# Organize Data

# Setting

**A manufacturing company XYZ depends upon 2 different raw materials (raw material A and B) which come from various companies.**
**Company A provides raw material A. This data is stored in on-prem MYSQL database.**
**Company B provides raw material B. This data is stored in DB2 on XYZ cloud provider.**

Both the Raw materials are required by the manufacturing company to manufacture a product XYZ.
How do we classify the following –

# Tasks

1. View the combined data of raw material A and raw material B as a single view for manufacturing XYZ product.
2. Some of the columns which appear in the view are redundant. We need to get rid of the column in the view without changing the data at physical location.
3. Calculate the total discount applied when the raw material A and B were purchased.
4. Above processing needs to be scheduled every fortnightly.
5. Visualise a bar chart of discounts applied for raw material A vs raw material B
6. Discount value applied on purchase of raw material A should be hidden as it is considered as confidential column.
7. The email id of the customer is sensitive data which needs to be masked.

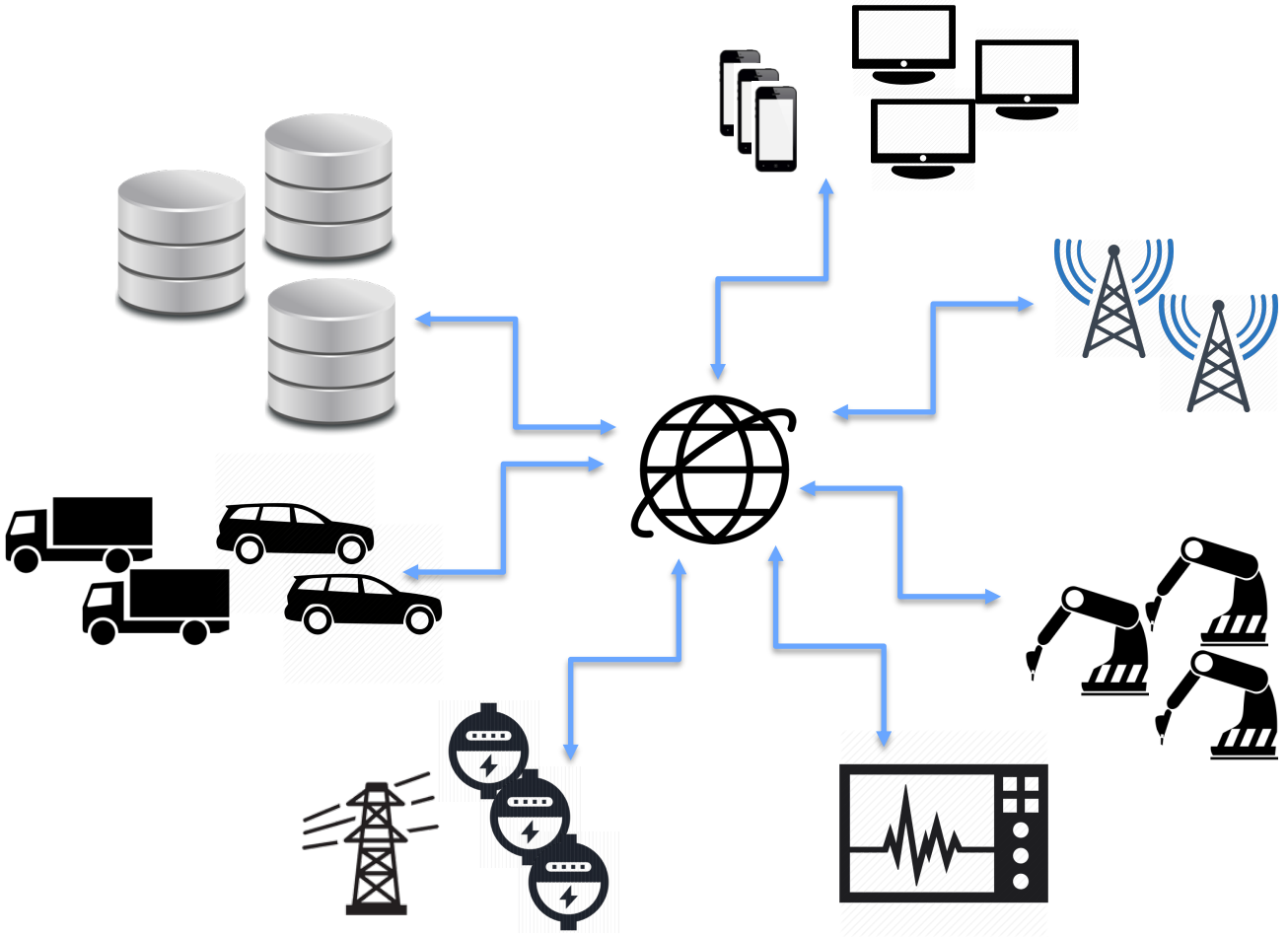
| Collect | Organize | Analyze |
| --- | --- | --- |
| 1,4 | 2,3,4,6,7 | 2,3,4,5 |

