# Power your journey to AI with DataStage

*IBM Data Integration – Vision and Roadmap*
September 10, 2020

***Data Integration Offering Management***
Scott Brokaw, Principle Offering Manager - slbrokaw@us.ibm.com
Upasana Bhattacharya, Senior Offering Manager – upasana.bhattacharya@ibm.com

# 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.

# **DataStage** – Reinventing and leading time and again

## 1997

1st commercial ETL tool on the market

## 2001

1st Parallel Execution Engine in a commercial ETL tool

## 2006

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

## 2015

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

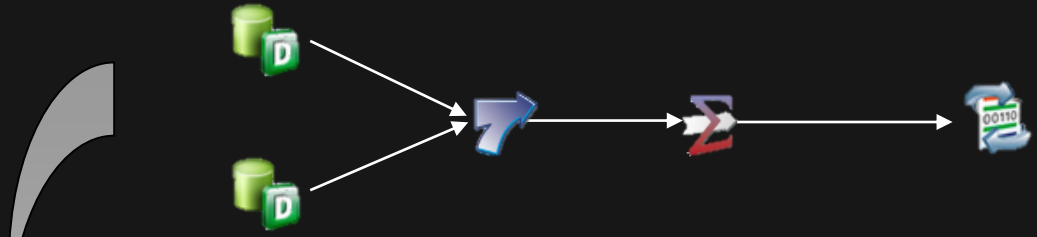
## 2019

1st Cloud native Data and AI container platform

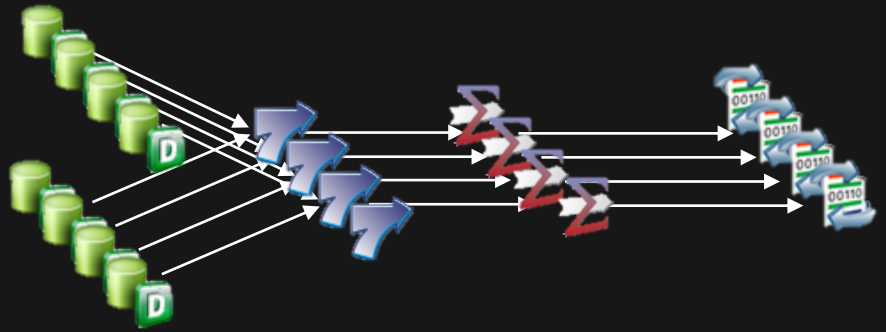
# *DataStage Parallel Engine*

## Job design versus execution

User assembles the flow using DataStage Designer

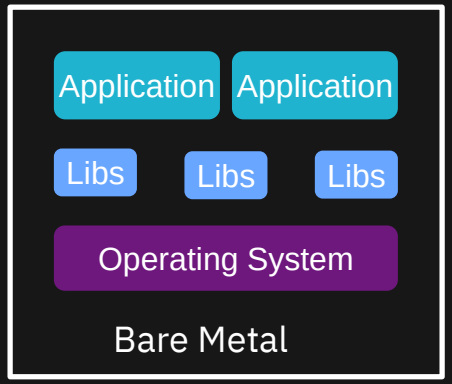


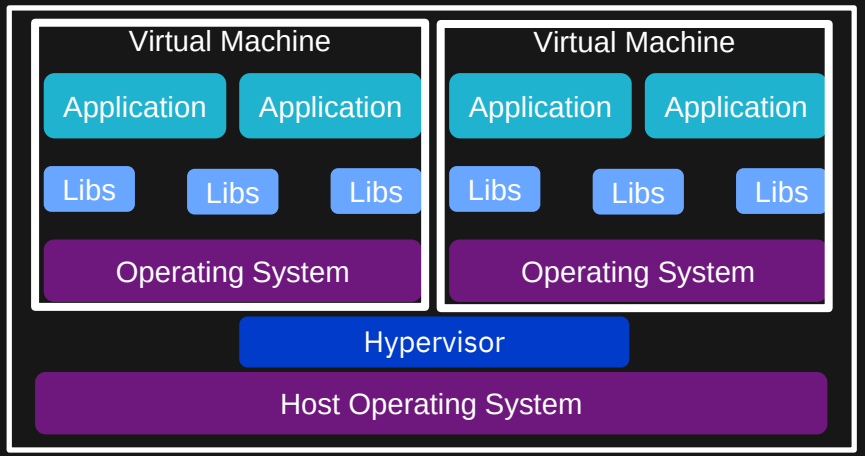
... at runtime, this job runs in parallel for any configuration
(1 node, 4 nodes, *N* nodes)

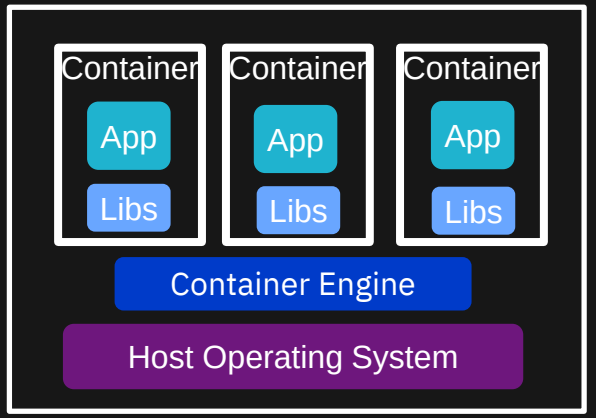


No need to modify or recompile the job design!

# Why Containers?



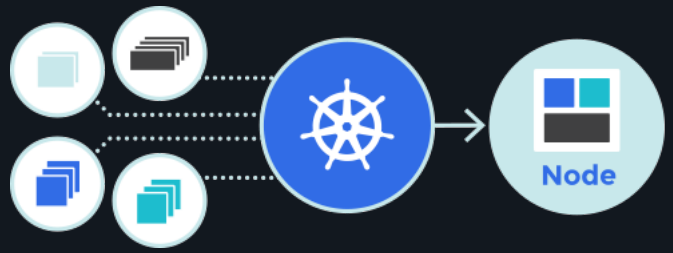
One Container...

