Transformación de Datos con IBM DataStage

Diseño, lógica empresarial y técnicas avanzadas

IBM Cloud / Cloud Pak for Data

¿Qué es la Transformación de Datos?

La transformación es la fase crítica del ETL donde los datos adquieren valor:

- **Se limpian:** Eliminación de ruidos y errores.
- **Se validan:** Verificación de integridad.
- **Se convierten:** Ajuste de formatos y tipos.
- **Se enriquecen:** Agregado de información externa.

Objetivo: Convertir datos crudos en información útil y confiable para el negocio.

Rol de DataStage en la transformación

Visual

Transformaciones sin programación compleja mediante diagramas.

Paralelo

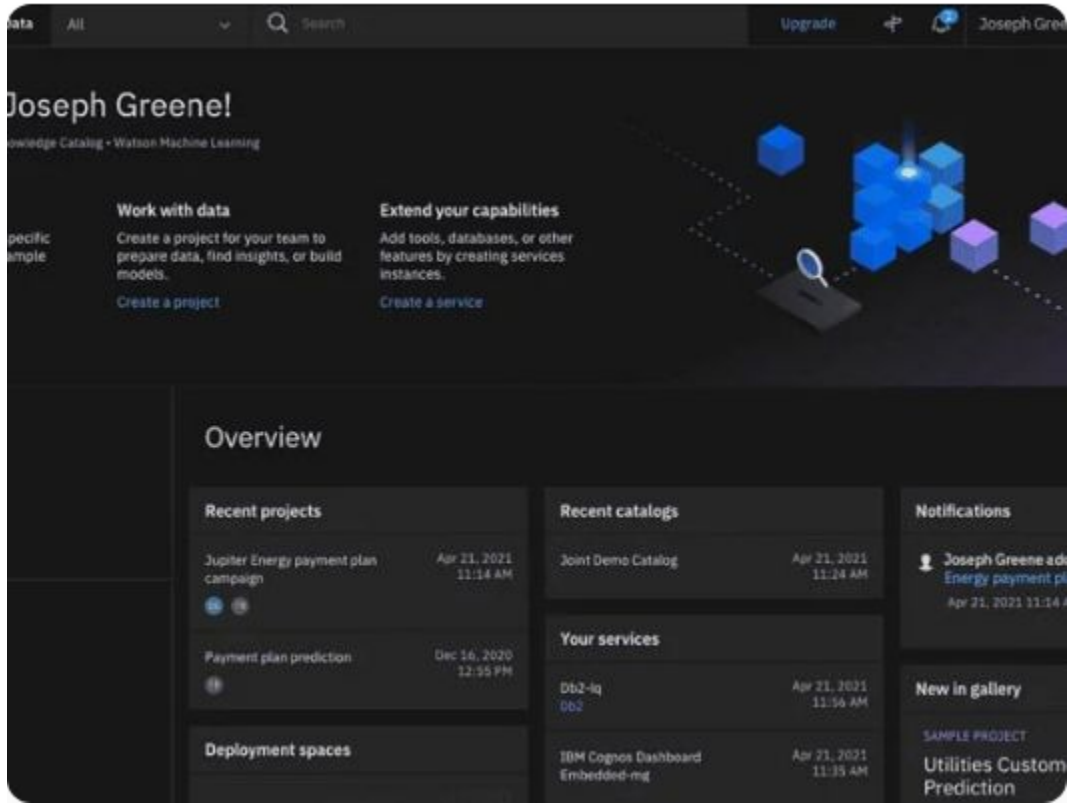
Procesamiento masivo de datos mediante motor paralelo.

Escalable

Manejo eficiente de grandes volúmenes y reglas complejas.

Todo se implementa mediante un diseño gráfico basado en el flujo de datos.

Comprensión del DataStage Designer



La herramienta central para el desarrollo ETL en IBM Cloud:

- **Diseño:** Crear Jobs y flujos visuales.
- **Configuración:** Definir lógica y parámetros.
- **Validación:** Compilar y verificar el proceso.

Accesible vía navegador web en Cloud Pak for Data.

Elementos del Designer

Palette

Lista de stages disponibles para arrastrar.

Canvas

Área central de diseño del flujo.

Links

Representan el movimiento de los datos.

Properties

Panel de configuración detallada.

Flujo básico en un Job

Fuente de Datos

Stages de Transformación

Validaciones y Limpieza

Destino Final

¿Qué es un Stage?

4

Categorías Base

Un Stage es un componente pre-construido que ejecuta una función específica dentro del flujo.

Entrada

Conexión a fuentes (Source).

