IBM DataStage and Data Virtualization : Improving the collaboration between business and IT

Cliff Candiotti, DTE Technical Specialist

Albert Liang, WW Technical Sales - Data Integration and Preparation Lead

Hebert Pereyra, STSM, Big SQL & Data Virtualization Chief Architect

Bharath Chari, WW Product Marketing Manager, IBM DataStage

There is no AI without an IA

“Information Architecture”

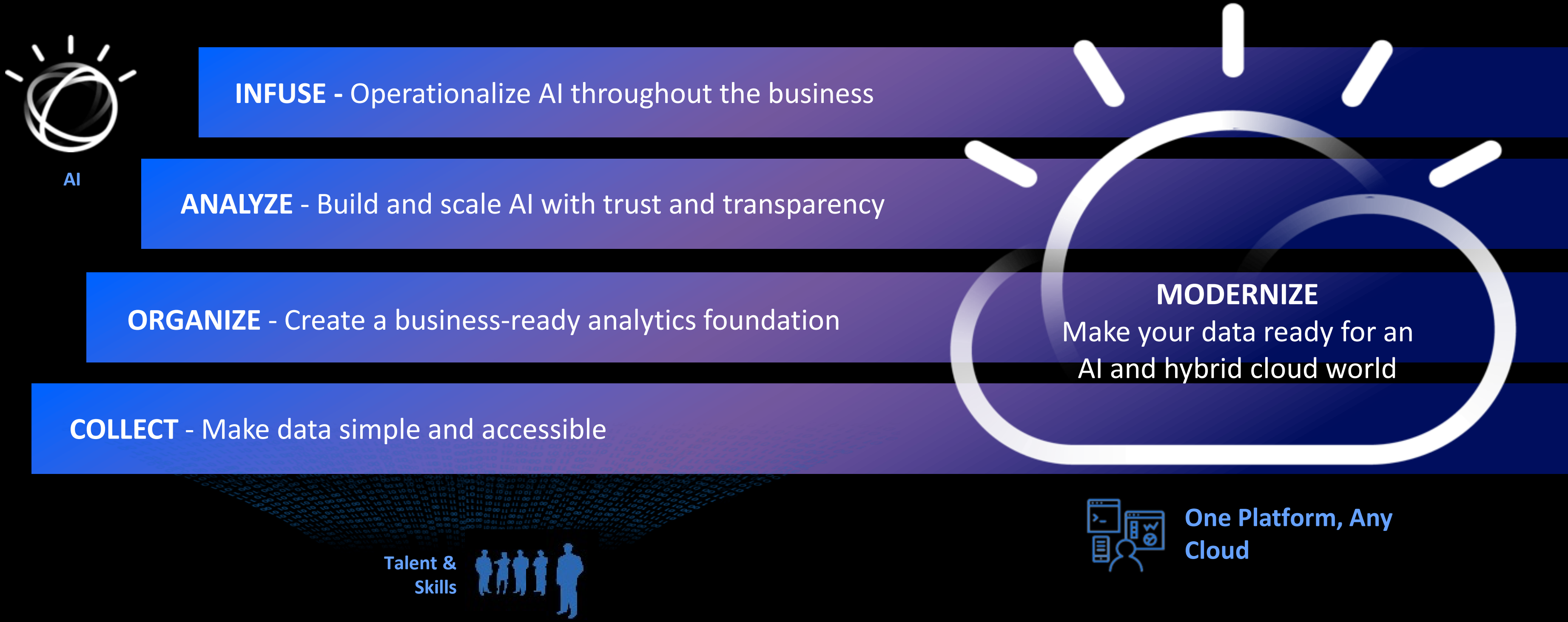
> *“ No amount of AI algorithmic sophistication will overcome a lack of data [architecture]*
>
> *Data collection & preparation is the most time consuming and difficult part of AI*

MITSloan

Sources: 2018 MITSlone ”Reshaping business with AI”

The AI Ladder

A prescriptive approach to the journey to AI



AI

INFUSE - Operationalize AI throughout the business

ANALYZE - Build and scale AI with trust and transparency

ORGANIZE - Create a business-ready analytics foundation

COLLECT - Make data simple and accessible

MODERNIZE
Make your data ready for an
AI and hybrid cloud world

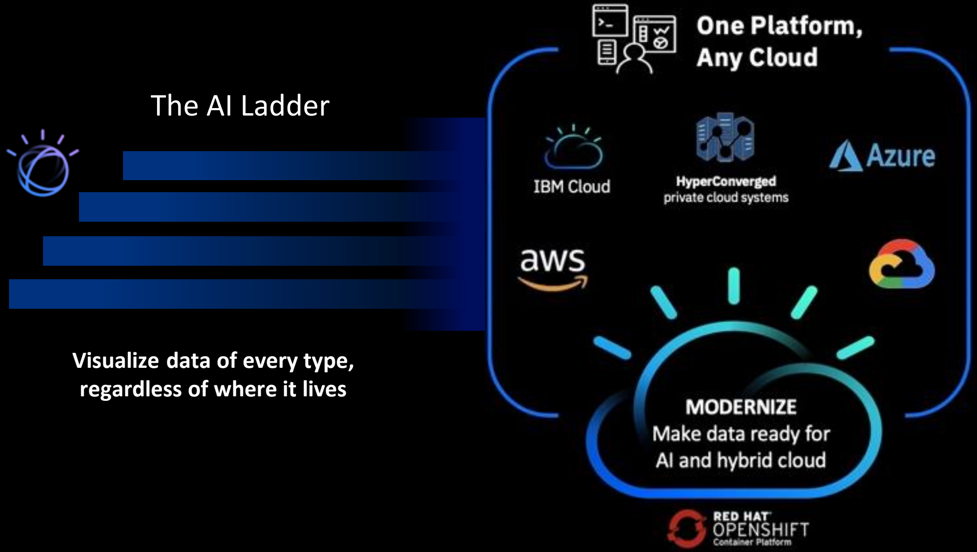
Talent & Skills

One Platform, Any Cloud

Cloud Pak for Data
Certified on Red Hat OpenShift

Delivers the foundational platform for deploying an information architecture for AI, on any cloud

- Eliminate data silos, connect all data
- Automate and govern the data & AI lifecycle
- Operationalize AI with trust & transparency
- Avoid lock-in, run anywhere with agility



The AI Ladder

Visualize data of every type, regardless of where it lives

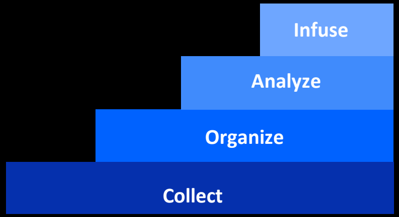
One Platform, Any Cloud

MODERNIZE
Make data ready for AI and hybrid cloud

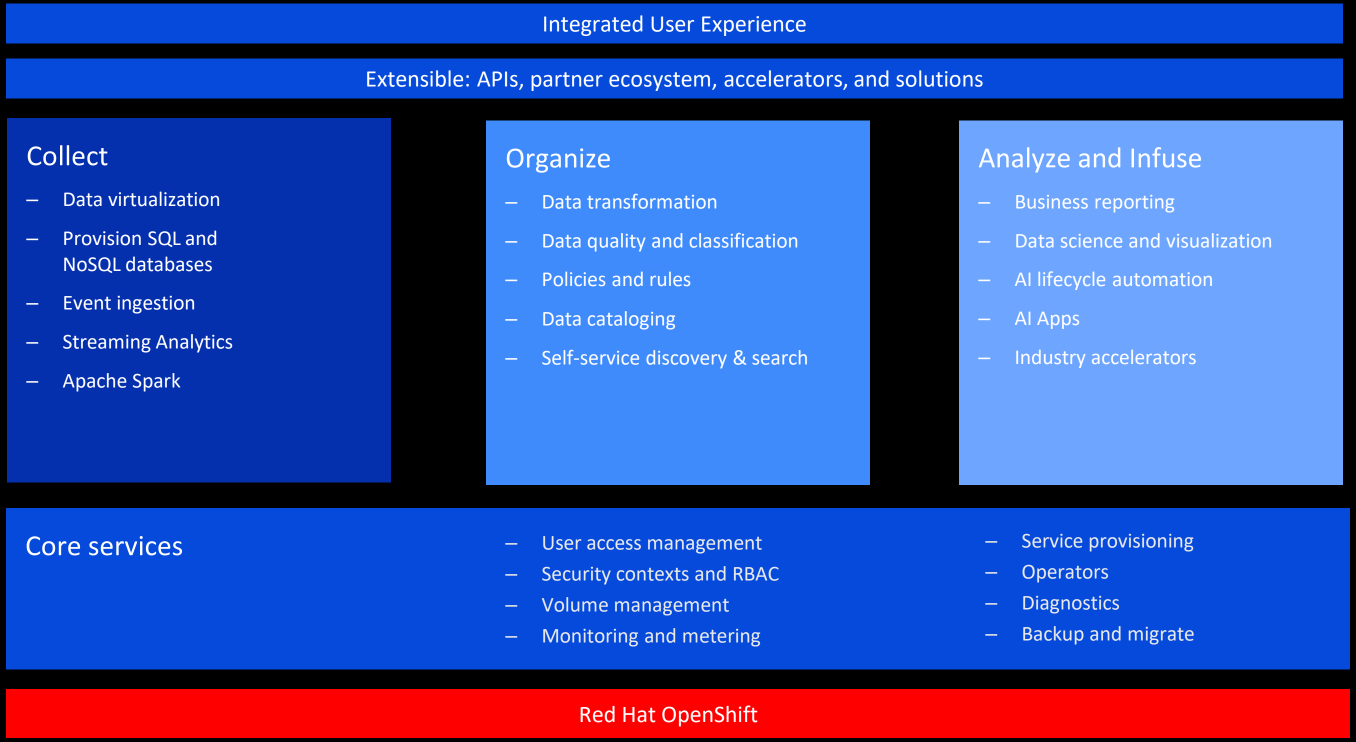
IBM Cloud Pak for Data

Unified, modular, deployable anywhere

The AI Ladder



App Developers and SREs| Business Partners | Data Engineers | Data Stewards | Data Scientists | Business Users



Integrated User Experience

Extensible: APIs, partner ecosystem, accelerators, and solutions

Collect

- Data virtualization
- Provision SQL and NoSQL databases
- Event ingestion
- Streaming Analytics
- Apache Spark

Organize

- Data transformation
- Data quality and classification
- Policies and rules
- Data cataloging
- Self-service discovery & search

