Power your journey to AI with IBM Cloud Pak for Data DataStage

Tech-talk: Operationalizing containers within your organization and speed up data pipelines using IBM DataStage

Scott Brokaw
Offering Management - Data Integration
slbrokaw@us.ibm.com

IBM DataOps/ © 2020 IBM Corporation

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.

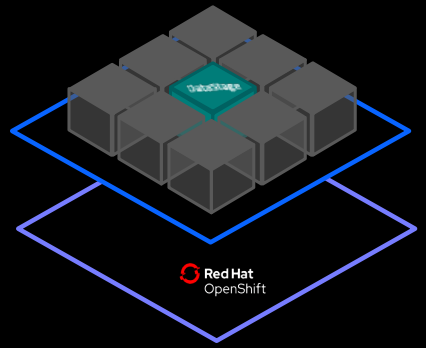
The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

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.

 / © 2020 IBM Corporation

DataStage – Available anywhere you need it

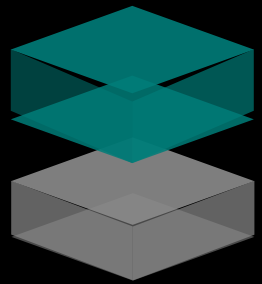
DataStage / Information Server on IBM Cloud Pak for Data



aws *On-Premises*

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

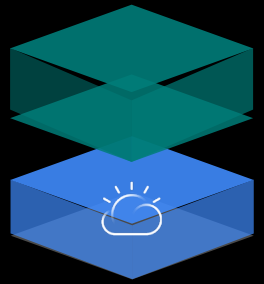
DataStage / Information Server (stand-alone)



aws *On-Premises*

- Traditional deployment on bare metal or virtual environments
- Deploy on-premises, private cloud, or any pubic cloud (BYOL)

DataStage / Information Server on IBM Cloud



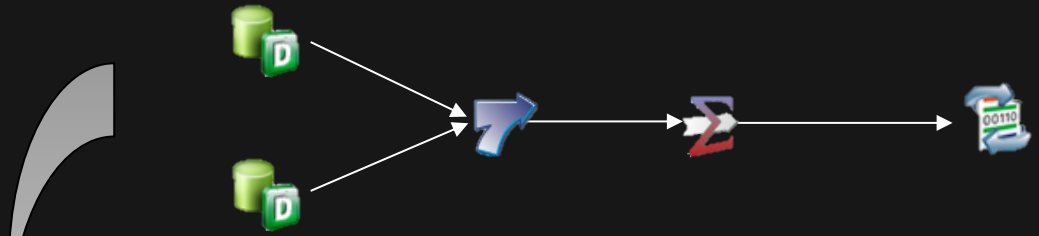
- DataStage fully managed and provisioned on IBM Cloud

DataStage Parallel Engine

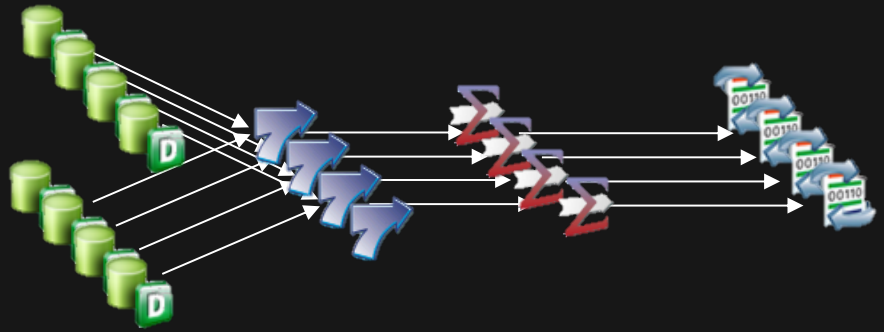
 / © 2020 IBM Corporation

Job design versus execution

User assembles the flow using DataStage Designer



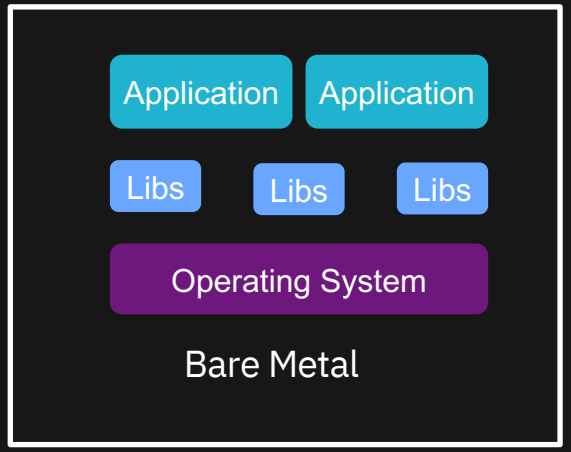
... at runtime, this job runs in parallel for any configuration
(1 node, 4 nodes, *N* nodes)

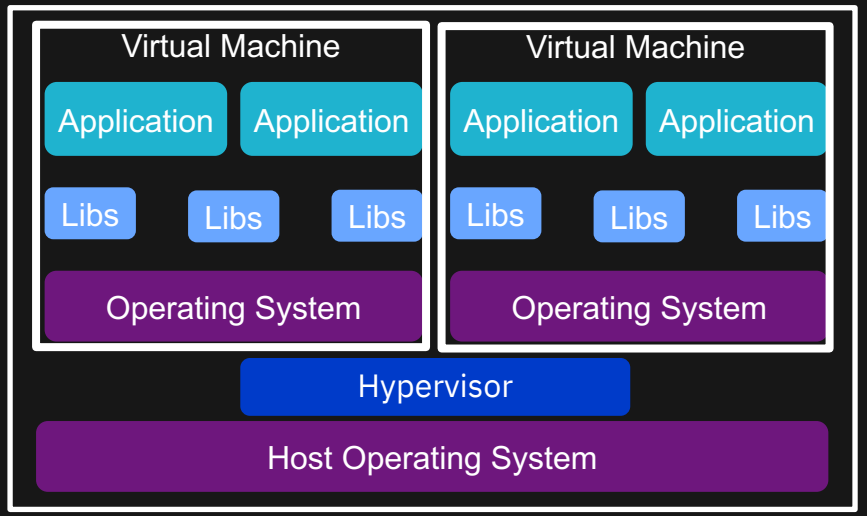


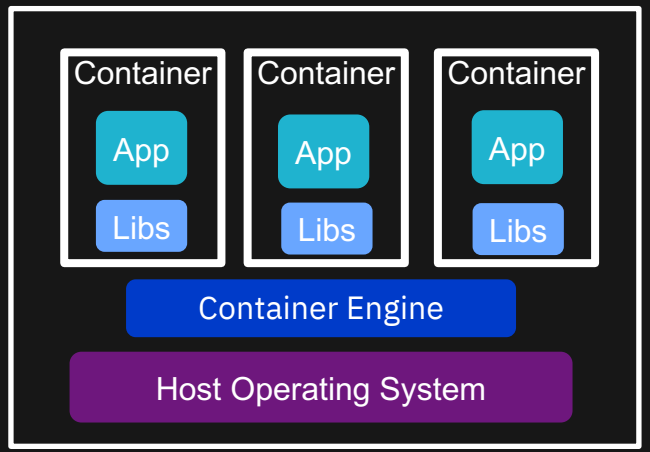
No need to modify or recompile the job design!

