

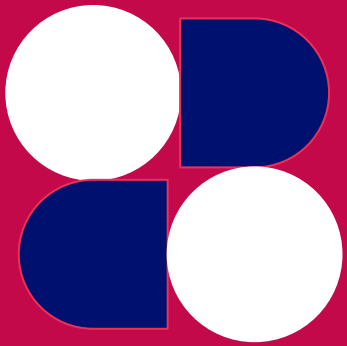


Programa PUCP DATA DRIVEN

Formación tecnológica

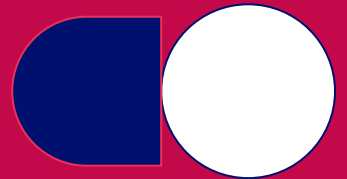
Dirección Académica de Planeamiento y Evaluación
Dirección de Gestión del Talento Humano





SESIÓN 3:

Introducción a arquitectura de ELT





Al finalizar la sesión, lograrás:

- Entender los conceptos básicos de ELT, su origen y su aplicación en la gestión de datos.
- Diferenciar entre ETL (Extract, Transform, Load) y ELT (Extract, Load, Transform).
- Identificar cuándo utilizar cada enfoque.
- Identificar buenas prácticas para implementar procesos ELT.



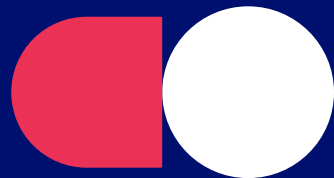
Contenido de la sesión:



- 1 Presentación del profesor, interacción con participantes y reglas de juego.
- 2 Introducción a ELT
- 3 Buenas Prácticas
- 4 Ejemplos



2. Introducción a ELT





INTRODUCCIÓN A ELT



Origen

- Big Data
 - Producción y recopilación de datos a un ritmo sin precedentes.
 - Muchas fuentes de datos en diferentes estructuras.
 - Estructurada, Semi Estructurada, No Estructurada.
 - Falta de análisis sobre los datos para generar valor.





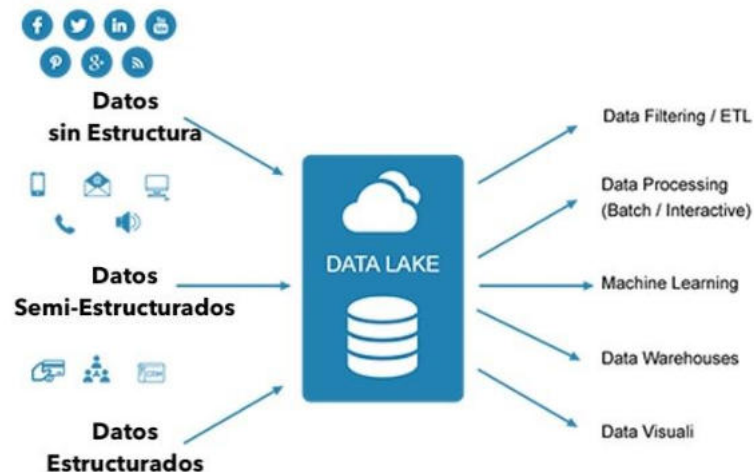
INTRODUCCIÓN A ELT



Origen

DATALAKE

- Repositorio de datos crudos.
- Estructurado y no estructurados.
- Sin esquema -> “Schema on read”
- **ELT**