Salida

Carga en destinos (Target).

Transform

Modificación de datos.

Process

Ordenamiento y unión.

Stages más utilizados

Transformer

Reglas de negocio y lógica compleja.

Filter

Filtrado selectivo de registros.

Join

Unión de múltiples fuentes de datos.

Lookup

Búsquedas en tablas de referencia.

Aggregator

Cálculos agrupados (Suma, Media).

Sort

Ordenamiento para procesamiento.

Logic Logic

Transformer Stage

El núcleo de las transformaciones:

- Crear columnas derivadas.
- Condiciones IF / THEN / ELSE.
- Manipulación avanzada de texto.
- Conversión de tipos y fechas.

Es el Stage más versátil y potente del Designer.

Transformaciones Comunes

- Conversión de formatos de fecha.
- Normalización de mayúsculas.
- Tratamiento de valores Nulos.
- Cálculo de tasas e impuestos.
- Estandarización de códigos.
- Concatenación de campos.

Implementación de Lógica Empresarial

La lógica empresarial define las reglas de supervivencia y calidad del dato:

Reglas de Negocio

Clientes activos solamente, cálculo de descuentos progresivos.

Validaciones

Formatos de identificación, rangos de edad, integridad referencial.

Expresiones en el Transformer

DataStage utiliza un lenguaje de expresiones potente:

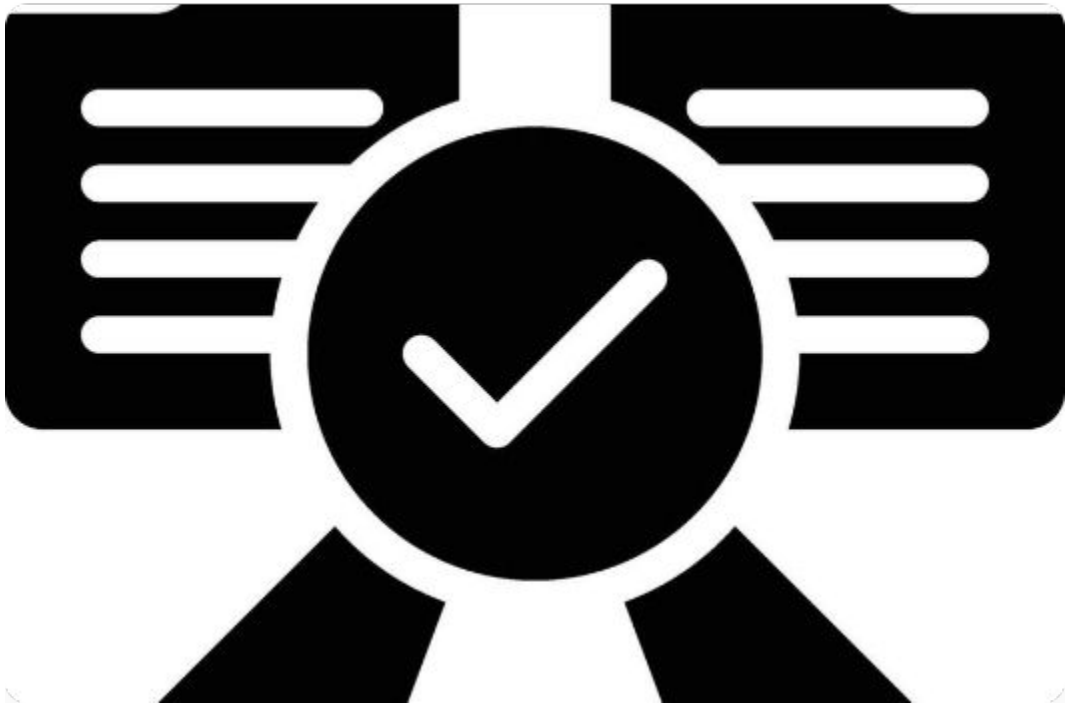
Logic: IF / ELSE, CASE.

Strings: UpCase, Trim, Field.

Dates: DateToString, DaysSince.

```
// Ejemplo Conceptual
IF edad >= 18 THEN
  "Adulto"
ELSE
  "Menor"
```

Manejo de datos inválidos



Estrategias para asegurar la calidad sin detener el flujo:

- **Reject Links:** Desviar registros erróneos.
- **Logging:** Registrar errores en el Job Log.
- **Continuidad:** Procesar datos correctos en paralelo.

Uso de Lookup Stage

Búsqueda rápida en memoria para enriquecer datos:

Enriquecimiento

Agregar nombres desde tablas maestras.

