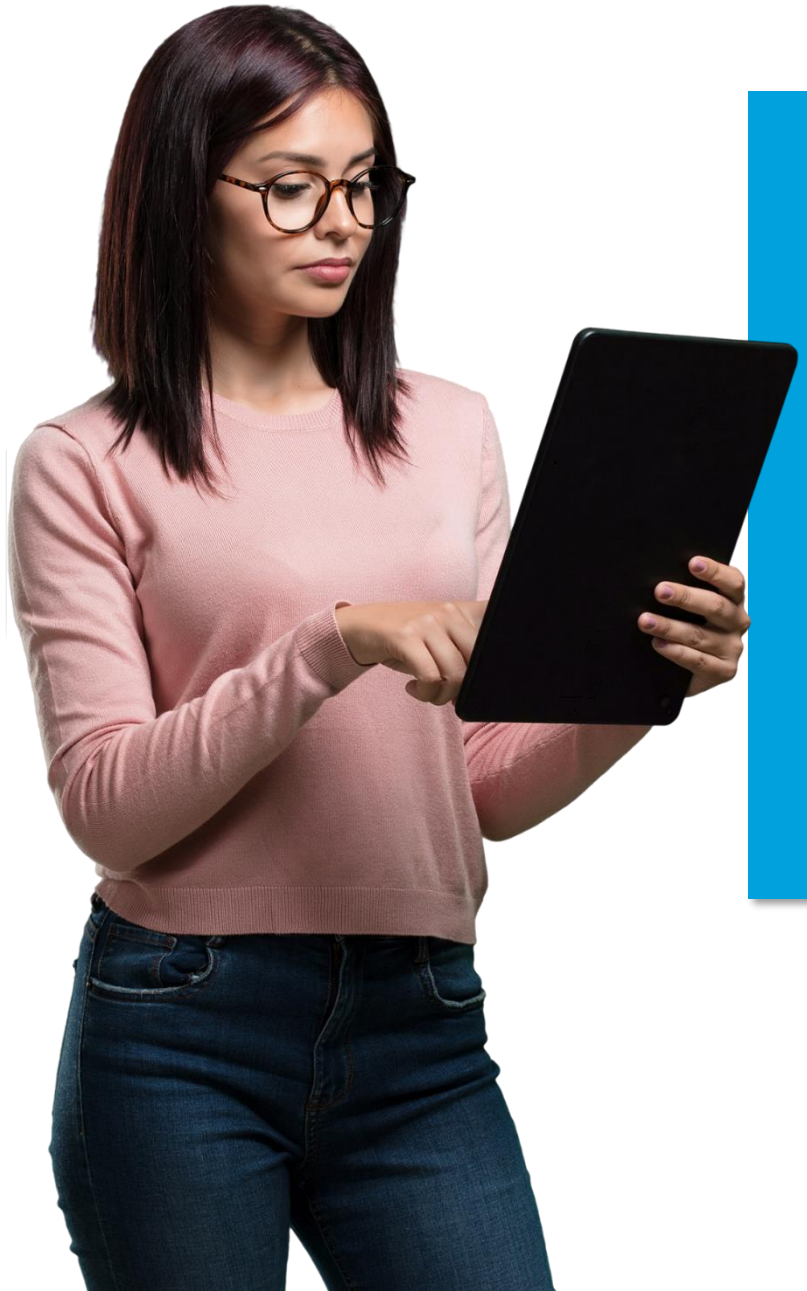


Unidad # 3
Almacenes de Datos y Minería de Datos

Procesos ETL/ELT y herramientas para almacenes de datos (Data Warehouse)

Administración de Base de Datos II

UNIVERSIDAD GERARDO BARRIOS
Líderes en Gestión del Conocimiento

Comisión de Acreditación de la Calidad de la Educación Superior
UNIVERSIDAD GERARDO BARRIOS (UGB)
ACREDITADA
2019- 2024

Contenidos a desarrollar

1. ¿Qué es ETL?
2. ¿Qué es ELT?
3. ETL vs. ELT
4. Herramientas (Software) para los almacenes de datos

¿Qué es un ETL?

