

Dramatically increased data speed

IBM's Chief Data Office modernizes data movement
IBM DataStage



< [Overview](#) [Modernizing data movement](#) [From platform to](#)



X Close

Hello! How can we help you?

In an earlier [case study](#), we shared how IBM's Global Customer Data Office (GCDO) faced the all-too-familiar challenge of data silos at the company, and how—with no commercially available solution in hand—it developed the Cognitive Enterprise Data Platform as a central source of governed data for users to load, transform, and analyze enterprise data. This use case continues our journey of digital modernization leveraging the [IBM Cloud Pak® for Data](#).

This time it's all about data movement.

The pain point was clear. Vast amounts of data needed to be ingested into the new platform, and it was going to take months to complete.

Initial Data Loads (IDLs) replicate data from one system to another. Change Data Capture (CDC) increases efficiency because after the first transfer, only data needs to be moved.

As that first transfer, IDLs are usually a huge amount of data, and the time needed to load was no exception: the largest of the dozens of transfers involved millions of records, weighing in at 186 GB. Loading data sets was taking weeks in some cases. Nearing the breaking point and essentially blocked from progress, they needed a new solution. They found it in the [IBM® DataStage®](#) for Data Cloud solution.



Hello! How can we help you?

Faster data loading

Initial Data Load in a fraction of the time, from 3 days to 3 hours

“After just a few days of highly successful testing, we incorporated DataStage for IBM Cloud Pak for Data ... IDLs of 60 million records that had taken three days were completed in just about three hours.”

Inderpal Bhandari
Global Chief Data Officer
IBM

Modernizing data movement

When GCDO started its data and AI journey, the IBM Cloud Pak for Data solution didn't exist. While the CEDP drove significant advancement, the development of the IBM Cloud Pak for Data solution gave GCDO a homefield advantage for taking its own platform to the next level.

As a suite of services and extensions that can be used as needed, the IBM Cloud Pak for Data solution gave GCDO the required flexibility to modernize in stages and start with highest needs first. There was no prescriptive order to adoption or deployment.

Hello! How can we help you?

GCDO first started using the AI suite of services within the IBM Cloud Pak for Data solution, including the [IBM Watson® Studio](#) solution. IBM Watson Studio technology runs on premises and in the cloud, analyzing data in the [IBM Db2® Big SQL](#) solution. The details of this part of GCDO's modernization journey are described in this [case study](#).

For the next step in the journey, GCDO turned to DataStage technology to dramatically increase the speed of ingesting vast amounts of data with stability and accuracy.

"After several months setting up servers, establishing database connections, and trial and error configuration and self-learning efforts, a 60 million record table would still take three days to replicate," says Frank Duffy, Senior Project Manager with GCDO Master Data. "Looking at those statistics, with approximately 20 large tables to go, we were looking at another 60 days just to migrate the data."

GCDO's Data Movement team tested the performance of DataStage and Spark technology in executing common data load use cases. In more than 75% of the cases, they achieved better performance with DataStage technology than with Spark technology. For the remaining 25%, the results were a close match.

Beyond performance, factors that attracted GCDO to the DataStage solution include:

- Integration with the IBM Cloud Pak for Data ecosystem, specifically related to the [IBM Watson Knowledge Catalog](#) and data lineage
- Breadth of supported sources, targets and intermediate stages that met current and forward-looking needs
- Custom stages to encapsulate needs into reusable units when necessary
- Capabilities that supported a pattern-based approach

The IBM Cloud Pak for Data solution is aligned with several industry data sources and is constantly evolving those sources to meet new technology. The DataStage for IBM Cloud Pak for Data solution comes bundled with a large inventory of industry connectors, representing most of the data stores that GCDO users wanted to work with. These connectors meant that GCDO could work with these different storage formats and systems without needing to write any code.

In those instances where a connector wasn't already available, custom connectors could be developed, deployed and dropped on to the canvas.

The DataStage for IBM Cloud Pak for Data solution also offers Runtime Column Propagation functionality, which appealed to GCDO engineers because it allowed a pattern-based approach to data movement. By expressing common data movement patterns as jobs, GCDO scaled up operations to support thousands of tables without needing to increase staffing.

