IBM DataStage

AI-powered data integration for all of your multicloud and hybrid cloud environments
Key benefits

- 50% lower cost of operations through simpler operations and maintenance.¹

- 30% savings on workload execution time due to elastic scaling and balancing.¹

- 50% or more reduction in development costs when running on multiple cloud environments thanks to the design once and run anywhere feature.²

- 87% savings in development cost when using visual and ML-assisted design, as compared to hand coding.³

- Existing customers can retain existing investments in skills and assets and save millions of dollars in license cost savings by eliminating the need to purchase Windows/Citrix thick client licenses.

Deliver business-ready data through data integration

Digital enterprises are creating and consuming data like never before—today’s data landscape looks significantly different than it did even a few years ago. Enterprise data is stored in disparate data stores—whether it’s transactional data representative of customers for an organization, product data from suppliers, or operational data generated about day-to-day operations. This data is further spread across multiple systems, including various multicloud and hybrid cloud environments, and data lakes. Moreover, organizations are using their enterprise data warehouses (EDWs) for purposes for which they were never originally intended, including storing large volumes of unused data and supporting unstructured data from social media and other data sources.

At the same time, CIOs, in partnership with their chief marketing officers, chief revenue officers, and chief data officers, are looking to leverage their data to increase share of wallet with their customers. They also want to seize market share, using AI to deliver differentiated and personalized experiences. According to a Forrester study, data scientists spend 80% of their time on preparing and managing data for AI initiatives.⁴ These results, coupled with an IBM survey that shows 91% of organizations aren’t using their data effectively, suggest that businesses are struggling to deliver value from data silos.

Trusted data delivery at scale across the hybrid cloud environment

IBM® DataStage® on IBM Cloud Pak® for Data delivers a modernized data integration solution to cleanse and deliver trusted data anywhere, at any scale and complexity, on and across multicloud and hybrid cloud environments. Save on data movement costs by bringing ETL to where your data is. Save on storage costs by only storing the data you trust and use. Increase the productivity of your business and IT users through automated job design and out-of-the box integration with Netezza®, Db2®, cloud data warehouses, data virtualization, and DataOps services.
Key differentiators for IBM DataStage

- Design once, run anywhere allows you to bring data integration or extract, transform and load (ETL) to where your data resides.
- A best-in-breed parallel engine processes huge data volumes and built-in workload balancing supports multicloud scalability and elasticity.
- DataOps is accelerated due to out-of-the-box data integration with data virtualization, governance, business intelligence (BI) and data science services on IBM Cloud Pak for Data.
- In-flight data quality and security helps ensure trusted data delivery to data lakes.
- Automated design templates, backwards compatibility and license cost savings benefit existing DataStage customers.

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 and bring data integration to where their data is. Design your ETL jobs once and deploy runtime components through containers on any cloud environment to save development costs on multiple clouds while eliminating data latencies and egress costs associated with moving data out from cloud environments.

Parallel processing and automatic workload balancing

With a fully cloud-native architecture, DataStage can dynamically scale workloads as well as optimize for large data sets with a best in breed parallel engine (PX). Users have the choice to create a parallel, sequence or an Apache Spark job in IBM DataStage Flow Designer.

Seamless integration with Netezza, Db2 Warehouse or cloud warehouses

Colocation with Netezza and Db2 on IBM Cloud Pak for Data system removes network bottlenecks and supports high-speed data delivery. Easily connect cloud data warehouses with pre-built connectors for Snowflake and Amazon Redshift.

Reduce operations and maintenance management

Meet mission-critical SLAs and reduce overall management by automating backup, recovery, and patch management. Free your personnel to work on high-value tasks. According to Forrester, IBM Cloud Pak for Data delivers container efficiencies resulting in a 65%–85% reduction in infrastructure management.

In-flight data quality and security for trusted data delivery.

Track lineage of your data in ETL jobs and automatically resolve quality issues using IBM InfoSphere® QualityStage® at the time data is ingested by target environments, such as data lakes. Provide metadata support for policy-driven access to sensitive data and prevent unauthorized users from getting access to your sensitive data. This concept of data quality can also be extended to support comprehensive data governance across the EDW.
Faster time to value with automated job design and out-of-the-box integration for DataOps and data science

Accelerate job design through automated integration templates and improve the collaboration between business and IT thanks to out-of-the-box connectivity with governance, business intelligence (BI), data virtualization, and data science services.

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

**Benefits**
**Backwards compatibility and license cost savings.** There’s 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, isn’t 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. Existing DataStage customers can eliminate the need to purchase Microsoft Windows or Citrix licenses and perform thick client upgrades with web-based DataStage Flow Designer and support an unlimited number of users.

**Increase in developer productivity.** IBM DataStage Flow Designer has features like built-in search, a quick tour to get users started, automatic metadata propagation, smart palette, and 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 up to 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 features drastically reduce the time developers spend on connecting to new cloud data sources and targets. With new operators added every few weeks, developer productivity is enhanced over time.

**Key use cases for IBM DataStage**

**Data modernization**
- Move your applications and data integration to the cloud.
- Integrate data across hybrid and multicloud environments.
- Support extract, transform and load (ETL) and data movement on a private cloud.

**Data warehouse modernization**
- Connect to Netezza or Db2 Warehouse on IBM Cloud Pak for Data.
- Move your warehouse to the cloud, using warehouses such as Snowflake and Amazon Redshift.

**Real-time analytics and data science with data lakes**
- Move data into data lakes on multicloud environments such as AWS, Google Cloud and Microsoft Azure.
- Prepare data for business intelligence tools.
- Integrate data as part of DataOps pipelines for AI.

Take a free guided demo to continue exploring IBM DataStage.
Why IBM?

As organizations accelerate their digital transformations to better predict and shape future outcomes, empower higher-value work and automate experiences, the need to embrace AI is growing ever more critical. But to implement AI successfully, companies need to overcome three major challenges: access to data at scale, skills, and trust. This process starts with good, clean, and secure data.

IBM is committed to helping clients drive digital transformation by leading with IBM Cloud Pak for Data, an open, Kubernetes-based, data and AI platform that integrates with an array of technology solutions that enhance an organizations’ ability to make their data ready for AI. Among those technologies core to the platform are DataOps capabilities, like IBM DataStage, a market leader in data integration to operationalize trusted data delivery across data lakes, multicloud or hybrid cloud environments for AI, with built-in capabilities for data quality.

To learn more about IBM DataStage, visit ibm.com/products/infosphere-datastage.

To find out which deployment option is right for you, visit ibm.com/products/infosphere-datastage/pricing.

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

IBM, the IBM logo, ibm.com, DataStage, Db2, IBM Cloud Pak, InfoSphere, Netezza and QualityStage are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml.

Microsoft and Windows are trademarks of Microsoft Corporation in the United States, other countries, or both.

The content in this document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

The client is responsible for ensuring compliance with laws and regulations applicable to it. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the client is in compliance with any law or regulation.


58031458USEN