- ETL es un proceso que extrae los datos de diferentes sistemas de origen, luego los transforma (como aplicar cálculos, concatenaciones, etc.) y finalmente carga los datos en el sistema de almacenamiento de datos. La forma completa de ETL es Extraer, Transformar y Cargar (**E**xtract-**T**ransform-**L**oad).
- No es tan simple pensar que la creación de un almacén de datos es simplemente extraer datos de múltiples fuentes y cargarlos en la base de datos de un almacén de datos. Esto está lejos de la verdad y requiere un proceso ETL complejo. El proceso ETL requiere aportaciones activas de varias partes interesadas, incluidos desarrolladores, analistas, probadores, altos ejecutivos y es un desafío técnico.
- ETL es una actividad recurrente (diaria, semanal, mensual) de un sistema de almacenamiento de datos y debe ser ágil, automatizada y bien documentada.

¿Cuál es su necesidad en una empresa?

- Ayuda a las empresas a analizar sus datos comerciales para tomar decisiones comerciales críticas.
- Las bases de datos transaccionales no pueden responder preguntas comerciales complejas que pueden responderse con el ejemplo de ETL.
- Un almacén de datos proporciona un repositorio de datos común
- ETL proporciona un método para mover los datos de varias fuentes a un almacén de datos.
- A medida que cambian las fuentes de datos, el almacén de datos se actualizará automáticamente.
- Un sistema ETL bien diseñado y documentado es casi esencial para el éxito de un proyecto de almacenamiento de datos.

¿Cuál es su necesidad en una empresa? (2)

- Permitir la verificación de las reglas de transformación, agregación y cálculo de datos.
- El proceso ETL permite la comparación de datos de muestra entre el sistema de origen y el de destino.
- El proceso ETL puede realizar transformaciones complejas y requiere un área adicional para almacenar los datos.
- ETL ayuda a migrar datos a un almacén de datos. Convierta a varios formatos y tipos para adherirse a un sistema consistente.
- ETL es un proceso predefinido para acceder y manipular datos de origen en la base de datos de destino.
- ETL en el almacén de datos ofrece un contexto histórico profundo para el negocio.
- Ayuda a mejorar la productividad porque codifica y reutiliza sin necesidad de conocimientos técnicos.

