Introducción a IBM DataStage

Panorama del proceso ETL y arquitectura

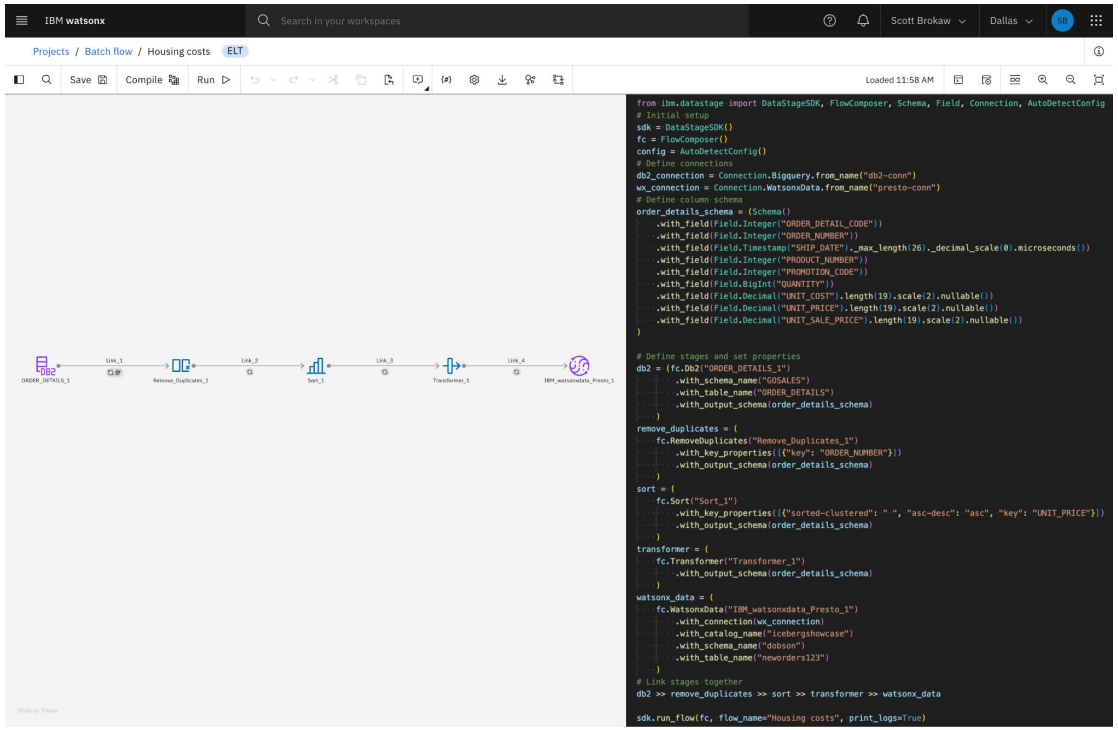
Expositor: Consultor Senior BI

Fecha: Febrero 2024

¿Qué es IBM DataStage?

IBM InfoSphere DataStage es una herramienta líder en integración de datos para:

- Diseñar procesos ETL complejos.
- Integrar múltiples fuentes heterogéneas.
- Cargar Data Warehouses y Data Lakes.
- Soporta procesamiento paralelo masivo.



¿Para qué se utiliza?

Integración Empresarial

Consolidación de datos de toda la organización en un punto único de verdad.

Migración de Datos

Movimiento seguro de información entre sistemas legacy y plataformas modernas.

Business Intelligence

Construcción de bases sólidas para reporting, analítica avanzada y Big Data.

Concepto Fundamental: ETL

E Extract (Extraer)

T Transform (Transformar)

L Load (Cargar)

Proceso clave para convertir **datos crudos** en **información confiable** para la toma de decisiones empresariales.

Fases del Proceso ETL

1. Extracción

- Bases de Datos (Oracle, DB2)
- Archivos (CSV, JSON)
- APIs y CRM/ERP

2. Transformación

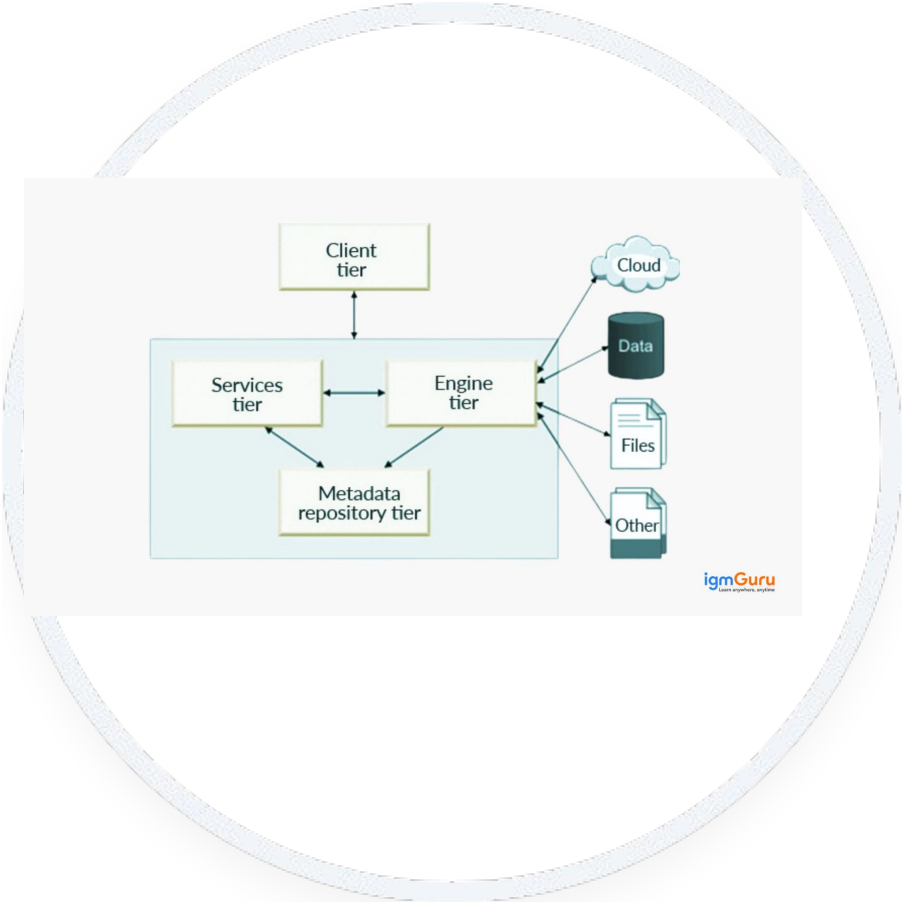
- Limpieza y Calidad
- Reglas de Negocio
- Uniones y Cálculos