Data is Everywhere

Number of Sources and volume rapidly increasing

More and more heterogeneous

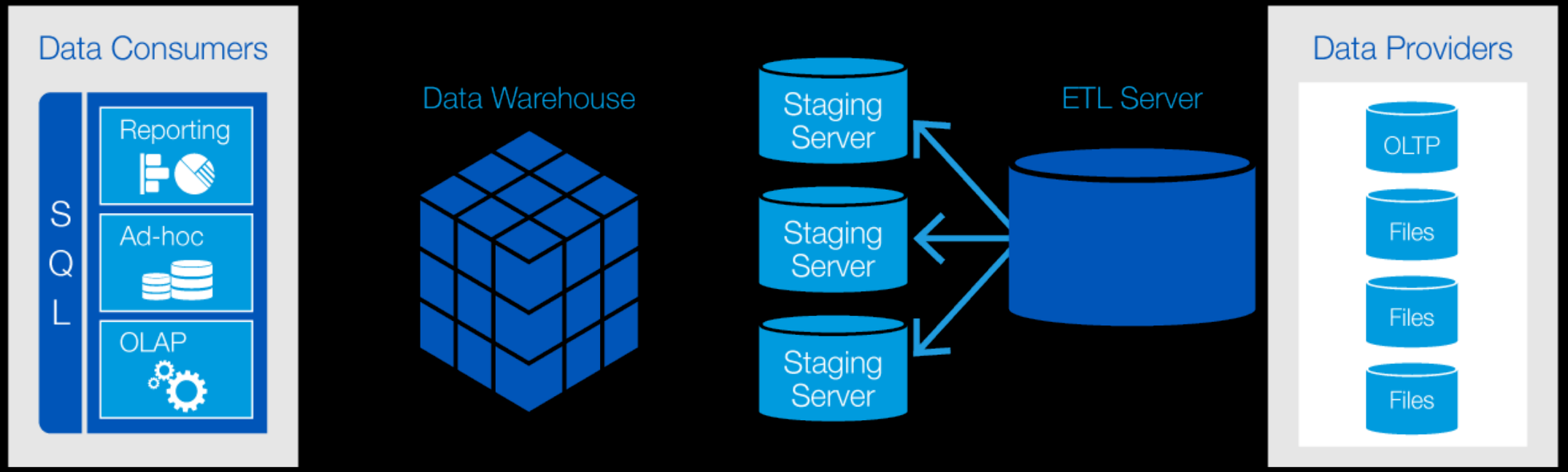
Highly distributed – internal and external



