

Carga e Integración de Datos con IBM DataStage

Carga, calidad de datos y manejo de errores

Fecha: Febrero 2024

Rol de DataStage en la Carga de Datos

- **Cargar grandes volúmenes:** Arquitectura MPP para procesamiento masivo.
- **Integrar múltiples fuentes:** Conectores nativos para DB2, Oracle, Snowflake, AWS.
- **Ejecución paralela:** Distribución de carga entre nodos para máximo rendimiento.
- **Validar datos:** Verificación de esquemas antes de la inserción.
- **Manejar errores:** Lógica automatizada de reintentos y rechazos.

Sistemas Objetivo (Targets)



Almacenes

Data Warehouses relacionales y
Bases de Datos relacionales
(RDBMS).



Cloud Lakes

Data Lakes y Cloud Object
Storage (S3, IBM COS, Azure
Blob).



Enterprise

Sistemas ERP/CRM y
Plataformas analíticas
avanzadas.

Tipos de Carga de Datos



Carga Completa

Reemplaza todos los datos existentes. Ideal para cargas iniciales o tablas maestras pequeñas.



Carga Incremental

Solo procesa registros nuevos o modificados (Delta). Optimiza tiempo y recursos.



Tiempo Real

Integración casi inmediata mediante Change Data Capture (CDC) para procesos críticos.

| Stages Utilizados para Carga

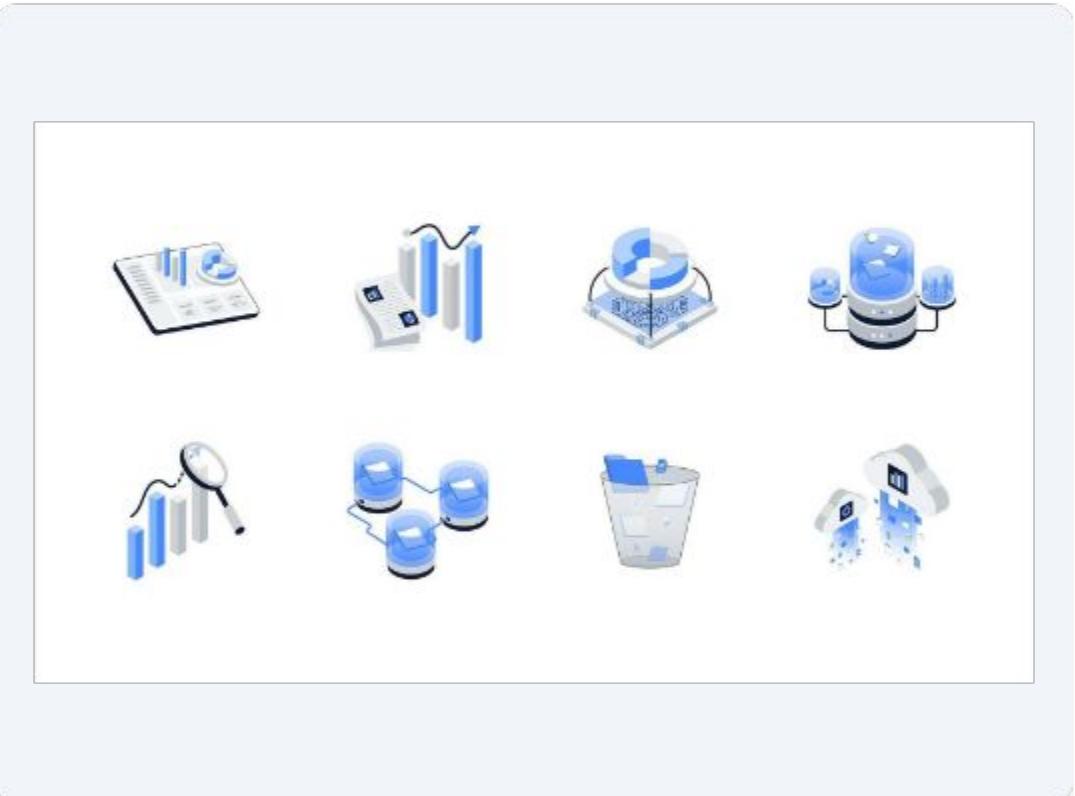
Database Connector: Conectividad nativa optimizada.

ODBC Connector: Estándar para bases de datos diversas.

Sequential File: Escritura en archivos planos/CSV.

Data Set: Almacenamiento interno de alta velocidad.

Cloud Storage: Integración con IBM Cloud Object Storage.



Métodos de Escritura en Bases de Datos

Operaciones Estándar

Insert: Agregar nuevos registros.

Update: Modificar registros existentes.

Delete: Eliminar datos obsoletos.

Estrategias Avanzadas

Upsert: Lógica inteligente de Insert o Update.

Bulk Load: Carga masiva omitiendo logs.

"La elección depende del volumen y las reglas de negocio."

| Carga Masiva (Bulk Load)

10x

Incremento de Velocidad

Uso de utilidades nativas de la base de datos (DB2 Load, SQL Loader).

Logging reducido: Menor sobrecarga en el sistema destino.

