

Power your journey to AI with IBM DataStage

Cost optimization on Multi cloud environments using IBM DataStage



Beate Porst – porst@us.ibm.com
Program Director Offering Management
Data and AI

July 1st, 2020



Multi and hybrid cloud deployments

90%

of enterprise customers
are hybrid cloud environments



Customers are moving
to a modular /
disaggregated
architecture

Performance is the
primary driver to move
from public to private
cloud

IDC Q1/202

Application
performance and
security are amongst
the key outcomes for
cloud investments

IDC Q1/202

By 2023, **75%** of all databases will be on a cloud platform, reducing the DBMS vendor landscape and increasing complexity for data governance and integration.

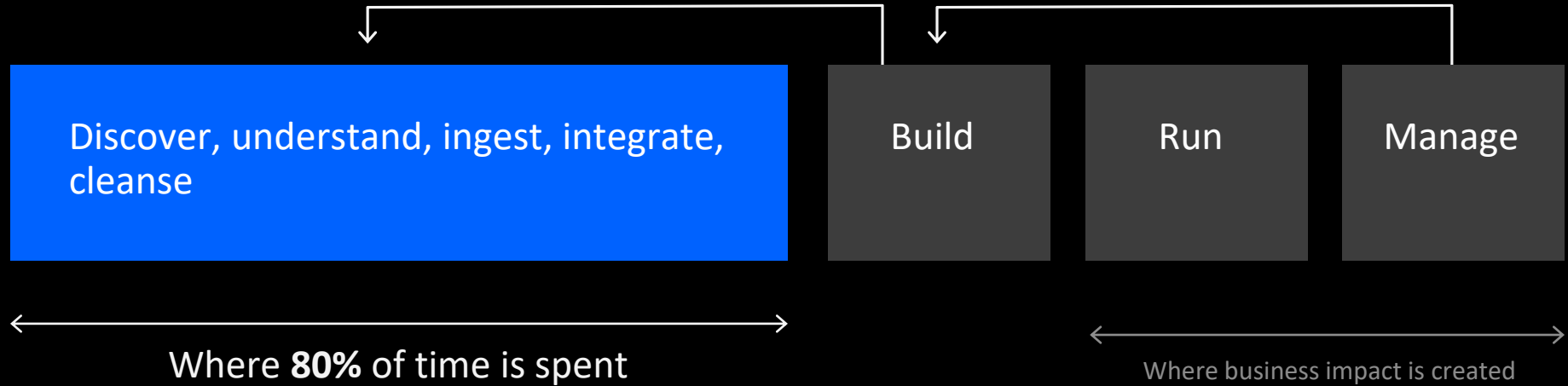
Gartner, Data & Analytics 2019

Why is cost optimization more important than ever?

- COVID-19 is having a major business impact
- 64% of Q1/2020 IDC Cloud survey respondents state:
 - COVID-19 impacting their Cloud Strategy in the short and mid term
 - with ***majority reducing spending*** both public and private cloud
 - Some will shift heavier to public cloud
 - Some will increase the number of cloud vendors

Getting data ready is hard

Source: Data scientist report, Figure Eight Inc



Transaction aware replication
and synchronization of data
between peered data sources

Data
Replication

Self-service style data
integration. Focusing on
simplicity for non technical
users