Analyze and Infuse

- Business reporting
- Data science and visualization
- AI lifecycle automation
- AI Apps
- Industry accelerators

Core services

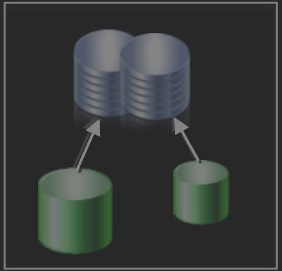
- User access management
- Security contexts and RBAC
- Volume management
- Monitoring and metering
- Service provisioning
- Operators
- Diagnostics
- Backup and migrate

Red Hat OpenShift

IBM Cloud | Amazon Web Services | Microsoft Azure | Google Cloud | Hyperconverged system

A Variety of Ways to Deliver Data to Applications and Systems

Use the optimum delivery approach for varied business requirements

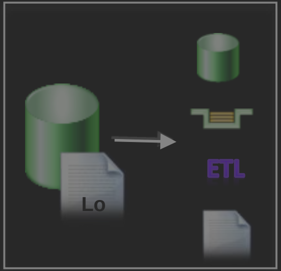


ETL - High speed bulk data delivery

IBM DataStage

Virtual Database - Multiple sources in one database

IBM Data Virtualization



Data Replication – Real time Change Data Capture

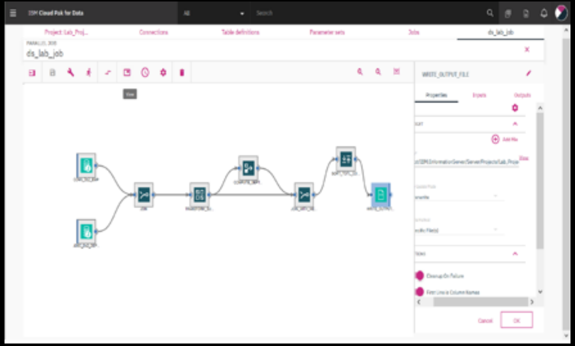
IBM
Data Replication

Self-service – light data integration for business users

IBM Watson Data Refinery

Extract Transform & Load (ETL)

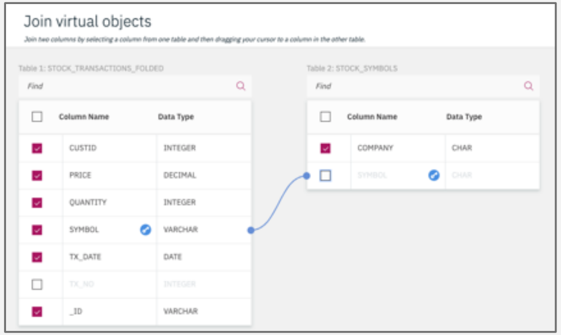
- Moves and transforms **large** volumes of data
  - Complex transformations easily done
  - Extremely fast, high performing
- Data gets tailored for new uses
  - Governed data can be trusted
  - Data quality can be applied
- ETL Jobs are developed in design tool
  - Executed in batch or real-time



DataStage
Flow Designer

Data Virtualization (DV)

- Leaves data where it is and simplifies access
  - Multiple sources appears as a single database
  - Combines heterogenous data efficiently
- Data obtained is **“as-is”**
  - Sources may not be governed
  - Data governed once published to central Catalog (Watson Knowledge Catalog)
- Uses standard SQL - no additional skills needed
  - Data accessed on demand in real time



Join virtual objects

Join two columns by selecting a column from one table and then dragging your cursor to a column in the other table.

Table 1: STOCK_TRANSACTIONS_FOLDED

Find

☐	Column Name	Data Type
☑	CUSTID	INTEGER
☑	PRICE	DECIMAL
☑	QUANTITY	INTEGER
☑	SYMBOL	VARCHAR
☑	TX_DATE	DATE
☐	TX_NO	INTEGER
☑	_ID	VARCHAR

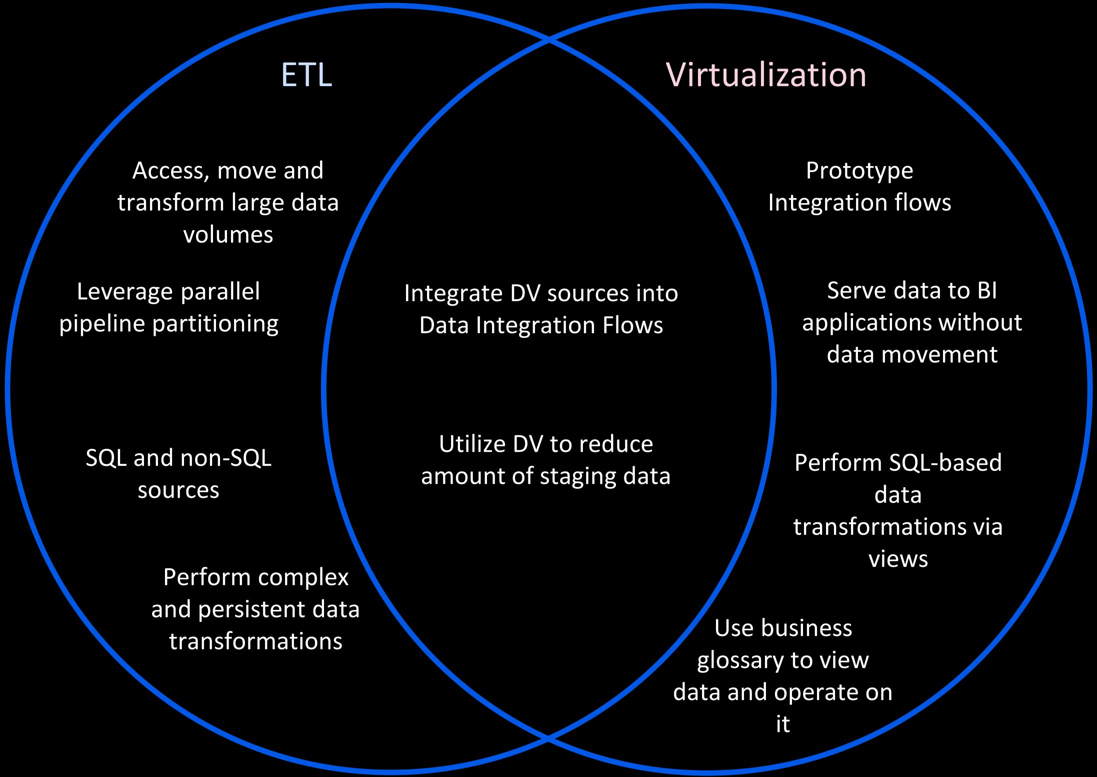
Table 2: STOCK_SYMBOLS

Find

☐	Column Name	Data Type
☑	COMPANY	CHAR
☐	SYMBOL	CHAR

Joining data
inside Data
Virtualization

DataStage | Data Virtualization
Better, together



ETL

Access, move and transform large data volumes

Leverage parallel pipeline partitioning

SQL and non-SQL sources

Perform complex and persistent data transformations

Integrate DV sources into Data Integration Flows

Utilize DV to reduce amount of staging data

Virtualization

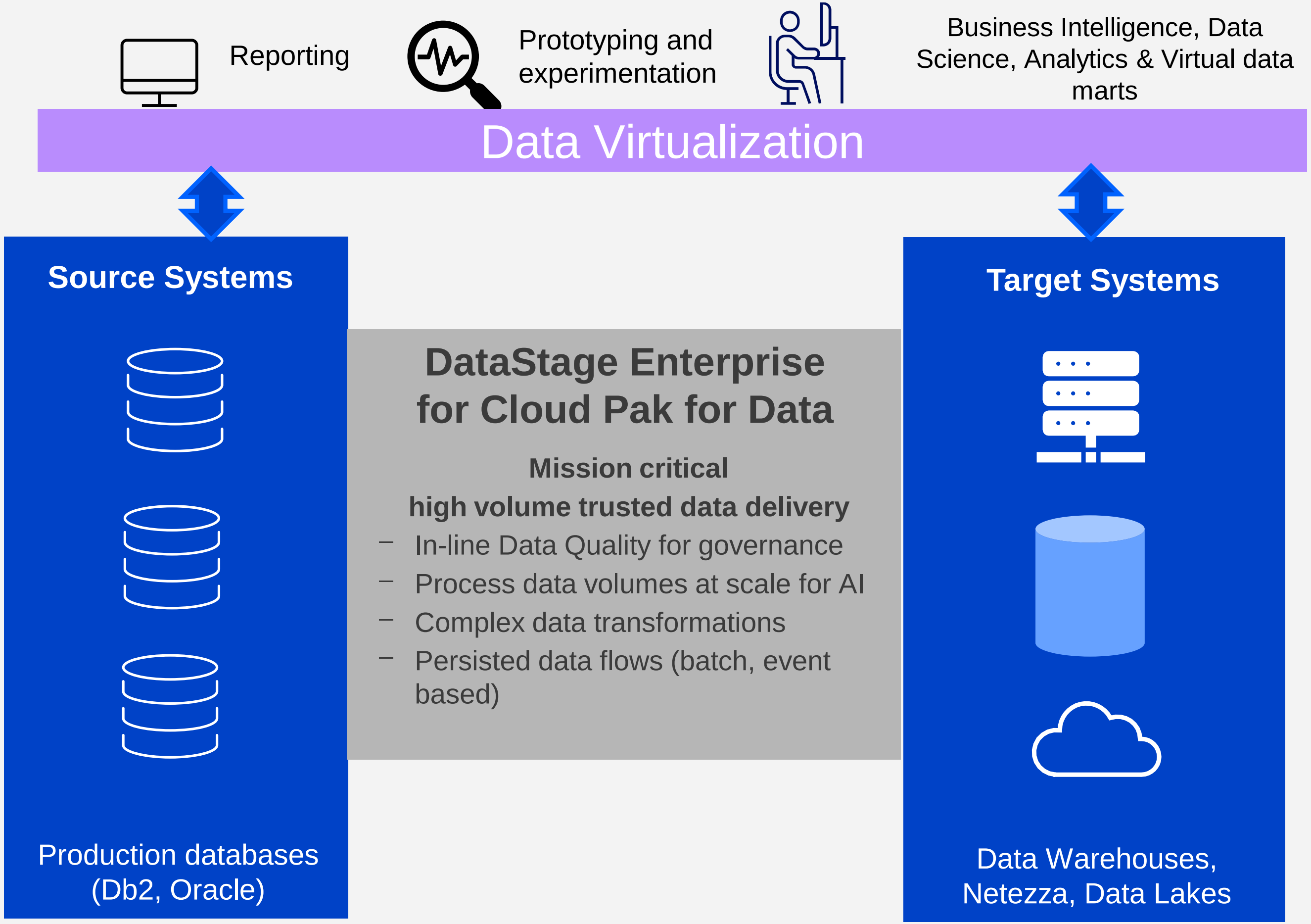
Prototype Integration flows

Serve data to BI applications without data movement

Perform SQL-based data transformations via views

Use business glossary to view data and operate on it

Agile & Accelerated Analytics with DataStage and Data Virtualization



Reporting

Prototyping and experimentation

Business Intelligence, Data Science, Analytics & Virtual data marts

Data Virtualization

Source Systems

Production databases (Db2, Oracle)