Surrogate Keys

Obtener claves sustitutas para el Warehouse.

Validación

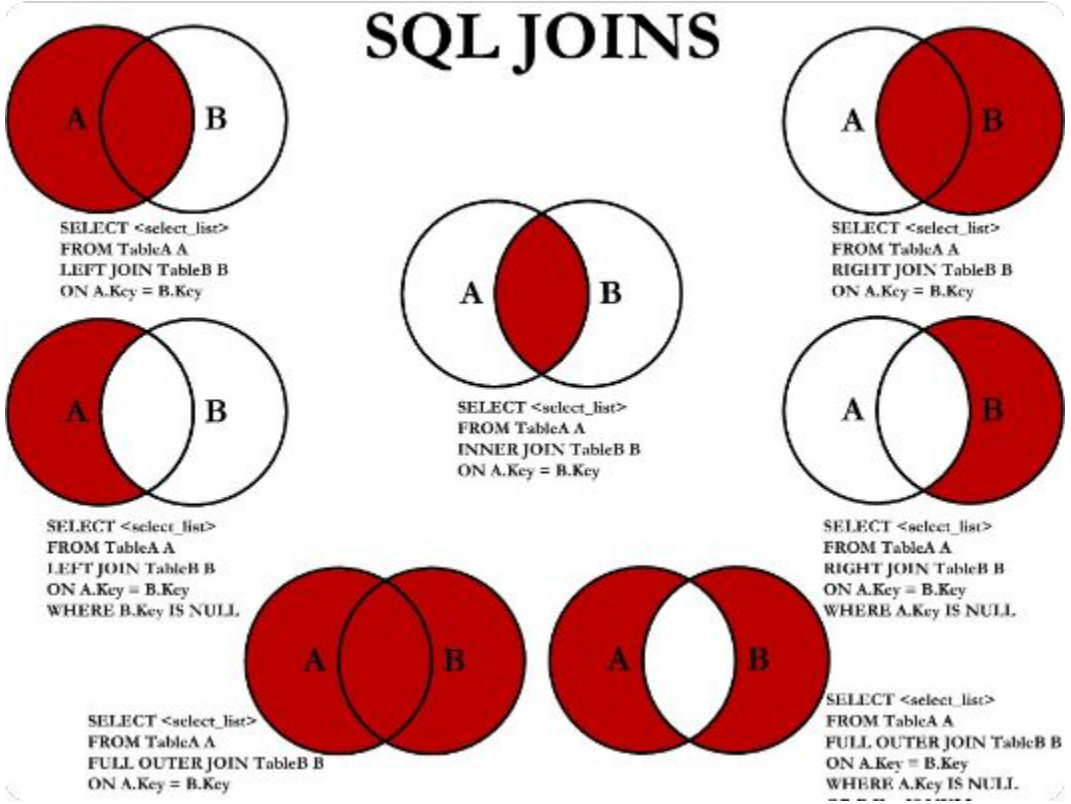
Verificar si un ID existe en el origen.

Join Stage

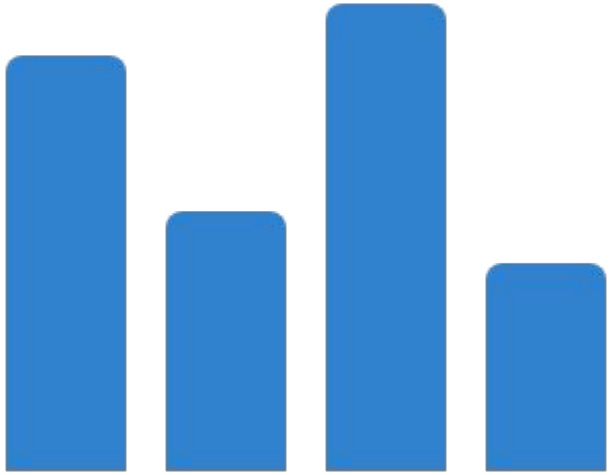
Permite combinar datasets masivos optimizado para paralelismo.

Tipos soportados:

- Inner Join
- Left Outer Join
- Right Outer Join
- Full Outer Join



Aggregator Stage



Consolidación de Datos

Realiza cálculos matemáticos agrupando por una o más claves (Group By).

Fundamental para la creación de reportes y tablas resumen en el Data Warehouse.

Técnicas Avanzadas

Particionamiento

Divide los datos para procesarlos en múltiples CPUs simultáneamente.

Pipelining

Los datos fluyen entre stages sin esperar a que el anterior termine.

Data Skew

Técnicas de balanceo para evitar cuellos de botella en nodos lentos.