Data
Preparation

Event-based triggering of
data processing. E.g. a sales
transaction

Event / Near
Realtime
Integration

Instant analysis of a
continuous stream of data
E.g. stock ticker analysis

Streaming
Analytics

Discrete mass data
movement and
transformation

Bulk / Batch
Integration

Instance access style data
integration avoiding data
persistence

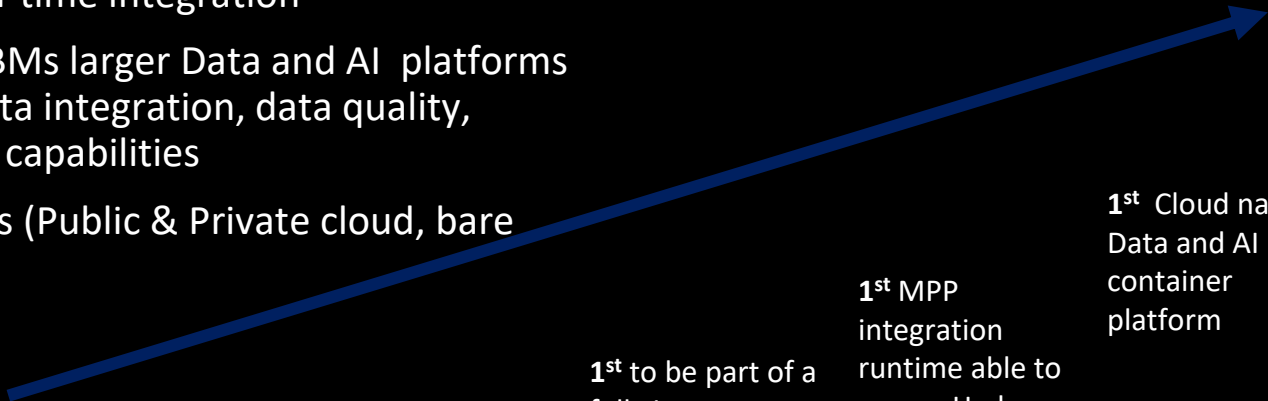
Data
Virtualization

Data
Integration



IBM DataStage – 20+ years invention and leadership in Data Integration

- Market leading Data Integration solution
- Supporting batch and real-time integration
- Natively integrates into IBM's larger Data and AI platforms to combine with other data integration, data quality, governance and analytics capabilities
- Many deployment choices (Public & Private cloud, bare metal, Hadoop)



1st commercial
ETL tool on the
market

1st Parallel
Execution Engine
in a commercial
ETL tool

1st to be part of a
fully integrate
Data
Management
Platform

1st MPP
integration
runtime able to
run on Hadoop
and stand alone

1st Cloud native
Data and AI
container
platform

IBM DataStage on IBM Cloud Pak for Data

Built for Hybrid Cloud Data Integration

Increased design productivity

.....

ML augmented design and autonomous data delivery

Versatile data delivery

.....

Supports multiple data delivery styles for in-time integration and data access

Multi-cloud flexibility

.....

Design once, deploy and run anywhere

Increased confidence and compliance

.....

Cleanse, standardize and protect data in flight to increase confidence and compliance

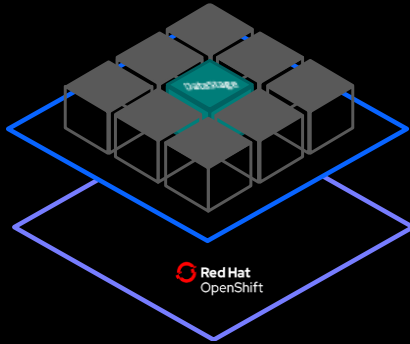
Faster workload execution

.....

Automatic workload balancing and best of breed parallel processing

DataStage – Available anywhere you need it

DataStage on *IBM Cloud Pak for Data*



- Fully containerized on a true multi cloud Hyperscale platform
- Run on any cloud including on managed container service

DataStage / Information Server (stand-alone)