 / © 2020 IBM Corporation

Why Containers?


Application
Application
Libs
Libs
Libs
Operating System
Bare Metal


Virtual Machine
Application
Application
Libs
Libs
Libs
Operating System
Virtual Machine
Application
Application
Libs
Libs
Libs
Operating System
Hypervisor
Host Operating System


Container
App
Libs
Container
App
Libs
Container
App
Libs
Container Engine
Host Operating System



One Container...



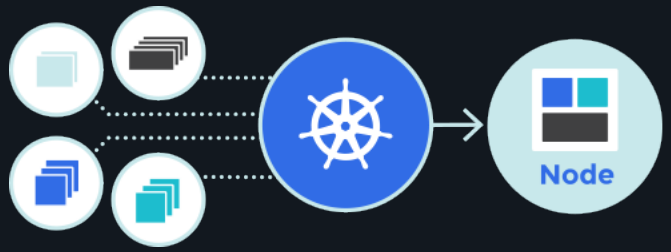
...leads to many applications and containers...

IBM DataOps / © 2020 IBM Corporation

Operationalizing Container Technology

As organizations grow their container strategy, orchestration and management are needed:

- Automated deployment, scaling, and management of containerized applications
- Self-healing
- Automated rollouts and rollbacks of applications

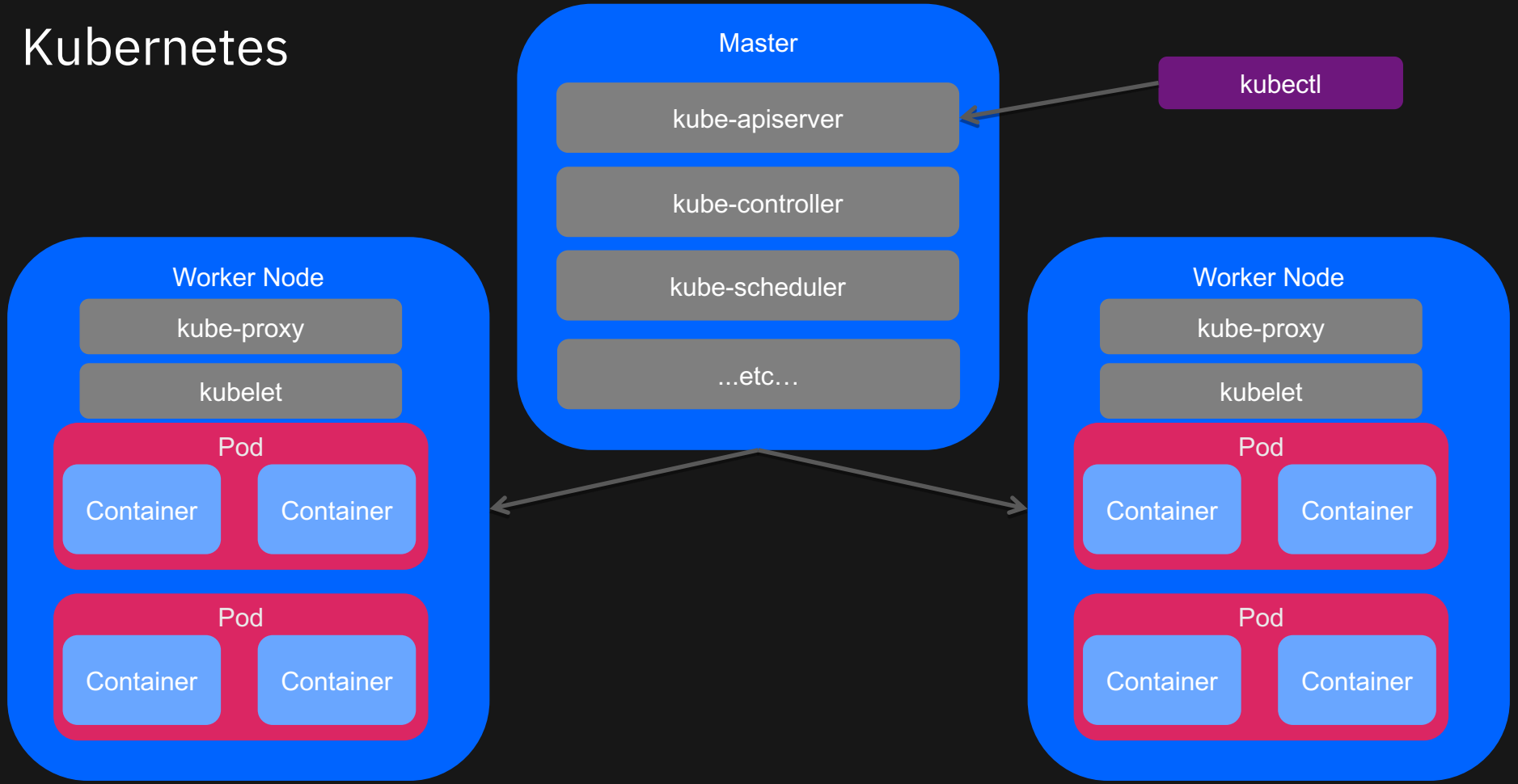


77% of containers are managed by Kubernetes

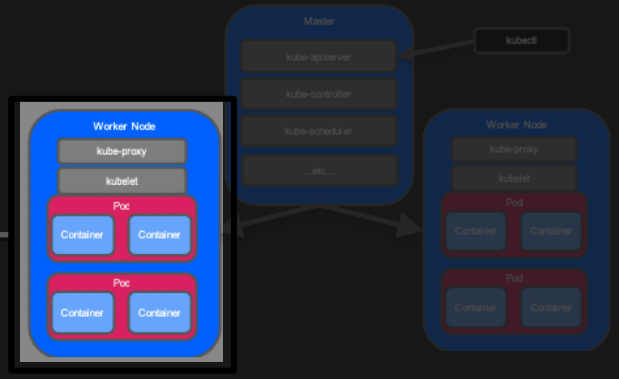
200% Increase in Kubernetes adoption since 2017