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

# 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

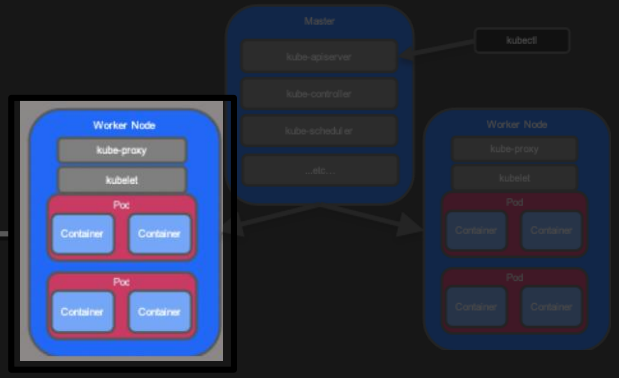



*77% of containers are managed by Kubernetes*

*200% Increase in Kubernetes adoption since 2017*

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

# Kubernetes




Worker Node

Namespace

Deployment

ReplicaSet

Pod

Container

Container

Rolling Update

ReplicaSet

Pod

Container

Container

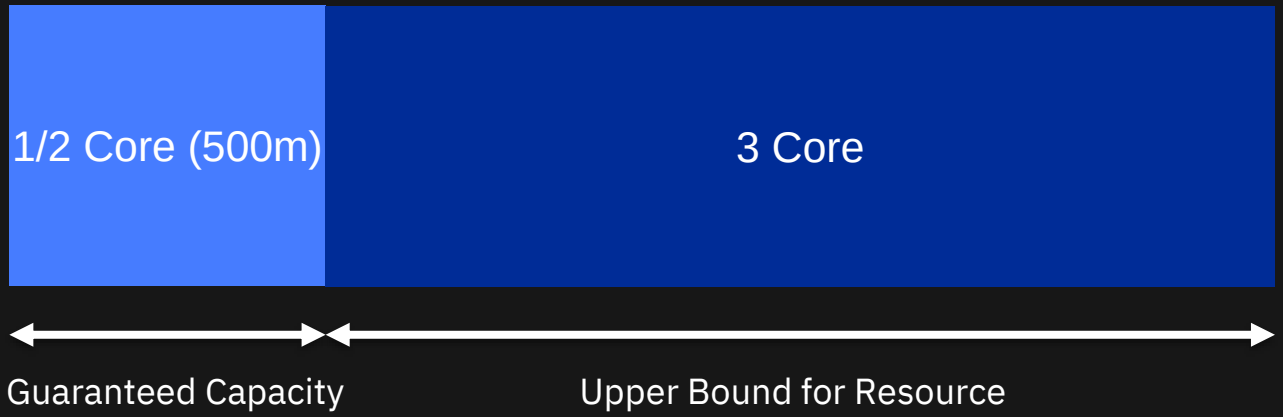
StatefulSet

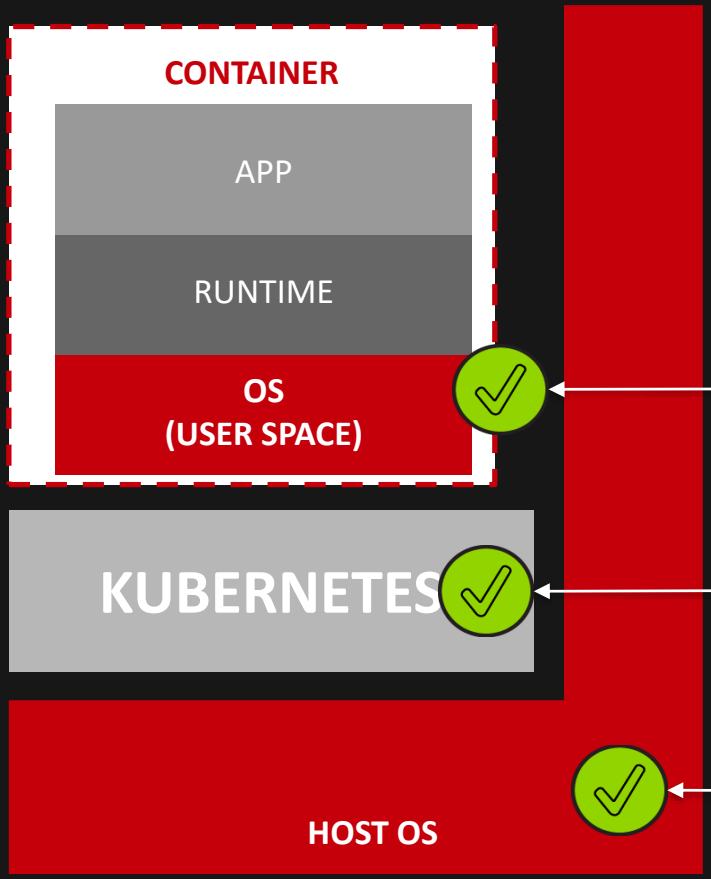
Pod

Container

Container

# Resource Requests/Limits





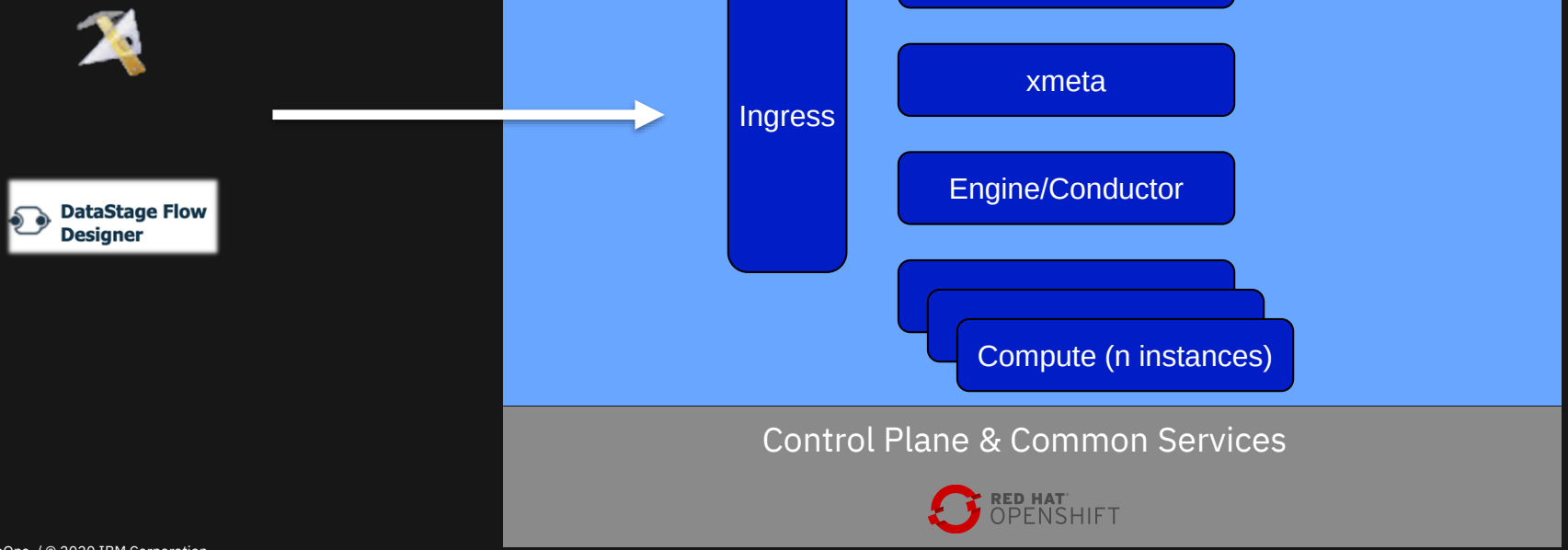

RED HAT® OPENSHIFT

