IBM DataOps

IBM DataStage

Provea datos preparados para su negocio en tiempo real
y para la IA con IBM Cloud Pak® for Data DataStage

Cómo proveer datos preparados para su negocio mediante la integración de datos

En la actualidad, las empresas digitales generan y consumen datos con una rapidez nunca vista. Estos datos contienen información sobre clientes, transacciones y empleados, los cuales acaban almacenados en múltiples sistemas y repositorios. Estos almacenes de datos se encuentran distribuidos en diferentes entornos multicloud y nube híbrida, así como en data lakes. Por este motivo, las organizaciones tratan de encontrar formas de unificar estas fuentes y entornos tan dispares con el fin de obtener información con mayor rapidez mediante **el uso de IA** y ofrecer experiencias diferenciadas y personalizadas a sus clientes. Según un estudio realizado por Forrester, los científicos de datos dedican el 80% de su tiempo a preparar y gestionar los datos para sus **iniciativas de IA**. Dichos resultados, junto con los de una encuesta elaborada por IBM, la cual reveló que el 91% de las organizaciones no utilizan sus datos de manera eficaz, indican que las empresas se esfuerzan por generar valor a partir de los silos de datos. El conjunto de técnicas, prácticas y herramientas de arquitectura de software que se emplean para obtener acceso a datos en tiempo real a partir de grandes cantidades de datos y proveer datos listos para su uso se denomina integración de datos. Con una tecnología de integración de datos flexible y escalable, las empresas pueden realizar analíticas para encontrar la mejor oferta, detectar y analizar la rotación de personal, hacer predicciones relacionadas con la cadena de suministro y ejecutar una detección instantánea del fraude mediante la extracción, transformación y carga de datos (ETL) en múltiples fuentes de datos.

IBM® InfoSphere™ DataStage, una solución de integración de datos líder en el mercado que ofrece capacidades de datos preparados para el negocio fiables y que van más allá del ETL, proporciona a los altos ejecutivos, arquitectos o líderes de operaciones de empresas que tienen dificultades para gestionar datos a través de múltiples nubes o data lakes y buscan reducir el tiempo necesario para desarrollar y actualizar los modelos y aplicaciones de IA, una solución escalable de integración y entrega de datos multicloud para garantizar que la información fiable y preparada para el negocio se utilice en tiempo real. Entre las principales capacidades de DataStage se incluye soporte de tiempo de ejecución multicloud para utilizar el diseño una sola vez y ejecutarlo en cualquier nube, al mismo tiempo que permite escalar las cargas de trabajo mediante el balanceo automático de las mismas y un motor paralelo de baja latencia. Además, también permite la disponibilidad de datos en tiempo real con tecnología de replicación incorporada, reducción del tiempo y los costes de DevOps con soporte para integración y entrega continuas (CI/CD), mayor rapidez en el desarrollo de modelos de IA con diseño de integración autónoma y reglas de validación para detectar y resolver automáticamente problemas de datos mediante la utilización de la calidad de los datos en línea.

DataStage forma parte de las capacidades de IBM DataOps con el fin de poner en funcionamiento datos continuos y de alta calidad para habilitar la IA y proporcionar una tubería de datos automatizada y de autoservicio a las personas correctas y en el momento adecuado, desde cualquier fuente de datos. IBM InfoSphere DataStage está disponible de forma local, en IBM Cloud™ y en plataformas hiperconvergentes como IBM® Cloud Pak™ for Data, que permiten realizar despliegues en cualquier lugar. IBM® Cloud Pak™ for Data es una plataforma de datos e inteligencia artificial totalmente integrada y desarrollada sobre Red Hat® OpenShift®, que ofrece una arquitectura de DataStage totalmente nativa de la nube que puede escalar con su negocio. También proporciona a las organizaciones una plataforma que soporta múltiples estilos de provisión de datos, entre los que se incluyen la integración, la replicación y la virtualización de datos, al mismo tiempo que los CDC capturan los cambios basados en registros a medida que se producen y entregan la información a las bases de datos de destino en la nube y los data lakes mediante colas de mensajes basadas en Kafka.

