IBM DataStage Webinar

Cloud Pak for Data Migration Best Practices

Jessie Snyder
Offering Manager, IBM Data and AI Data Integration

Bala Vaithyalingam
Principal, IBM Data and AI Expert Labs and Learning

Scott Brokaw
Offering Manager, IBM Data and AI Data Integration

IBM

Please Note

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM's sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

IBM Cloud Pak for Data

Simplifies, unifies and automates your journey to AI

Analyze & Infuse

Plug and play 45+ data, analytics and AI apps. Manage your favorite open source capabilities along side IBM's market leading differentiators.

Organize

Catalog and govern all enterprise data, models, rules, and insights through a common experience

Collect

Virtually connect, manage and query data & AI assets no matter where they live.

Run On-Premises or on ANY Cloud

Decoupling enterprise data, analytics and AI will prevent lock in and accelerate polyglot eco-systems.

OpenShift

Leverage the leading open source hybrid cloud platform to SCALE data & AI workloads.

openstack.

On Premise

aws

Why *DataStage on IBM Cloud Pak for Data?*

Future proofing your Data Integration

– Design with Speed through smart automation and high level of reusability

– No additional development cost when scaling out to new environments

– Runtime independence through Enterprise container platform foundation

– Cost, speed and scale optimized data integration

– Increased compliance and deploy @ scale operate through full CI/CD integration

– Significantly reduce effort in management and operation
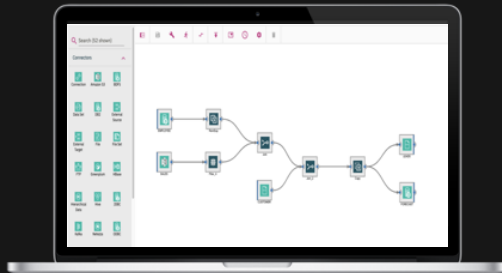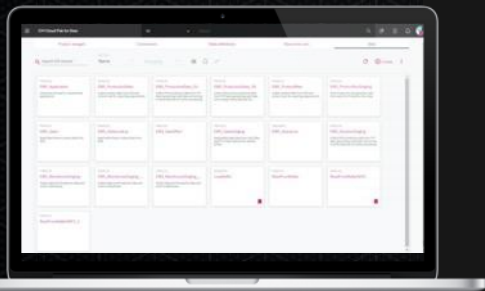
Key customer benefits

5x

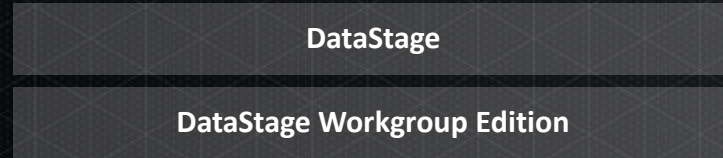Faster execution than on Spark parallel engine