"The DataStage for IBM Cloud Pak for Data pattern capability allowed us to have one job that could run thousands of ways," says Rick McCall, GCDO Technical Lead for the Data Movement Tool. "In some cases, we had upwards of 8,000 jobs — pages and pages of them — that could be associated to a single pattern and run as a single job. That meant one set of code, optimized performance, and source control all rolled into one super-f super-reliable solution."

Hello! How can we help you?

Another benefit of the DataStage for IBM Cloud Pak for Data solution is that it integrates seamlessly with [RedHat® OpenShift®](#). It also offers API support so users can build

custom workflows around it if needed.

“DataStage for IBM Cloud Pak for Data was a game changer for our data ingestion,” says Peter Herr, Global Leader for Client Master Data. “Our team had tried everything within the constraints of our existing system and were still at an impasse for acceptably accomplishing the massive amount of data migration we required. When Rick and team showed us the speed and power of DataStage, we were productive within weeks instead



“DataStage for IBM Cloud Pak for Data was a game changer for our data ingestion. The team had tried everything within the constraints of our existing system and were still at an impasse for acceptably accomplishing the massive amount of data migration we required. When Rick and team showed us the speed and power

Hello! How can we help you?

DataStage, we were productive within weeks instead of months.”

Peter Herr

Global Leader for Client Master Data
IBM Global Chief Data Office

From platform to privacy

Once GCDO chose the DataStage for IBM Cloud Pak for Data solution, positive results rapidly followed. In the pilot phase alone, huge tables with billions of rows were loaded in hours instead of days. Smaller tables were migrated within minutes. Further, regardless of the size of the table, data ingestion was error-free and highly stable.

“Needless to say, data movement powered by DataStage was a godsend to help rescue our data migration effort and move us from a blocked state to a production-ready state in a matter of weeks,” says Duffy.

“IBM Cloud Pak for Data continues to advance CEDP,” says Inderpal Bhandari, IBM’s Global Chief Data Officer. “DataStage for IBM Cloud Pak for Data as the engine for our data movement strategy saved us literally weeks and brought new levels of efficiency and flexibility in serving our users. Next, we have our sights set on leveraging IBM Cloud Pak for Data as we build out our enterprise-wide privacy capabilities.”

GCDO is partnering with the IBM Chief Privacy Office to build the engine to power an end-to-end hybrid cloud system that will dramatically enhance the efficiency of our regulatory compliance. The current roadmap for privacy capabilities features Watson Knowledge Catalog, [IBM Knowledge Accelerators](#) and [IBM OpenPages® with Watson](#) from the IBM Cloud Pak for Data solution.

[Discover solutions to help you eliminate data silos](#)



Hello! How can we help you?

About the IBM Global Chief Data Office

The IBM Global Chief Data Office develops data strategies and platforms that include governance and management systems, deep data and analytics partnerships. The strategy transforms business data into business value. These platforms become the central data source for business analytics across the enterprise and for developing and scaling talent. Together these innovative capabilities use analytic insights to enable growth and productivity.

Solution components

IBM Cloud Pak® for Data	→	IBM® DataStage® for IBM Cloud Pak for Data	→
IBM Db2® Big SQL	→	IBM Knowledge Accelerators	→
IBM OpenPages® with Watson	→	IBM Watson® Knowledge Catalog	→

Take the next step

To learn more about the IBM solutions featured in this story, please contact your IBM representative or IBM Business Partner.

- [View more case studies](#) →
- [Contact IBM](#) →

Insights you can't miss

Subscribe to curated newsletters for the latest in technology, business and thought leadership.

[Get the best of IBM in your inbox](#) →

IBM Global Chief Data Office

A unified data and AI platform adds business benefits

[Read the case study](#)

Hello! How can we help you?

Legal

© Copyright IBM Corporation 2022. IBM Corporation, IBM Watson, New Orchard Road, Armonk, NY 10504

Produced in the United States of America, March 2022.

IBM, the IBM logo, ibm.com, DataStage, Db2, IBM Cloud Pak, OpenPages and IBM Watson 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 ibm.com/legal/copyright-trademark.

Red Hat® and OpenShift® are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

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 performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions. 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.

Hello! How can we help you?