3. Carga

- Data Warehouses
- Data Marts
- Sistemas Destino

Flujo Lógico del Proceso

El Rol de DataStage



- **Diseño Gráfico:** Interfaz visual para procesos complejos.
- **Procesamiento Paralelo:** Optimización de recursos.
- **Automatización:** Programación de cargas recurrentes.
- **Monitoreo:** Control total de la ejecución.

Arquitectura

Componentes, Motor y Repositorio

Vista Lógica de la Arquitectura

El flujo de información en la infraestructura:

Usuarios
↓
Cliente DataStage
↓
Servidor DataStage
↓
Engine (Motor)
↓
Origen/Destino

Componentes Principales

1. Capa Cliente

Interfaz gráfica para diseño de Jobs, configuración de parámetros y etapas.

2. Capa Servidor

Administración de la ejecución, seguridad y gestión de conexiones concurrentes.

3. Engine (Motor)

Núcleo que ejecuta las transformaciones con procesamiento paralelo optimizado.

Tipos de Jobs

Parallel Jobs

Diseñados para alto rendimiento. Dividen la carga en múltiples nodos de procesamiento.

Server Jobs

Arquitectura secuencial más antigua. Útiles para procesos ligeros o legados.

Sequence Jobs

Orquestadores que controlan el flujo y las dependencias entre múltiples jobs.

Unidades de Trabajo

- **Jobs:** La unidad básica de ejecución ETL.
- **Stages:** Bloques de función específica (Join, Sort, etc).
- **Links:** Conexiones que transportan los datos entre Stages.
- **Repository:** Almacén central de metadatos y versiones.
- **Director:** Herramienta de monitoreo y logs.

¿Qué es un Stage?

Componentes modulares dentro de un Job que ejecutan acciones sobre los datos:

- Sequential File
- Transformer
- Aggregator
- Join
- Lookup
- DB Connector

Stage: Transformer

Es el corazón de la lógica de negocio en DataStage.

Permite:

- Cálculos matemáticos y lógicos.
- Aplicación de condiciones IF/THEN.
- Creación de columnas derivadas.
- Conversión de tipos de datos complejos.

Combinación de Datos

Join Stage

Combina múltiples flujos de grandes volúmenes. Equivale al JOIN de SQL. Requiere datos ordenados.

Lookup Stage

Ideal para validar contra tablas de referencia en memoria. Optimizado para "Left Outer Joins".

Metadatos: Repository

El repositorio central almacena:

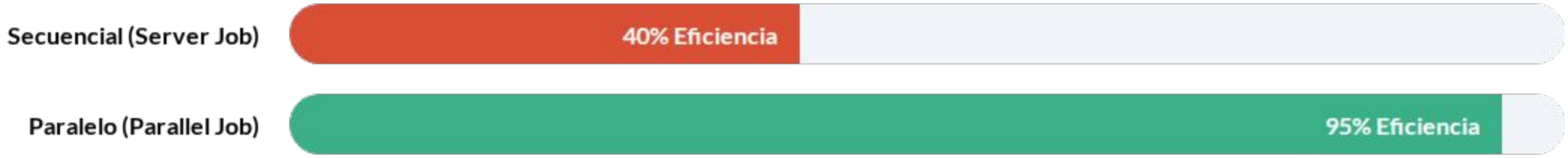
- Definiciones de Tablas.
- Diseños de Jobs y Rutinas.
- Control de Versiones.
- Parámetros Compartidos.

Operación: Director

Herramienta fundamental para el equipo de Operaciones:

- Ejecución manual o programada.
- Análisis profundo de logs.
- Diagnóstico y resolución de errores.
- Reinicio selectivo de procesos.

Procesamiento Paralelo



DataStage utiliza **Particionamiento de Datos** para ejecutar tareas simultáneas, garantizando escalabilidad lineal en entornos de Big Data.

¿Por qué elegir DataStage?

- **Escalabilidad:** Crece con tus volúmenes de datos.
- **Conectividad:** Soporte nativo para nubes y Big Data.
- **Estabilidad:** Estándar de la industria financiera.
- **Gobernanza:** Control centralizado de metadatos.

Líder en el Cuadrante de Gartner

Reconocido mundialmente por su robustez.

Conclusiones

✔ Herramienta ETL empresarial robusta.

✔ Arquitectura escalable Cliente-Servidor-Motor.

✔ Dominio del paralelismo para grandes volúmenes.

✔ Basado en Jobs, Stages y Repositorio.

¿Preguntas?

Muchas gracias por su atención.

www.grupcief.com