9x

Faster design than hand-coding

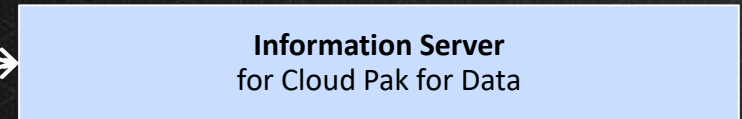**85%**

Reduction in infrastructure management time & effort

Paths to Modernize

Today

DataStage

DataStage Workgroup Edition

DataStage Enterprise
for Cloud Pak for Data

DataStage

QualityStage

Information Server for Data Integration

DataStage Enterprise Plus
for Cloud Pak for Data

Information Server Enterprise Edition

Information Server for Data Quality

Information Server
for Cloud Pak for Data

Capability Comparison Cloud Pak for Data

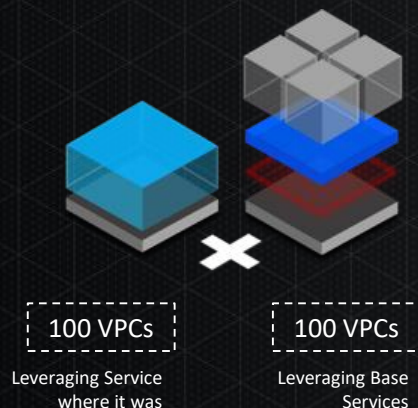|	DataStage (on-prem)	DataStage Enterprise for Cloud Pak for Data	DataStage Enterprise Plus for Cloud Pak for Data	Information Server for Cloud Pak for Data
Business Glossary				✓
Business & Technical Data Lineage				✓
Self-service Data Preparation		**Add by licensing**		✓
Search & find relevant data		**Watson Knowledge Catalog (WKC) via Cloud**		✓
Review, Rate & Share data		**Pak for Data base**		✓
Data Profiling				✓
Sensitive data discovery				✓
Data Cleansing and Enrichment			✓	✓
Data Quality Validation & Monitoring			✓	✓
Service enablement			✓	✓
Data Specification Mapping			✓	✓
Extract, transform, load data	✓	✓	✓	✓
Metadata Management	✓	✓	✓	✓
Common Cloud Pak for Data Platform Management		✓	✓	✓
Unlimited users		✓	✓	✓
Automatic Workload Balancing		✓	✓	✓

How entitlements traded up to Modernization Upgrade can be allocated

Scenario: Your existing DataStage

- Today: 7000 PVUs of DataStage Standalone (Prod)
- At renewal: trade-up to DataStage Enterprise Upgrade
- Get at a minimum: 100 VPCs of DataStage Enterprise + 100VPCs of Cloud Pak for Data
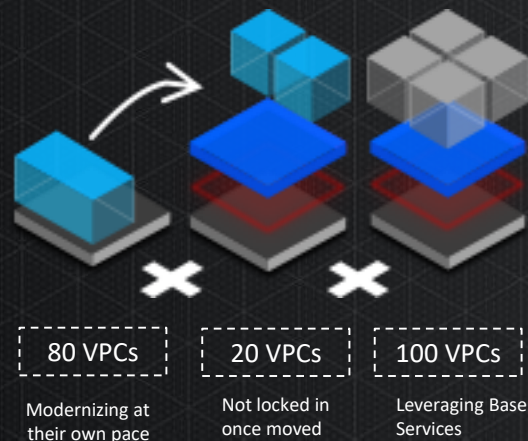- Once you trade-up, you can allocate this entitlement in an infinite number of ways, some shown below:

Example
Trade-up license entitlement but workload still runs on stand-alone offering

Example
Trade-up license entitlement and move workload to extension gradually

100 VPCs	100 VPCs
Leveraging Service where it was	Leveraging Base Services

80 VPCs	20 VPCs	100 VPCs
Modernizing at their own pace	Not locked in once moved	Leveraging Base Services

Synergies and Benefits of an integrated Data & AI Platform

The Whole is Greater than the sum of the parts
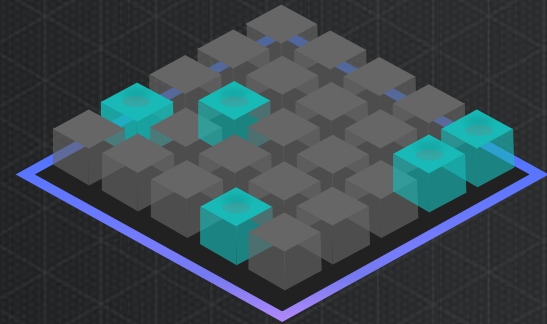
DataStage taking advantage of co-located Netezza or Db2 WH for *ultra highspeed* data load

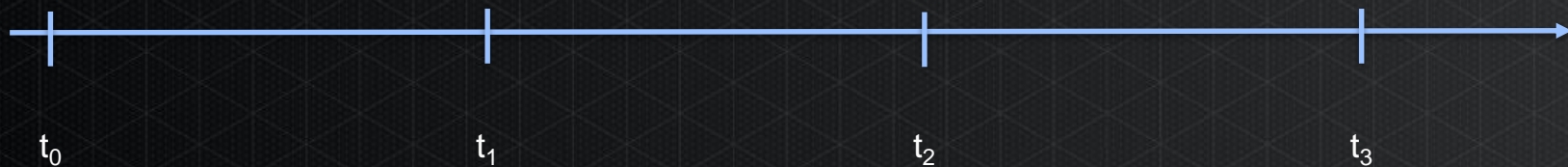Utilize implicit Data Discovery, Lineage, *Governance* and Metadata Management

Optimized and agile data processing when combing DataStage with *Data Virtualization*

DataStage being the vital link of the *Edge to Analytics value chain* on Cloud Pak for Data

DataStage Environment Modernization

Getting Started - Lift and Shift

A schematic project plan:

t_0 t_1 t_2 t_3

Installation	Migrate (Import / Export)	Test	Go live !
Setup	assets		
Config	*Optionally*: Convert Server to		
Network	Parallel jobs		

- Lift and Shift (CPD) takes place in parallel to regular DataStage operation.
- Scenario assumes no change in use cases
- Perform asset conversion if needed – can be performed on the legacy environment
- Migration can be done directly from current version of DataStage / QualityStage
- Migration period managed by waiver.