**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

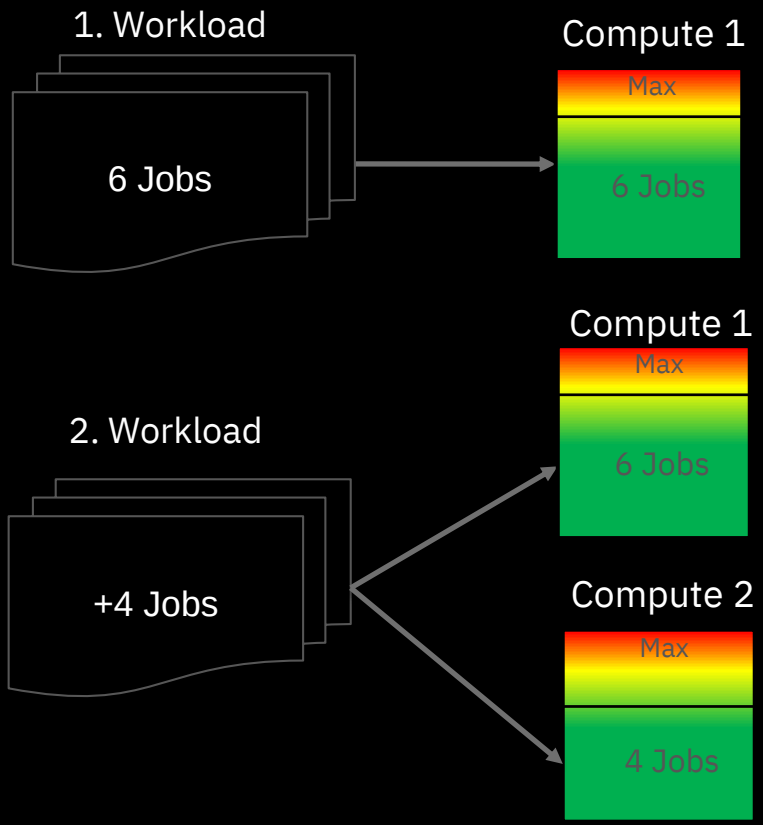
**TRUSTED HOST**
OpenShift runs on Red Hat Enterprise Linux, the most deployed commercial operating system in the public cloud, trusted by more than 90% of the Fortune 500

# DataStage for Cloud Pak for Data

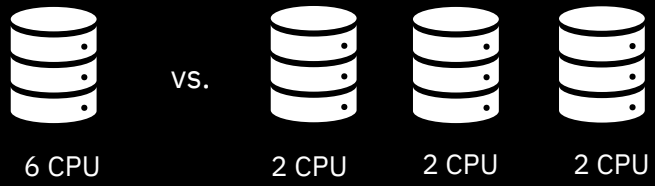


# Built-in automatic workload balancing and best of breed parallel engine

- 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



# **Performance** of DataStage for Cloud Pak for Data

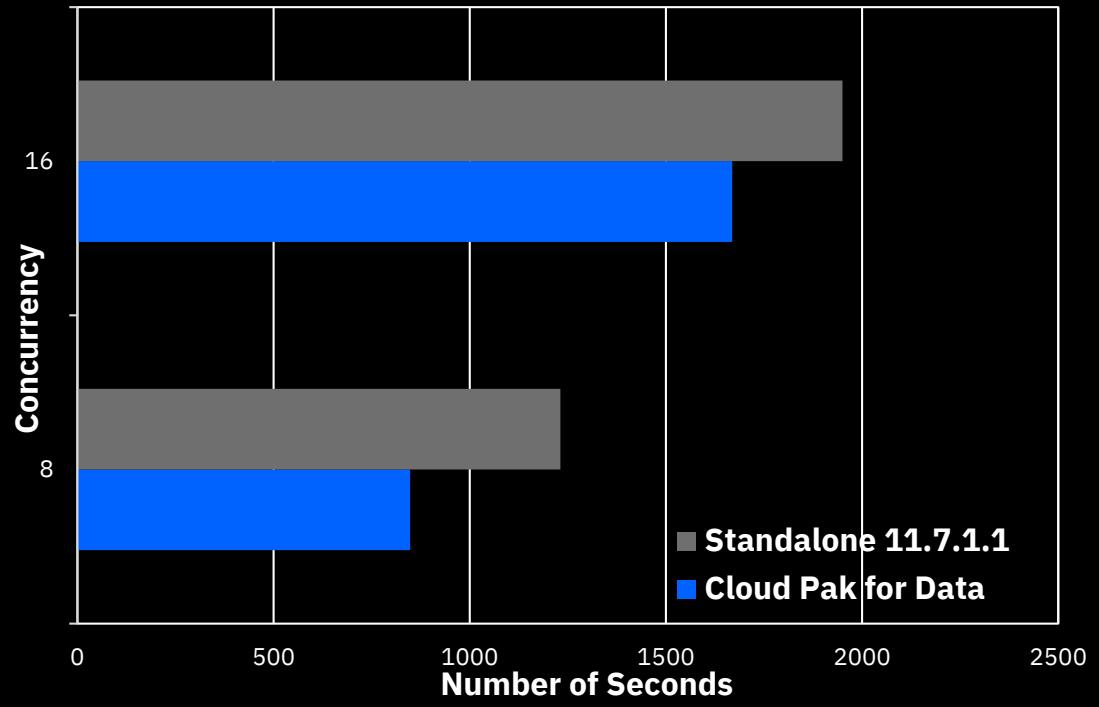



## Objective:

- Validate performance during execution windows of resource contention
- Demonstrate value of default execution of Massively Parallel Processing (MPP)