Industry has aligned itself with Kubernetes: IBM, Microsoft, Google, RedHat, Amazon

Kubernetes


Master
kube-apiserver
kube-controller
kube-scheduler
...etc…
kubectl
Worker Node
kube-proxy
kubelet
Pod
Container
Container
Pod
Container
Container
Worker Node
kube-proxy
kubelet
Pod
Container
Container
Pod
Container
Container

Kubernetes



Worker Node

Namespace

Deployment

ReplicaSet

Pod

Container

Container

Rolling Update

ReplicaSet

Pod

Container

Container

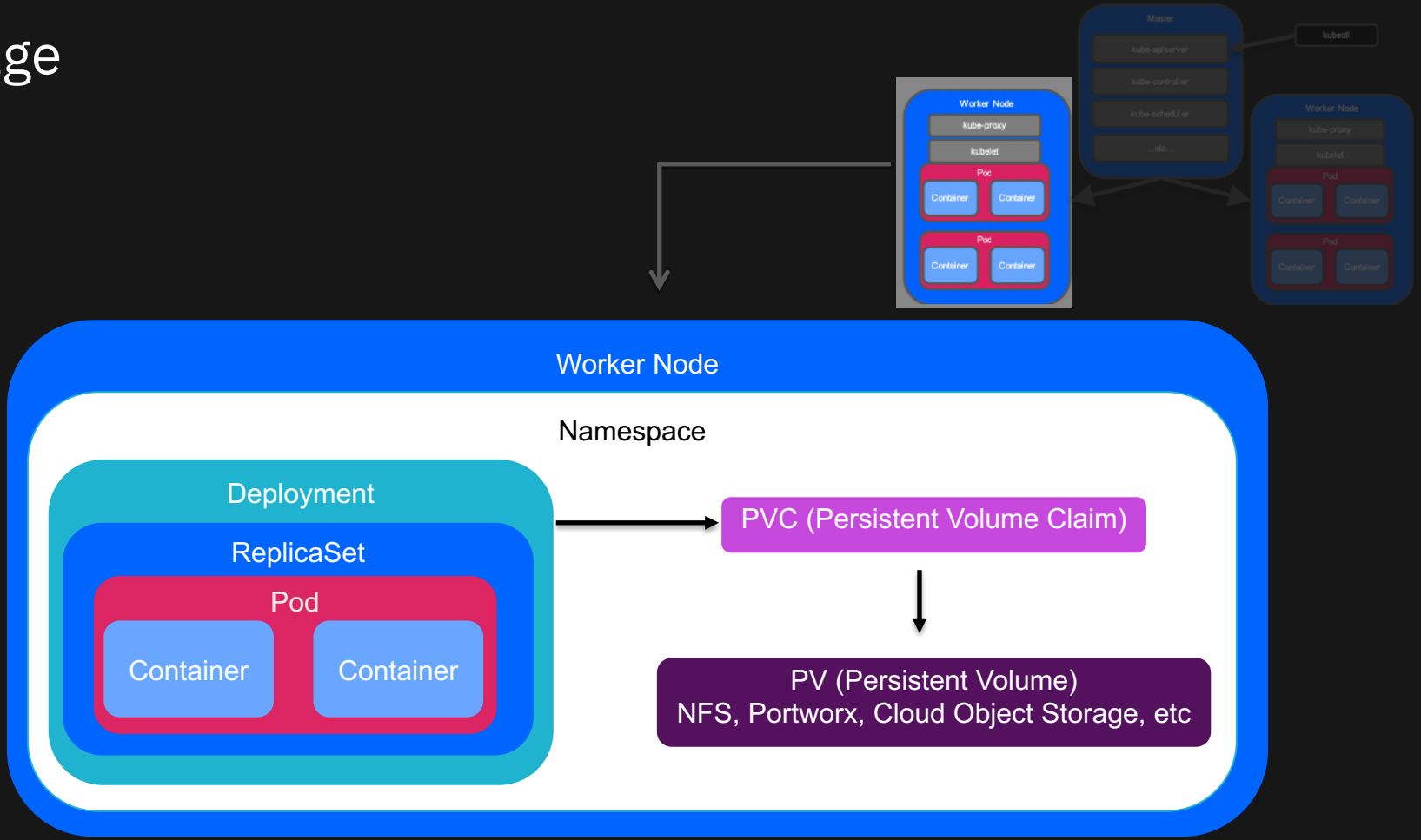
StatefulSet

Pod

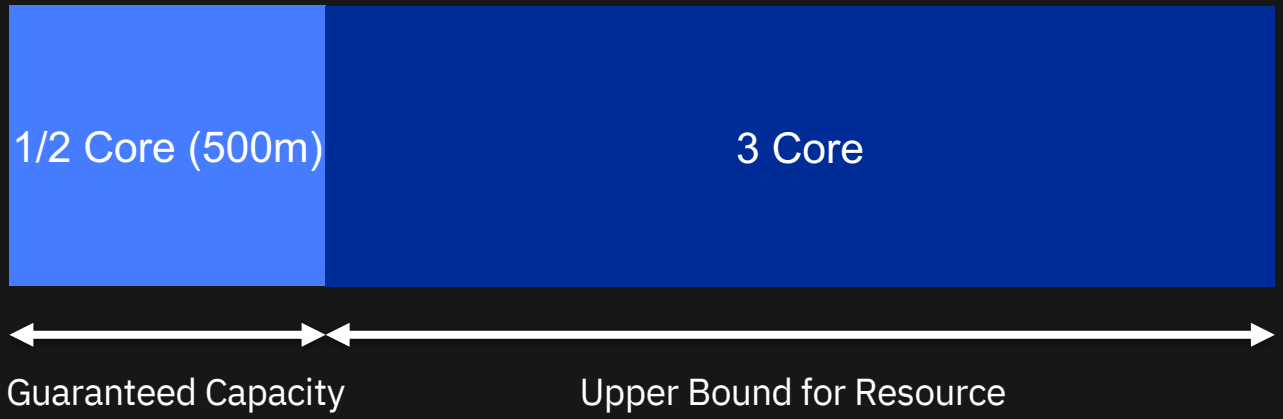
Container

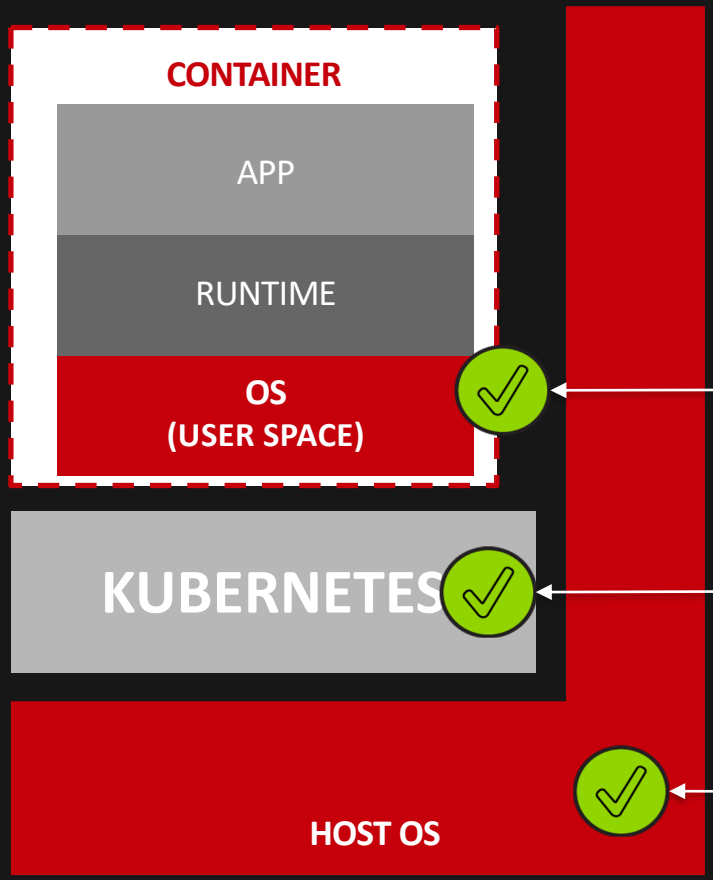
Container

Storage


Master
kube-apiserver
kube-controller
kube-scheduler
...etc...
kubectl
Worker Node
kube-proxy
kubelet
Pod
Container
Container
Pod
Container
Container
Worker Node
kube-proxy
kubelet
Pod
Container
Container
Pod
Container
Container
Worker Node
Namespace
Deployment
ReplicaSet
Pod
Container
Container
PVC (Persistent Volume Claim)
PV (Persistent Volume)
NFS, Portworx, Cloud Object Storage, etc

Resource Requests/Limits





RED HAT® OPENSHIFT

TRUSTED CONTENT