Proceso ETL

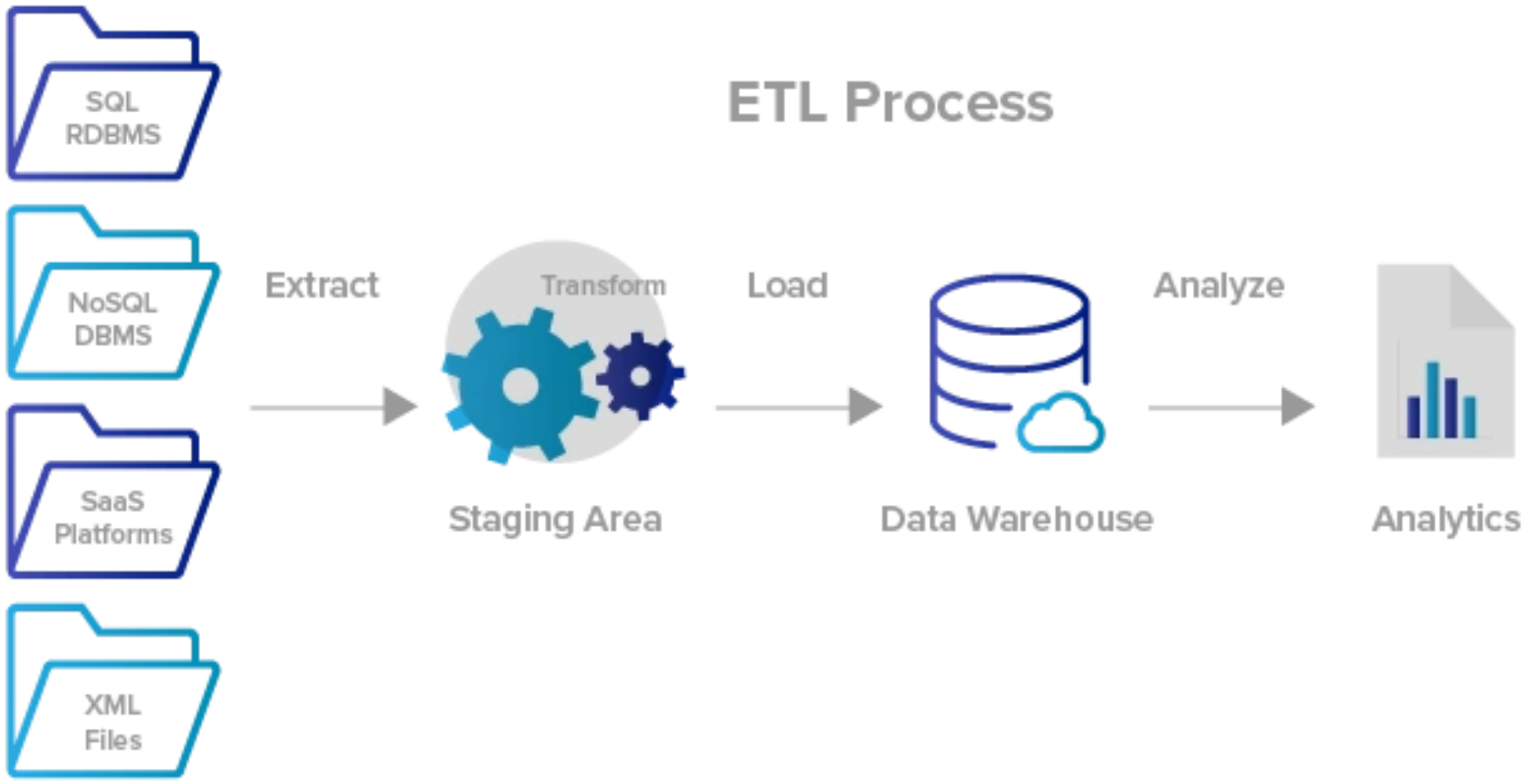


Figura 1. proceso ETL. Fuente: https://cdn.filestackcontent.com/auto_image/compress/LjY9fP8fQWyBQ4aVlLy7

Paso 1) Extracción

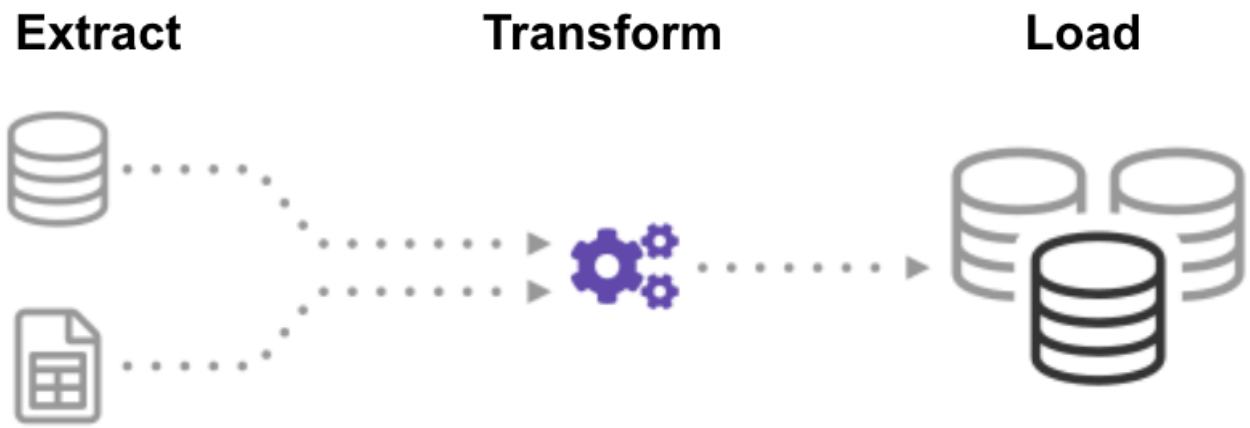
- En este paso de la arquitectura ETL, los datos se extraen del sistema de origen al área de preparación. Las transformaciones, si las hay, se realizan en el área de preparación para que el rendimiento del sistema fuente no se degrade. Además, si los datos dañados se copian directamente desde la fuente a la base de datos del almacén de datos, la reversión será un desafío. El área de preparación brinda la oportunidad de validar los datos extraídos antes de que se muevan al almacén de datos.
- El almacén de datos necesita integrar sistemas que tengan diferentes DBMS, hardware, sistemas operativos y protocolos de comunicación. Las fuentes podrían incluir aplicaciones heredadas como Mainframes, aplicaciones personalizadas, dispositivos de punto de contacto como cajeros automáticos, conmutadores de llamadas, archivos de texto, hojas de cálculo, ERP, datos de proveedores, socios, entre otros.