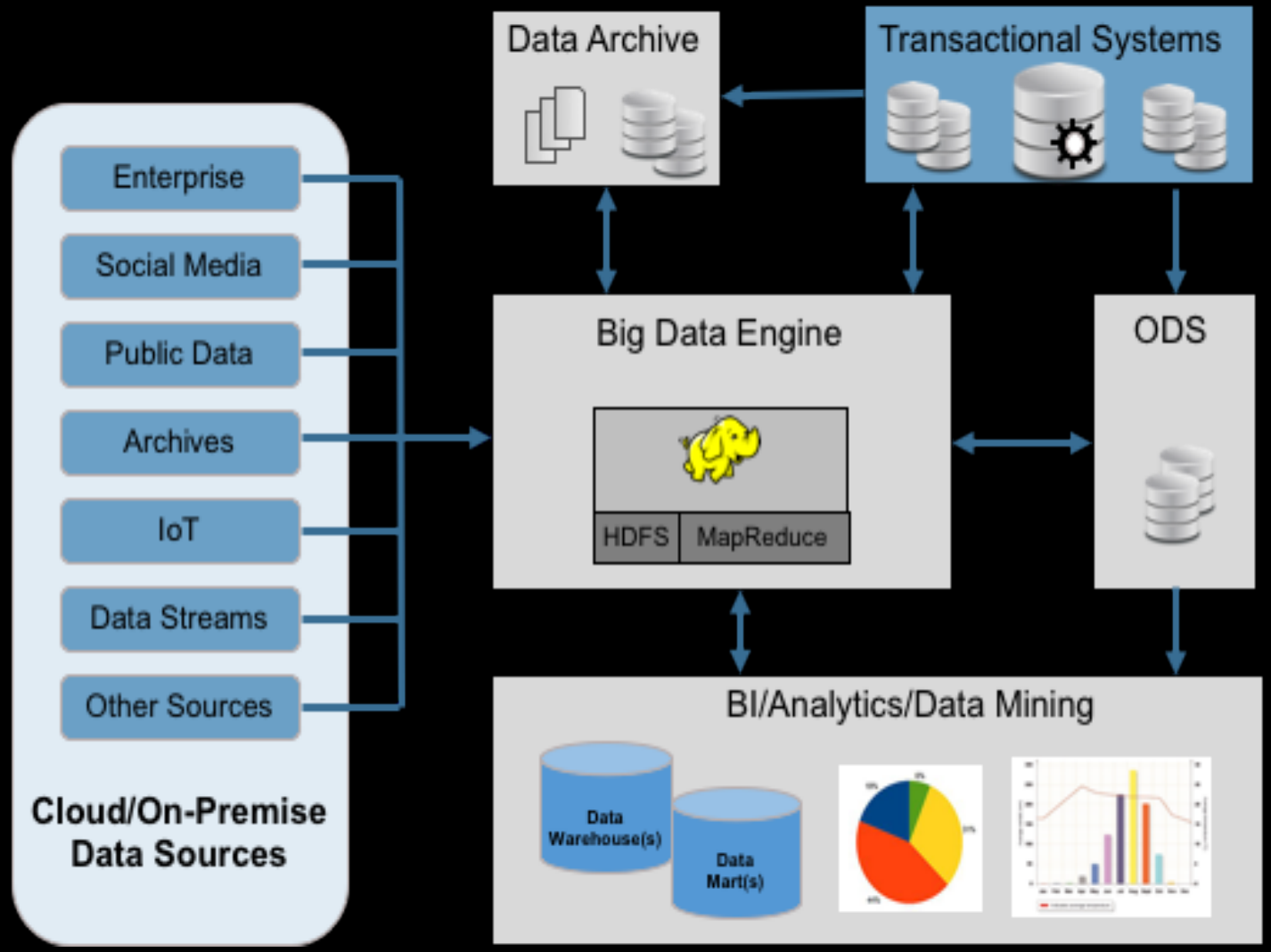
# Traditional Data Integration

Not Viable to continue to Move or Copy *all* Data
(using extract, transform, load ETL)

- Risk to data security
- Data inconsistency & error prone
- Rigid and limits business agility
- High cost and latency
- Finite scalability



# Current Data Architectures




- Numerous ETLs
- Unnecessary duplication and data replication as business users demand more data views
- Data governance issues accelerating across the enterprise

**Physics have driven Cost & Complexity**
**And impede Productivity**

# A Single View for Self-Service

Study Reference : “From Data to Insight: Work Practices of Analysts in the Enterprise”, Kandogan, Balakrishnan, Haber, Pierce, submitted to IEEE Computer Graphics and Applications, Special Issue on Business Intelligence Analytics

The growth of Big Data compounds complexity, resulting in an increase in *friction* that already exists between IT and business consumption

Studies report **friction between elements** of the ecosystem lead to major inefficiencies.