Diseñe una vez y ejecute en cualquier nube

Según un estudio de IDC, el 90% de los clientes empresariales utilizan múltiples nubes. Con la integración de datos multicloud, los usuarios pueden separar el diseño del tiempo de ejecución, es decir, pueden diseñar sus trabajos ETL una vez e implementar los componentes de tiempo de ejecución a través de contenedores en cualquier entorno de nube para reducir la latencia producida por el procesamiento de grandes volúmenes de datos. Puede crear y probar un trabajo de forma local y luego ejecutarlo en un entorno de nube, como una instancia de Microsoft Azure, mediante la utilización del lago de datos Azure en la nube. Los parámetros y valores del trabajo se pasan a una instancia remota de DataStage a través de un mensaje de Kafka.

La integración de datos multicloud ofrece las siguientes ventajas:

- la capacidad de integrar los datos de forma local y en entornos de nube;
- una experiencia de diseño de trabajos automatizados para simplificar el proceso de diseño;
- la ejecución de trabajos de forma remota para minimizar los costes de salida del traslado de los datos;
- el cumplimiento de requisitos geopolíticos;
- una reducción de la latencia para el procesamiento de grandes conjuntos de datos, ya que no es necesario transferir los datos a otro lugar.

Balanceo automático de cargas de trabajo y procesamiento paralelo

A través de una arquitectura totalmente nativa de la nube, puede utilizar contenedores locales o contenedores compartidos para DataStage para escalar sus cargas de trabajo de forma dinámica además de optimizar para grandes conjuntos de datos con un motor paralelo (PX) de última generación. Con IBM DataStage Flow Designer, los usuarios tienen la opción de crear trabajos en paralelo, en secuencia o de Apache Spark.

Se pueden ejecutar trabajos de DataStage Flow Designer en dos motores de tiempo de ejecución:

- Los trabajos de tipo paralelo o secuencial solo se pueden ejecutar en un motor paralelo. Generalmente, los trabajos que requieren muchos recursos se ejecutan en el motor paralelo y, por lo tanto, el tiempo medio para completar trabajos complejos mediante el procesamiento paralelo es de dos minutos.
- Los trabajos de tipo Apache Spark únicamente pueden ejecutarse en un motor Spark.

Provisión de datos en tiempo real

DataStage, en combinación con la tecnología Change Data Capture (CDC) para la captura en tiempo real desplegada como contenedores, puede proporcionar lo mejor de los mundos de la integración y la replicación de datos. DataStage permite realizar transformaciones complejas con grandes conjuntos de datos. A su vez, CDC captura los cambios basados en registros a medida que se producen, los transforma a través de transformaciones complejas y los provee a las bases de datos de destino en la nube y los data lakes mediante colas de mensajes basadas en Kafka. DataStage también permite que los trabajos de transformación de datos en lotes e incluso en grandes cantidades se alimenten en los almacenes de datos.

Reducción del tiempo y los costes de DevOps para soporte CI/CD

Para hacer frente al desafío de gestionar la cantidad de aplicaciones en contenedores distribuidas en diferentes sistemas operativos, las organizaciones necesitan disponer de una sólida herramienta de código abierto, como Red Hat OpenShift, disponible en Cloud Pak for Data. La plataforma Cloud Pak for Data permite escalar y suministrar contenedores para apoyar iniciativas clave de TI como microservicios y estrategias de migración a la nube. Los contenedores de DataStage permiten crear y automatizar pipelines de integración y entrega continuas (CI/CD) para los trabajos, desde la fase de desarrollo hasta la de prueba y producción, y permiten utilizar una tubería de CI/CD mediante herramientas de control código fuente como GitHub para publicar los trabajos con frecuencia y entregarlos a la producción de forma periódica.

Diseño de integración autónomo para alimentar la IA

Acelere la recopilación e integración de datos para la IA con mayor rapidez y a escala mediante el descubrimiento y la clasificación automática de los activos, la generación de flujos de integración basados en transformaciones personalizadas e integradas, así como la detección y protección de la información de carácter confidencial.

Rápida rentabilidad gracias a un diseño de trabajo automatizado

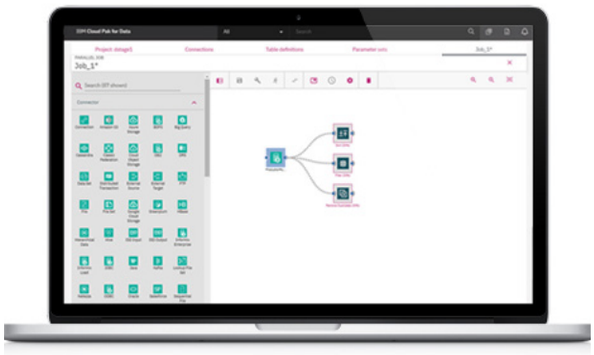

Figura 1. Datastage Flow Designer con capacidades de diseño automatizado.

