# Técnicas de Extracción de Datos en IBM DataStage

Conexión a fuentes de datos y mejores prácticas

**IBM Cloud / Cloud Pak for Data**

# La Extracción de Datos en el Flujo ETL

### Objetivos Principales

Obtener datos desde múltiples fuentes manteniendo la integridad absoluta de la información recopilada.

Minimizar el impacto en los sistemas de origen para asegurar la continuidad operativa del negocio.

### El Valor de DataStage

Ofrece una extracción altamente configurable y escalable, permitiendo preparar los datos de manera óptima para la transformación.

# Capacidades Críticas de DataStage

### Escalabilidad

Extracción de grandes volúmenes mediante procesamiento en paralelo masivo.

### Control

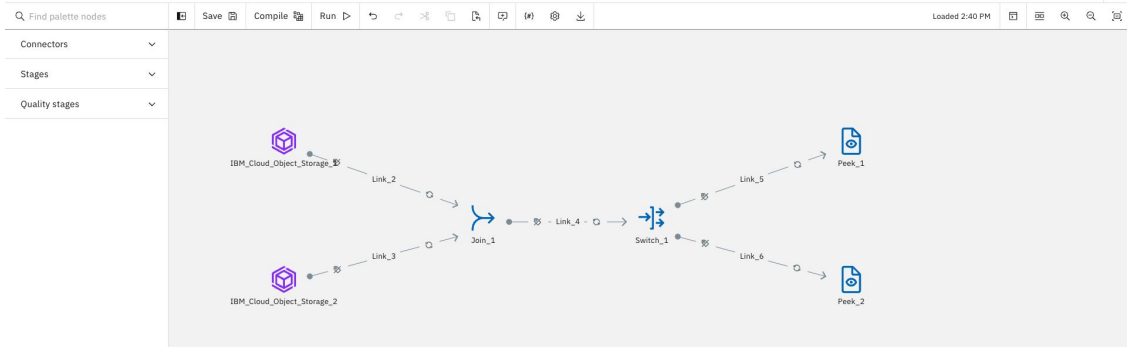
Capacidad para aplicar filtros avanzados desde el origen y manejar errores con precisión.

### Flexibilidad

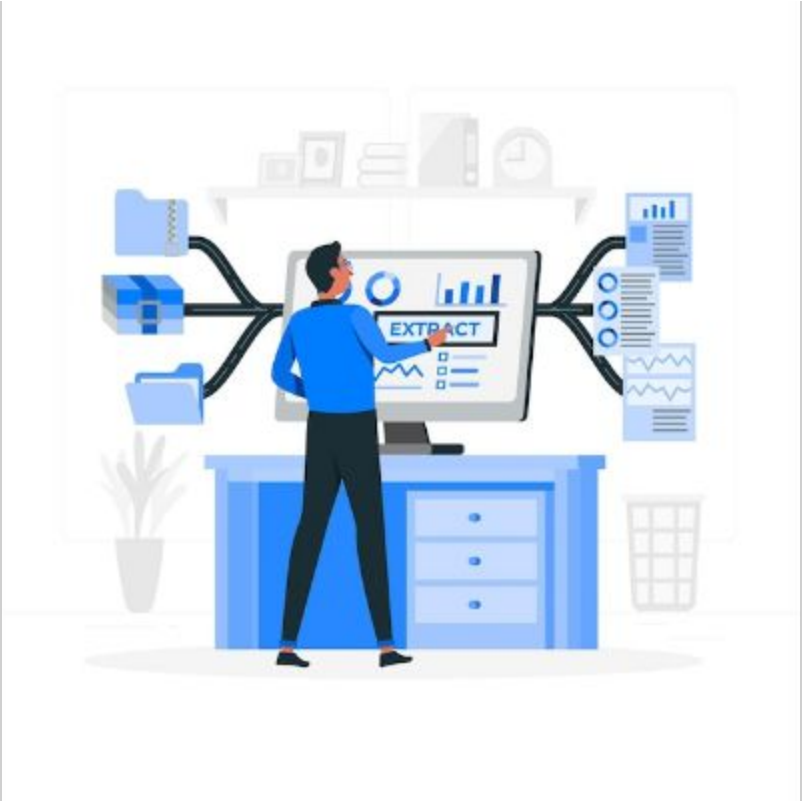
Conexión nativa a múltiples tipos de fuentes mediante conectores optimizados.

# Ecosistema de Conectores

- **Bases de Datos:** Oracle, DB2, SQL Server, PostgreSQL, MySQL.
- **Archivos:** Sequential File, File Set, Excel optimizados.
- **Cloud:** REST APIs, Cloud Object Storage (IBM COS), FTP/SFTP.
- **Big Data:** Gestión avanzada mediante conectores nativos.



# Extracción desde Bases de Datos




**Métodos Comunes**

- **SELECT Directo:** Uso de queries personalizadas para precisión.
- **Parametrización:** Consultas dinámicas basadas en variables.
- **Particionamiento:** División por clave para mejorar el rendimiento paralelo.

*"Optimiza el motor de la BD filtrando siempre en el origen."*

# Estrategias de Extracción Incremental

## Timestamp

Filtrado por fecha de última actualización para obtener solo registros modificados.

## Columnas de Control


Uso de flags o IDs secuenciales para identificar nuevos datos.