Paso 1) Extracción

- Por lo tanto, se necesita un mapa de datos lógico antes de extraer y cargar físicamente los datos. Este mapa de datos describe la relación entre las fuentes y los datos de destino.

Métodos de extracción de datos:

- **Extracción completa**
- **Extracción parcial**: sin notificación de actualización.
- **Extracción parcial**: con notificación de actualización



Paso 1) Extracción

Algunas validaciones que se realizan durante la extracción:

- Conciliar registros con los datos de origen
- Se debe asegurar de que no se carguen spam o datos no deseados
- Verificación del tipo de datos
- Eliminar todo tipo de datos duplicados / fragmentados
- Comprobar si todas las llaves están en su lugar o no

Paso 2) Transformación

- Los datos extraídos del servidor de origen están sin procesar y no se pueden utilizar en su forma original. Por lo tanto, necesitan ser limpiados, mapeados y transformados. De hecho, este es el paso clave en el que el proceso ETL agrega valor y cambia los datos de manera que se puedan generar informes detallados.
- Es uno de los conceptos ETL importantes donde aplica un conjunto de funciones en datos extraídos. Los datos que no requieren ninguna transformación se denominan datos de transferencia directa o transferencia .
- En el paso de transformación, puede realizar operaciones personalizadas en los datos. Por ejemplo, si el usuario desea ingresos por suma de ventas que no están en la base de datos. O si el nombre y el apellido en una tabla están en columnas diferentes. Es posible concatenarlos antes de cargarlos.

Problemas de integridad de datos

- Diferente ortografía de la misma persona como Jon, John, etc.
- Hay varias formas de denotar el nombre de una empresa como Google, Google Inc.
- Uso de diferentes nombres como Cleaveland, Cleveland.
- Puede darse el caso de que varias aplicaciones generen diferentes números de cuenta para el mismo cliente.
- En algunos datos, los archivos requeridos permanecen en blanco.
- Un producto no válido recogido en el punto de venta como entrada manual puede provocar errores.

Las validaciones se realizan durante esta etapa

- **Filtrado:** seleccione solo ciertas columnas para cargar
- Usar reglas y tablas de búsqueda para la estandarización de datos
- Conversión de juegos de caracteres y manejo de codificación
- Conversión de unidades de medida como conversión de fecha y hora, conversiones de moneda, conversiones numéricas, etc.
- Comprobación de la validación del umbral de datos. Por ejemplo, la edad no puede tener más de dos dígitos.
- Validación del flujo de datos desde el área de preparación a las tablas intermedias.
- Los campos obligatorios no deben dejarse en blanco.
- Limpieza (por ejemplo, asignar NULL a 0 o Sexo masculino a "M" y femenino a "F", etc.)
- Divida una columna en múltiplos y combine varias columnas en una sola columna
- Transposición de filas y columnas
- Utilice búsquedas para fusionar datos
- Usar cualquier validación de datos compleja (p. Ej., Si las dos primeras columnas de una fila están vacías, automáticamente se rechaza el procesamiento de la fila)

Paso 3) Cargando

- La carga de datos en la base de datos del almacén de datos de destino es el último paso del proceso ETL. En un almacén de datos típico, es necesario cargar un gran volumen de datos en un período relativamente corto (noches). Por lo tanto, el proceso de carga debe optimizarse para el rendimiento.
- En caso de falla de carga, los mecanismos de recuperación deben configurarse para reiniciarse desde el punto de falla sin pérdida de integridad de los datos. Los administradores del almacén de datos deben monitorear, reanudar y cancelar cargas según el rendimiento del servidor prevaleciente.
- Tipos de carga:
    1. **Carga inicial** : completando todas las tablas del almacén de datos
    2. **Carga incremental** : aplica cambios continuos cuando sea necesario periódicamente.
    3. **Actualización completa**: borra el contenido de una o más tablas y vuelve a cargar con datos nuevos.