IBM DataStage Flow Designer es una interfaz de usuario para DataStage basada en web con capacidades de machine learning (ML), que facilita a los usuarios, incluso a los no técnicos, la creación de flujos y etapas dentro de un trabajo.

DataStage Flow Designer ofrece las siguientes ventajas:

- Compatibilidad retroactiva. No es necesario migrar los trabajos. Muchas empresas tienen miles de trabajos en un mismo proyecto, y dependen de estos para funcionar las 24 horas del día, los 7 días de la semana. La migración de trabajos, junto con la posibilidad de que se produzcan errores e interrupciones, no es una opción válida. Estas empresas pueden utilizar cualquier trabajo de DataStage existente y renderizarlo en IBM DataStage Flow Designer, por lo que no necesitan migrar esos trabajos a un nuevo lugar.
- Aumento de la productividad de los desarrolladores. IBM DataStage Flow Designer dispone de funciones como la búsqueda integrada, una guía rápida para hacer que las empresas se pongan en marcha de forma inmediata, propagación automática de metadatos, paleta inteligente, etapas sugeridas y resaltado simultáneo de todos los errores de compilación. Los desarrolladores pueden utilizar estas características para ser más productivos en sus diseños, de modo que su productividad puede aumentar y llegar a ser hasta nueve veces mayor que la de los proyectos de programación tradicionales.
- Amplio número de operadores y conectividad. Además de las capacidades de diseño y desarrollo, DataStage permite disponer de cientos de operadores previamente desarrollados y listos para usar. Estos reducen drásticamente el tiempo que los desarrolladores dedican a preparar los datos para las acciones de analítica. Gracias a la incorporación de nuevos operadores cada pocas semanas, la productividad de los desarrolladores aumenta con el tiempo.

Calidad y seguridad de los datos durante el proceso para la provisión de datos de confianza

DataStage ofrece una experiencia de usuario única para la integración de datos mediante la utilización de DataStage Flow Designer. Esta herramienta permite ejecutar la validación de los datos, la normalización y las reglas de congruencia en el momento en que los datos se proveen a los entornos de destino (como los data lakes), a fin de evitar problemas de calidad y posibles problemas de seguridad al proporcionar a usuarios no autorizados acceso a sus datos de carácter confidencial. Este concepto de calidad de los datos también puede ampliarse para dar soporte al gobierno integral de los datos en todo el almacén de datos (DWH).

Resumen

DataStage proporciona:

- la posibilidad de diseñar una vez y ejecutar en cualquier lugar gracias al balanceo automático de las cargas de trabajo, el paralelismo y la escalabilidad;
- la captura de actualizaciones en tiempo real o con estilos de entrega por lotes;
- resiliencia incorporada, facilidad de uso y CI/CD;
- una integración de datos optimizada para la IA;
- el diseño de trabajos automatizados mediante la utilización de las capacidades de ML;
- la calidad de los datos en tránsito y seguridad de estos para una provisión de datos fiable.

IBM ofrece una amplia gama de capacidades de integración de datos a través de entornos híbridos multicloud, de forma local o en sistemas hiperconvergentes como IBM Cloud Pak for Data, o en cualquier otra plataforma de nube. Estas diferentes capacidades proporcionan una solución de integración de datos flexible y escalable que permite acceder rápidamente a volúmenes de datos de alta calidad para la IA, en el modelo de despliegue de su elección.

Realice una demostración guiada gratuita para obtener más información sobre
IBM InfoSphere DataStage

¿Por qué IBM?
Las capacidades de IBM DataOps ayudan a crear una base analítica

preparada para el negocio gracias a una tecnología líder en el mercado que funciona en combinación con una automatización habilitada por la IA, un gobierno infundido y un potente catálogo de conocimientos con el fin de poner en funcionamiento datos continuos y de alta calidad en toda la empresa. Aumente la calidad de los datos para proporcionar una tubería de datos eficiente y de autoservicio a las personas correctas en el momento adecuado, desde cualquier fuente.

Para obtener más información sobre DataOps, visite ibm.com/dataops
Para obtener más información sobre IBM InfoSphere DataStage, visite ibm.com/productos/infosfera-datastage
Visite el centro de datos y analítica de Big Data and Analytics en ibmbigdatahub.com

IBM España
Santa Hortensia, 26-28
28002 Madrid
España
Abril de 2020