Ideal para grandes volúmenes de datos históricos.

Beneficio Cloud: Rendimiento extremo en entornos distribuidos.

Integración de Datos

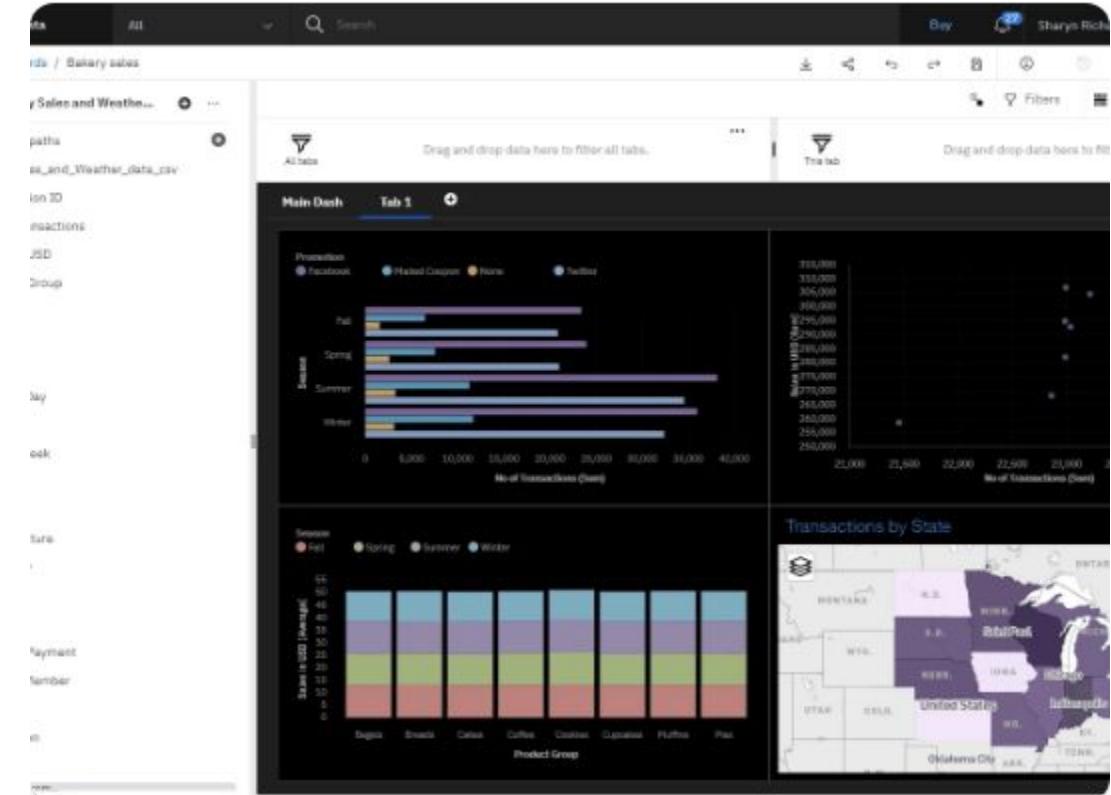
DataStage actúa como plataforma central de integración:

Combinar datos de múltiples sistemas heterogéneos.

Unificar formatos y estándares de negocio.

Resolver duplicados mediante lógica avanzada.

Crear una Visión Única del negocio.



Garantía de Calidad de Datos

Validaciones críticas antes de persistir la información:



Tipos de datos correctos



Campos obligatorios (Not Null)



Detección de duplicados



Integridad referencial

Técnicas de Validación de Datos

Transformer Validations: Reglas condicionales integradas.

Lookup: Verificación contra tablas maestras.

Reglas Condicionales: Filtros de calidad granulares.

Rechazo de registros: Separación de datos inválidos en tiempo de ejecución.



Integridad de Datos

Integridad de Entidad

Garantizar que las Claves Primarias (PK) sean únicas y no nulas.

Integridad Referencial

Relaciones válidas entre tablas (Padre-Hijo) mediante Claves Foráneas (FK).

Integridad de Dominio

Validar que los valores se encuentren dentro de los rangos o listas permitidas.

Manejo de Duplicados

Estrategias en DataStage

Aggregator Stage: Consolidación por claves.

Remove Duplicates: Limpieza directa de registros idénticos.

Hash Keys: Comparación rápida mediante algoritmos de hash.

El manejo correcto evita inconsistencias críticas en el **Data Warehouse** y garantiza la veracidad de los reportes.

Manejo de Errores en DataStage

- ! **Capturar errores por registro:** Sin detener el proceso completo.
- ⤵ **Redirigir registros:** Uso de "Reject Links" para análisis posterior.
- ▶ **Continuidad:** Configuración de umbrales para continuar la ejecución.
- 📄 **Logs detallados:** Mensajes técnicos precisos para depuración.

Tipos de Errores Comunes

Datos

Formato incorrecto, valores nulos no permitidos, violación de tipos.

Técnicos

Conexión fallida, Timeouts de base de datos, problemas de red/firewall.