- Real-time data access for real-time decisions
- Finding data is hard
- Need metadata for lineage, quality, currency
- Need for virtualized access to persistent data

# What is Data Virtualization?

The ability to view, access, manipulate and analyze data without the need to know or understand its physical format or location, and without having to move or copy it.

# What is the value proposition ?

What if you could tap into all of your critical data assets no matter where they physically are?

What if you could query 2 or 2,000 data systems with a single query?

# Data virtualization key Use Cases

Driven by patterns needing low data latency, high flexibility with transient schema

- On-demand Virtual Data Marts to save cost of standing up EDW
- EDW prototyping and migration (mergers/acquisitions)
- Virtualization with Big Data (Hadoop, NoSQL, and Data Science)
- EDW augmentation (offload workload)
- Data discovery for ”what if” scenarios across hybrid platforms
- Data Caching of combined data for frequently accessed data
- Combine MDM with IoT for Systems of Insight (IT/OT)
- Data integration preparation tool to complement ETL
- Master data hub extension to enrich 360 View (e.g. multi-channel CRM)

✓ single View across your business
✓ real-time analytics without moving data
✓ fast TTV with high ROI

# How does it work ?

# Some Definitions

*Data Federation - technology*

Federation is the underlying approach for defining access/authorization to logically mapped remote data sources and query technology used for the execution of distributed query processing against multiple data sources.

*Data Virtualization - platform*

Any approach to data management that allows an application to retrieve and manipulate data without requiring technical details about the data, such as how it is formatted at source, or where it is physically located, and can provide a **single customer view** of the overall data. Typically necessitates a platform or information architecture. (Data Federation is embedded)

*Data Integration enhanced*

Data integration solutions traditionally **move a copy of the data** from disparate sources into a new consolidated source. Data Virtualization complements by providing a view of the integrated data while **leaving the source data exactly where it is**.

# IBM Data Virtualization

## Query across multiple data sources:

- Oracle, Db2, SQLServer, Informix, Netezza, MySQL, PostgreSQL, Big SQL, Apache Hive, HDP Hive, Cloudera Impala and more!

## Rich application capabilities

- Connect to Data Virtualization with your favorite SQL apps and tools
- RStudio, Jupyter Notebook, Cognos, Tableau, Microstrategy