Modernizing with MettleCI

IBM is planning to provide the entire MettleCI tool set to any DataStage or Information Server on Cloud Pak for Data licensee*

This tool set supports both upgrade/migration of DataStage as well as CICD

*OEM of MettleCI expected to be complete in Q4 2020
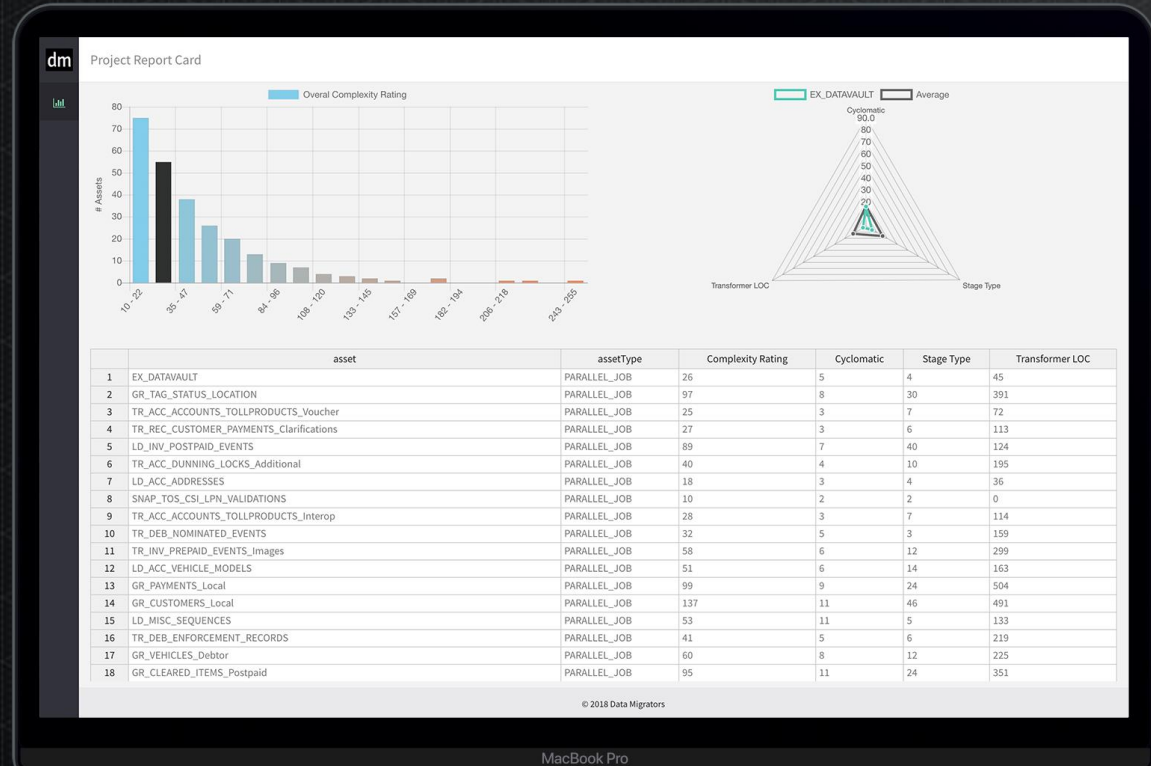
MettleCI Features & Functions
- Automated Peer Review (Compliance)
- Automated Unit Testing (generation & management)
- Universal Git integration (Jobs and Unit Tests)
- Continuous Integration
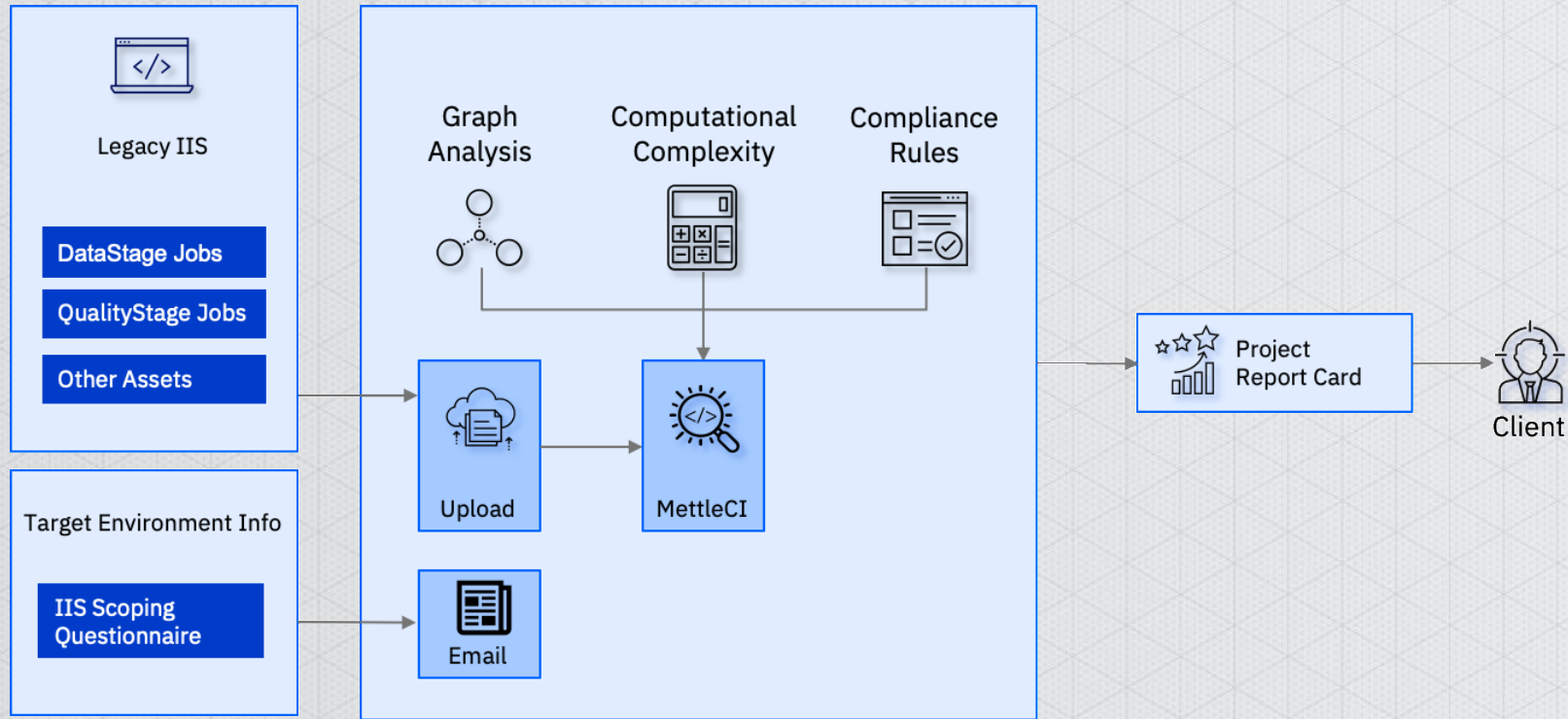- Continuous Delivery

