IBM DataStage

Deliver business-ready data in real time for
AI with IBM Cloud Pak for Data DataStage

Delivering business-ready data through data integration

Today's digital enterprises are creating and consuming data like never before. This includes data on customers, transactions and employees that are stored across multiple systems and repositories. These data stores are spread across various multicloud, hybrid cloud environments and data lakes, so organizations are looking at ways to bring these disparate sources and environments together to derive faster insights **using AI** help deliver differentiated and personalized experiences to their customers. According to a Forrester study, data scientists spend 80% of their time on preparing and managing data for **AI initiatives**. These results coupled with an IBM survey where 91% of organizations are not using their data effectively means that businesses are struggling to deliver value from data silos. The architectural techniques, practices and tools that are used to achieve real-time data accessibility from vast amounts of data and deliver business-ready data, is called data integration. With a flexible and scalable data integration technology, businesses can perform analytics for the next-best offer, churn detection and analysis, supply chain forecasting and execute instant fraud detection through extract, transform and load (ETL) data on multiple data sources.

For CXOs, Enterprise Architects or Operations leaders who struggle with managing data across multiple clouds or data lakes and want to shorten the time it takes to build and update AI models and applications, IBM® InfoSphere™ DataStage, a market leading data integration solution that delivers trusted business-ready data capabilities going beyond ETL provides a scalable multicloud data integration and delivery solution to ensure trusted business-ready information is being used in real time. The key capabilities in DataStage include multicloud runtime support that uses design once and run on any cloud while being able to scale workloads with automatic workload balancing and low latency parallel engine. In addition, it also features real-time data delivery with built in replication technology, reduced time and cost for DevOps with Continuous Integration and Continuous Delivery (CI/CD) support, fast time to build AI models with Autonomous Integration Design and validation rules to detect and resolve data issues automatically by using in-line data quality.

DataStage is part of IBM DataOps capabilities to operationalize continuous, high-quality data to enable AI and provide an automated, self-service data pipeline to the right people at the right time from any data source. IBM InfoSphere DataStage is available on-prem, on IBM Cloud and hyper-converged platforms such as IBM® Cloud Pak™ for Data that can deploy anywhere. IBM® Cloud Pak™ for Data is a fully-integrated data and AI platform that is built on Red Hat® OpenShift® which offers a fully cloud native architecture of DataStage that can scale with your business. It also provides organizations with a platform to support multiple data delivery styles including data integration, data replication and data virtualization while CDC captures log-based changes as they occur and delivers the information to target databases on the cloud and data lakes using Kafka-based message queues.

Design once, run on any cloud

According to an IDC study, 90% of enterprise customers are using multiple clouds. With multicloud data integration, users can separate the design from runtime—you can design your ETL jobs once and deploy runtime components through containers on any cloud environment to reduce latency due to processing large data volumes. You can create and test a job on premises then run it in a cloud environment, such as Microsoft Azure instance, making use of the on-cloud Azure data lake. The job parameters and their values are passed to a remote instance of DataStage by way of a Kafka message.

Multicloud data integration offers the following benefits:

- Ability to integrate data across on-premises and cloud environments
- Automated job design experience to simplify the design process
- Remote job execution to minimize egress cost of moving data out
- Fulfillment of geo-political requirements
- Reduced latency for processing large datasets, as data doesn't need to be moved and stays where you have it

Automatic workload balancing and parallel processing

With a fully cloud native architecture, you can use local containers or shared containers for DataStage to scale out your workloads dynamically as well as optimize for large datasets with a best in breed parallel engine (PX). Users have the choice to create a parallel, sequence or Apache Spark job in IBM DataStage Flow Designer.

You can run DataStage Flow Designer jobs on two run time engines:

- Jobs with the job type parallel or sequence can only be run on a parallel engine. Typically resource intensive jobs are run on the parallel engine, and as a result, the average time to complete complex jobs using parallel processing is two minutes.
- Jobs with the job type Spark can only be run on a Spark engine

Real time data delivery

DataStage with fully built-in Change Data Capture (CDC) technology for real time capture deployed as containers can provide the best of both the data integration and data replication worlds. DataStage allows for complex transformation with large datasets while CDC captures log-based changes as they occur, transforms them using complex transformations and delivers to target databases on the cloud and data lakes using Kafka-based message queues. DataStage also allows for batch-based and even-based bulk data transformation jobs to be fed to data warehouses.