Red Hat provides up-to-date base container images and validated content from dozens of ISV partners

TRUSTED PLATFORM
OpenShift extends Kubernetes with built-in authentication and authorization, secrets management, auditing, logging, and container registry for granular, centralized control

TRUSTED HOST
OpenShift runs on Red Hat Enterprise Linux, the most deployed commercial operating system in the public cloud, trusted by more than 90% of the Fortune 500

IBM DataOps / © 2020 IBM Corporation



Portability

- Often you can often run binaries built for one Linux distribution on another distribution of the same architecture
- Image (OCI standard) can move between different container engines (Docker, CRI-O, containerd, RKT, etc)

Portability ≠ Compatibility

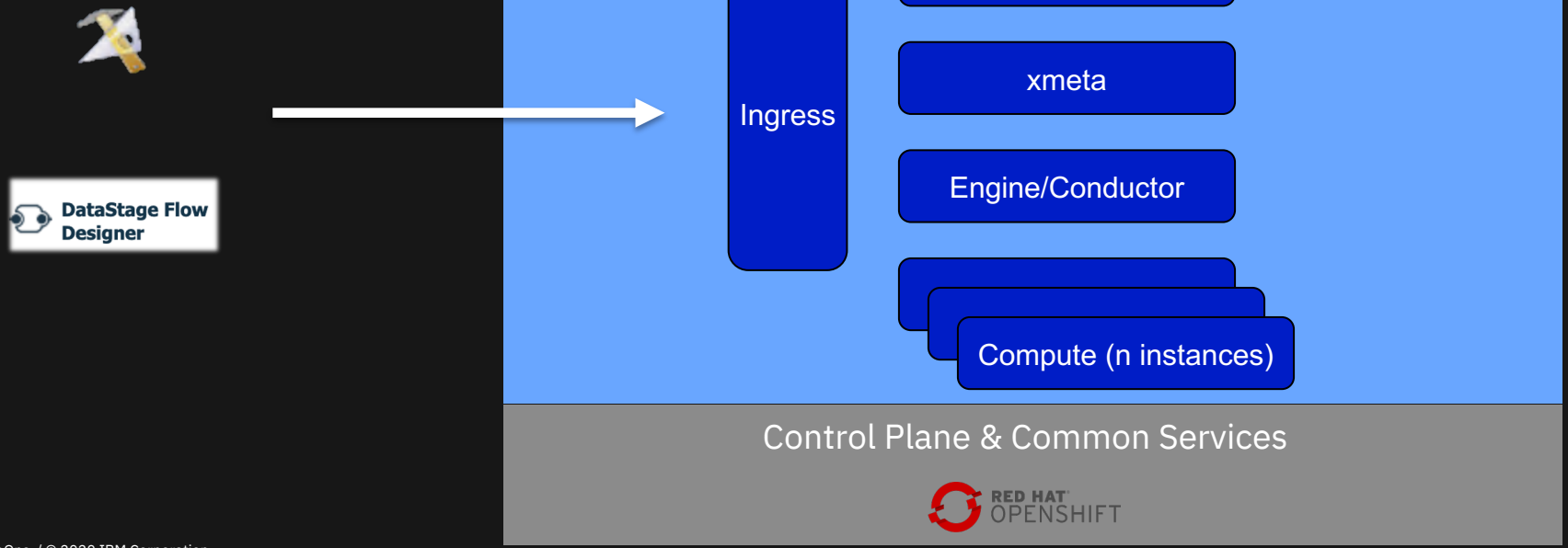
Compatibility

- Images are designed to work with a container engine (UBI)

Supportability

- Containers are Linux

Cloud Pak for Data DataStage



Cloud Pak for Data

1. Services Ecosystem

With a click, access and deploy an ecosystem of 45+ analytics services and templates from IBM and third parties.