MettleCI Benefits
- Shorter time to delivery
- Lower cost of maintenance
- Higher performance Jobs
- Earlier, cheaper defect discovery
- Lower testing costs
- More reliable E2E execution
- Zero-effort work item traceability / auditability
- Release management with minimal effort
- Visibly more productive DataStage teams
- Higher utilisation of Information Server platform
- Demonstrable alignment with organisation strategy

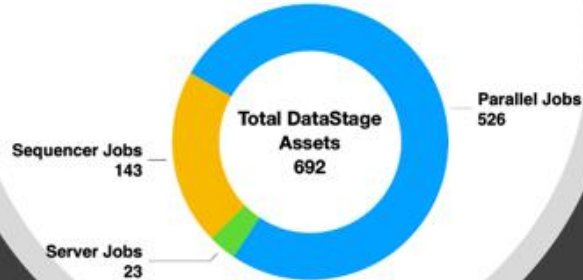# DataStage Upgrade Report Card

- Provides job assessment to help determine complexity for migration and upgrade compatibility

- IBM's Rapid DataStage Upgrade Assessment (free) additionally provides estimates of upgrade duration, costs, and approach

- Minimizes manual effort, cost, risk, and elapsed time associated with upgrading your DataStage environment

Rapid DataStage Upgrade Assessment Process
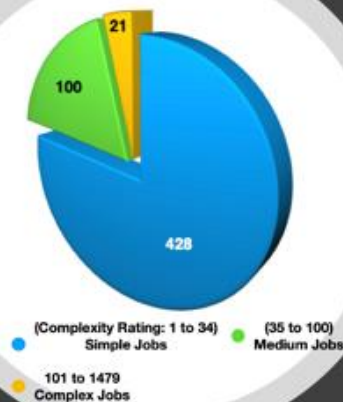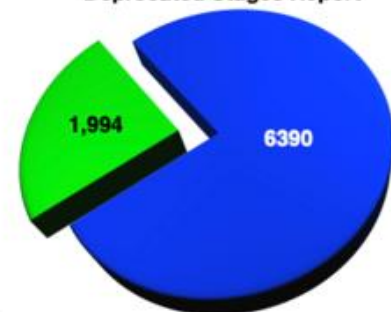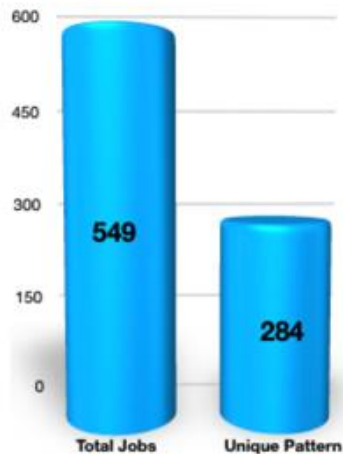
DataStage Compliance Report Summary

Compliance Type / Count

Compliance Type	Count
★ Adjacent Transformers	19
★ Deprecated Stages	40
★ Database Row Limit	5
★ Debug Row Limit	25
★ Default Naming	173
★ Hardcoded File Paths	10
★ Job Naming	1
★ Link Sort	207
★ Lookup Failure	3
★ More than One Data Flow	18
★ Range Lookup without allow duplicates set	3
★ Too Many Stages	11
★ Unique Sort	3
★ Invalid Asset Model	0
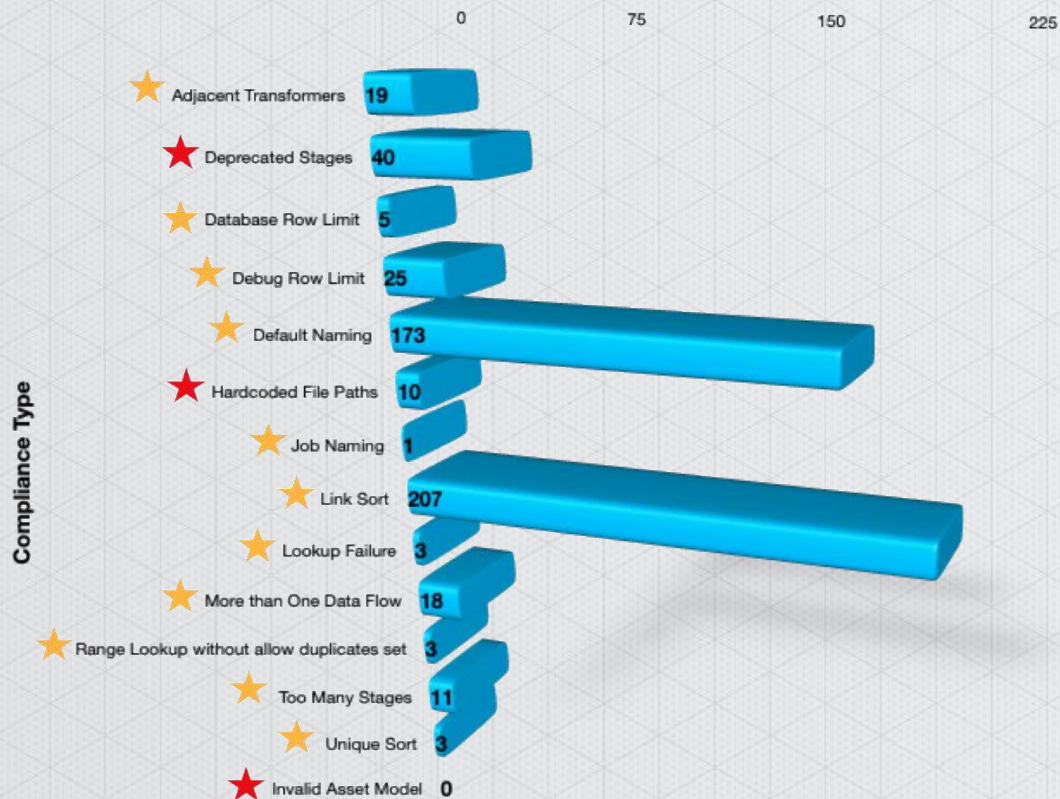
LEGEND
★ Warnings
★ Blockers