DataStage Enterprise for Cloud Pak for Data

Mission critical
high volume trusted data delivery

- In-line Data Quality for governance
- Process data volumes at scale for AI
- Complex data transformations
- Persisted data flows (batch, event based)

Target Systems

Data Warehouses, Netezza, Data Lakes

ETL Use Case Explained

Acme Company

New Customer

Self-Trade System

Business Scenario:

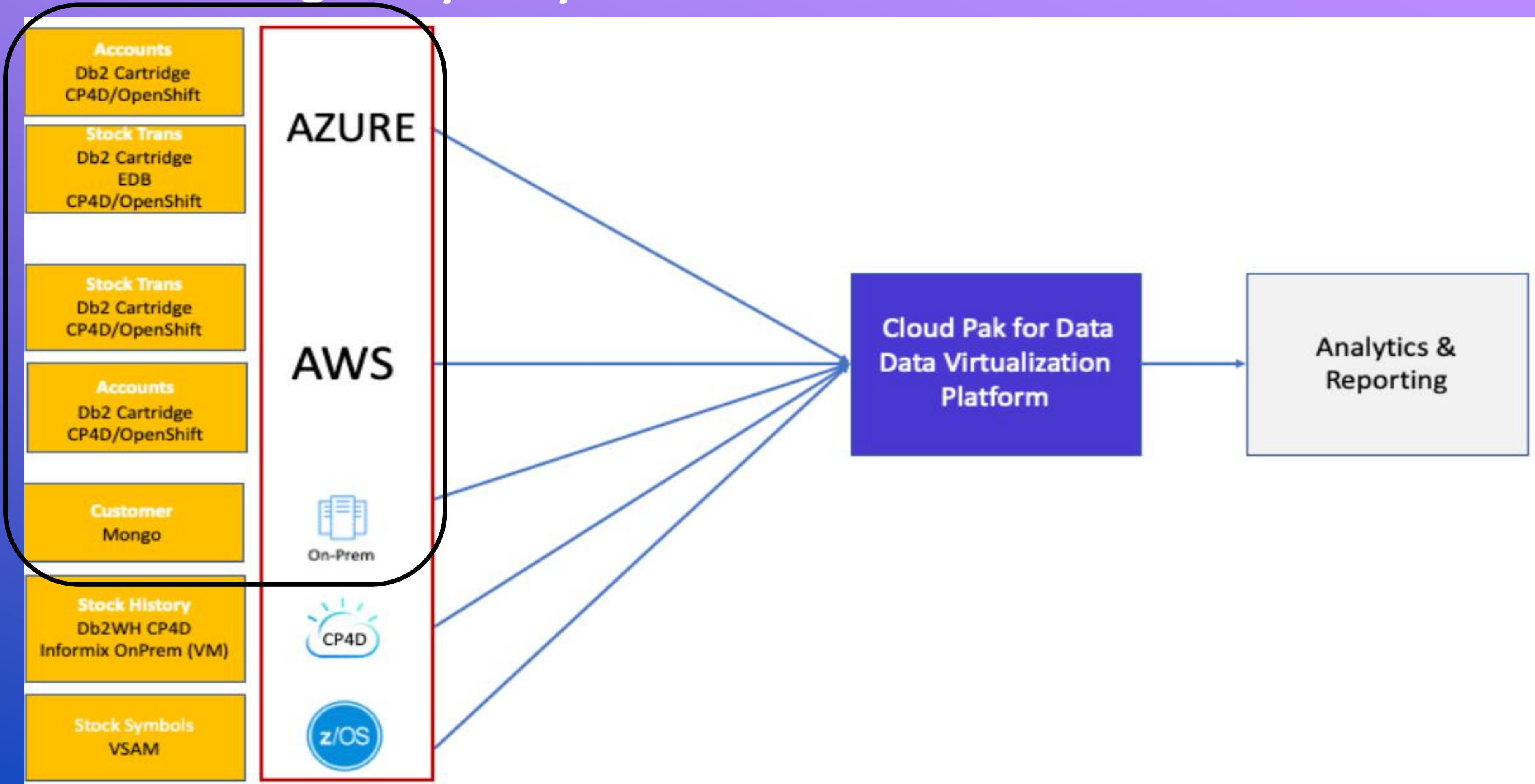
- Stock brokerage firm with large portfolio customers
    - Currently no Self-Trading options for clients - future planning

- Current "work-from-home" mandate preventing clients meeting with FA's
    - Revenue and commissions are down

- Jump start completion and roll-out of new Customer Self-Trade System
    - Cloud based application developed in Python and Java
        - Uses cloud-native databases and storage
    - Does not update enterprise reporting data

- ETL process built to update the reporting systems with Self-Trade transactions
    - Simplified converting non-SQL object types to relational format
    - Required high performance for processing to finish between run intervals

Stock Trading Analysis System Architecture



Accounts
Db2 Cartridge
CP4D/OpenShift

Stock Trans
Db2 Cartridge
EDB
CP4D/OpenShift

Stock Trans
Db2 Cartridge
CP4D/OpenShift

Accounts
Db2 Cartridge
CP4D/OpenShift

Customer
Mongo

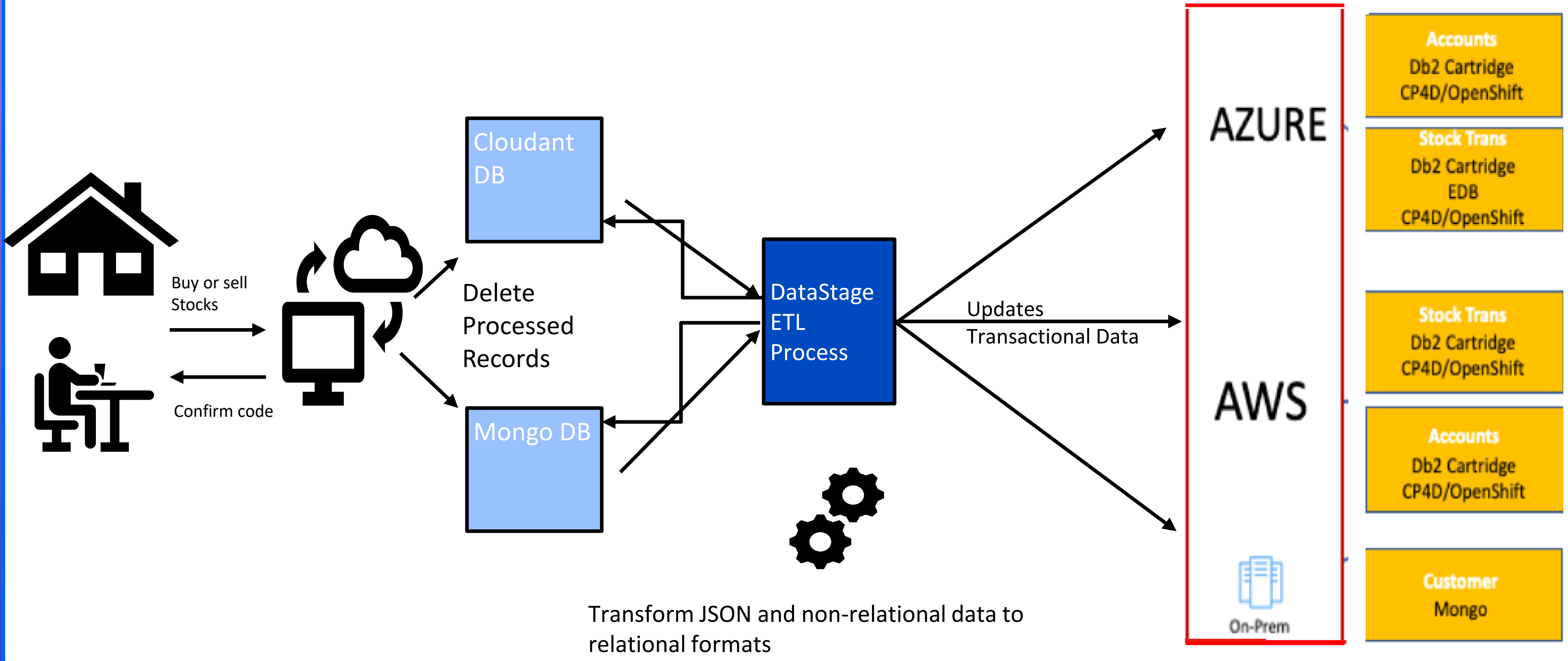
Stock History
Db2WH CP4D
Informix OnPrem (VM)

Stock Symbols
VSAM

Cloud Pak for Data
Data Virtualization
Platform

Analytics &
Reporting

Self-Trading System Architecture



Transform JSON and non-relational data to relational formats

DataStage

Industry Leading Data Integration Tool. Simple to design - Powerful to deploy

Rich capabilities spanning six critical dimensions

1 **Developer Productivity**

Rich user interface features that simplify the design process and metadata management requirements

2 **Transformation Components**

Extensive set of pre-built objects that act on data to satisfy both simple & complex data integration tasks

3 **Connectivity**

Native access to common industry databases and applications exploiting key features of each

4 **Runtime Scalability & Flexibility**

Performant engine providing unlimited scalability through all objects tasks in both batch and real-time

5 **Operational Management**

Simple management of the operational environment lending analytics for understanding and investigation

6 **Enterprise Class Administration**

Intuitive and robust features for installation, maintenance, and configuration

Connectivity

Native access to common industry databases and applications exploiting key features of each.