- Traditional deployment on bare metal or virtual environments

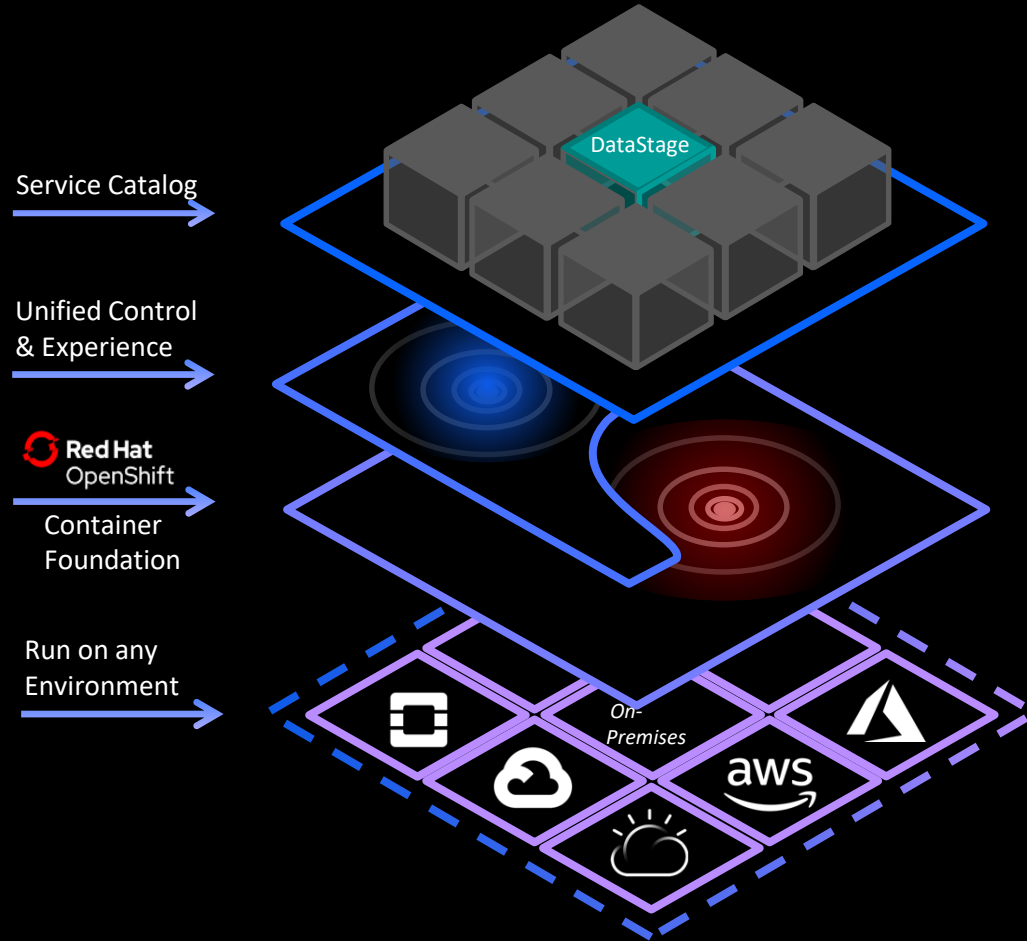
DataStage on IBM Cloud



- DataStage (PaaS) fully managed and provisioned on IBM Cloud

IBM DataStage on IBM Cloud Pak for Data...

- ... is a Service
- on a hyper-scaled Data & AI container platform
- designed for Hybrid Cloud
- utilizing a shared foundation and unified user experience
- and providing state of the art multi-style *Data Integration and Quality* capabilities



Let's look at 5 areas to optimize Data Integration cost

1. Design
2. Runtime & Execution
3. I/O & Storage Reduction
4. Orchestration
5. Management & Operations

1. Design

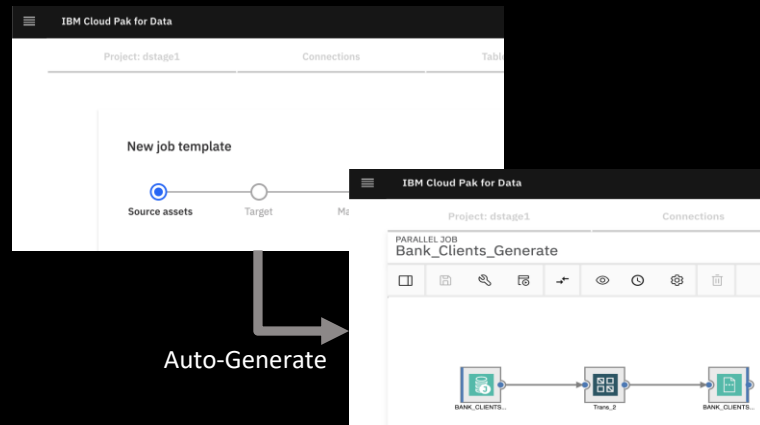
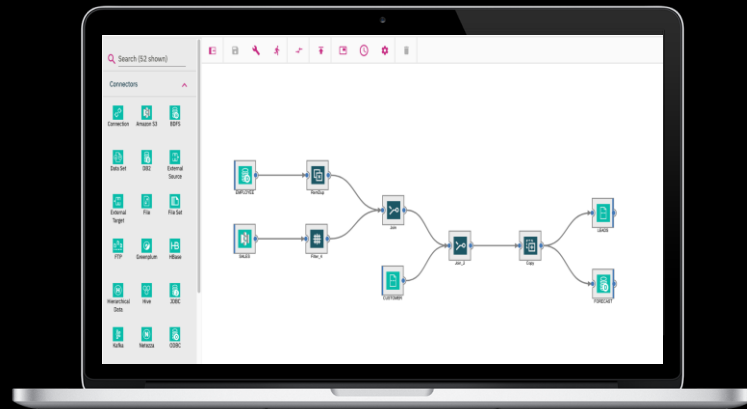
Increased Productivity and Efficiency