Lógica

Transformaciones incorrectas, errores en expresiones o joins fallidos.

Estrategias de Manejo de Errores

Reject Links: Implementar siempre salidas de rechazo en conectores.

Tablas de Auditoría: Registrar el error y el registro original.

Notificaciones: Alertas automáticas vía Email o Slack ante fallos críticos.

Reintentos: Configuración de reintentos automáticos para errores temporales.

Registro y Monitoreo (Logging)

DataStage registra eventos vitales:

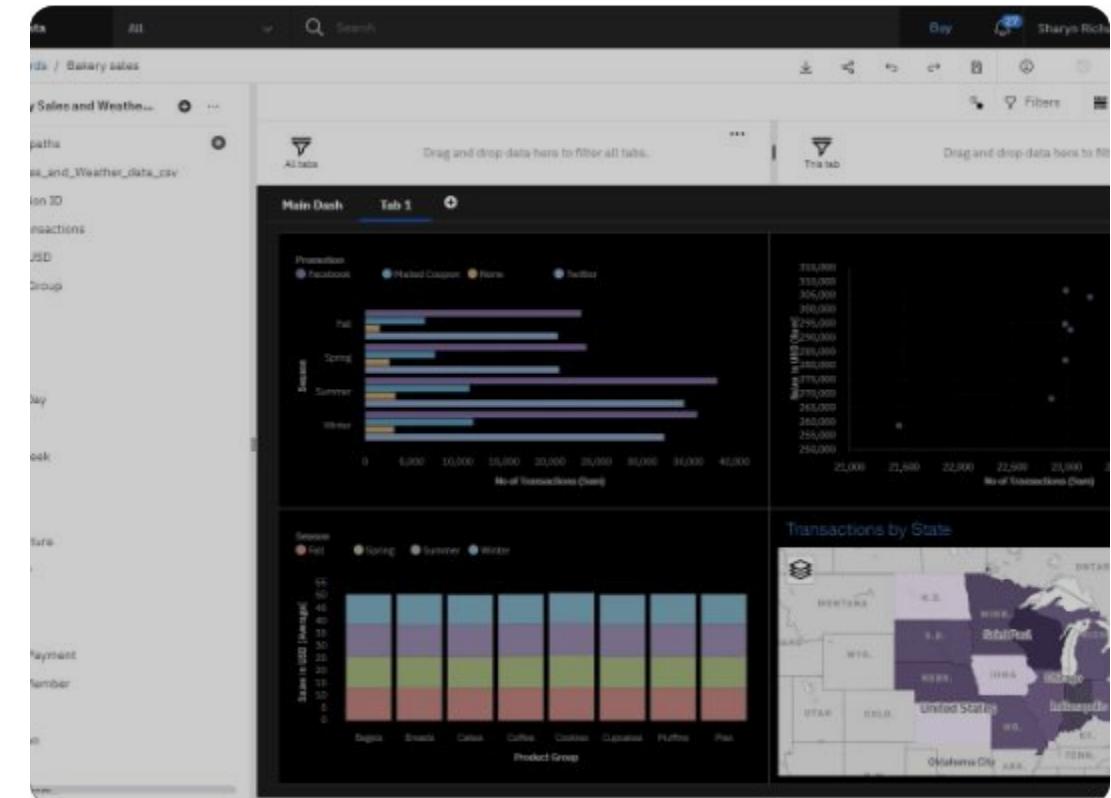
Timestamp de inicio y fin del Job.

Contador de filas procesadas por link.

Advertencias (Warnings) y Errores fatales.

Métricas de rendimiento de CPU y memoria.

Consola Web: Centralización de logs en Cloud Pak for Data.



Auditoría de Cargas

Métrica de Auditoría	Propósito
Registros Leídos	Verificar la integridad de la fuente.
Registros Cargados	Confirmar el éxito de la inserción en destino.
Registros Rechazados	Cuantificar la pérdida de calidad de datos.
Tiempo de Ejecución	Monitorear el SLA y rendimiento del pipeline.

Control Transaccional

Commit por Lotes

Define cada cuántas filas se confirma la escritura.

Balance entre velocidad y seguridad.

Rollback Automático

Ante un error fatal, deshace los cambios para evitar cargas incompletas o corruptas.

Optimización de Cargas en IBM Cloud

- ✓ **Usar paralelismo:** Configurar el grado de particionamiento adecuadamente.
- ✓ **Implementar Bulk Load:** Minimizar el overhead de red y logging.
- ✓ **Ajustar tamaño de Commit:** Optimizar el rendimiento de IOPS.
- ✓ **Staging intermedio:** Usar archivos temporales en COS para transformaciones pesadas.

Buenas Prácticas Generales



Validar antes de cargar para proteger el destino.



Parametrizar conexiones para portabilidad entre entornos.



Documentar reglas de carga y lógica de negocio.

Casos de Uso Reales



Data Warehouse

Carga diaria de billones de transacciones financieras.



MDM

Consolidación de clientes (Visión 360) de múltiples fuentes.



Reporting

Actualización de tableros analíticos en tiempo real para ejecutivos.