Verificación de carga

- Asegúrese de que los datos del campo clave no falten ni sean nulos
- Pruebe las vistas de modelado basadas en las tablas de destino
- Compruebe los valores combinados y las medidas calculadas
- Verifique los datos en la tabla de dimensiones y en la tabla de historial
- Consulte los informes de BI en la tabla de hechos y dimensiones cargada

¿Qué es ELT?

- ELT es un método diferente de analizar el enfoque de la herramienta para el movimiento de datos. En lugar de transformar los datos antes de que se escriban, ELT permite que el sistema de destino realice la transformación. Los datos se copiaron primero en el destino y luego se transformaron en su lugar.
- ELT generalmente se usa con bases de datos sin SQL como el clúster Hadoop, el dispositivo de datos o la instalación en la nube..

Proceso ELT

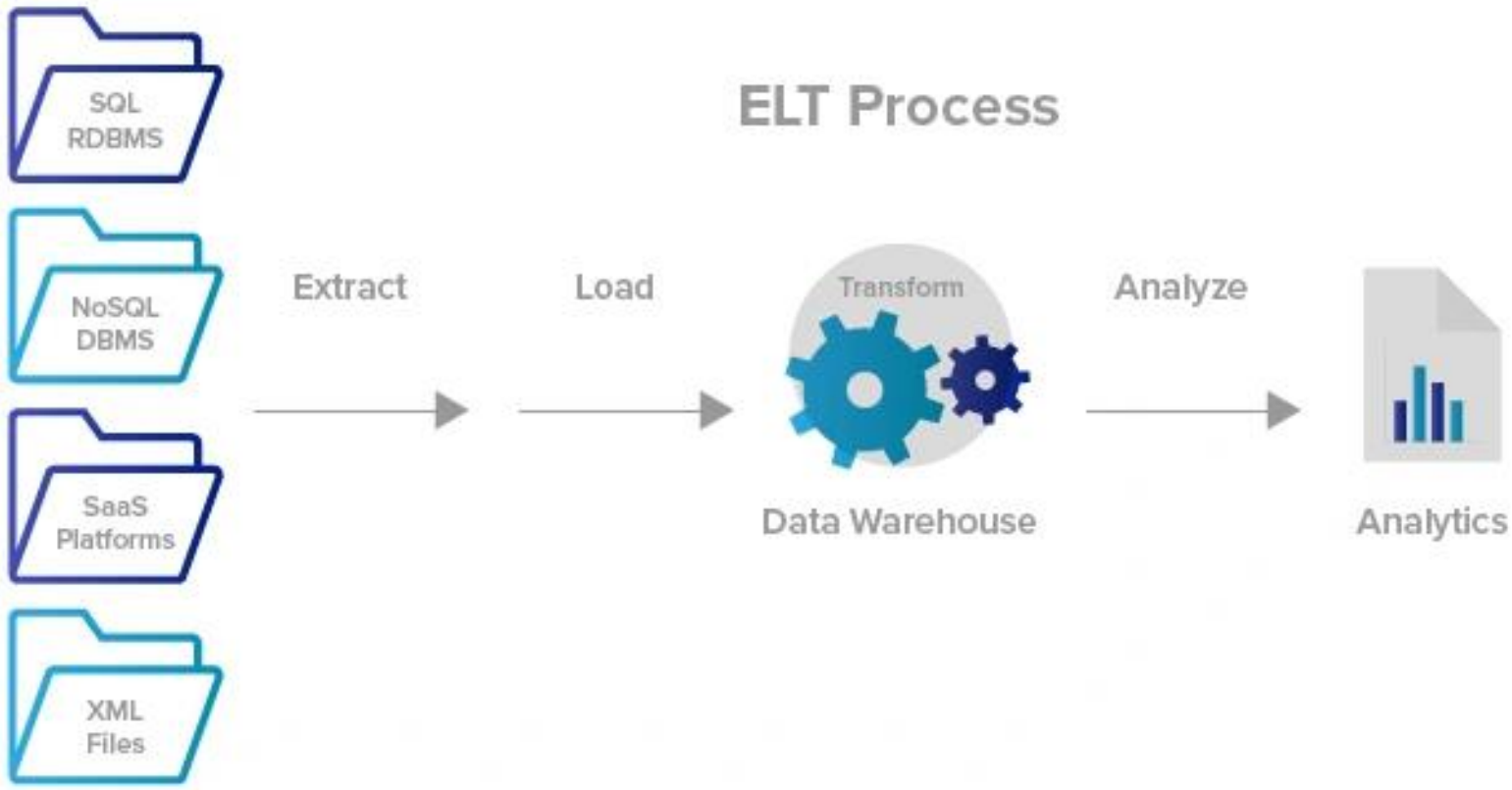


Figura 2. proceso ELT. Fuente: https://cdn.filestackcontent.com/auto_image/compress/esfcZz4QqSrriADPU2i7

DIFERENCIA CLAVE

- ETL significa Extract, Transform and Load, mientras que ELT significa Extract, Load, Transform.
- ETL carga los datos primero en el servidor de ensayo y luego en el sistema de destino, mientras que ELT carga los datos directamente en el sistema de destino.
- El modelo ETL se usa para datos locales, relacionales y estructurados, mientras que ELT se usa para fuentes de datos estructuradas y no estructuradas en la nube escalables.
- ETL se usa principalmente para una pequeña cantidad de datos, mientras que ELT se usa para grandes cantidades de datos.
- ETL no proporciona soporte de datos, mientras que ELT sí lo proporciona.
- ETL es fácil de implementar, mientras que ELT requiere habilidades específicas para implementar y mantener.

Diferencia entre ETL y ELT

Parámetro	ETL	ELT
Proceso	Los datos se transforman en el servidor de ensayo y luego se transfieren a DB Data Warehouse.	Los datos permanecen en la base de datos del Data Warehouse.