- Relational
  - DB2 LUW
  - DB2 z/OS
  - Db iSeries
  - Db2 Warehouse
  - Netezza / PDA
  - IBM Integrated Analytics Sy
  - Oracle DB
  - Oracle EXA data
  - Oracle PDB
  - SQL Server
  - Informix
  - Teradata
  - Greenplum
  - Postgres
  - MySQL
  - Sybase ASE
  - Sybase IQ
  - Universe
  - SAP Hana
  - EnterpriseDB
  - Stored Procedures
- Hadoop
  - Hive
  - HBase
  - Cassandra
  - Big SQL
  - Impala
  - Presto
  - HDFS
  - MongoDB
  - Spark
  - HAWQ
- Real time / Files
  - IBM MQ
  - Kafka
  - XML
  - JSON
  - FTP / SFTP
  - Simple Files
  - Complex-structure Files
  - File Sets
  - Mainframe Files
- Cloud
  - AWS S3
  - AWS Redshift
  - AWS RDS
  - AWS Aurora
  - IBM Cloud Object Storage
  - IBM Db2 Warehouse on Cloud
  - Azure File/Blob Storage
  - Azure Data Lake Storage
  - Azure SQL Server
  - Azure SQL Data Warehouse
  - Google BigQuery
  - Google Cloud Object Storage
  - Snowflake
  - Salesforce.com
- Generic / External
  - ODBC
  - JDBC
  - REST API
  - Webservices
  - external command
  - external program
  - Java Application
  - C/C++ Plugins
- Applications
  - IBM MDM
  - IBM ILOG
  - IBM Streams
  - IBM Cognos TM1
  - SAP ERP / R3 / CRM
  - SAP BW
  - Oracle App
  - Peoplesoft
  - Siebel
  - JD Edwards
  - Hyperion
  - SAS
  - iWay
  - Excel

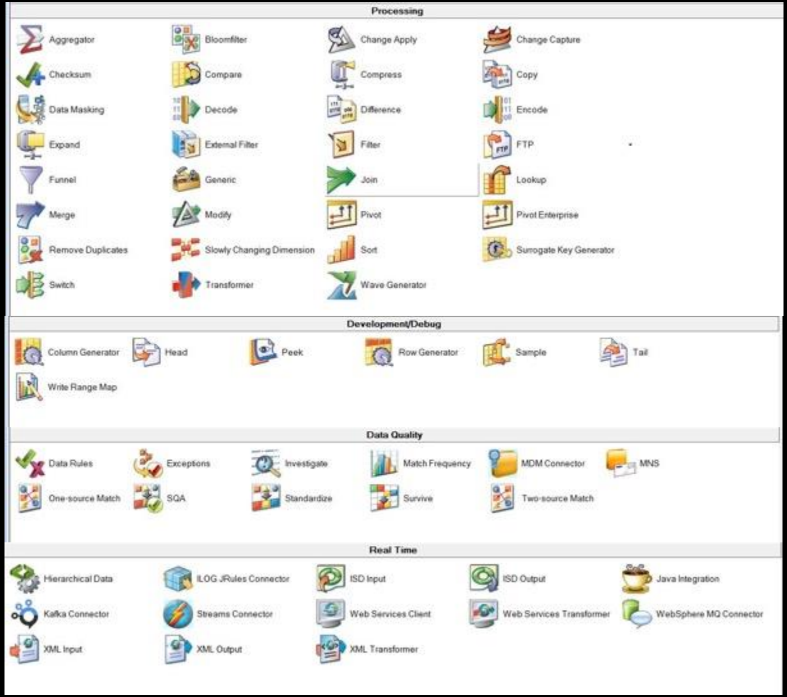
Transformation

Extensive set of pre-built objects that act on data to satisfy both simple & complex data integration tasks

Transformation Features for “any” Data

- Integration of Quality & Transformation Components
- Extensive Library of ready to use operations for:
    - Simple & Complex integration task
    - Hierarchical and relational transformations
    - Warehouse-specific features
    - Data Cleansing features
    - Development & Testing

- Extensibility to include your custom operations

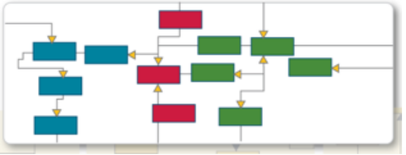


Integrated Data Quality

Single user experience for data integration, designing & running data validation, standardization & matching rules

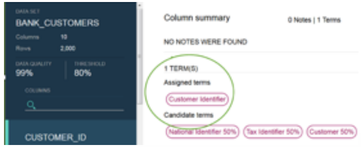
Discover

Discovery of Business entities across heterogeneous sources



Assess

Automatic Data Classification, Data Quality Analysis and linkage to business rules / terms



DATA SET
BANK_CUSTOMERS
Columns 10
Rows 2,000
DATA QUALITY 99%
THRESHOLD 80%
COLUMNS
CUSTOMER_ID

Column summary
0 Notes | 1 Terms
NO NOTES WERE FOUND
1 TERM(S)
Assigned terms
Customer Identifier
Candidate terms
National Identifier 50%
Tax Identifier 50%
Customer 50%

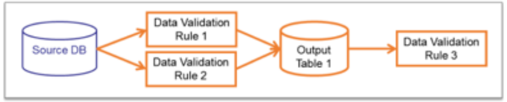
Cleanse

Business-driven Data Standardization & Matching

Drag the values from the example record to the output columns. Right-click a value to choose actions for that value.

Validate

Rule-based data validation to ensure complete & consistent data



Source DB
Data Validation Rule 1
Data Validation Rule 2
Output Table 1
Data Validation Rule 3

Monitor & Remediate

Enterprise-wide DQ Exception Monitoring and collaborative remediation

Life Cycle Governance

Ownership and management of Policies & Rules

Policy Details
Customer Data Integrity

New DataStage Flow Designer (DFD)

A New Integration Experience

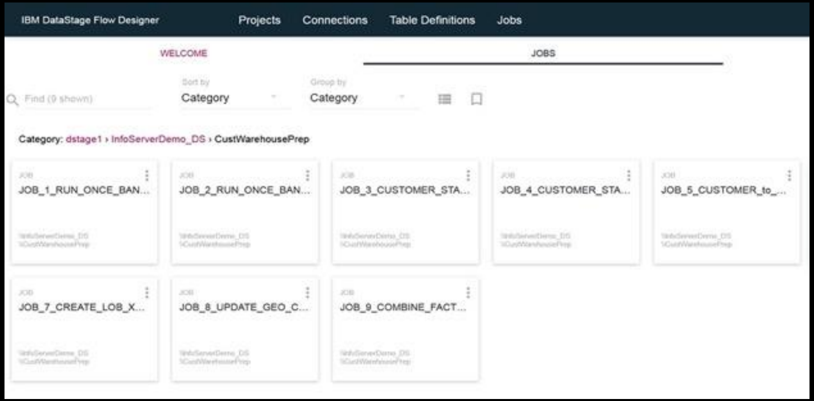
Intuitive, browser-based (no-install) experience

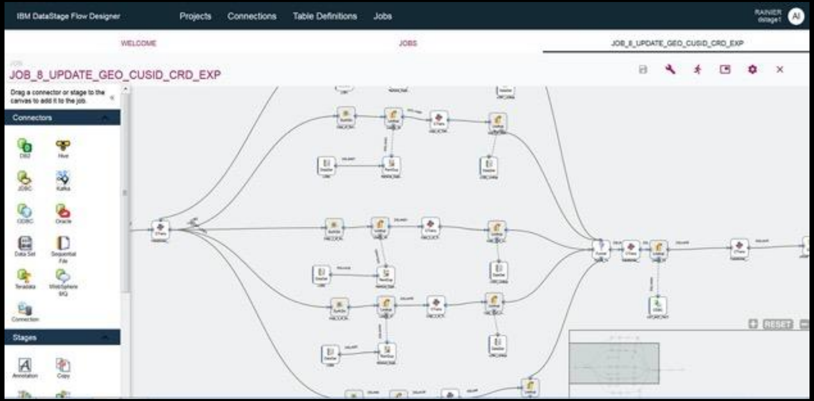
- Reducing total cost of ownership

Full backwards compatibility

Accelerated productivity through:

- Automatic schema propagation
- Git Source code control Integration
- Highlighted errors
- Powerful type-ahead search
- Server-side compilation





Cloud Pak for Data DataStage

Multi-cloud scalability and elasticity

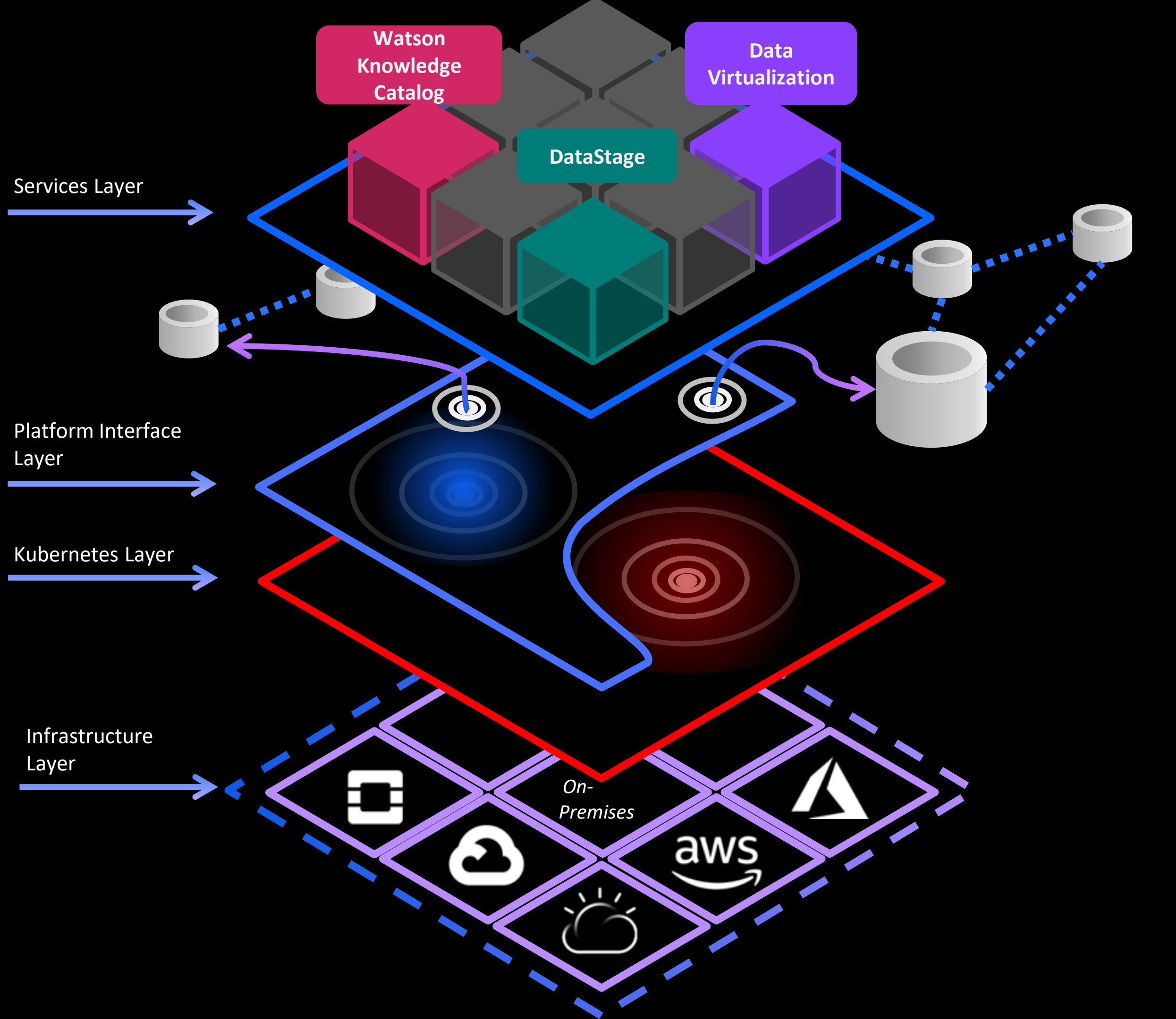
- Design once, dynamically run anywhere with built-in automatic workload balancing, parallelism and dynamic scalability