## Confirmed Result:

- Significant reduction in runtime on DataStage Cloud Pak for Data
- Delivers more evenly balanced and distributed workload



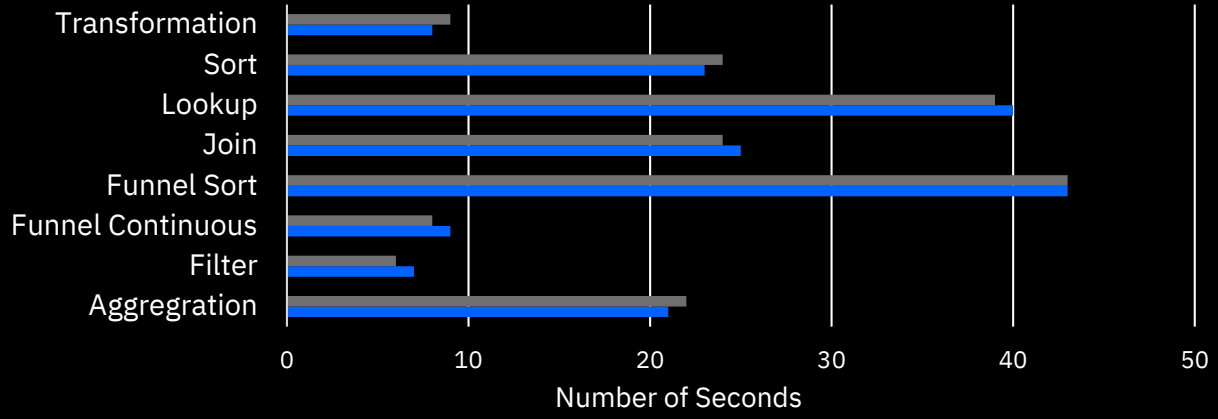
# **Comparing** Core Function Performance

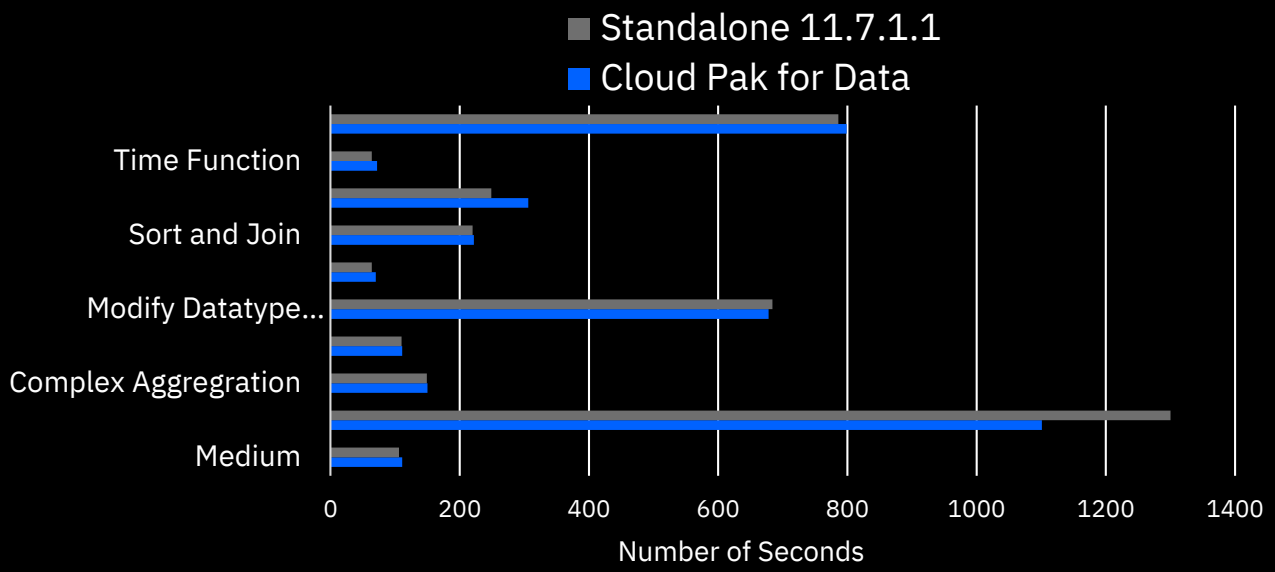
## Objective:

- Validate no difference in performance behavior of core operations/patterns
- Standalone binaries vs. Binaries deployed via Containers
- Each function was compared/isolated in a single job

## Confirmed Result:

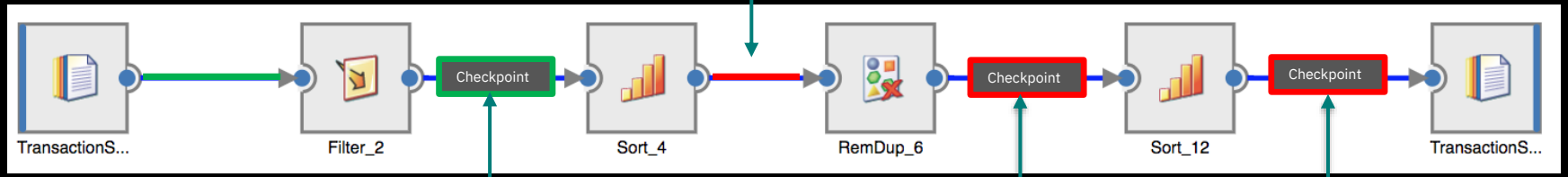
- No discernible difference in performance
- Validates expected behavior and provides proof-point via lab testing
- Provides confidence for running critical DataStage workload in containers