Uso de código	• Transformaciones intensivas en computación • Pequeña cantidad de datos	Usado para grandes cantidades de datos
Transformación	Las transformaciones se realizan en el servidor ETL / área de ensayo.	Las transformaciones se realizan en el sistema de destino
Carga de tiempo	Los datos se cargaron primero en la preparación y luego se cargaron en el sistema de destino. Tiempo intensivo.	Los datos se cargan en el sistema de destino solo una vez. Más rápido.
Transformación del tiempo	El proceso ETL debe esperar a que se complete la transformación. A medida que aumenta el tamaño de los datos, aumenta el tiempo de transformación.	En el proceso ELT, la velocidad nunca depende del tamaño de los datos.
Tiempo- Mantenimiento	Necesita un alto mantenimiento, ya que necesita seleccionar datos para cargar y transformar.	Bajo mantenimiento ya que los datos siempre están disponibles.

Diferencia entre ETL y ELT (2)

Parámetro	ETL	ELT
Complejidad de implementación	En una etapa temprana, más fácil de implementar.	Para implementar el proceso de ELT, la organización debe tener un conocimiento profundo de las herramientas y las habilidades de los expertos.
Soporte para almacenamiento de datos	Modelo ETL utilizado para datos locales, relacionales y estructurados.	Se utiliza en una infraestructura de nube escalable que admite fuentes de datos estructuradas y no estructuradas.
Soporte de Data Lake	No lo soporta	Permite el uso de Data Lake con datos no estructurados.
Complejidad	El proceso ETL carga solo los datos importantes, identificados en el momento del diseño.	Este proceso implica el desarrollo desde la salida hacia atrás y la carga solo de datos relevantes.
Costo	Altos costos para pequeñas y medianas empresas.	Bajos costos de entrada utilizando software en línea como plataformas de servicio.