2. Platform Interface

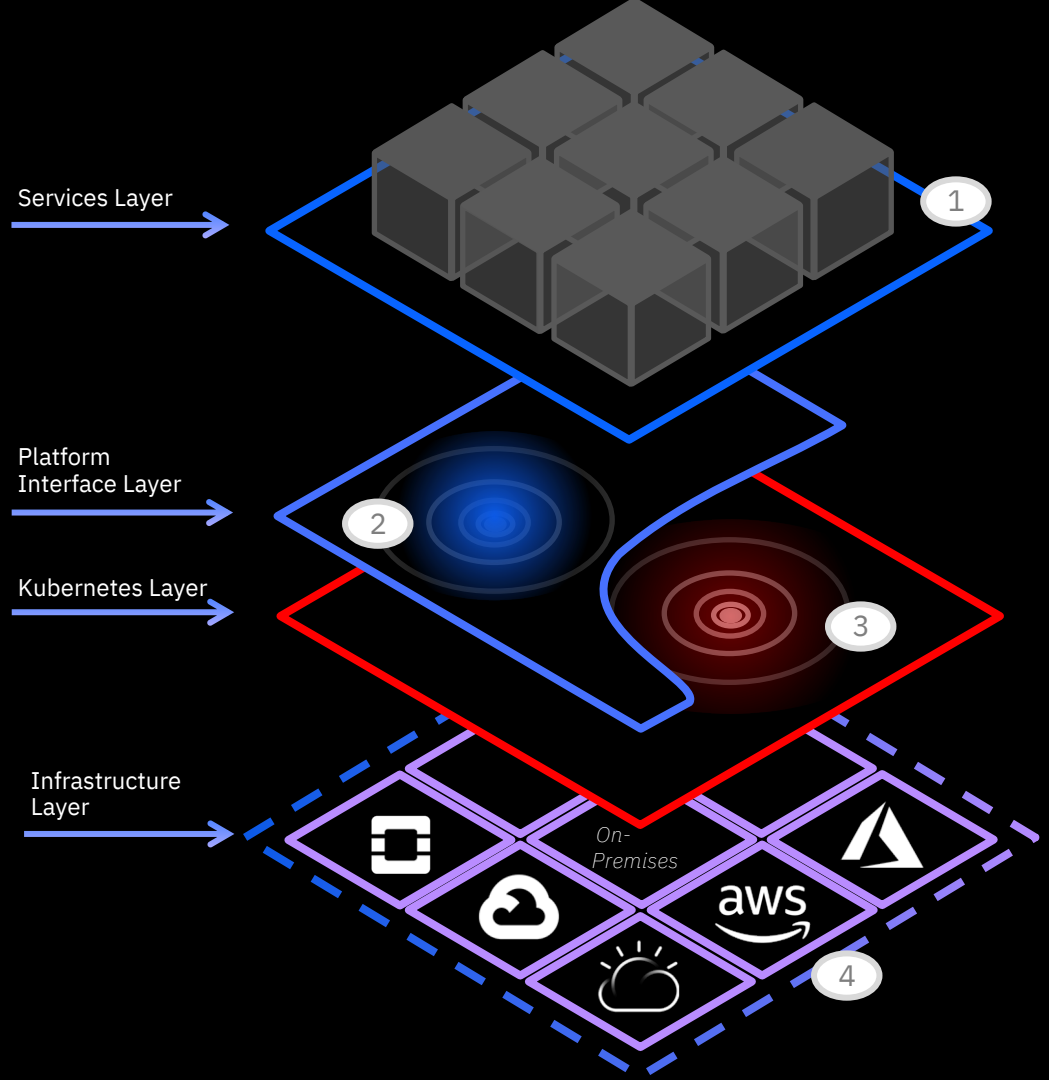
Speed time-to-value with a single user experience that integrates data management, data governance and
analysis for greater efficiency and improved use of resources.

3. Red Hat **OPEN**SHIFT®

Leverage the leading hybrid cloud, enterprise container platform for an innovative and fast deployment strategy

4. Any Cloud

Avoid lock-in and leverage all cloud infrastructures with our multi-cloud approach.



Cloud Pak for Data DataStage

Multi-cloud scalability and elasticity

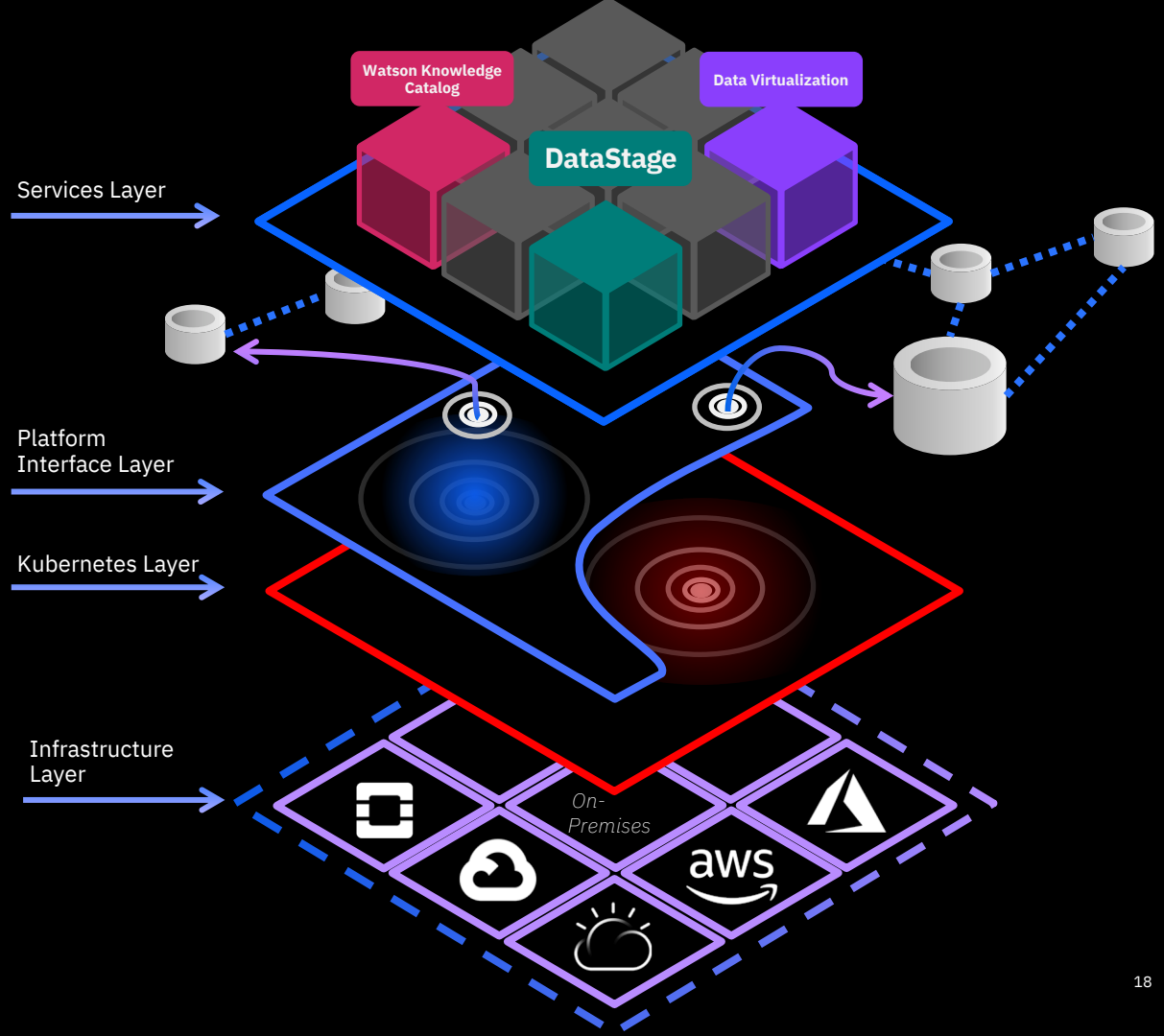
- Design once, dynamically run anywhere with built-in automatic workload balancing, parallelism and dynamic scalability