What

- “No code” methodology, platform-neutral drag and drop design
- Utilize design automation, shared container, parameterization, schema at runtime
- Utilize pre-built & shared connections and operations
- Implicitly integration logic optimization

Benefits

- Fast design, easy understanding of design logic
- Increase failure resiliency and testing efforts
- Separate Design from Runtime (design once, run anywhere)
- High re-usability & compliance
- Ability to focus on higher level integration logic.
- *Remote* (Data co-located) job execution to minimize egress cost



Designing Integration Flows made easy

Hundreds of pre-built, pluggable Connectors and Operations



Easily connect to:

- Cloud hosted sources
- Hadoop sources/services
- Relational / noSQL databases
- Enterprise and Web Apps
- Realtime / Streaming / Files
- Generic Interfaces, Custom programs

Ready to use Transformations:

- Simple and complex operations
- Warehouse specific operations
- Data Quality, Cleansing and Business logic
- Logical, String, Date, Time, Math operations
- Aggregations
- Hierarchical transformation
- Data security and obfuscation
- Development & Testing

2. Optimizing *Runtime* & *Execution*

What

- Automatic partitioning or repartitioning and pipelining data
- Removal / reduction of un-necessary blocking operations (e.g. sort)
- I/O reduction through pushdown
- Ability to enable restarting checkpoints
- Automatic, elastic workload balancing
- Container-based foundation enabling easier dynamic invocation to support data gravity

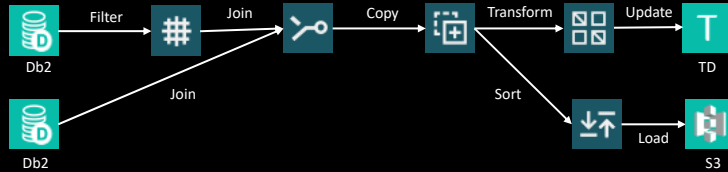
Benefits

- Virtually unlimited scaling (horizontal, vertical)
- Maximizing throughput and minimizing resource congestion
- High resiliency through automatic restart at point of failure capability
- Platform independence, HA and dynamic auto scaling