Diferencia entre ETL y ELT (3)

Parámetro	ETL	ELT
Búsquedas	En el proceso ETL, tanto los hechos como las dimensiones deben estar disponibles en el área de preparación.	Todos los datos estarán disponibles porque la extracción y la carga ocurren en una sola acción.
Agregaciones	La complejidad aumenta con la cantidad adicional de datos en el conjunto de datos.	La potencia de la plataforma de destino puede procesar rápidamente una cantidad significativa de datos.
Cálculos	Sobrescribe la columna existente o necesita agregar el conjunto de datos y enviarlo a la plataforma de destino.	Agregue fácilmente la columna calculada a la tabla existente.
Madurez	El proceso se utiliza desde hace más de dos décadas. Está bien documentado y las mejores prácticas están fácilmente disponibles.	Concepto relativamente nuevo y complejo de implementar.
Hardware	La mayoría de las herramientas tienen requisitos de hardware únicos que son costosos.	El costo del hardware SaaS no es un problema.
Soporte para datos no estructurados	Soporta principalmente datos relacionales	Soporte para datos no estructurados fácilmente disponible.

Herramientas (Software) de almacenes de datos

El almacén de datos es el futuro de todas las empresas. Por lo tanto, antes de elegir una herramienta final, uno debe asegurarse de que la herramienta sea capaz de satisfacer los requisitos integrales y de crecimiento de la organización en el presente y en el futuro.

Herramientas ETL

1. **MarkLogic**. es una solución de almacenamiento de datos que hace que la integración de datos sea más fácil y rápida utilizando una variedad de funciones empresariales. Se puede consultar diferentes tipos de datos como documentos, relaciones y metadatos. (Enlace)

2. **Oracle.** Base de datos líder en la industria. Ofrece una amplia gama de opciones de soluciones de almacenamiento de datos tanto en las instalaciones como en la nube. Ayuda a optimizar las experiencias de los clientes aumentando la eficiencia operativa. (Enlace)

3. **Amazon RedShift**. Es una herramienta de Datawarehouse, sencilla y rentable para analizar todo tipo de datos utilizando SQL estándar y herramientas de BI existentes. También permite ejecutar consultas complejas contra petabytes de datos estructurados. (Enlace)

4. **Otras herramientas**: CData Sync, BiG EVAL, QuerySurge, Xplenty, Panoply, Domo, Teradata, SAP, SAS

Herramientas ELT/ETL

1. **IBM Infosphere**. (Licencia). Excelente herramienta ETL que utiliza notaciones gráficas para ejecutar actividades de integración de datos.

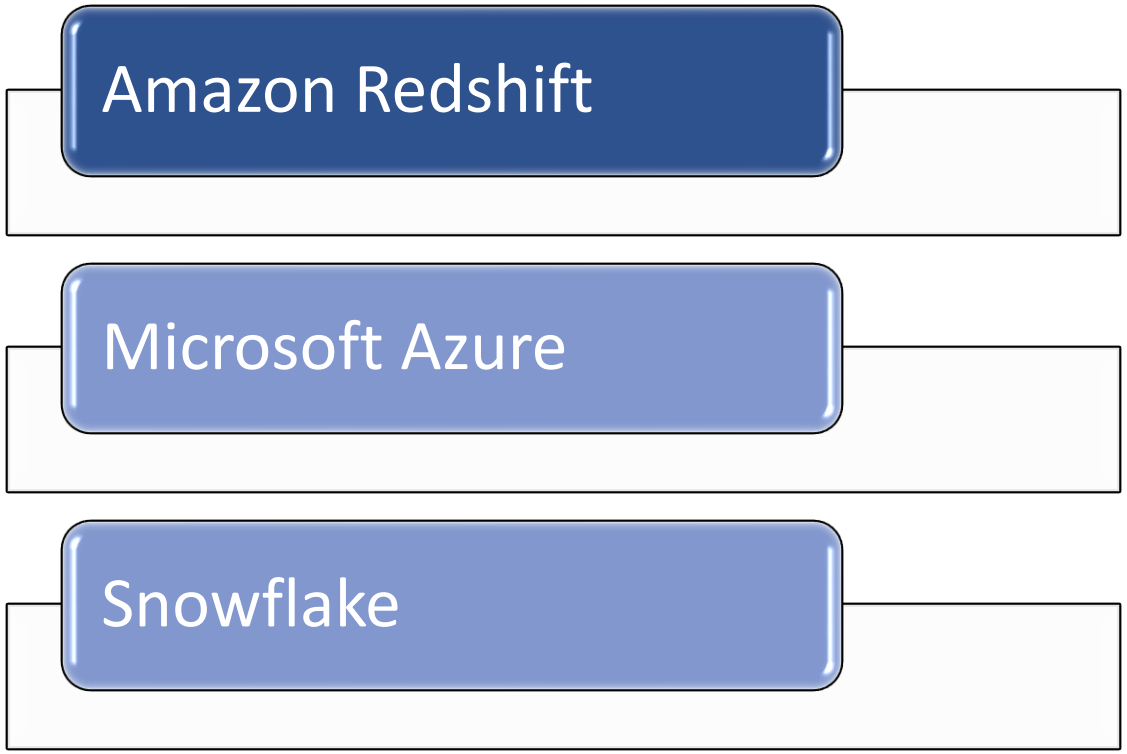
2. **Informática**. (Licencia). El centro de poder de Informatica consta de tres componentes principales: Herramientas de cliente, repositorio de Power Center y servidor Power Center