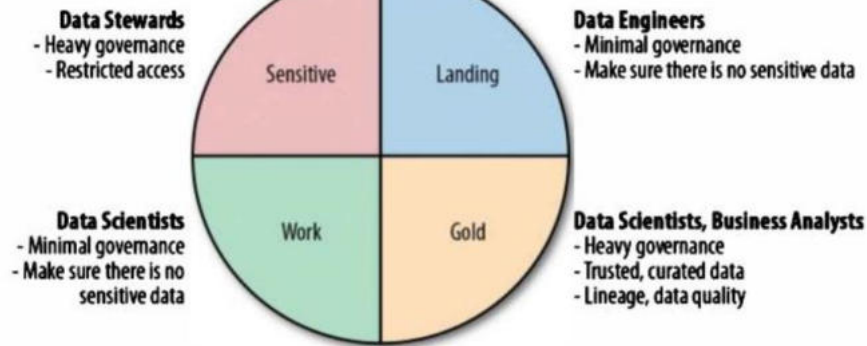
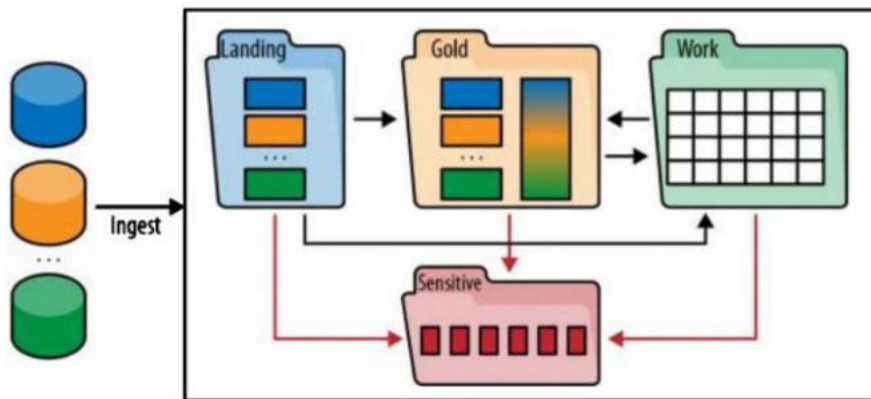