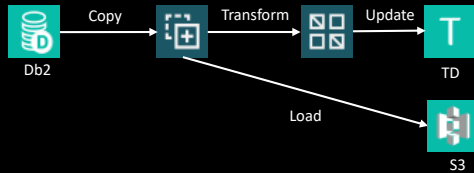
Optimizing for best runtime Performance

Reducing operations and partitioning

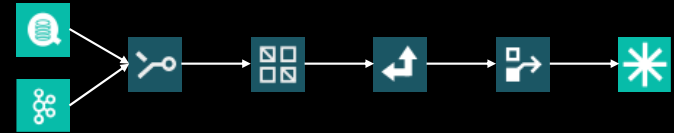
At Design Time



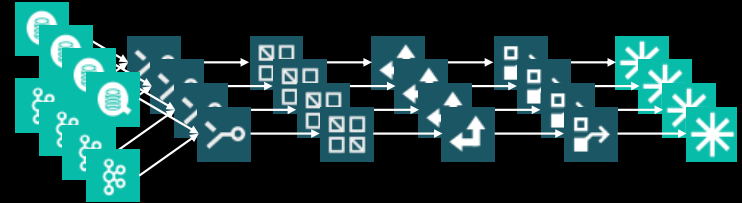
At Runtime



Design Sequential

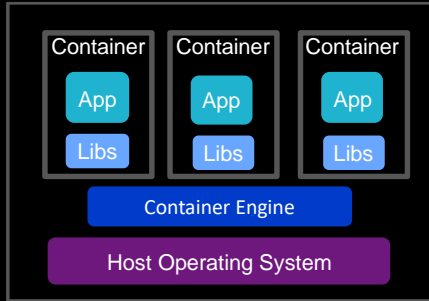


Run Parallel



Optimizing for best runtime Performance

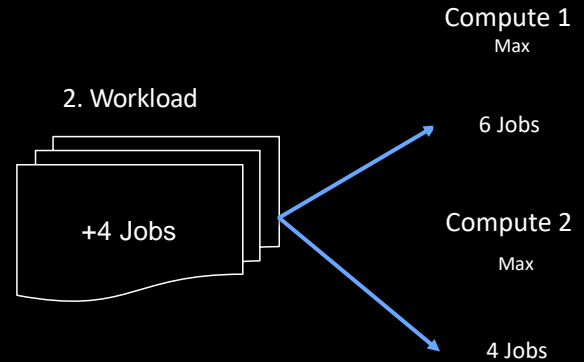
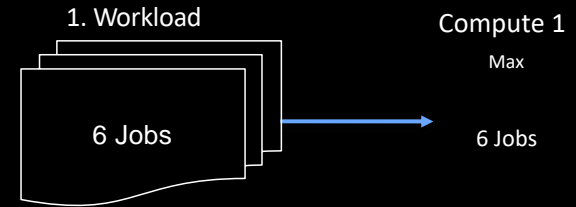
Containerization & Elastic scale



One Container...

...leads to many applications and containers...

Auto Scaling based on workload demand



3. I/O & Storage optimization

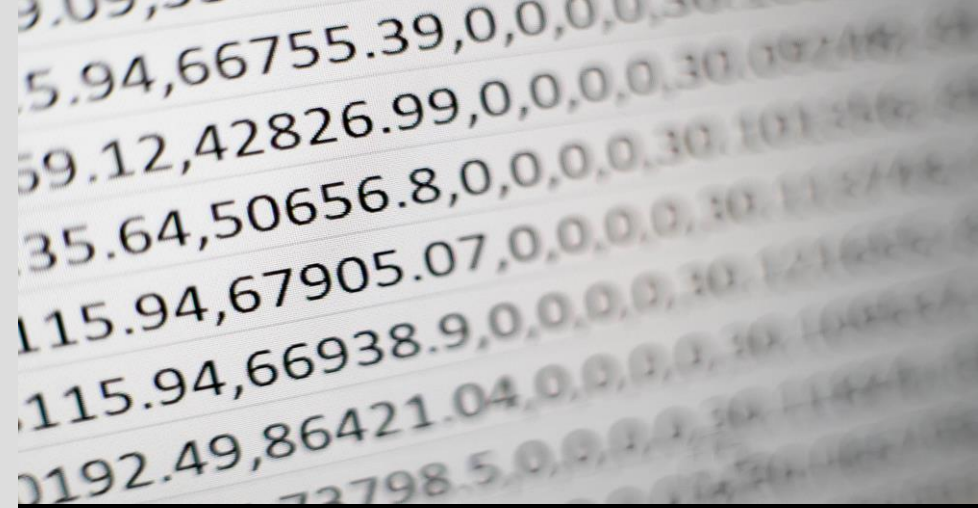
Avoid “Data at Rest” between task

What

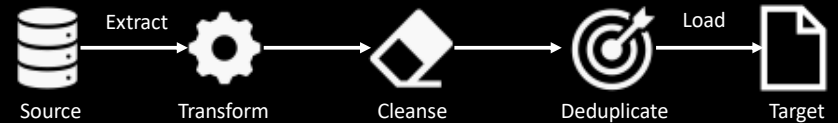
- Avoid break up of integration logic into micro steps
- Utilize DataStage’s power to combine integration with other operations such as Data Quality into a *single in-memory* flow

Benefits

- Reduce storage requirements
- Minimize slow I/O operations
- Simplified data orchestration
- Avoidance of data collisions



No Temporary Staging

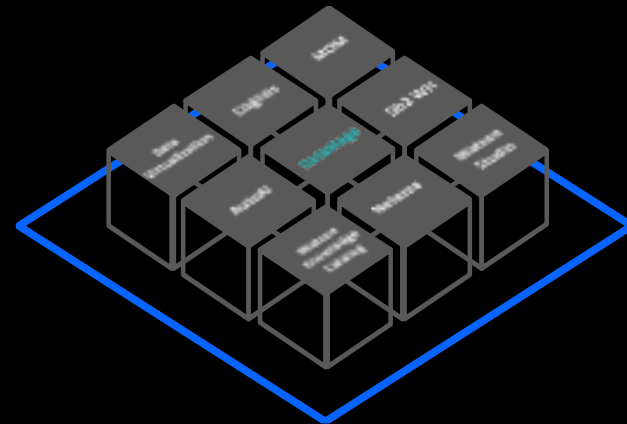


4. Data & Service Orchestration

Combining DataStage with other Cloud Pak for Data Services

What

- Compose and orchestrate data delivery workflows across a broad range of integration services to leverage strength in each
 - *Built in support for Batch, Virtualization, Preparation, Streaming, Event & Real time integration*
- Tie integration services to other Data and AI services
- Simply orchestrate services based on SLA or use case
- Built on a common control plane supporting common asset, metadata and governance