## Scale!  1 or 1,000 at once

- More than 10x better for several important use cases

## Secure!

- Strict access controls
- Fully encrypted communications

## Schema discovery and folding

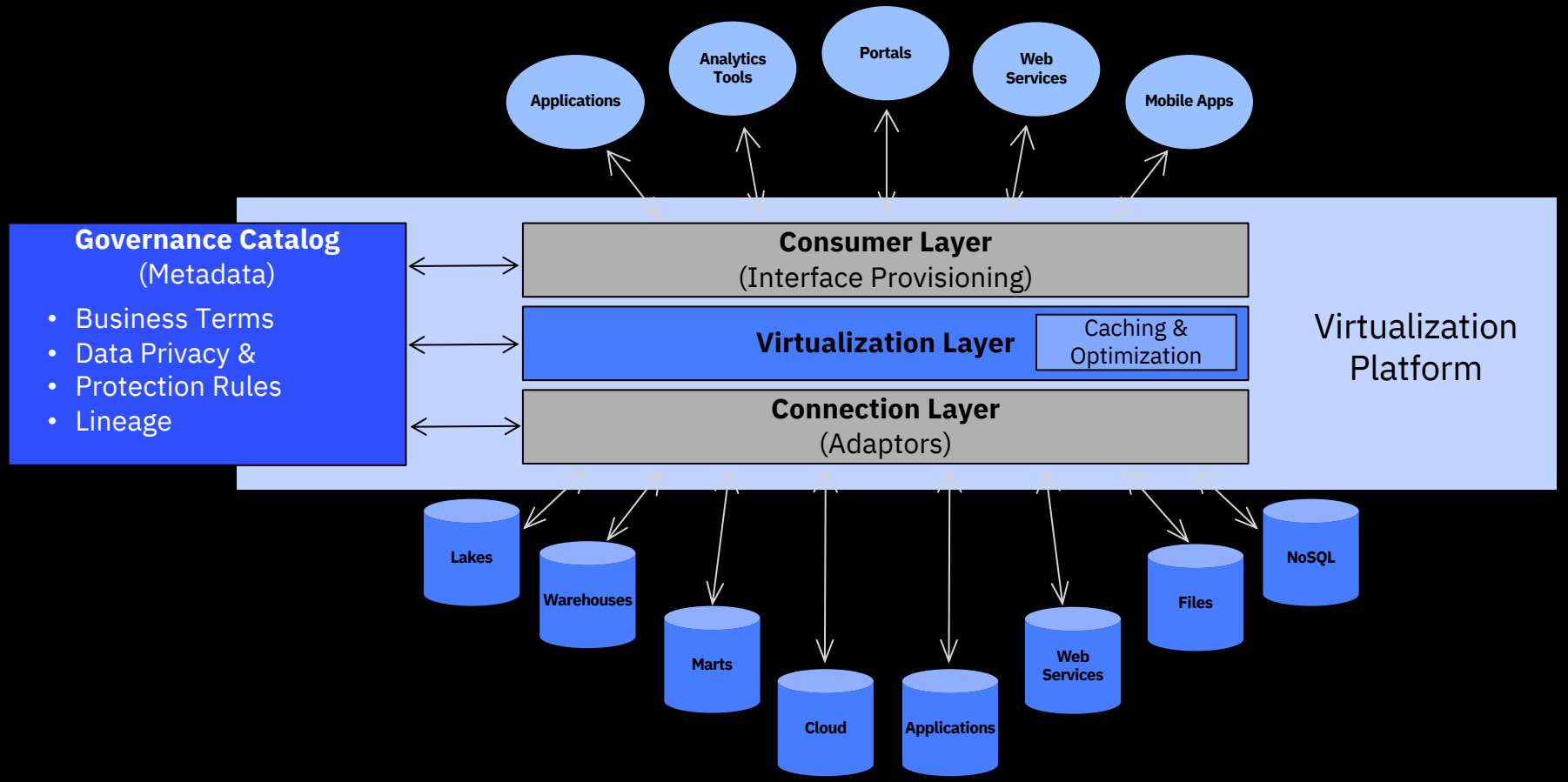
- Automatically find and match tables across systems so you can query them as a single virtual table.

## Deeply integrated w Cloud Pak for Data

- Enterprise Data Catalog, governance and security. e.g. Automatic publishing of virtualized data into the data catalog.
- Immediate access via Cognos and Watson Studio

# Data Virtualization in Cloud Pak for Data

The ability to view, access, manipulate and analyze data without the need to know or understand its physical format or location, and without having to move or copy it.