Reduced time and costs for DevOps with CI/CD support

To address the challenge of managing the number of containerized applications across different operating systems, organizations need a robust open source tool such as Red Hat OpenShift, available on Cloud Pak for Data. The Cloud Pak for Data platform helps them scale and provision containers to support key IT initiatives such as microservices and cloud migration strategies. DataStage containers allow for creation and automation of Continuous Integration/ Continuous Delivery (CI/CD) pipelines for jobs from dev to test to production and support a CI/CD pipeline by supporting source control tools such as GitHub to frequently publish jobs and release to production.

Autonomous integration design to fuel AI

Accelerate collecting and integrating data for AI at a faster rate and at scale by automatically discovering and classifying assets, generating integration flows based on built-in custom transformation and quality rules and detecting and protecting sensitive information.

Fast time to value with automated job design

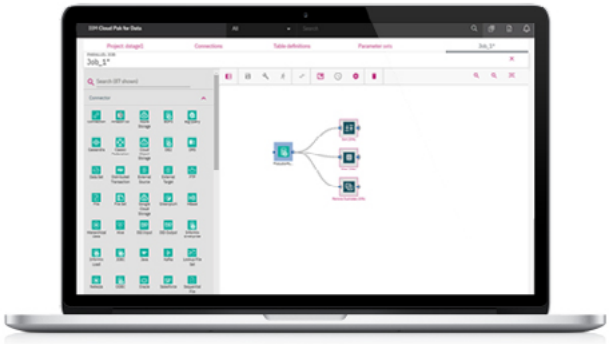

Figure 1. DataStage Flow Designer with automated design capabilities

IBM DataStage Flow Designer is a web-based UI for DataStage with machine learning (ML) capabilities to assists users, even non-technical ones, to build flows and stages within a job.

The DataStage Flow Designer offers the following benefits:

- **Backwards compatibility.** No need to migrate jobs. Many companies have thousands of jobs in a single project, and they depend on these jobs to run 24 hours a day, 7 days a week. Migration, with the likely possibility of errors and outages, is not an option for them. These companies can take any existing DataStage job and render it in IBM DataStage Flow Designer, so there's no need to migrate those jobs to a new location.
- **Increase in developer productivity.** IBM DataStage Flow Designer has features like built-in search, a quick tour to get companies jump-started, automatic metadata propagation, smart palette, suggested stages and simultaneous highlighting of all compilation errors. Developers can use these features to be more productive while designing jobs, and their productivity can increase to be as much as nine times faster than traditional hand coded jobs.
- **Extensive operators and connectivity.** In addition to the design and development capabilities, DataStage offers hundreds of out-of-the-box pre-built, ready-to-use operators. These drastically reduce the time developers spend on preparing data for analytics actions. With new operators added every few weeks, developer productivity is enhanced over time.

In-flight data quality and security for trusted data delivery

DataStage offers a single user experience for data integration using DataStage Flow Designer for running data validation, standardization and matching rules at the time data is being delivered to target environments such as data lakes to prevent quality and potential security issues with giving unauthorized users access to your sensitive data. This concept of data quality can also be extended to support comprehensive data governance across the data warehouse (DWH).

Summary

DataStage provides:

- Design once, run anywhere with built-in automatic workload balancing, parallelism and scalability
- Capture updates in real time or with batch based delivery styles
- Built-in resiliency, easy operation and CI/CD
- Optimized data integration for AI
- Automated job design using ML capabilities
- In-flight data quality and data security for trusted data delivery

IBM offers a breadth of data integration capabilities across hybrid multicloud environments, on premises or hyper-converged systems like IBM Cloud Pak for Data, or on any cloud platform of choice. These different capabilities provide a flexible and scalable data integration solution to quickly access volumes of high quality data for AI, on the deployment model of their choice.

Take a free guided demo to learn more about
IBM InfoSphere DataStage

Why IBM?
IBM DataOps capabilities help create a business-ready analytics foundation by providing market-leading technology that works together with AI-enabled automation, infused governance, and a powerful knowledge catalog to operationalize continuous, high-quality data across the business. Increase data quality to provide an efficient, self-service data pipeline to the right people at the right time from any source.

To learn more about DataOps visit
ibm.com/dataops
To learn more about IBM InfoSphere DataStage visit
ibm.com/products/infosphere-datastage
Visit the Big Data and Analytics hub at
ibmbigdatahub.com

IBM Corporation
New Orchard Road, Armonk, NY 10504
Produced in the United States of America
April 2020