DataStage on IBM Cloud Pak for Data

An automated data integration solution
on a multicloud data platform

IBM

Contents

The rise of a new AI fueled data integration strategy

According to IDC, the worldwide amount of stored data will grow nearly 17% in 2020 to 6.8 zettabytes (ZB), with compound annual growth rate of nearly 18% through 2024. This dramatic growth in data increases the amount of time and money it takes to ingest and manage enterprise-wide data, and this starts to hinder users' productivity and client satisfaction. But with the rise of artificial intelligence (AI) technology, there are new solutions to combat these problems. AI technology accelerates the pace of data discovery, broadens the range of data that can be leveraged and automates tasks that previously required human expertise. Gartner even states that by the end of 2024, 75% of enterprises will shift from piloting to operationalizing AI, driving a 5x increase in streaming data and analytics infrastructures.

That being said, AI can only be effective if the full range of data is trustworthy, accessible and compatible. The increased use of AI highlights weaknesses and limitations that have long existed in data systems, so enterprises must turn to new, modern strategies. Such agility requires a new information architecture, one that allows for seamless integration and operation across the entire data lifecycle. Which is why IBM clients are modernizing and transitioning away from legacy systems to move to a modern cloud-based architecture: IBM Cloud Pak® for Data. This data and AI platform provides improved scalability and elasticity for varying workloads and lowers operating costs while being able to connect to cloud data warehouses and real time analytical applications.

There are many factors that contribute to a major shift in how data integration tools are deployed and used with the rise of AI. These could be anything from high data variety in an enterprise to data users' needs, and because of the many factors, companies need to adopt a process-oriented approach to manage the data lifecycle with DataOps, improve business performance, and increase competitiveness. Companies embracing AI for their products and processes will require a highly flexible and scalable data integration technology embedded in the market-leading data integration tool IBM® DataStage® on IBM Cloud Pak for Data. It is equipped with features that improve the productivity of your business and IT users:

- A best-in-breed parallel engine and automatic workload balancing to elastically scale your workloads up to 30% faster than DataStage on-premises
- Design once, run anywhere capabilities to bring data integration to your data
- Automated job design and integration with Netezza®, IBM Db2® or cloud data warehouses, data virtualization or DataOps services

By integrating this technology seamlessly with other services on the data platform, enterprises can benefit from comprehensive and automated data provisioning while maintaining the performance, security and governance they need. Containerized architectures—specifically those deployed on a cloud-enabled platform such as IBM Cloud Pak for Data—are key to this transformation.

Using containers for your data integration tool

Cloud native by design, IBM Cloud Pak for Data unifies market-leading services spanning the entire data and analytics lifecycle. This includes the capabilities previously provided by the IBM InfoSphere® Information Server platform which are now available as DataStage and IBM Watson Knowledge Catalog cloud-ready services on IBM Cloud Pak for Data. With IBM Cloud Pak for Data you can streamline your data integration at a lower cost on a unified, cloud-native platform, and with the automation capabilities included on the service, your organization can gain business insights from your data in near real-time.

Often the IBM DataStage and Information Server platforms have historically been deployed to handle large scale enterprise workloads and perform mission critical functions. To ensure a seamless move to a modernized AI and cloud ready architecture, the DataStage and Information Server modernization upgrades provide an easy migration that delivers access to the capabilities on the platform as well as providing an even higher level of resiliency, scalability, automation and operational efficiency.

By deploying DataStage and Watson Knowledge Catalog services on IBM Cloud Pak for Data, enterprises can leverage all of the powerful features that make up an industry-leading data platform.

Built for AI

- In-line data quality and metadata exchange with Watson Knowledge Catalog for improved data governance
- Out-of-the-box integration with data science, event messaging, data virtualization and data warehousing services on IBM Cloud Pak for Data

Powered by AI

- Increased user productivity through built in design accelerators such as stage suggestions, schema propagation and automatic job template generation

IBM DataStage on IBM Cloud Pak for Data is the containerized version of IBM InfoSphere DataStage, based on a microservices architecture and optimized for Kubernetes. Through IBM Cloud Pak for Data, DataStage can run natively on Red Hat® OpenShift®, the world's leading container orchestration platform.

By breaking down the DataStage capabilities into microservices instead of a monolithic stack, you gain several opportunities:

- Deploy within minutes; enable standard deployment and management while retaining flexibility to modify parameters as needed.
- Gain reliability due to out-of-the-box enhanced Kubernetes availability and support for high availability/disaster recovery (HADR) automated failover.
- Reduce management burden with automated updates. Service packs, versions and mods can be deployed with one click.
- Automate management by "application group" so administrators can use namespaces to manage access control and provisioning options.
- Monitor and manage at an application level thanks to platform and service-level features.
- Scale microservices independently to respond to changing needs.

Containerizing your data integration technology enables you to run DataStage as part of a hybrid cloud environment (combination of cloud and non-cloud platforms) or multicloud environment (clouds from different providers) that uses the appropriate infrastructure for each type of data.

These advantages may account for the recent popularity of containers. According to the Red Hat Global Customer Tech Outlook 2019, 57% of organizations are already using containers and container usage is also expected to increase by 89% in the next 2 years. With IBM Cloud Pak for Data, you can more easily access the full scale of IBM services to design, deploy and manage advanced analytics that help you deliver business value.