DataOps and DevOps enabled

- Built-in resiliency, easy operation and CI/CD

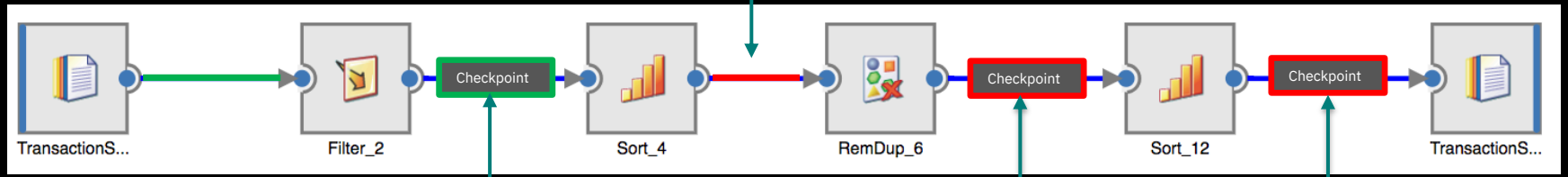
Accelerate AI initiatives

- Automating Data Integration for faster ROI



DataStage: Checkpoint/Restart

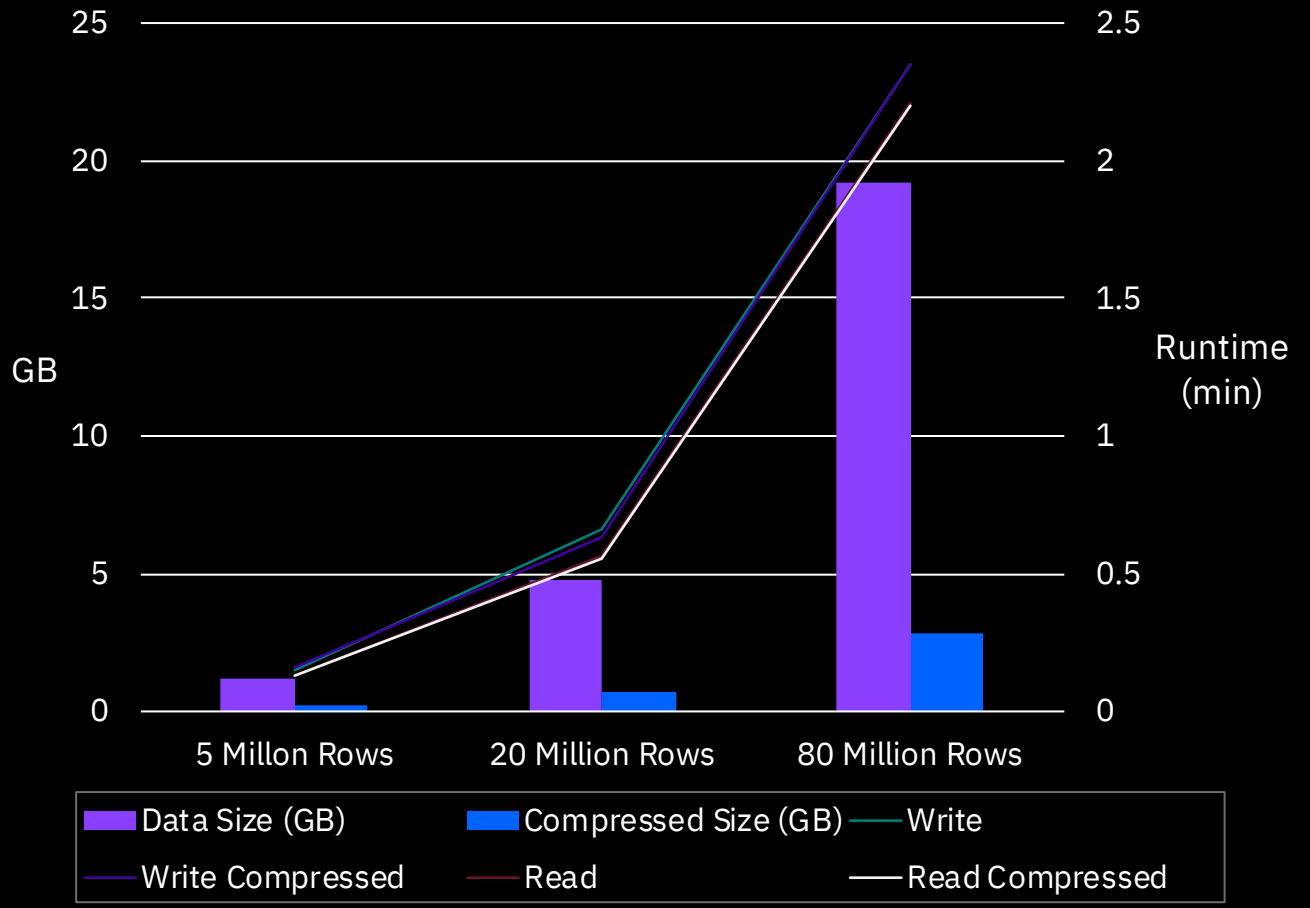
- Failure occurs while the link in red is processing data



- First checkpoint is complete
- Second and third checkpoints are not complete
- The job automatically restarts using data from first checkpoint