Benefits

- Provides support for entire spectrum of Edge to Analytics use cases
- Provides ability to further minimize I/O operations
- Inherent, out of box data lineage to support governance & compliance

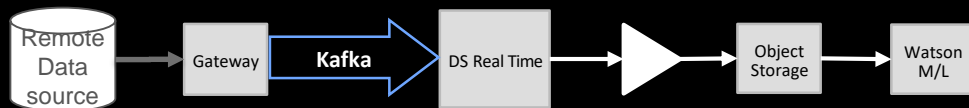
The Whole is greater than the sum of the parts

4 Examples to leverage platform synergies

1

Edge to Analytics -- Creating Data Sets for Model Training

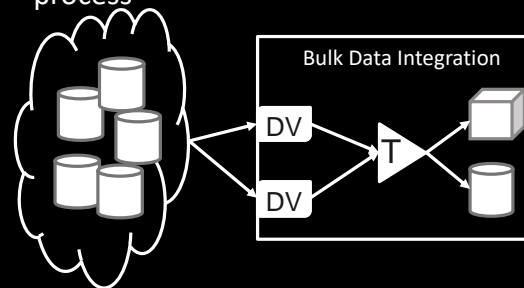
- Capture events or database updates in real time
- Apply complex transformation and enrichment
- Use the enriched data for model training or other AI based analytics



2

Combining Data Virtualization and DataStage (ETL)

- For optimized and simplified data delivery
- Reduction in I/O
- Increased independence within the ETL process



3

Automatic Data Discovery and registration of Data Sources

- Automatically determine data classes, understand schemas and potential data quality issues
- Utilize that information in down stream integration processes .
 - E.g.Run Data Quality rules during the integration process to only process data that meeting specific rule criterias

4

Operationalize Data Preparation or deliver data assets for Data Preparation

- Integrate Data Preparation recopies into a DataStage sequence
- Deliver cleanse & trusted data into a Data Lake ready for LoB users to further refine

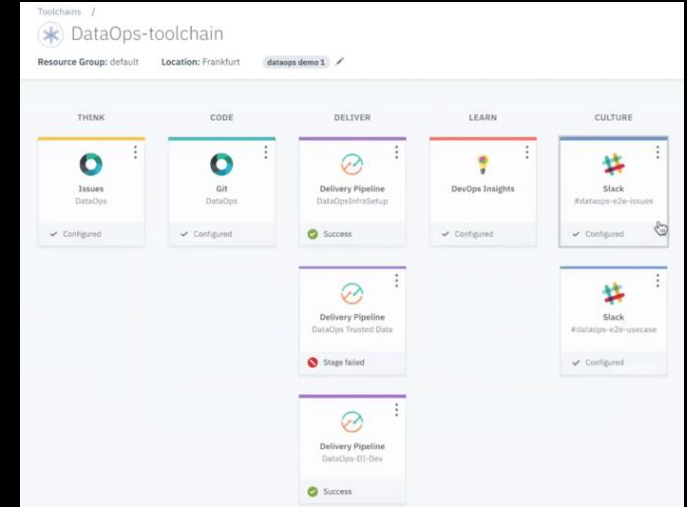
5. Simplifying and automating management and operations

What

- Integration into DataOps / DevOps Tool chains for task automation
- Implicit HA and failover*
- OTA version upgrades, patching and ability for rollbacks*
- Common control plane for managing multiple services at once