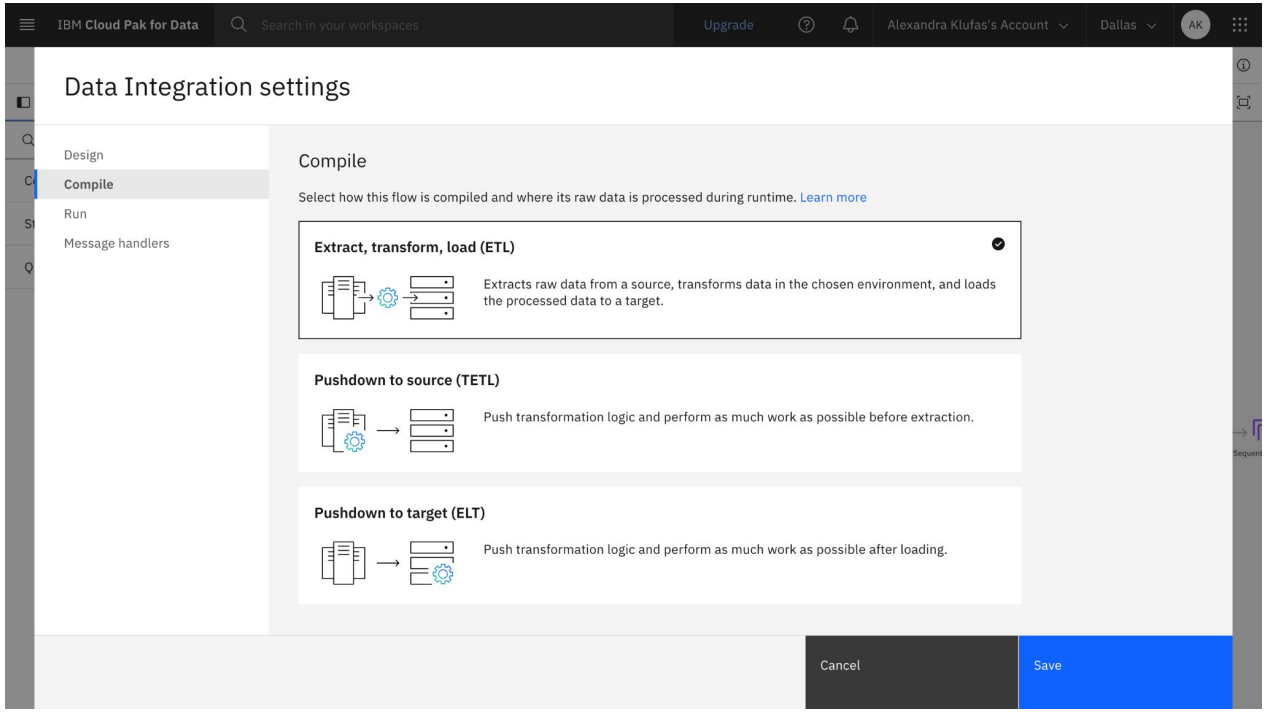
## Change Data Capture (CDC)

Captura de cambios en tiempo real sin sobrecargar el sistema origen.

# Manejo de Archivos Planos

## Tipos y Consideraciones

- **Formatos:** CSV, TXT, Delimitados y de longitud fija.
- **Codificación:** Validación estricta de UTF-8 y ASCII.
- **Validación:** Manejo de encabezados y layouts detallados.



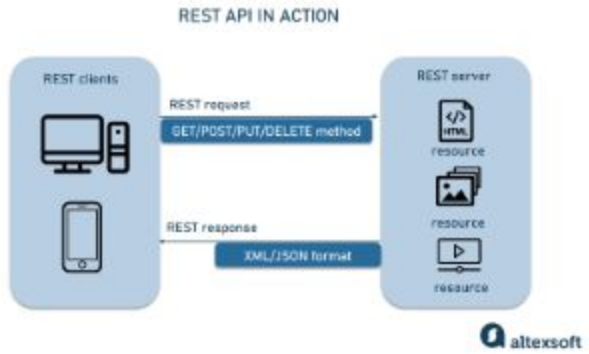
# Fuentes Modernas: APIs y Cloud

### Extracción desde APIs REST

Soporte nativo para autenticación OAuth/Token, paginación y formatos JSON/XML complejos.

### IBM Cloud Object Storage (COS)

Integración nativa para staging de datos, alta disponibilidad y escalabilidad automática en la nube.



# Rendimiento y Optimización

100%

**Paralelismo Máximo**

En IBM Cloud, el escalado mejora el rendimiento mediante el uso eficiente de nodos de computación distribuidos.

- Minimiza la latencia de red.
- Optimiza el tamaño de los buffers.

# Resumen de Mejores Prácticas

| Área | Acción Recomendada | Valor Agregado |
|---|---|---|
| Conexión | Usar conexiones parametrizadas | Reutilización y Seguridad |
| Bases de Datos | Evitar SELECT * / Usar índices | Velocidad y Bajo Impacto |
| Archivos | Validar estructura antes de procesar | Integridad de Datos |
| Seguridad | Encriptación en tránsito (SSL) | Cumplimiento y Privacidad |

# Casos de Uso Comunes

Cargas diarias transaccionales, integración con proveedores externos y procesamiento de archivos masivos en Cloud Pak for Data.

# ¿Dudas o Preguntas?

Técnicas de Extracción de Datos en IBM DataStage

**¡Gracias por su atención!**