DataOps and DevOps enabled

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

Accelerate AI initiatives

- Automating Data Integration for faster ROI



Watson Knowledge Catalog

Data Virtualization

DataStage

Services Layer

Platform Interface Layer

Kubernetes Layer

Infrastructure Layer

On-Premises

Built-in automatic workload balancing

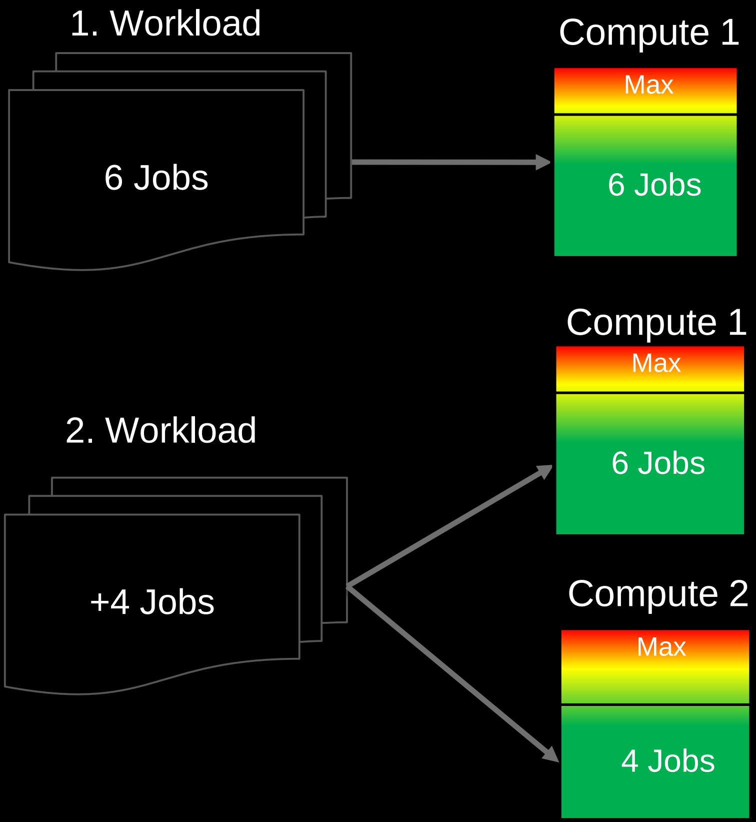
Best of breed parallel engine

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



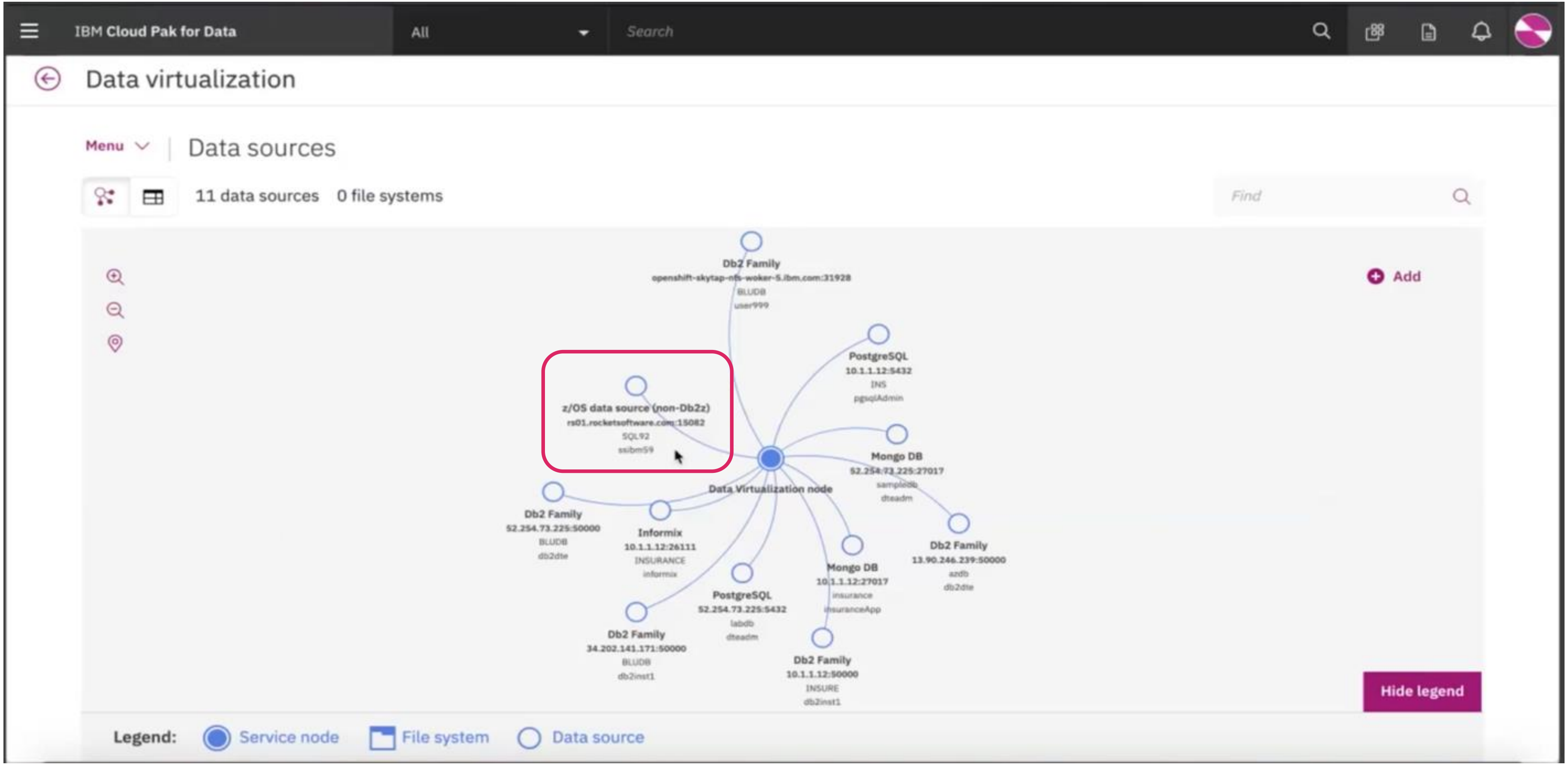
What is Data Virtualization?

Definition (Gartner):

“Data virtualization technology is based on the execution of distributed data management processing (primarily for queries) against multiple data sources, federation of query results into virtual views, and consumption of these views by applications, query/reporting tools or other infrastructure components. It can be used to create virtualized and integrated views of data in-memory (rather than executing data movement and physically storing integrated views in a target data structure), and provides a layer of abstraction above the physical implementation of data.”

Cloud Pak for Data

View virtualization constellation on the Platform



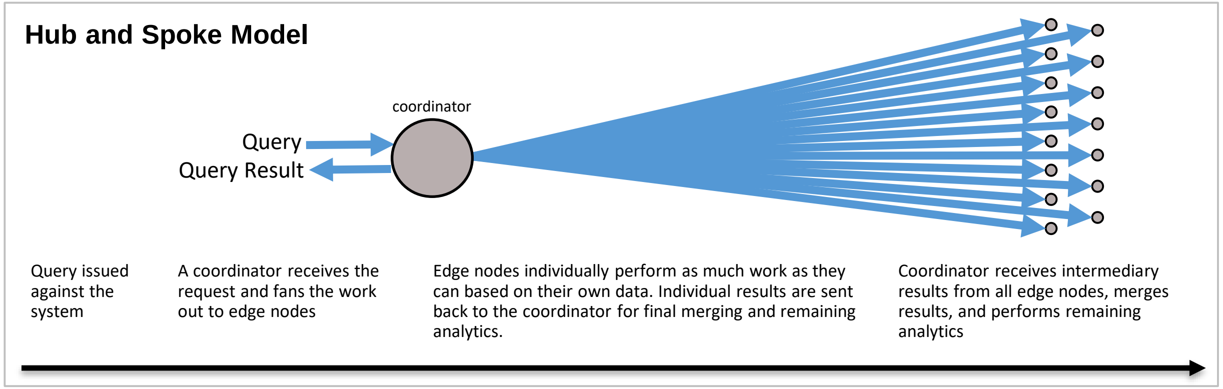
Key Architectural Differentiation

Hub and spoke execution models:

- Lacks scalability
- Performance constrained
- Basis for Federation and our competitors

IBM is first to market with a parallel processing model:

- Theoretically unlimited scalability
- Ease of addition/removal of sources
- Execution pushed down into the constellation mesh



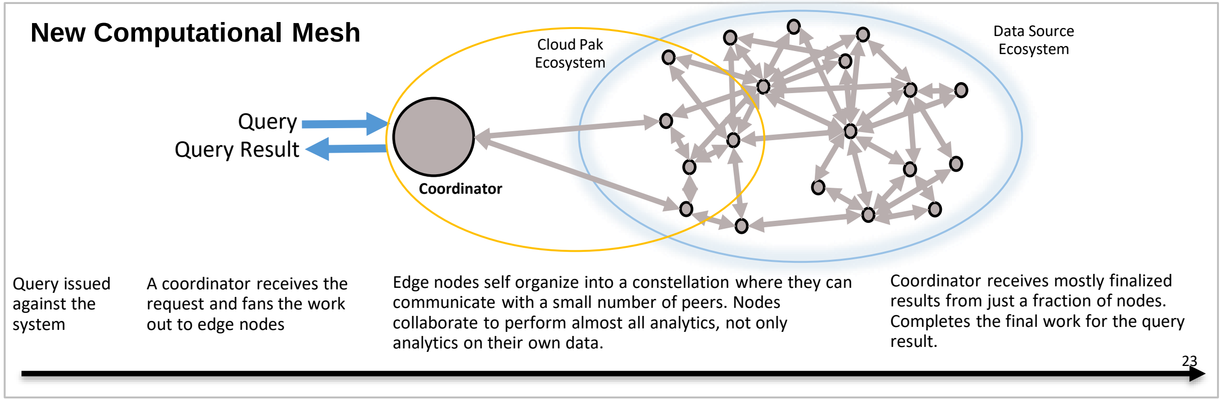
Hub and Spoke Model

Query issued against the system

A coordinator receives the request and fans the work out to edge nodes

Edge nodes individually perform as much work as they can based on their own data. Individual results are sent back to the coordinator for final merging and remaining analytics.