# DataStage: Checkpoint/Restart

- Failure occurs while the link in red is processing data

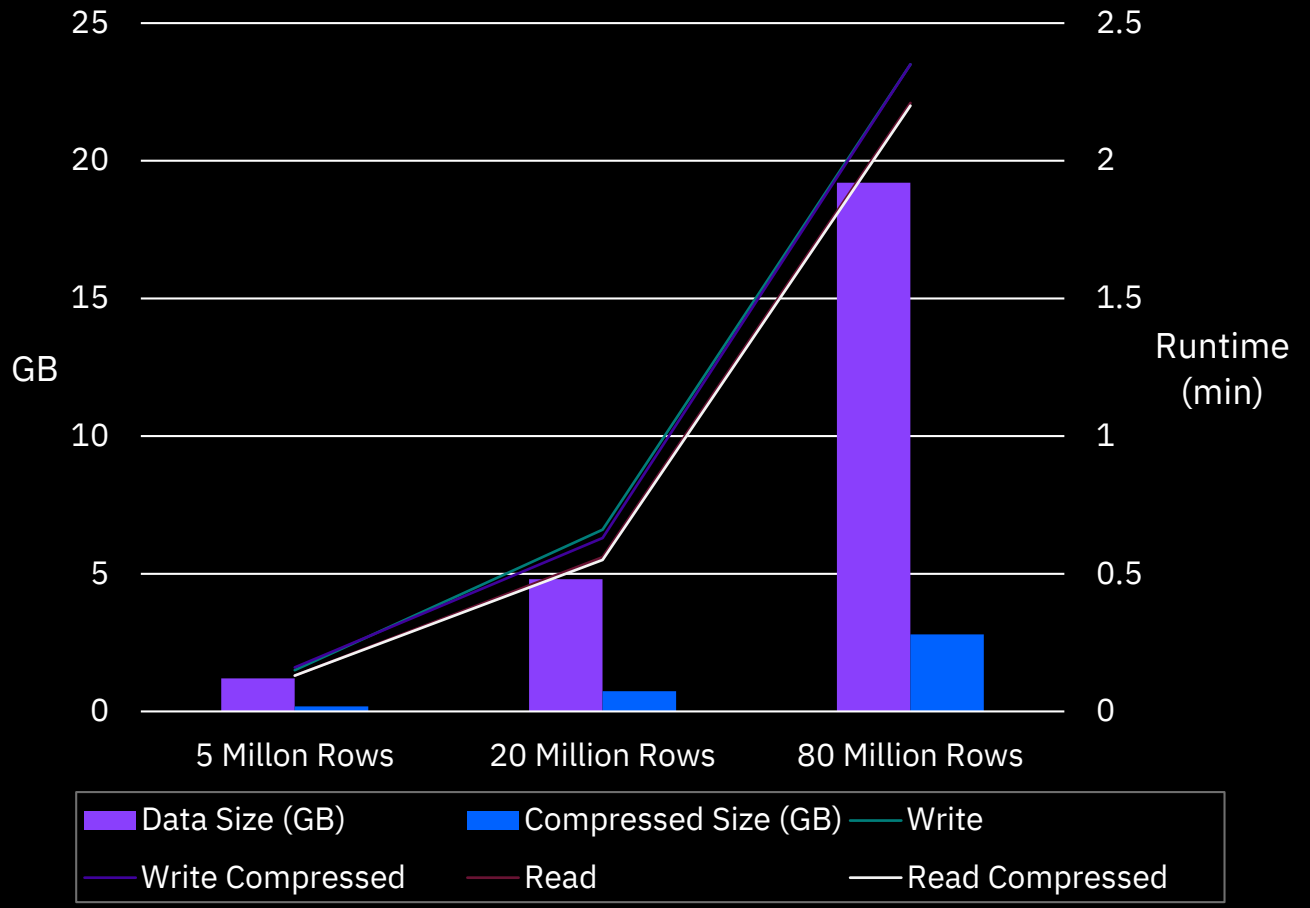



- First checkpoint is complete
- Second and third checkpoints are not complete
- The job automatically restarts using data from first checkpoint