DataStage Upgrade through Continuous Delivery Pipeline

Legacy Environment

DataStage

Git

Testing

5. Upgrade
Jobs are automatically upgraded to replace deprecated stages

6. Testing
Verify baseline tests from legacy still pass in the new target environment

7. Target Ready
All jobs are imported, converted, compiled, and unit tested

Deploy

1. Source Delta
Jobs and Unit Test specifications and data are managed using DataStage Designer and MettleCI Workbench

2. Git Check-in
All code is promoted via build events triggered by an automatically-managed Git repository.

3. Testing
Legacy testing for job compliance with coding standards, as well as automated unit testing

4. Deploy
Automated import of all artifacts into the new target

Upgrade

Testing

DataStage

Target Environment

Automated

DataStage Modernization/Migration FAQs

What would a *Lift & Shift* migration to Cloud Pak for Data look like?

- Lift and Shift from stand alone DataStage to Cloud Pak for Data is very similar to Lift and Shift when doing a version to version migration.

- Users will use similar tools / techniques to move assets (import / export) between a stand-alone DS instance and DS on CP4D as if they would do in a previous version migration

- Users can additionally use MettleCI, to be packaged with any DataStage/Information Server for Cloud Pak for Data extension, to assist with moving assets

Are there any jobs that will not run on Cloud Pak for Data?

- For a Lift and Shift approach:

  - Most jobs are going to work unchanged on Cloud Pak for Data v3.5.

  - Some functionality available on Cloud Pak for Data (v3.5), is in deprecated state and should therefore be appropriately migrated to supported functionality

DataStage Modernization/Migration FAQs

What should clients do with their Server jobs / Server routines?

- IBM plans to provide a Server to Parallel accelerator tool – planned in early 2021 -- to help clients convert Server jobs to Parallel jobs. The Server to Parallel job migration can be started prior to migrating to Cloud Pak for Data.

What should clients do if they are on DataStage V11.5 ?

- Clients should work with their sales representative to create a modernization path to the DataStage cartridge, as there are various routes that can be taken depending on the environment in question.

What help/assistance is IBM providing?

Description		When Available
Migration License Waiver	Users can request a migration waiver - up to a total of 12 month – during which they can continue to run their excising DataStage / IS environment while migrating to DataStage on CP4D	Now
Job assessment	Users can utilize the job assessment tool from Data Migrators https://upgrade.mettleci.com/ to understand job compatibility and upgrade complexity	Now
CI/CD Tooling	IBM is planning to provide the entire MettleCI tool set (including the above assessment component) to any DataStage / Information Server on Cloud Pak for Data licensee. This tool set supports both upgrade/migration of DataStage as well as CICD	Q4/2020 – Q1/2021
Server to Parallel Job conversion accelerator	IBM is planning to provide users with an accelerator tool to migrate DS Server jobs to Parallel jobs	Q1/2021
Service-led deployment / migration	IBM has a full range of service offerings. Starting from a JumpStart deployment services all the way to complete end to end migration engagements	Now

One of England's major sport leagues uses IBM technology to increase ticket sale and fan engagement

Transformation:

- Creation of a centralized customer hub
- IBMs Integration and Governance capabilities provide the backbone to cleanse, integrate, merge, manage and govern the CRM
- Now providing a 360-degree view of each fan's individual behaviors and preferences based on their interactions across all its channels

Result:

Increased ticket sale by 50%

43% increase in email marketing campaign success

360 View

Business Goal:

- Boost fan engagement on the digital channel to drive investment and participation in the sport

Challenge:

- Lacking a single view of a customer (fan) due to siloed data spread across many separate systems
- High risk of customer dissatisfaction due to inconsistent communication

A state judiciary in the US accelerates its case review process by 98 percent

Transformation:

- A risk assessment tool keeps people from being jailed while awaiting trial

- Utilizing IBMs Data Integration and MDM solution, the Judiciary established an automated risk assessment system that takes less than three minutes per case. This translates into a 98 percent time savings compared to three hours required for manual information gathering and records analysis for an expected savings of USD 10 million annually.