INTRODUCCIÓN A ELT



Origen

DATALAKE - ELT



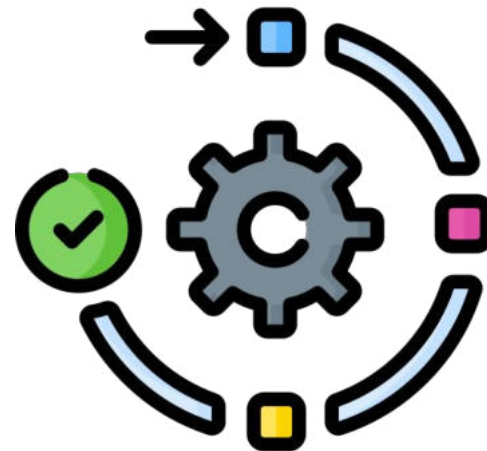


INTRODUCCIÓN A ELT



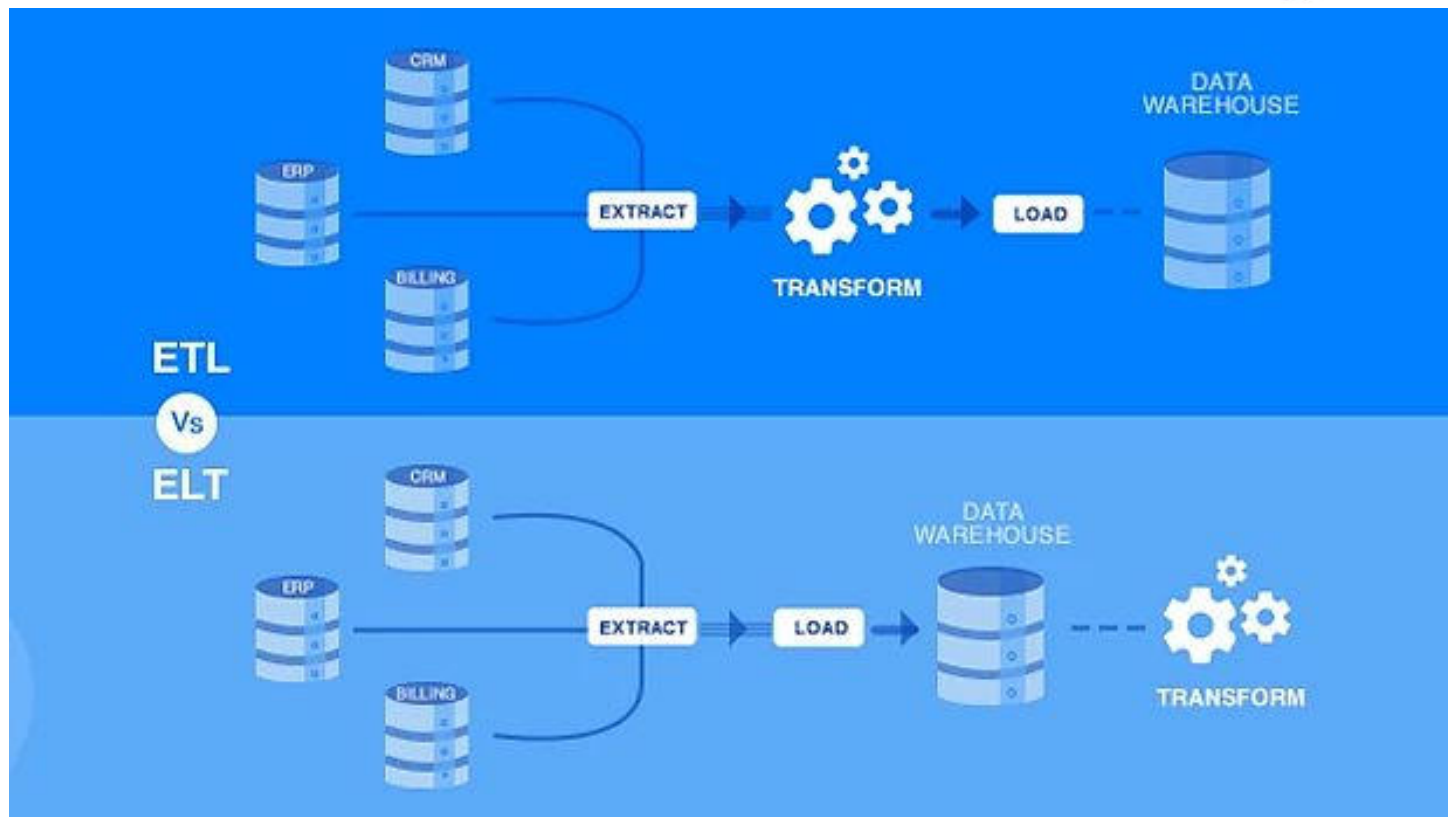
Definición

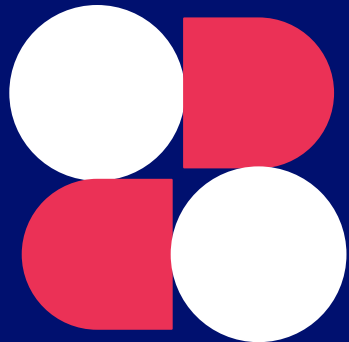
- **Extract (Extracción):** Proceso de extracción de datos desde una o más fuentes, como bases de datos, archivos planos, aplicaciones de terceros, etc.
- **Load (Carga):** Los datos extraídos se cargan directamente en un data warehouse o un ***data lake*** en su formato bruto (sin transformaciones significativas).
- **Transform (Transformación):** Las transformaciones se realizan dentro del entorno de almacenamiento (por ejemplo, en una base de datos o data warehouse), utilizando el poder de procesamiento del sistema de almacenamiento.