# Compression Performance

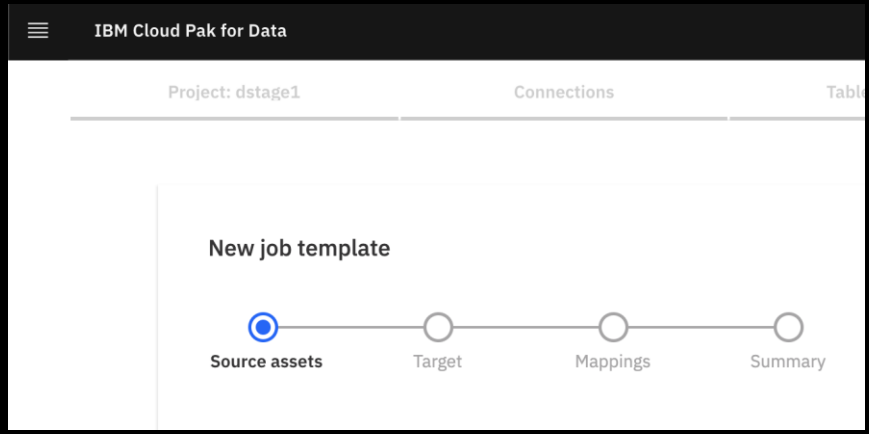
- Sorting
- Datasets
- Checkpoints

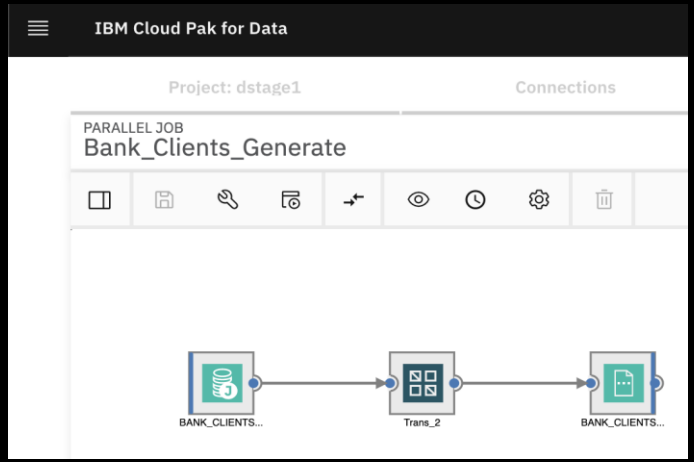


# Job Templates

*accelerating data integration for AI and analytics*

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




Auto-Generate

# DataStage : Broader, Faster, Safer Connectivity

IS 11.7.1+ / CPD 3.0

## Hadoop

- **HBase** connector
- **Hadoop File** Connector
- **Kafka** Connector enhanced
- **Hive** Connector enhanced (Write to Hive Partitioned Tables)
- MongoDB support
- **Cassandra** connector (incl. Data Lineage and metadata import)
- BDFS Kerberos improvements for non Hadoop environments
- Apache Sequential File support for File Connector
- HA support for HDFS/File Connector
- Presto
- Up to Cloudera up to 6.3.2
- Up to HDP up to 3.1

## File

- XML connector enhanced
- INT96 for Parquet file

## Cloud

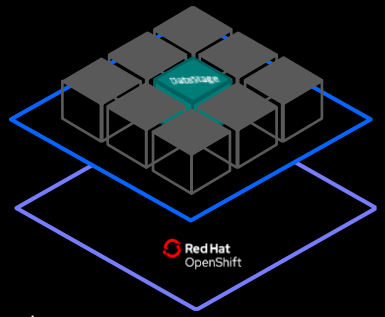
- Amazon EMR/Hive
- Amazon Redshift
- Amazon S3 KMS Support
- Amazon S3 Parquet and ORC support
- AWS Aurora PostgreSQL
- AWS Dynamo DB (limited)
- **Snowflake** connector enhanced
- **Azure Cloud Storage** connector
- Azure CosmosDB support via Cassandra connector
- **Azure Data Lake Storage** Connector (Gen 1 and Gen 2)
- **Salesforce** (PK Chunking) API 47 support
- **IBM Cloud Object Storage** connector
- **Google BigQuery** Connector
- **Google Cloud Storage** Connector
- SAP Odata support
- Oracle Autonomous Data Warehouse Cloud
- EoW Implementation for Azure, Cloud Object Store, S3, File Connector (Replication)

## Enterprise

- Oracle 19c (incl. CBD and PDB)
- Siebel 8.2.2.4 certification
- Sybase datatype enhancement & IQ 16.1 support
- New SAP BW & ERP Ppacks
- Data Masking ODPP v11.3 support and expanded Data masking policy support
- DTS Connector: MQ Client mode
- MQ Connector version update
- ILOG Connector Decision Engine
- Db2 v12 z/OS certification
- Greenplum v5.4 certification
- Teradata Connector V16.2 (Big Buffer Support, passthrough support)
- **SAP** ERP Pack V8.1 (Delta extract stage , contenerized delivery)
- Db2 connector support for External Tables
- RJUT usability improvements for easy PDA→IIAS migration
- Filter condition push-down
- FTP support for customizable /tmp and FTPS
- Teradata Connector enhanced
- Netezza Performance Server V11
- Security enhancement

# DataStage – Available anywhere you need it

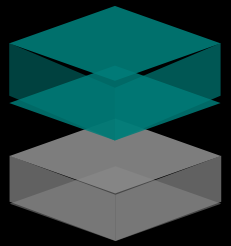
## DataStage / Information Server *on IBM Cloud Pak for Data*




aws *On-Premises*

- Fully containerized microservices
- Run on any cloud with Red Hat OpenShift to manage containers
- Subscription and perpetual license models
- *For existing customers:* multiple routes to upgrade existing entitlements

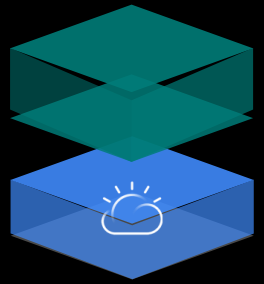
## DataStage / Information Server *stand-alone*



aws *On-Premises*

- Traditional deployment on bare metal or virtual environments
- Deploy on-premises, private cloud, or any public cloud (BYOL)
- Perpetual license based on PVU

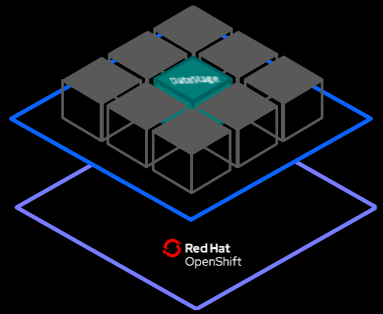
## DataStage / Information Server *on IBM Cloud*



- Information Server Enterprise Edition – traditional install provisioned and managed on IBM Cloud
- DataStage hosted on IBM cloud

# DataStage – Available anywhere you need it

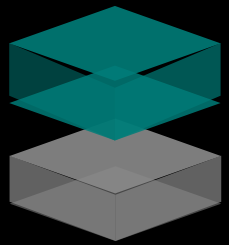
## DataStage / Information Server *on IBM Cloud Pak for Data*




aws *On-Premises*

- Fully containerized microservices
- Run on any cloud with Red Hat OpenShift to manage containers
- Subscription and perpetual license models
- *For existing customers:* multiple routes to upgrade existing entitlements

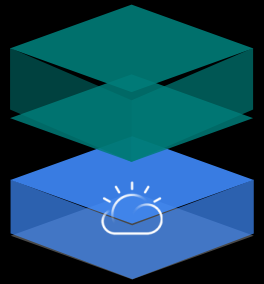
## DataStage / Information Server *stand-alone*



aws *On-Premises*

- Traditional deployment on bare metal or virtual environments
- Deploy on-premises, private cloud, or any public cloud (BYOL)
- Perpetual license based on PVU

## DataStage / Information Server *on IBM Cloud*



- Information Server Enterprise Edition – traditional install provisioned and managed on IBM Cloud
- DataStage hosted on IBM cloud