Result

98%-time savings during case review

up to $10M annual cost savings

Compliance & Risk Management

Challenges

Jailing people often makes their circumstances worse by keeping them from working and losing their jobs causing defaults on home mortgages, family stress and more.

This judiciary is finding ways to eliminate the need to jail people while waiting for their day in court.

Learn more about DataStage

Performance Tech Paper: Up to 30% Faster Execution Time

Video: Auto-scaling and workload management

Blog: Data Integration: The vital baking ingredient in your AI strategy

Tech Talks: Community webinars

Solution brief: IBM DataStage

Join the online DataStage community: bit.ly/datastage-community

Appendix

Canadian Health Provider improves quality of care through analytics

Transformation:

- Creating and utilizing a single integrated analytics platform

- End to end process from high speed integration of source data into the warehouse to sub-second response for complex analytics driving the insights dashboard and applying optimization

Result

Clinicians and managers Instant insights into operations

Improve efficiency and reduced cost

Analytics & Insights

Business Challenge

Decision makers must monitor key metrics that influence the hospital's care processes and funding reimbursement.

The hospital needed an analytics architecture that could provide dynamic insight into large volumes of data.

US Bank leverages IBMs Cloud Pak for Data platform to accelerate AI

Transformation:

- Using IBM Cloud Pak for Data System for rapid deployment and scaling of AI.

- Single interface platform for end-to-end enterprise analytics and easy creation of a customer 360 system

- Utilizing an integrated stack of services for data acquisition, cleansing, cataloguing, collaboration and data science

- Supports mandates for compliance with privacy regulations

Result

- Accelerated AI time to value

- Improved Client experience and analytics results

Data Hub / Data Exploration

Benefits of an integrated Stack

"…The integrated stack contains what we need to improve data quality, catalog our data assets, enable data collaboration, and build/operationalize data sciences. We're able to move quickly with design, test, build and deployment of new models and analytical applications."

Expert Lab Services
Modernization Launch Package

Assessment

- Assessment tool for sizing and estimating DataStage & QualityStage Job Migration:

  - Graph Analysis

  - Computational Complexity

  - Compliance Rules

Activities

- REMOTE ASSESSMENT | Review high level architecture, business objectives & prerequisites

- DISCOVERY WORKSHOP | Migration Planning; Determine sample set for migration

- INSTALL | Installation + configuration of OpenShift + Cloud Pak for Data + DataStage Enterprise (or Plus)

- OPERATIONAL WORKSHOP | Walk client through the platform & demonstrate basic functionality

- DATASTAGE CODE MIGRATION | Migrate DataStage project(s) to Cloud Pak for Data environment & perform sample testing

- MIGRATION GUIDANCE | 1 Month of Expertise Connect Advanced resource to support DataStage migration

DataStage Upgrade Process - Comparison

	Traditional Upgrade (Manual)	Automated Upgrade
Upgrade Approach	Reactive (Trial & Error Method)	Proactive (Upfront Analysis & Visibility to Ongoing upgrade Process)
Complexity	Mostly Unknown (Upgrade blockers unknown)	Upfront Complexity Analysis Report available. Also, provides Compliance summary to adopt best practices.
Upgrade Duration	Longer Timeline • Unit Test must be performed on almost all jobs due to job dependency • Late Discovery of Risks and Issues therefore more time spent on troubleshooting and issue remediation • Manual migration of project settings and environment configurations. • Data Comparison is manual and more time consuming to identify which jobs produces inconsistent results	Accelerated Timeline • Unit Test can be limited to jobs high complex jobs, unique patterns, jobs that has upgrade impacts without worry about dependency. • on almost all jobs due to job dependency • Risks / Issues known based on upfront analysis of code • Automated migration of project settings and environment configurations. • Data Comparison is automatic and quicker as it provides visual representation of discrepancies at field level as well as tied to a given DataStage job
Agile Method	Very limited due to jobs dependency	Agile based upgrade based on Integration with DevOps and OSH intercept capability
Monitoring & Tracking Upgrade Process	Manual reporting required	Provides Live dashboard, Report cards, Workflows and monitoring (consolidated view of multi-source SDLC)