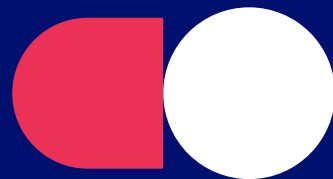


INTRODUCCIÓN A ELT





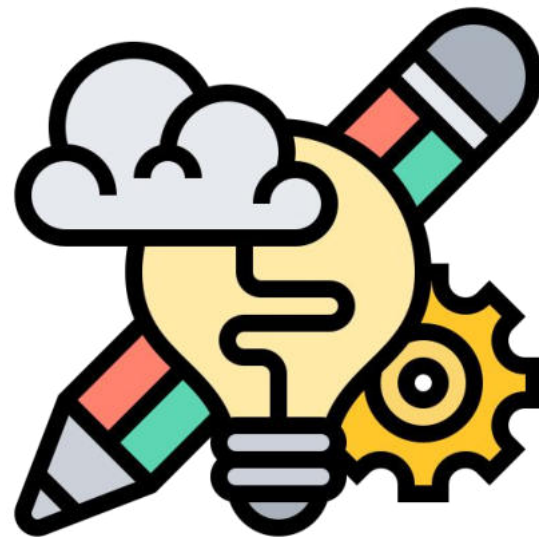
3. Buenas prácticas





Diseño Eficiente – Recursos

1. **Evaluación de las necesidades de datos**
 - Identificar los tipos de datos a procesar
 - Identificar fuentes de datos
2. **Selección de Herramientas para cada fase**
 - Una herramienta para cada letra (E L T)
 - Contratar una *suite* integradora de servicios (AWS, Google, etc)
3. **Consideración de costos**
 - Tecnológicos
 - Tiempo
 - Dinero





Diseño Eficiente – Rendimiento

1. Orden de Transformaciones

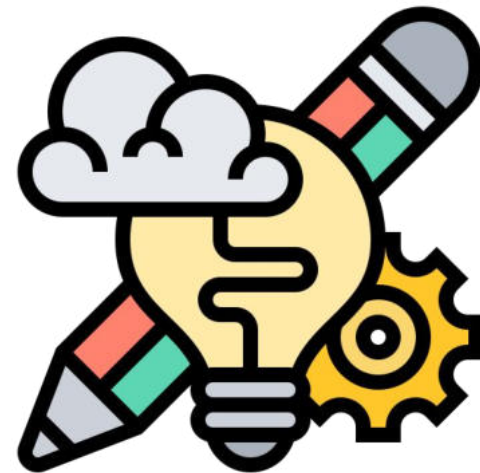
- Priorizar las transformaciones que filtran o reducen significativamente el volumen de datos antes de aplicar operaciones más costosas

2. Transformaciones Incrementales

- Procesar solo los datos nuevos o modificados

3. Particionar

- Dividir grandes volúmenes de datos en particiones más pequeñas y manejables, basadas en criterios como fecha, ubicación, o tipo de datos





Diseño Eficiente – Seguridad

1. Cifrado de Datos

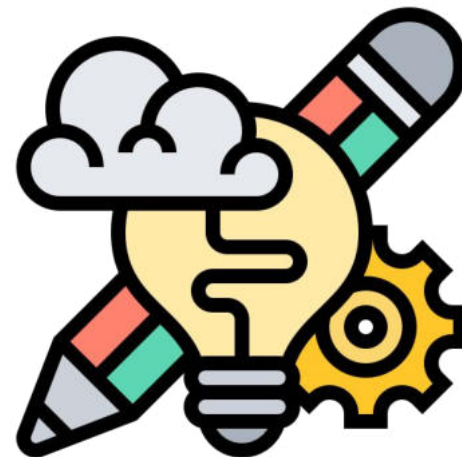
- Aplicar cifrado tanto en tránsito como en reposo para proteger los datos sensibles

2. Control de Acceso Basado en Roles

- Definir roles y permisos para controlar quién puede acceder y manipular diferentes tipos de datos en cada fase del proceso ELT

3. Auditoría y Registro de Actividades

- Implementar registros detallados de todas las actividades del proceso ELT para rastrear el acceso y cambios a los datos