# Data Integration Vision and Roadmap

***Powered by DataStage***

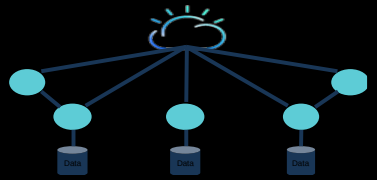
# Project Tahoe: Next gen DataStage

*modern architecture designed on native-cloud principles*

- **Agility**
- **Efficiency**
- **Cost Savings**

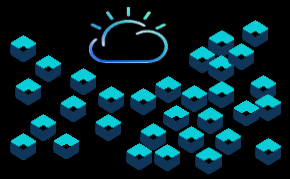
## Build for Agility & Scalability

### Loosely Coupled Services



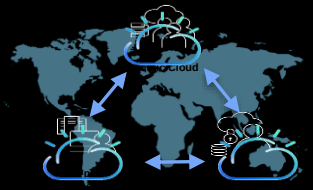
An architecture of loosely coupled data services, easily refactored to create containerized workloads

### Containerized Workloads



Stand-alone workloads composed of micro-services & data that are flexibly deployed, orchestrated and managed

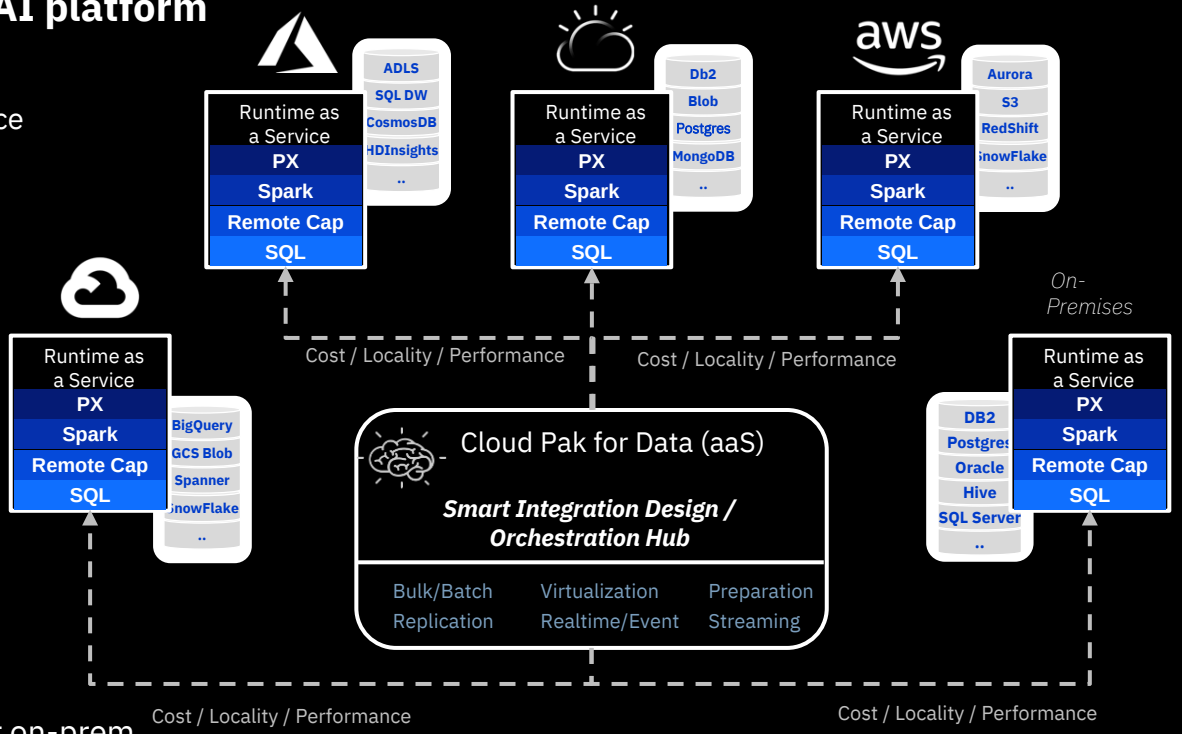
### Multi-Cloud Provisioning /Execution



Agile provisioning of containerized workloads in multi-Cloud environments and consumption of Cloud services

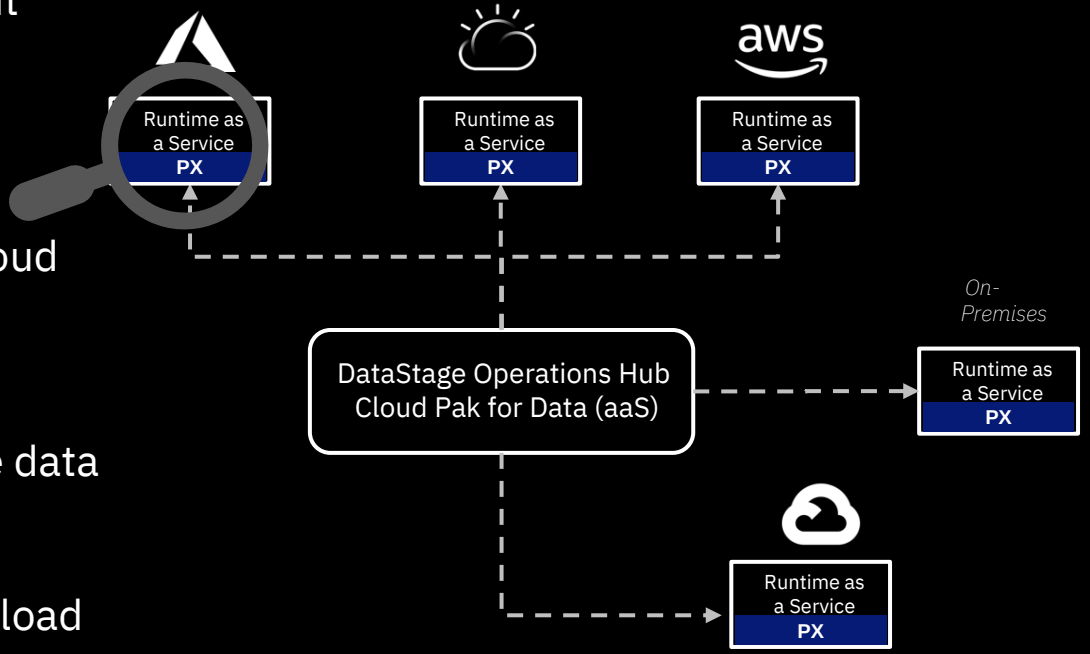
# Project Tahoe: Reinventing DataStage
*upon cloud native values*