Compression Performance

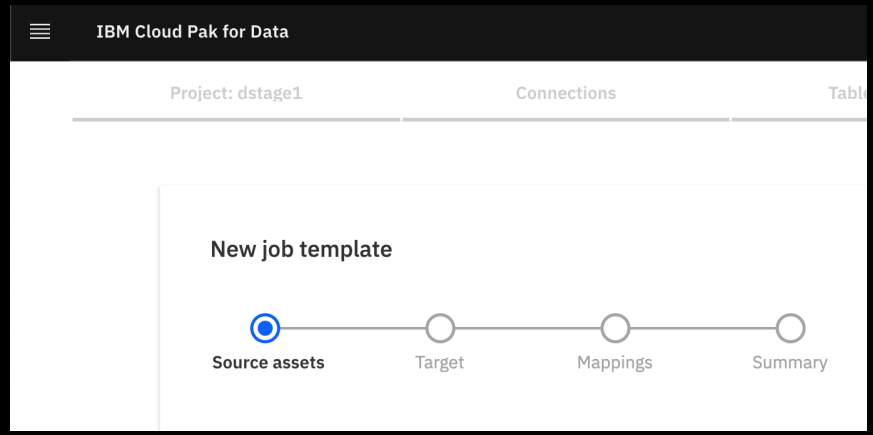
- Sorting
- Datasets
- Checkpoints

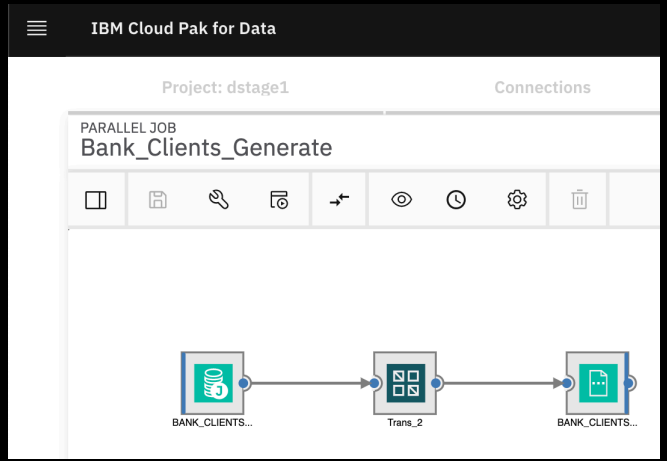


IBM DataOps / © 2020 IBM Corporation

Job Templates – Accelerating ETL for AI

- Reusable Job Templates to auto-generate ETL job(s)
- Rule sets to enforce patterns
- Simplify metadata mappings





Auto-Generate

PXRuntime

- New microservice to modernize DataStage

Why?

- Allow the DataStage stack to evolve and take advantage of new technologies concepts
- Fit better into the containerization model

Benefits?

- Clear text logs...ability to easily integrate into your own ELK stack, etc.
- No more log corruption or Universe related issues
- Autoscaling containers based on workload
- Compatibility

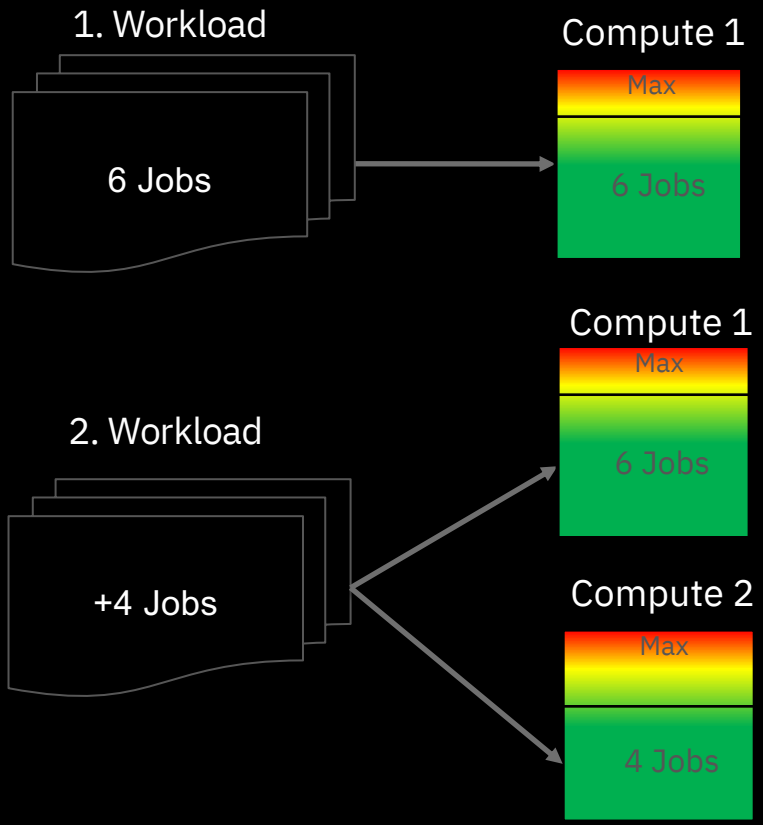
Built-in automatic workload balancing and best of breed parallel engine

Unlimited scaling (horizontal, vertical) using PX engine

Automatic load balancing to maximize throughput and minimize resource congestion

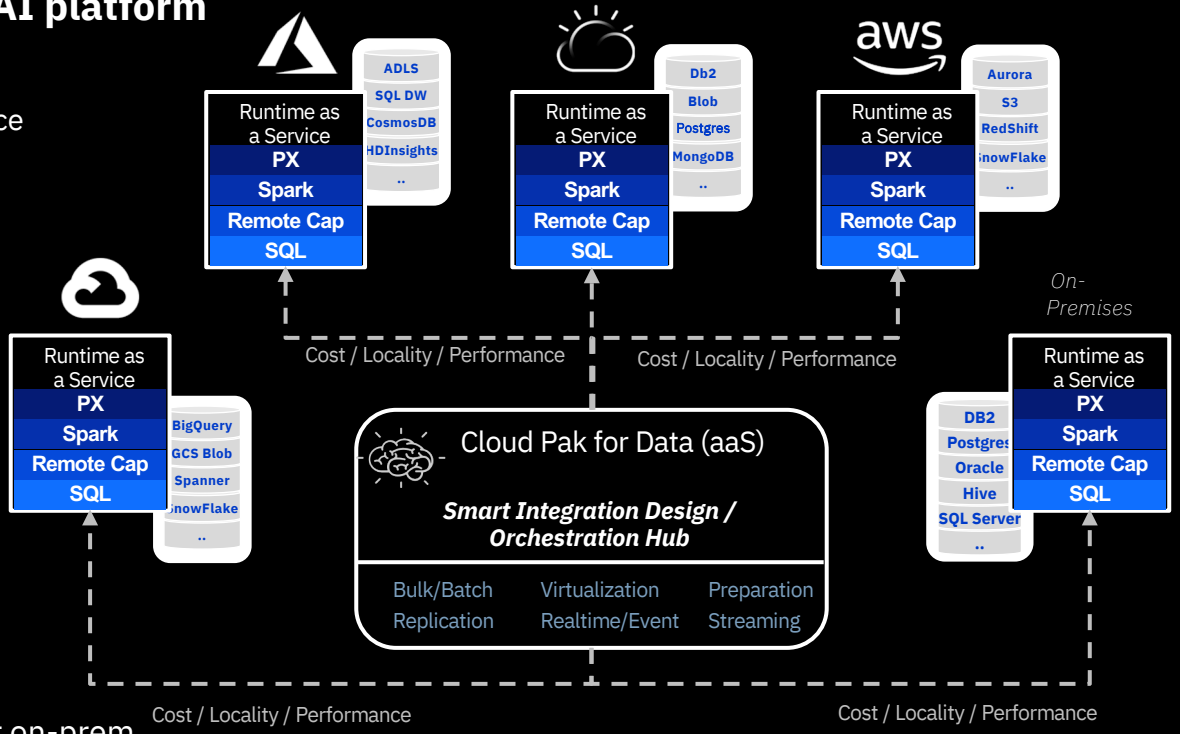
Supports to run resource intensive workloads in parallel pipelining