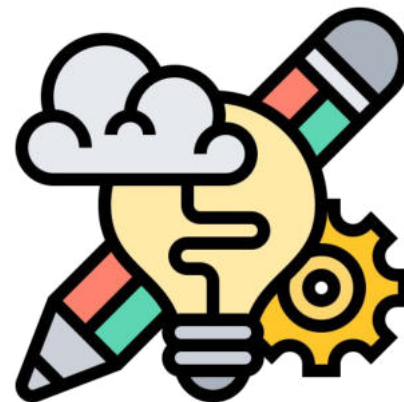
Diseño Eficiente – Gobernanza

1. Políticas de Retención y Eliminación de Datos

- Establecer políticas claras de tiempo de eliminación y backup

2. Trazabilidad y Lineage de Datos

1. Mantener un registro claro de la procedencia de los datos y las transformaciones que han sufrido a lo largo del proceso ELT





BUENAS PRÁCTICAS



Diseño Eficiente – Otros

1. Validación de Datos

- Implementar validaciones de datos en cada fase del proceso ELT

2. Manejo de fallos

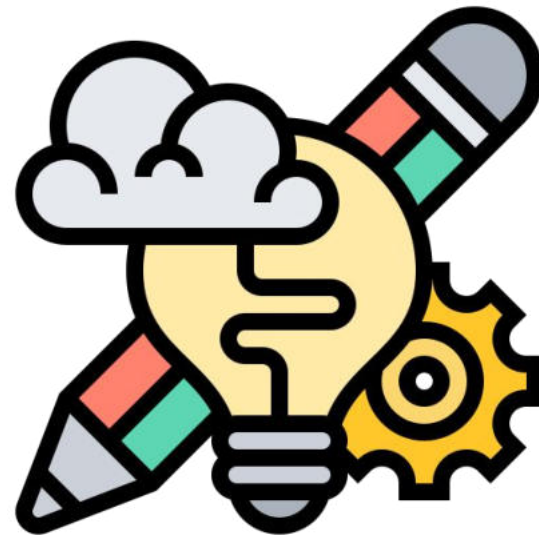
- Implementar mecanismos de recuperación ante fallos, como la reejecución de cargas o transformaciones fallidas sin necesidad de repetir todo el proceso

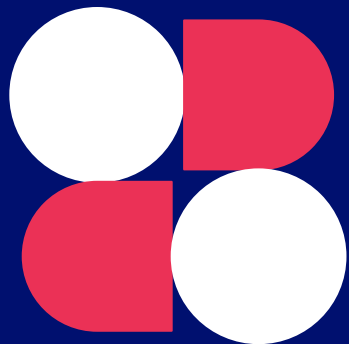
3. Automatización con Flujos de Trabajo

- Utilizar herramientas de orquestación de flujos de trabajo para automatizar y programar cargas y transformaciones de datos

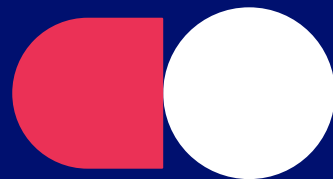
4. Monitorización

- Implementar monitorización en tiempo real del proceso





4. Ejemplos





Ejemplos

- I. Las universidades manejan grandes volúmenes de datos de inscripciones y matrículas de estudiantes en cada periodo académico. Estos datos suelen provenir de múltiples fuentes como sistemas de inscripción en línea, registros de papel, y bases de datos internas.
- I. Seguimiento del rendimiento académico de los estudiantes a lo largo de sus estudios. Esto incluye notas de exámenes, proyectos, y evaluaciones continuas, provenientes de diferentes sistemas de gestión académica.
- I. Monitoreo y optimización del uso de recursos como aulas, equipos de laboratorio, y recursos digitales (licencias de software, bases de datos de investigación).



BIBLIOGRAFÍA



- Kimball, R., & Caserta, J. (2011). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Wiley.
- Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Wiley.
- Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2009). *Conceptual Modeling for ETL Processes*. Springer.
- Rainardi, V. (2008). *Building a Data Warehouse: With Examples in SQL Server*. Apress.



Gracias