Coordinator receives intermediary results from all edge nodes, merges results, and performs remaining analytics



New Computational Mesh

Query issued against the system

A coordinator receives the request and fans the work out to edge nodes

Edge nodes self organize into a constellation where they can communicate with a small number of peers. Nodes collaborate to perform almost all analytics, not only analytics on their own data.

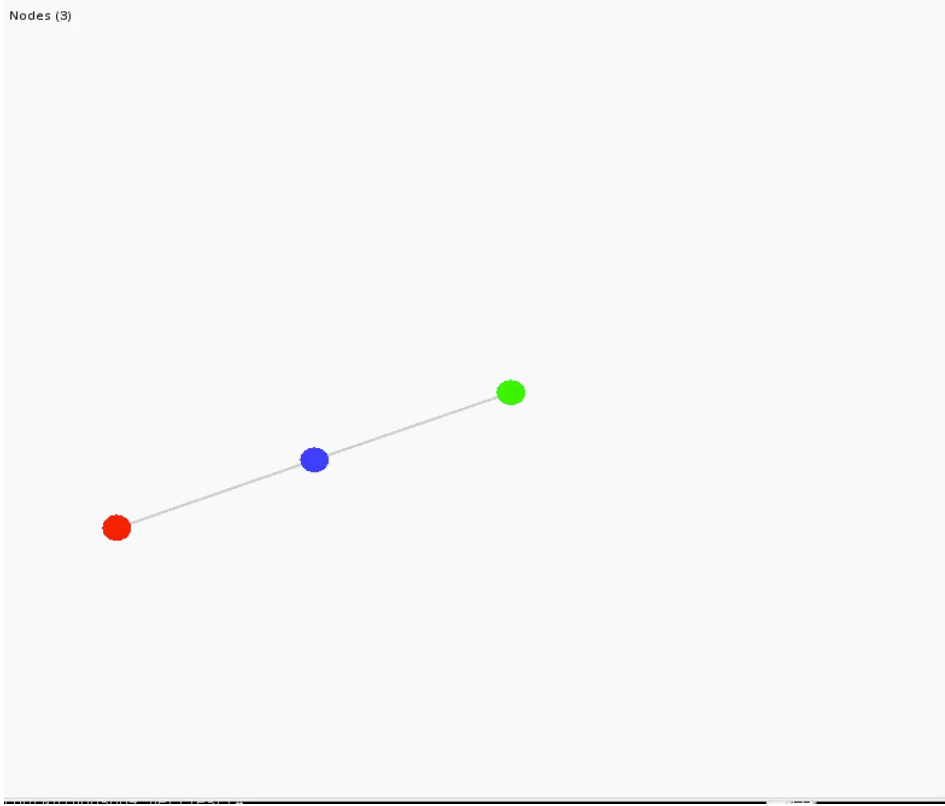
Coordinator receives mostly finalized results from just a fraction of nodes. Completes the final work for the query result.

Real System test

Growing a Constellation

- Video of constellation growing to 349 Nodes.
    - Network stays compact.
    - 2 and 10 links between nodes
    - No manual configuration.

- Latency aware connection between nodes
    - Which nodes connect to which others?
    - Fastest reply strategy
- Diameter of the constellation (i.e. the number of hops between the two furthest nodes) grows logarithmically. Small diameter is ideal for communications.

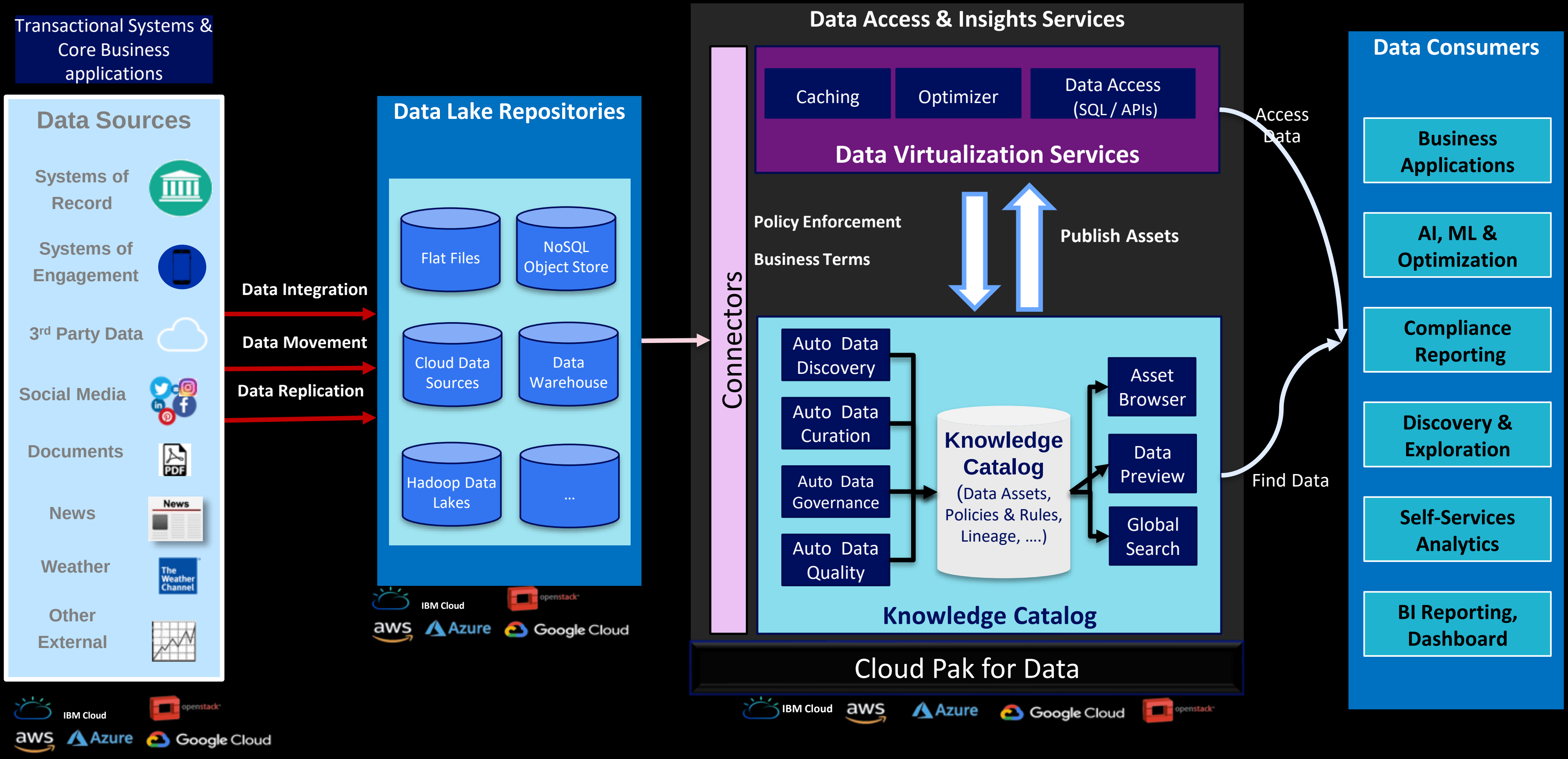


Actual system test performed by Emerging Technology Services, IBM Hursley, United Kingdom

Governed Data Access Plane - Across clouds

Data Virtualization with the Knowledge Catalog

Development Partnership
with Large EU Bank



Transactional Systems &
Core Business
applications

Data Sources

- Systems of Record
- Systems of Engagement
- 3rd Party Data
- Social Media
- Documents
- News
- Weather
- Other External

IBM Cloud
openstack
aws
Azure
Google Cloud

Data Integration

Data Movement

Data Replication

Data Lake Repositories

- Flat Files
- NoSQL Object Store
- Cloud Data Sources
- Data Warehouse
- Hadoop Data Lakes
- …

IBM Cloud
openstack
aws
Azure
Google Cloud

Data Access & Insights Services

Connectors

- Caching
- Optimizer
- Data Access (SQL / APIs)

Data Virtualization Services

Policy Enforcement

Business Terms

Publish Assets

- Auto Data Discovery
- Auto Data Curation
- Auto Data Governance
- Auto Data Quality

Knowledge Catalog
(Data Assets, Policies & Rules, Lineage, ….)