- **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



# PX as a Service
## *Lightweight, elastic Data Gravity support*

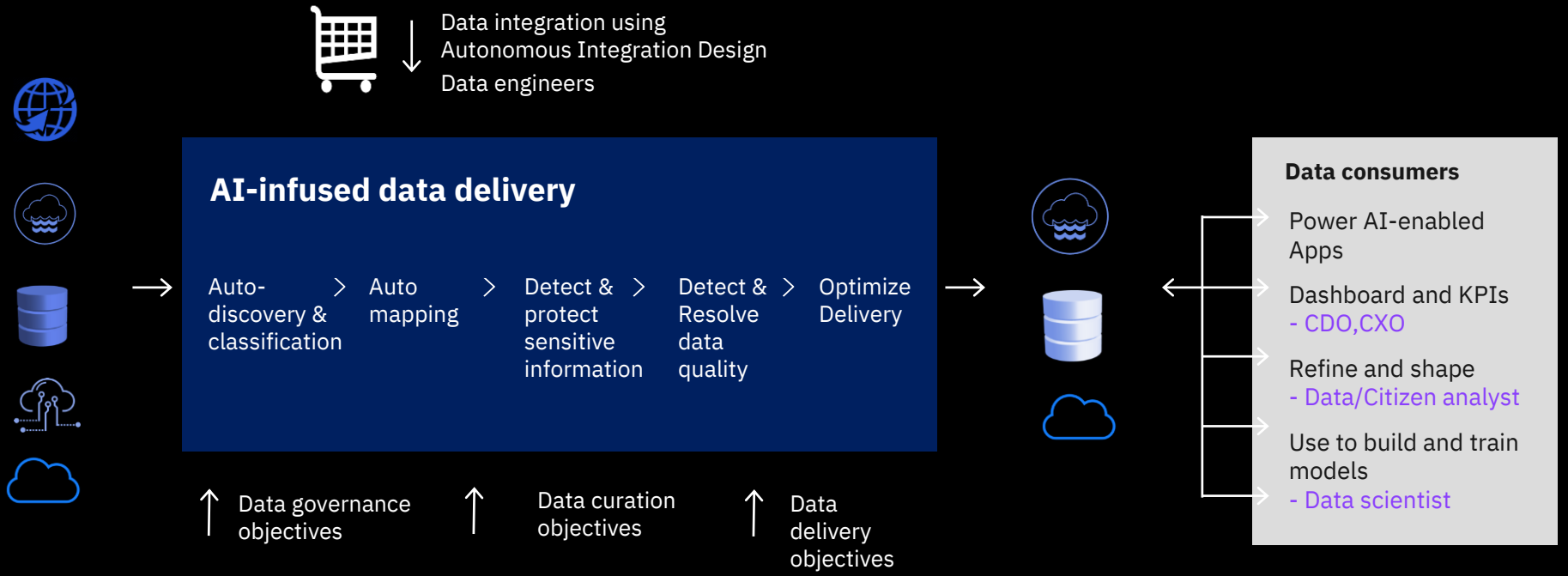
- Lightweight Engine Service to be used on any cloud or on premises environment
- Supports elastic scaling based on workload requirements
- *Operations Hub* supported on IBM Cloud (Cloud Pak as a Service) or on client's environment
- Pushing workload execution to where data resides (data gravity)
- Cloud-based licensing based on workload execution



# Deeply integrated with  Cloud Pak for Data

1. Design/Generate flows on Cloud Pak for Data's Common Canvas
    - Fully wired into Cloud Pak for Data
      → easy sharing or utilization of common assets
    - Built on a runtime neutral *canonical design model*
      → allows to translates into any possible runtime logic
    - Utilize and enhance on pre-existing flow design experience
      → One design canvas experience for the entire platform
2. Dynamically execute flows on supported *built-in* or  SaaS-based Runtime services
3. Built-in dynamic scaling and workload management
4. Utilizing common platform management and operations
5. The flow designs are using a publicly available JSON schema (that has been open sourced)
6. All the APIs used by the UI are a set of publicly documented micro-services

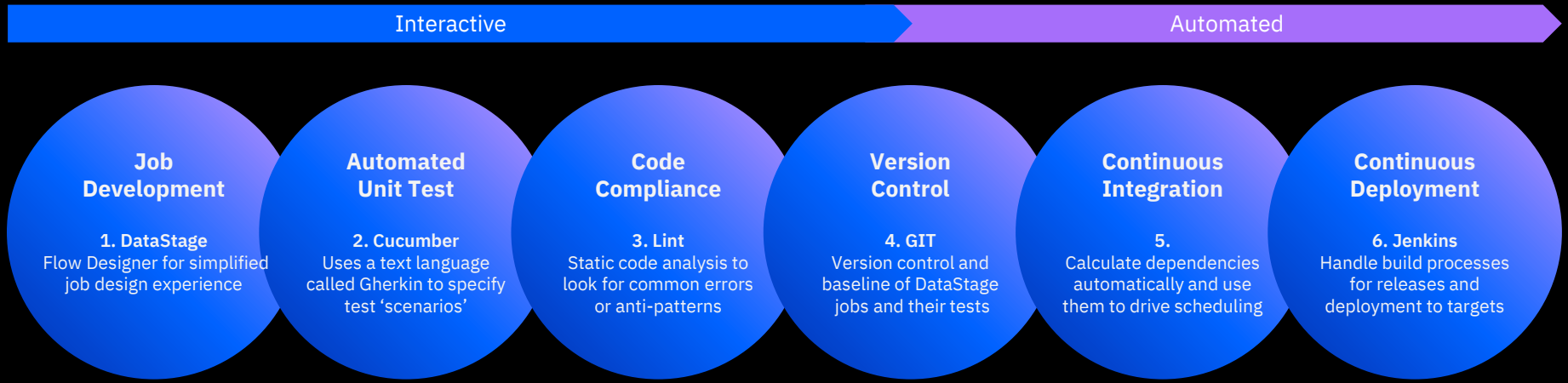
# AutoDI
## Autonomous Data Integration to accelerate time to value



# DevOps Support for Agility

*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.