The five major benefits of deploying DataStage on IBM Cloud Pak for Data

1. Ease of enabling hybrid cloud and multicloud on a single, unified platform

According to Gartner, the majority of enterprises use more than one cloud provider, and historically, from the context of data integration, the challenge has been that enterprises need to incur data latencies and data egress costs while moving data between different cloud platforms and their on-premises data sources. Organizations often had to run individual applications across multiple providers to execute their data integration jobs, and it took up more time and costs than should have been necessary. But now with IBM DataStage on IBM Cloud Pak for Data, users have the freedom to choose any cloud provider with one solution. With design once, run anywhere features within DataStage, users can design their jobs once on-premises, move runtimes to where their data resides, and thereby avoid data latencies and millions of dollars in egress costs. There's no added need to move your data out of where it's already housed.

2. 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 or an Apache Spark job in IBM DataStage Flow Designer.

Moreover, customers can expect up to around a 30% decrease in execution time with IBM DataStage on IBM Cloud Pak for Data compared to traditional DataStage on-premises. These performance improvements are particularly apparent during execution windows of resource contention due to the automatic workload balancing that distributes workloads across the worker nodes in the OpenShift cluster and maximizes throughput.

3. Savings on development time and costs thanks to automated job design and DevOps 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 IBM Cloud Pak for Data. The IBM 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. They also help streamline CI/CD pipelines by supporting source control tools such as GitHub to frequently publish jobs and release to production.

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

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. Rewriting these jobs, with the likely possibility of errors and outages, is not an option for them. Using the DataStage Flow Designer on IBM Cloud Pak for Data, these companies can take any existing DataStage job and render it in the thin client so there's no need to rewrite those jobs. Moreover, clients can save millions on license costs by eliminating the need for purchasing thick clients for job design, by instead using the DataStage Flow Designer thin client.

In addition to the design and development capabilities, DataStage offers hundreds of out-of-the-box, pre-built, ready-to-use connectors for Amazon S3, Azure, Db2, Hive and Kafka, and it also offers stages such as transformer, encode, annotate, tail and merge. These drastically reduce the time developers spend on preparing data for analytics actions. With new operations added every few weeks, developer productivity is enhanced over time.

4. Built-in integration with data and AI services

With DataStage on IBM Cloud Pak for Data, it is easy to leverage capabilities from the broader IBM and open source ecosystems. The platform includes many core services ranging from data warehouses, Watson Knowledge Catalog, data science and data virtualization to event messaging. 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 to access and transform data, no matter which cloud platform the data resides on.

To prevent data lakes from turning into "data swamps" with ungoverned data, you can simultaneously track data lineage in ETL jobs with IBM InfoSphere QualityStage® while data is ingested by target environments, such as data lakes, to automatically resolve quality issues. You can also 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 be extended to support comprehensive data governance across the enterprise data warehouse (EDW).

With the included data virtualization capabilities on IBM Cloud Pak for Data, business users can discover data, query data, and experiment with flows for data warehouses while also performing simple SQL-based data transformations, running development and testing, and managing both structured and unstructured data.

5. The value of Red Hat in IBM Cloud Pak for Data

The hybrid cloud and multicloud options are enhanced by the advantages of Red Hat OpenShift, upon which IBM Cloud Pak for Data is based. The Red Hat stack, OpenShift and Kubernetes operating together, is particularly beneficial. It allows you to develop secure and scalable Kubernetes applications without being overwhelmed by the complexities of large-scale manual Kubernetes administration. Using Kubernetes Operators, Red Hat OpenShift offers automated installation, upgrades, and lifecycle management for every part of the container stack: the operating system, Kubernetes and cluster services, applications, and persistent data storage.

OpenShift provides a comprehensive platform that enables automated operations and provides out-of-the-box support for languages such as Java, Node.js, Ruby and Python. OpenShift also provides supporting services such as monitoring, authentication and authorization and network management. These OpenShift features are not in the open source version of Kubernetes.

In addition, the included Kubernetes distribution is enterprise-grade, and benefits from hundreds of security, defect and performance fixes in each release. Validated popular storage and networking plug-ins for Kubernetes are also available. And finally, open source Red Hat tools provide additional functionality options, such as Apache Spark for streaming data, or the popular Python and R languages for machine learning applications. The additional functionality ensures that enterprises leverage essential open source tools necessary to develop, deploy and run applications through the OpenShift platform. When these varied resources are all part of a single, unified platform in IBM Cloud Pak for Data, they are easier to integrate and manage than they would otherwise be.

Next steps

When deployed via IBM Cloud Pak for Data, DataStage is more than a robust data integration tool that can process data at scale. It becomes part of a microservices-based data platform that also helps you organize and analyze your data, infusing AI capabilities throughout your enterprise.

DataStage on IBM Cloud Pak for Data provides:

1. AI capabilities, built for AI projects
2. Up to 50% lower cost of operations due to automatic failure resolution and automation of operational tasks such as backup, recovery, and patch management
3. 30% faster workload execution compared to traditional DataStage thanks to built-in workload balancing and best-in-breed parallel runtime that optimize workload execution
4. 87% savings in development cost when using visual and ML-assisted design, as compared to hand coding
5. Savings on data movement costs by bringing integration workloads to the data using design once, run anywhere
6. Pre-built integrations with data science, data warehouse and data virtualization services using a common UI

Existing customers can retain their investments in skills and assets and save millions of dollars in license costs by eliminating the need to purchase Windows or Citrix thick client licenses.

DataStage on IBM Cloud Pak for Data offers a unique combination of containerized architecture, Red Hat infrastructure, data connectivity and a broader IBM capability ecosystem, making it a compelling choice for enterprises that want to prepare their data foundations for the opportunities ahead.

To get started try IBM Cloud Pak for Data for free
Schedule a free one-on-one consultation with a data integration expert.

IBM Corporation
New Orchard Road, Armonk, NY 10504

Produced in the United States of America
October 2020

7EB2XONR