- Asset Browser
- Data Preview
- Global Search

Knowledge Catalog

Cloud Pak for Data

IBM Cloud
aws
Azure
Google Cloud
openstack

Access Data

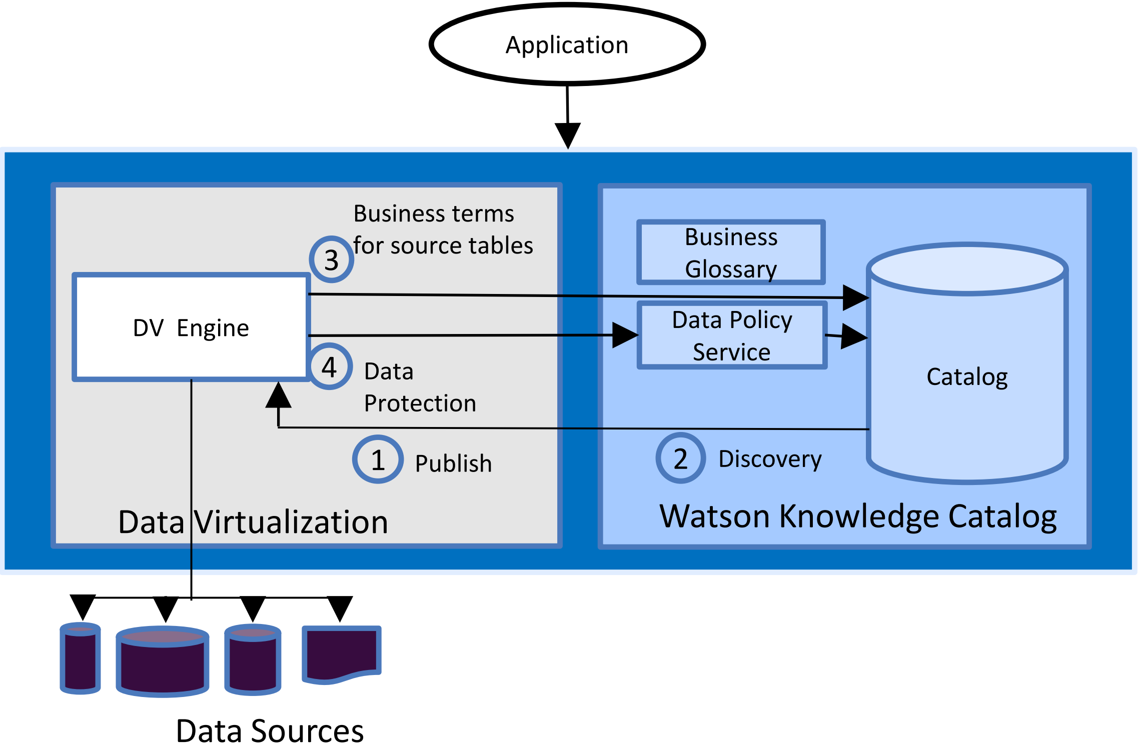
Find Data

Data Consumers

- **Business Applications**
- **AI, ML & Optimization**
- **Compliance Reporting**
- **Discovery & Exploration**
- **Self-Services Analytics**
- **BI Reporting, Dashboard**

Integrated Governance

1. DV Publishes remote assets to WKC.
2. WKC Discovery performs classification, scoring, term association
3. DV retrieves Business Terms for the source tables to give the users a common understanding of the data.
4. DV obtains information about the policies it needs to enforce upon data access.



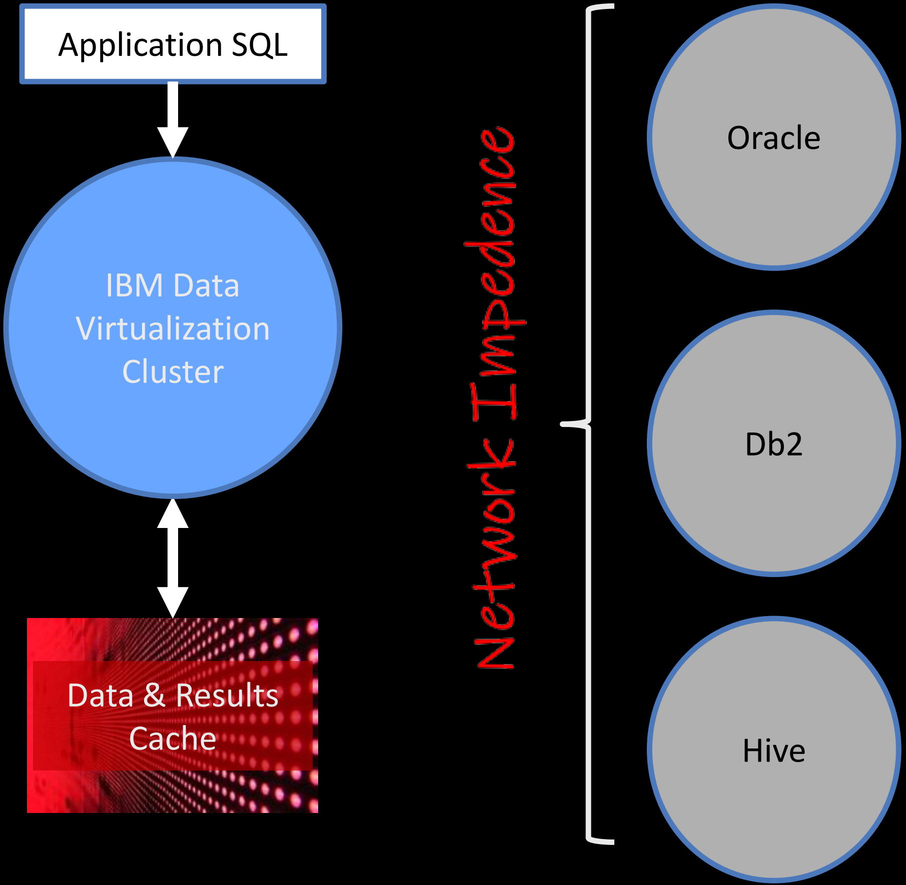
Data and Result Caching

Powerful

- Cache results (common SQL statements)
- Cache data (data or aggregates, etc).
- Define refresh rate
- Monitor usage/effectiveness

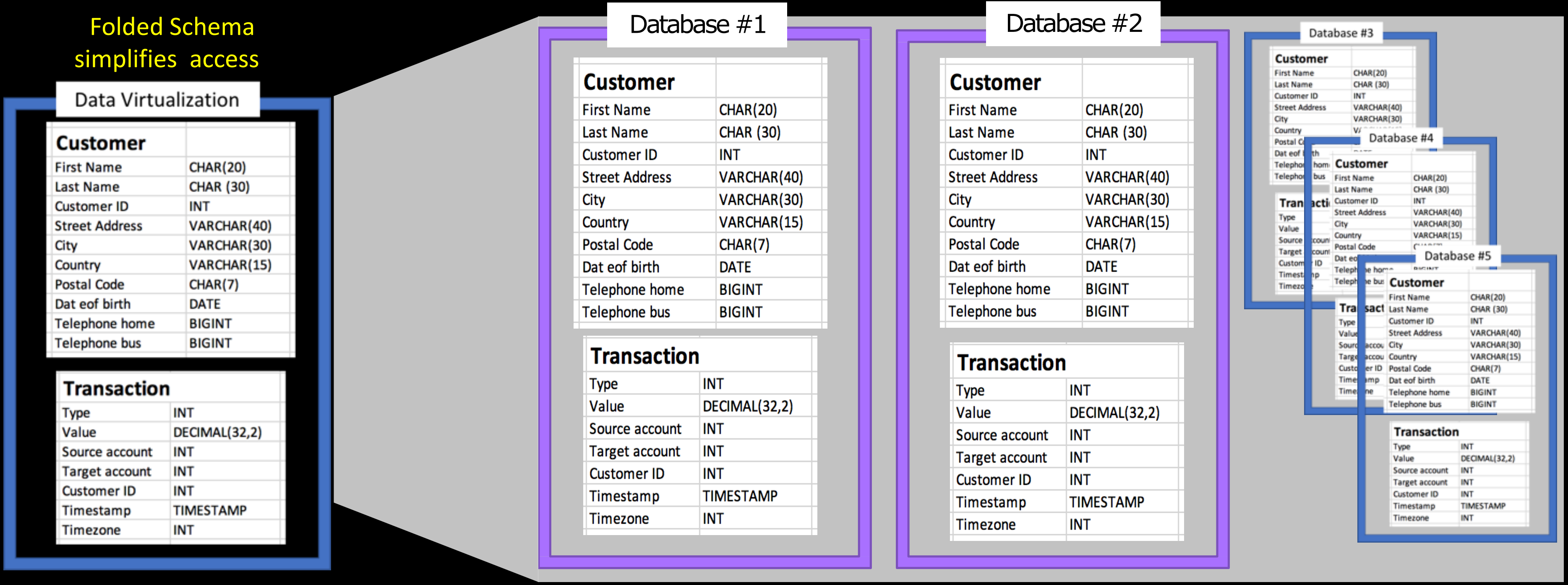
Under the hood

- Advanced query compiler determines whether to use cached data and results for part or all of a query result.



Schema Folding – Simplify Your Data

- Common or similar schemas appear in multiple databases.
  - E.g. branch database for a bank or retailer.



Folded Schema simplifies access

Data Virtualization

Customer	
First Name	CHAR(20)
Last Name	CHAR (30)
Customer ID	INT
Street Address	VARCHAR(40)
City	VARCHAR(30)
Country	VARCHAR(15)
Postal Code	CHAR(7)
Dat eof birth	DATE
Telephone home	BIGINT
Telephone bus	BIGINT

Transaction	
Type	INT
Value	DECIMAL(32,2)
Source account	INT
Target account	INT
Customer ID	INT
Timestamp	TIMESTAMP
Timezone	INT

Database #1

Customer	
First Name	CHAR(20)
Last Name	CHAR (30)
Customer ID	INT
Street Address	VARCHAR(40)
City	VARCHAR(30)
Country	VARCHAR(15)
Postal Code	CHAR(7)
Dat eof birth	DATE
Telephone home	BIGINT
Telephone bus	BIGINT

Transaction	
Type	INT
Value	DECIMAL(32,2)
Source account	INT
Target account	INT
Customer ID	INT
Timestamp	TIMESTAMP
Timezone	INT

Database #2

Customer	
First Name	CHAR(20)
Last Name	CHAR (30)
Customer ID	INT
Street Address	VARCHAR(40)
City	VARCHAR(30)
Country	VARCHAR(15)
Postal Code	CHAR(7)
Dat eof birth	DATE
Telephone home	BIGINT
Telephone bus	BIGINT

Transaction	
Type	INT
Value	DECIMAL(32,2)
Source account	INT
Target account	INT
Customer ID	INT
Timestamp	TIMESTAMP
Timezone	INT