Transformaciones de Alto Nivel

- **SCD (Slowly Changing Dimensions):** Manejo de historial de cambios en dimensiones.
- **Deduplicación:** Eliminación inteligente de registros duplicados.
- **Validaciones Cruzadas:** Lógica que depende de múltiples registros o fuentes.
- **Multinivel:** Procesos anidados y modularizados.

Uso de Parámetros y Variables

100%

Reutilización

Permiten desacoplar la lógica del entorno de ejecución:

Ejemplos comunes:

- Fecha de ejecución del proceso.
- Rutas de archivos de entrada/salida.
- Credenciales y conexiones a BD.

Optimización de Procesos

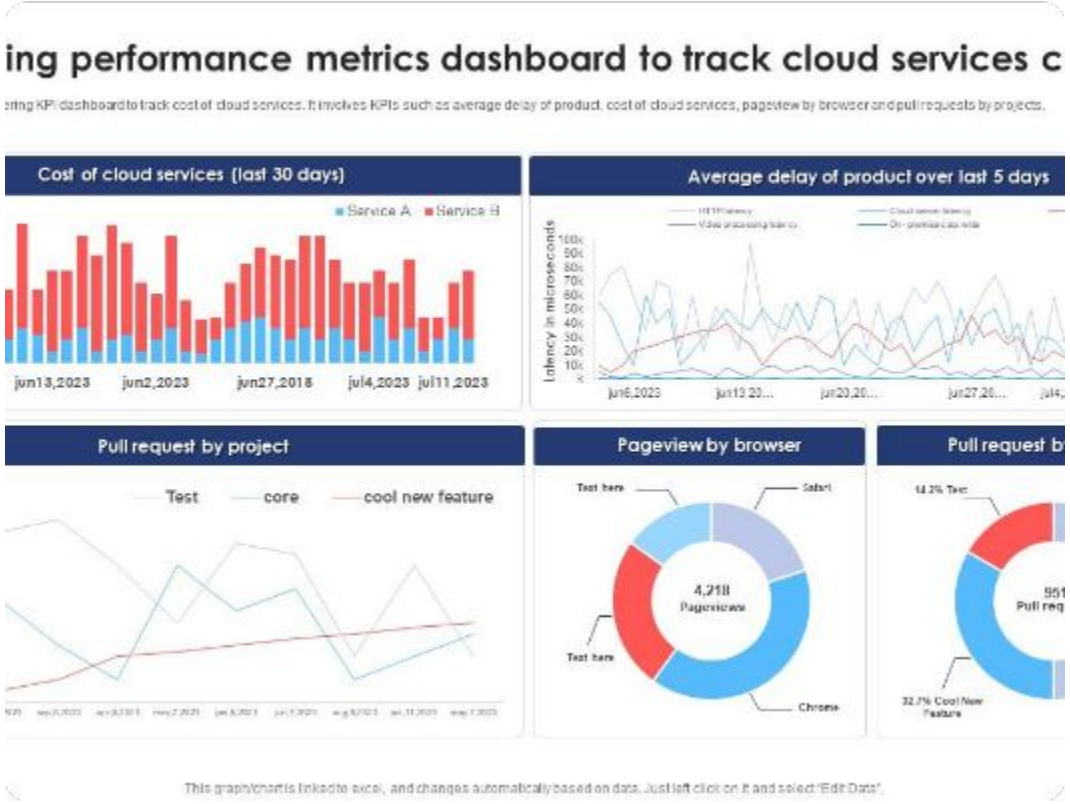
Filtrar pronto: Reduce el volumen de datos en los primeros pasos.

Lookup vs Join: Usar Lookup para tablas pequeñas que quepan en memoria.

Paralelismo: Configurar el número óptimo de particiones según el HW.

Evitar conversiones: No repetir conversiones de tipos innecesarias.

Rendimiento en IBM Cloud



Consideraciones clave para Cloud Pak for Data:

Auto-scaling: Escalado automático del cluster según carga.

Gestión de Memoria: Configuración eficiente de buffers.

Red: Minimizar el movimiento de datos entre regiones.

Buenas Prácticas Generales

Modularidad

Diseñar Jobs pequeños y reutilizables.

Nomenclatura

Nombres claros para stages y links.

Documentar

Anotar las reglas de negocio en el canvas.

Resumen Final

- El Designer permite un diseño visual potente y eficiente.
- El **Transformer Stage** es el corazón de la lógica.
- El paralelismo es la clave del rendimiento en DataStage.
- Las técnicas avanzadas aseguran la calidad empresarial.