3. **SAP.** (Licencia). Este entorno de almacén está completamente integrado en el entorno de SAP.

4. **Talend**. Herramienta de código abierto propiedad de la organización Talend para el almacenamiento de datos. Sencillo de usar y también han atraído a muchos usuarios.

5. **Teradata.** (Licencia). Una característica interesante de este almacén de datos es su segregación de datos en datos calientes y fríos. Aquí, los datos fríos se refieren a datos utilizados con menos frecuencia y esta es la herramienta en el mercado en estos días.

Herramientas de almacenamiento de datos populares:



Notas adicionales

- Hay varias opciones que están disponibles para las empresas en herramientas de almacenamiento de datos. Esto, a su vez, enfatiza la importancia de un análisis adecuado de los requisitos y necesidades de la organización antes de elegir cualquier herramienta.
- Siempre es mejor estar preparado con una imagen clara de los requisitos actuales y los patrones futuros de antemano. Al ser el repositorio central, el almacén de datos es extremadamente importante para cualquier organización en cualquier sector y, por lo tanto, la elección de la herramienta correcta es imprescindible.

Tabla de actividades	
Nombre de la actividad	Herramientas para almacenes de datos
Tipo de actividad	Equipos (3 integrantes)
Competencias especifica de la asignatura	Manipular bases de datos para asegurar la disponibilidad y seguridad de los datos, utilizando entornos web o locales, implementando Data Warehouse y minería de datos, trabajando de manera individual o colaborativa.
Instrucciones	1. Seleccionar una herramienta ETL 2. Una vez seleccionada, realice la selección deberá investigar el uso de la herramienta 3. Elabore una presentación con la información recolectada 4. Realice un ejemplo práctico del uso de la herramienta 5. Finalmente, deberá compartir la presentación en el espacio de tarea en canvas y realizar una exposición durante el desarrollo de clase práctica
Fecha de entrega	Durante sesión práctica
Instrumento de evaluación	Buzón de Tareas
Criterios de evaluación	• Trabajo colaborativo (2.0 puntos) • Participación y ejemplos prácticos (4.0 puntos) • Fundamentación y dominio del tema (4.0 puntos)
Ponderación	100% Laboratorio I

Recursos Complementarios		
Recurso	**Título**	**Cita Referencial**
Sitio Web	ETL vs ELT: ¿Cuál es la diferencia?	https://bit.ly/3aHZ26V
Sitio Web	¿Qué es ELT y cuáles son sus diferencias con ETL?	https://blog.bismart.com/es/que-es-elt-diferencias-etl
Video	Data Warehouse de POWER BI	https://www.youtube.com/watch?v=iynUAMKLxkE

¿Preguntas?

¡Muchas gracias!

UNIVERSIDAD GERARDO BARRIOS
Líderes en Gestión del Conocimiento

Comisión de Acreditación de la Calidad de la Educación Superior
UNIVERSIDAD GERARDO BARRIOS (UGB)
ACREDITADA
2019- 2024