Database #3

Database #4

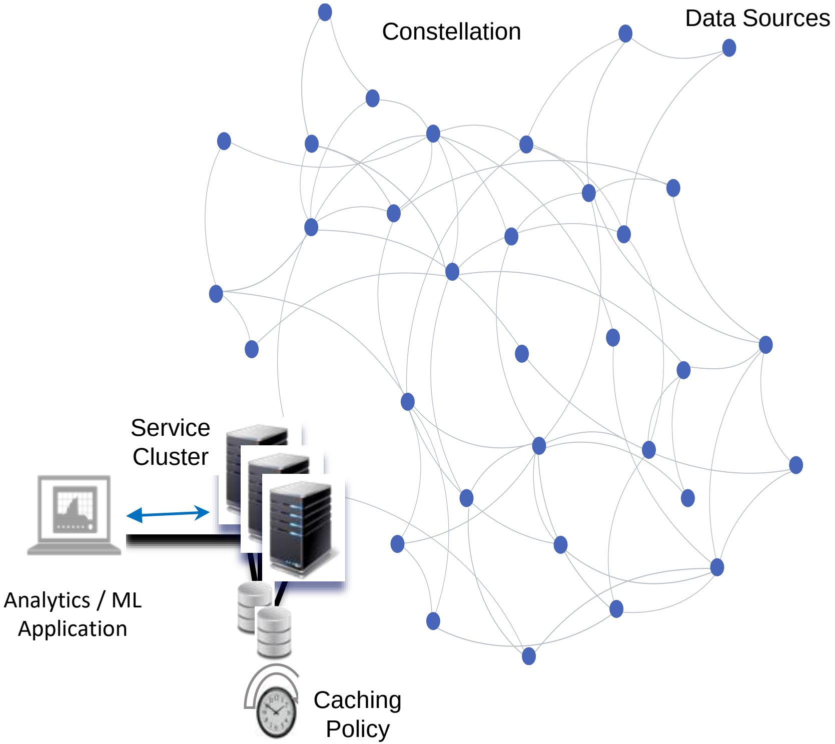
Database #5

Customer	
First Name	CHAR(20)
Last Name	CHAR (30)
Customer ID	INT
Street Address	VARCHAR(40)
City	VARCHAR(30)
Country	VARCHAR(15)
Postal Code	CHAR(7)
Dat eof birth	DATE
Telephone home	BIGINT
Telephone bus	BIGINT

Transaction	
Type	INT
Value	DECIMAL(32,2)
Source account	INT
Target account	INT
Customer ID	INT
Timestamp	TIMESTAMP
Timezone	INT

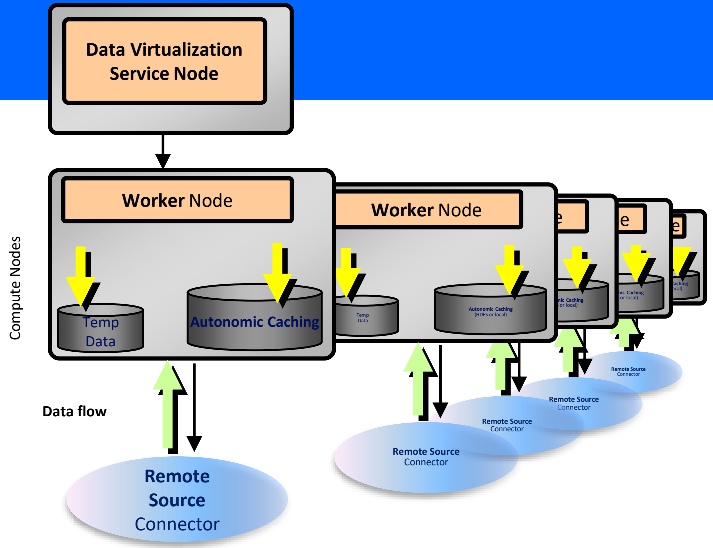
Remote Connectors and Data Discovery

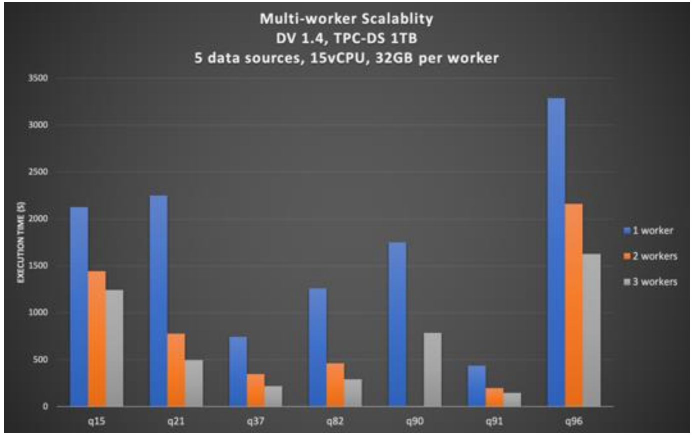
- Parallel processing mesh providing execution performance and scalability:
  - *Quickly deliver analytics results and easily evolve with new data source demands*
- Provides resilient connections between data sources:
  - *Reliability and ability to quickly adapt to increasing business demand*
- Scales seamlessly as new sources are added:
  - *New remote connectors and data sources can be added to the processing mesh without interruption of the service.*
- Provides data source discovery for data sources outside of CP4D
  - *Access to files on disk*
  - *Data sources outside of the cluster.*
- Richness of automation:
  - *Automatic formation and reorganization for best performance.*
- Highly distributed processing:
  - *Parallel access to data and query results.*
  - *Distributed algorithms to improve query performance.*



Service Compute Scalability

- Data Virtualization Compute scalability supported in Cloudpak for Data 3.0.1.
- Supporting multiple worker nodes to improve query performance for large workloads and data sets.
- Parallelized data fetch from data sources.
- When using remote connectors, additional advantages for network parallelism as data can flow on separate network routes to reach the cluster.



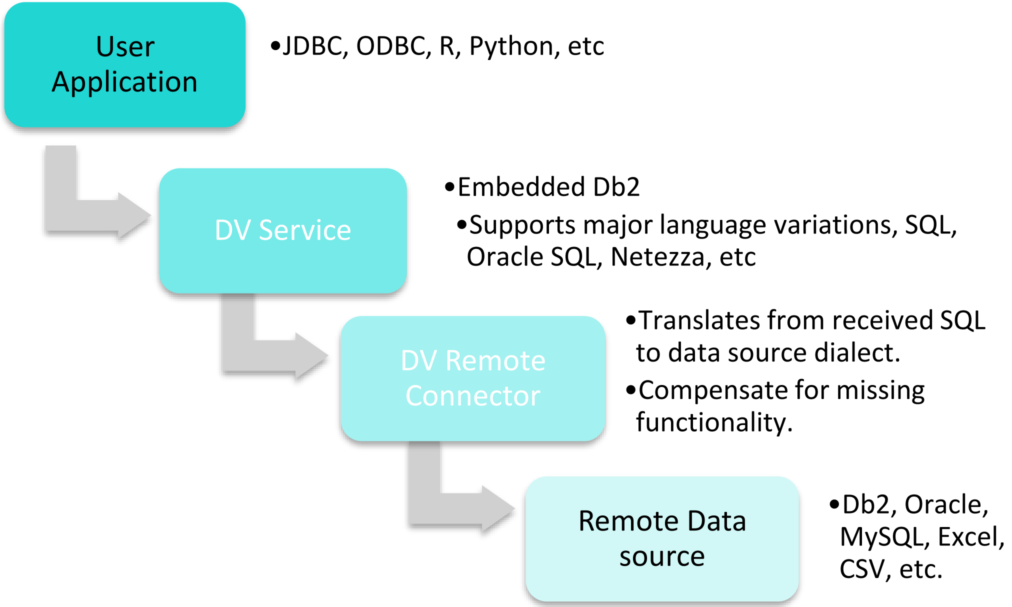


Language Translation in Data Virtualization

Broad set of data sources supported by Data Virtualization each with unique syntax variations.

Constellation is not limited only a single data source type. A logical schema is created across all connected sources.

Multiple levels of translation as we move from the applications through the constellation down to the data source.



User Application

- JDBC, ODBC, R, Python, etc

DV Service

- Embedded Db2
  - Supports major language variations, SQL, Oracle SQL, Netezza, etc

DV Remote Connector

- Translates from received SQL to data source dialect.
- Compensate for missing functionality.

Remote Data source

- Db2, Oracle, MySQL, Excel, CSV, etc.

H&R Block

Adopting a unified enterprise Data and AI platform

Business Challenge

- Lack of centralized enterprise-wide Data and Analytics strategy support processes
- Lack of modern infrastructure and automated analytics work process
- Need for an enterprise data inventory and improved data governance
- US and international data compliance requirements
- Timely access to quality data

Solution

Once the IBM Expert Lab team started working on solving the client challenges, H&R Block saw that Cloud Pak for Data was the perfect solution for many reasons. IBM provided a unified enterprise platform where all types of personas from the data ops value chain can collaborate to respond to business demands.

An engagement with the IBM Data Science and AI Elite team will get them moving fast on their data science projects. With the unification of models and access to central computing power that Cloud Pak for Data brings, H&R Block will see their projects speed up from days to just hours.

Outcome

- Unified experience for all personas in the Data Development Lifecycle
- Automated data discovery and classification plus the ability to automate routine and mundane data tasks
- Data virtualization speeds the access to governed data

Solution Components

Data Modernization and DataOps

- IBM Cloud Pak for Data System (on premise) with
  - IBM DataStage
  - IBM Watson Knowledge Catalog
  - IBM Watson Studio
  - IBM Watson Machine Learning
  - IBM SPSS Modeler

- IBM Cloud Pak for Data Initiate Services and Digital Learning from IBM's Expert Labs team
- Services from IBM Data Science and AI Elite team

[Listen to the 'H&R Block + IBM: Journey to AI' Audiogram]

Industry: Banking and Financial Markets
Geography: North America

A North American Retail Company

The largest employee owned supermarket chain in the United States looks to the future with IBM Cloud Pak for Data with IBM DataStage

Industry: Retail
Geography: North America

Business Challenge

Complex integrations between IBM z and numerous distributed POS, inventory, and operations databases drive the daily operations at this large-scale retailer. Hundreds of IBM DataStage create those integrations with speed and efficiency. Nevertheless, the client is pursuing a platform modernization away from the mainframe and on-premise workloads in lieu of cloud-ready platforms.

Solution

Our IBM Cloud Pak for Data modernization program focused on DataStage provides the client with the flexibility they need to maintain their existing DataStage workloads, all while preparing for a future where DataStage jobs can run across a hybrid cloud architecture. Additionally, data virtualization and governance features within Cloud Pak for Data provide net new capabilities that are on the client's technology roadmap to reduce data movement and provide more self-service data access and provisioning.

Outcome

A future proof hybrid cloud solution that support the growth of DataStage workloads and provides the client with several “default choice” capabilities for next gen capabilities like IBM Data Virtualization and IBM Watson Knowledge Catalog.

Solution Components

DataOps

- IBM Cloud Pak for Data on-premise with
  - IBM DataStage

Learn more about DataStage

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

Learn more about Data Virtualization on CPD

youtube: IBM Cloud Pak for Data - Intro to Data Virtualization

Blog: IBM Big Data & Analytics Hub

Solution brief: IBM Data Virtualization

Documentation: Knowledge Center (latest release)

Q&A