Built on container architecture to allow for handling of any data volume and execution on any environment



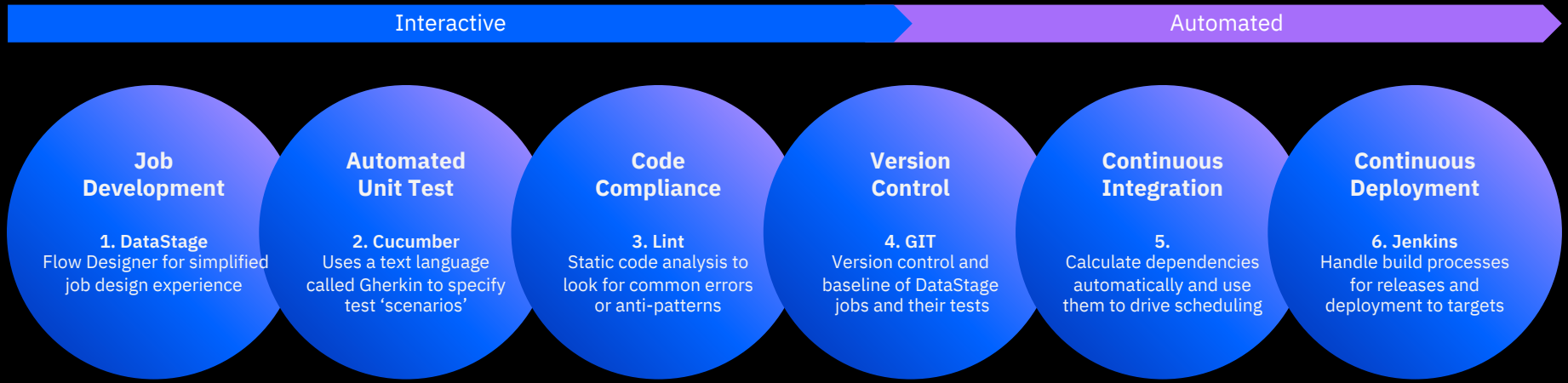
Project Tahoe: Reinventing DataStage
upon cloud native values

- **Integrated with the IBM data and AI platform**
  - Cloud Pak for Data and IBM Cloud
  - Common canvas on Cloud Pak for Data
  - Data integration, machine learning, data science
- **Design Automation**
  - Accelerate well known pattern
  - Automated workflows
- **Governance infused**
  - Catalog integration
  - Policy integration
- **Polyglot Execution Engines**
  - Spark, IBM PX, Virtualization, replication
- **Smart and optimized data flows**
  - Data Gravity
  - Distribute processing to multiple clouds or on-prem



IBM Cloud Pak for Data DataStage has built-in resiliency and supports 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 MettleCI offering.

Product_supplier_Join

SPECIFICATION

Product_supplier_Join

DATA

Product-Product IN

Product_Supplier-Product_Supplier IN

Supplier-supplier IN

Output-Output OUT

```
---
given:
- 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
```

Test 'Product_supplier_Join' failed

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	PI
+++		100-100-01	Snow Shovel, Basic 22 inch	9.99	7.25	20
+++		100-103-01	Snow Shovel, Super Deluxe 26 inch	49.99	39.99	20
	100	100-101-01	Snow Shovel, Deluxe 24 inch	19.99	15.99	20
...	...	...	...	...	...	...

dstage1

DataStage Project

dstage1

DataStage Asset

Product_supplier_Join

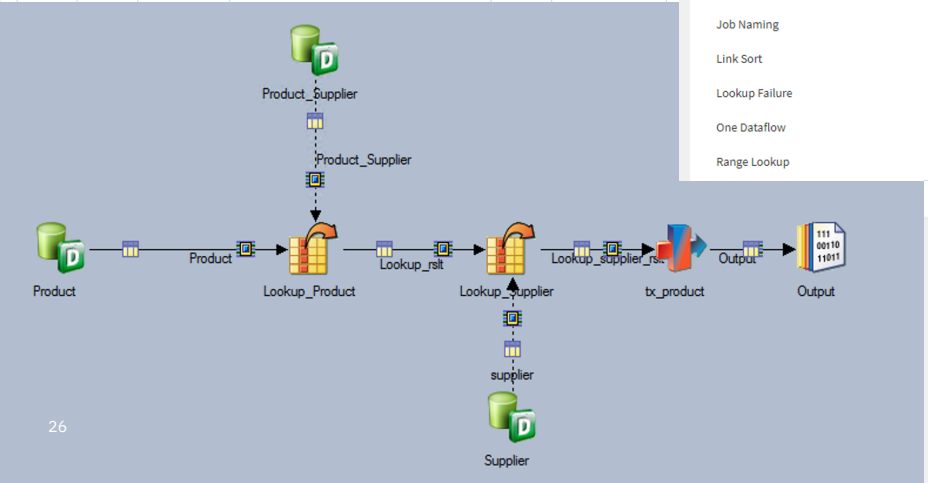
FAILURE

13 Rules

12 Passed Rules

1 Failed Rules

Rule	Duration	Status	Message
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	Stage Output of type PxDataSet has a hardcoded path: /tmp/Product_Supplier_Join.
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	



Output - Data Set

Stage | Input

Input name: Output

Columns... View Data...

General | Properties | Partitioning | Columns | Advanced

Target
File = /tmp/Product_Supplier_Join
Update Policy = Overwrite

File: /tmp/Product_Supplier_Join

Information:

Switch to multiline editor
Insert job parameter...
Browse for file...