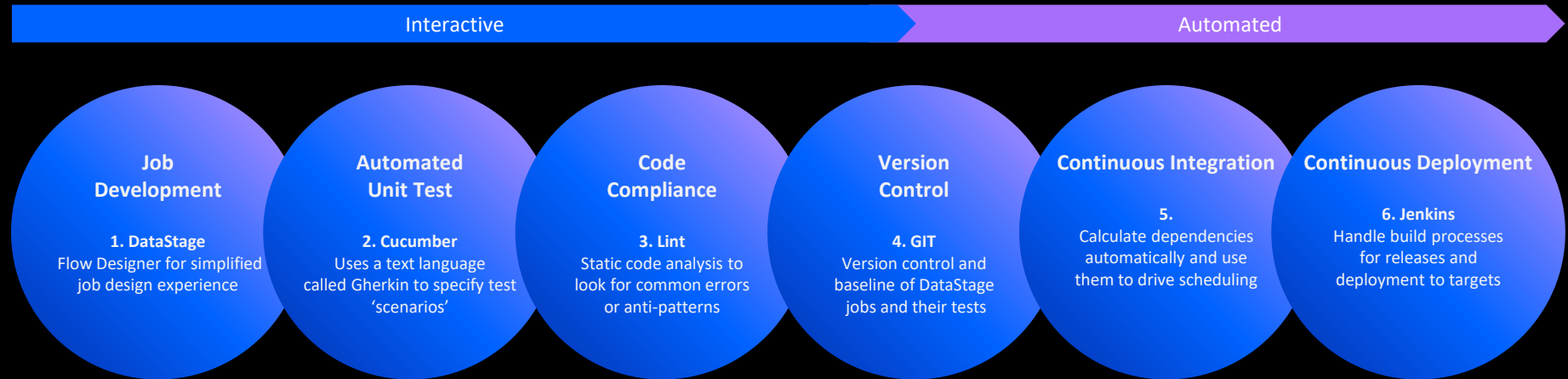
Benefits

- End to end (Design to production) automation
- Ensure high level of compliance
- Automatic platform maintenance



Tightly integrated CI/CD*

An idealized automated delivery system pipeline for workload designed with DataStage



* At present IBM offers CI/CD support direct from IBM's third party solution provider Data Migrators via its MettletCI offering.

dstage1

DataStage Project

dstage1

DataStage Asset

Product_supplier_Join

FAILURE

13 Rules

✓

12 Passed Rules

Rule	Duration	Status
Adjacent Transformers	0.002	SUCCESS
CCMigrateTool Stages	0.003	SUCCESS
Database Row Limit	0.043	SUCCESS
Debug Row Limit	0.007	SUCCESS
Default Naming	0.006	SUCCESS
Hardcoded File Paths	0.005	FAILURE
Job Naming	0.002	SUCCESS
Link Sort	0.003	SUCCESS
Lookup Failure	0.004	SUCCESS
One Dataflow	0.003	SUCCESS
Range Lookup	0.003	SUCCESS

Product_supplier_Join

SPECIFICATION

Product_supplier_Join

DATA

Product-Product

Product_Supplier-Product_Supplier

Supplier-supplier

Output-Output

```

---
- stage: "Supplier"
  link: "supplier"
  path: "Supplier-supplier.csv"
- stage: "Product_Supplier"
  link: "Product_Supplier"
  path: "Product_Supplier-Product_Supplier.csv"
- stage: "Product"
  link: "Product"
  path: "Product-Product.csv"
when:
  job: "Product_supplier_Join"
  parameters: {}
then:
- stage: "Output"
  link: "Output"
  path: "Output-Output.csv"
ignore: null

```

Stage Output of type PxDataSet has a hardcoded path:
/tmp/Product_Supplier_Join.

Test 'Product_supplier_Join' failed

Mar-09 08:37:17

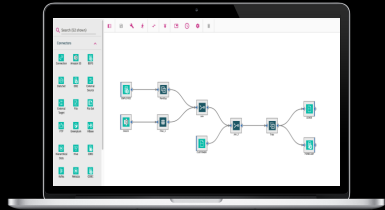
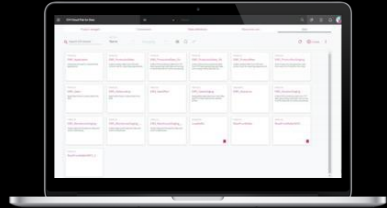
1 output(s) failed to matched expected results while running 'Product_supplier_Join' test.

Output.Output

2 row(s) added to expected output

	SID	PID	NAME	PRICE	PROMOPRICE	PROMOSTART	PROMOEND
+++		100-100-01	Snow Shovel, Basic 22 inch	9.99	7.25	2004-11-19	2004-12-19
+++		100-103-01	Snow Shovel, Super Deluxe 26 inch	49.99	39.99	2005-12-22	2006-02-22
	100	100-101-01	Snow Shovel, Deluxe 24 inch	19.99	15.99	2005-12-18	2006-02-28
---	---	---	---	---	---	---	---

IBM DataStage on Cloud Pak for Data



Future proofing Cloud 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 same operation on Spark parallel engine

9x

Faster design speed than hand-coding

85%

Reduction in infrastructure management time & effort