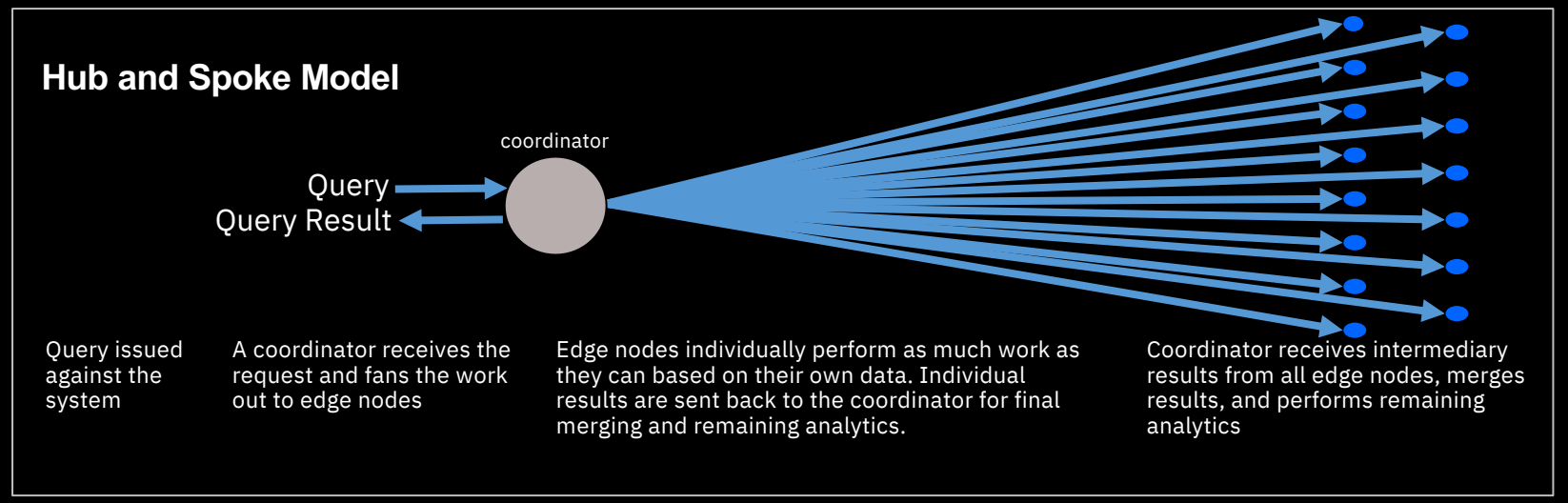
# Key Architectural Differentiation

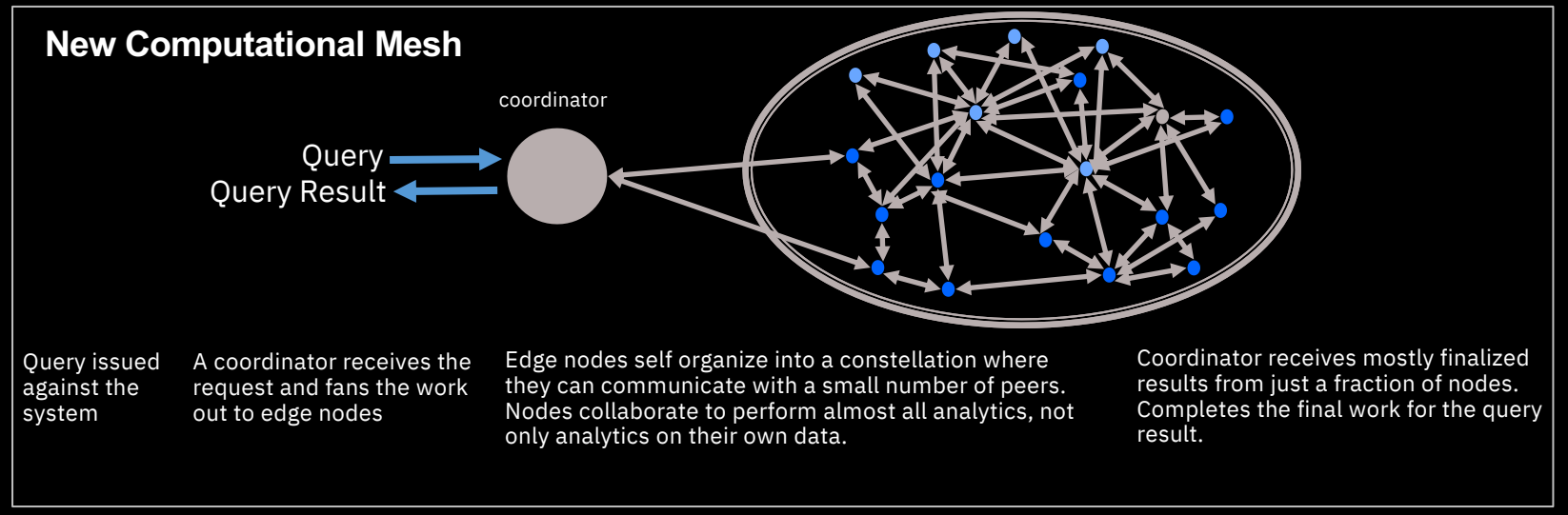
Hub and spoke execution models:

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

IBM is first to market with a parallel processing model:

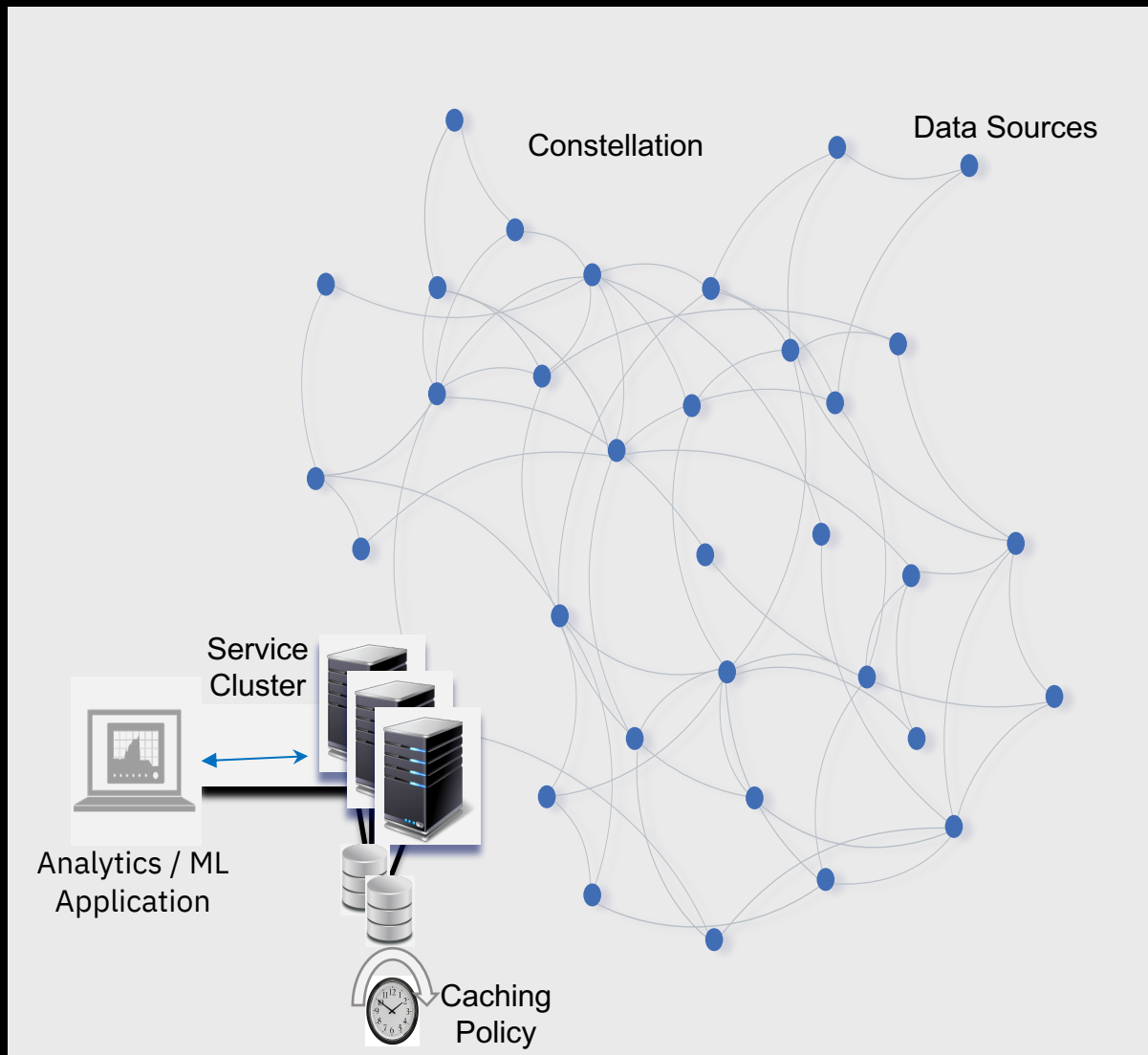
- Theoretically unlimited scalability
- Ease of addition/removal of sources
- Execution pushed down into the constellation mesh



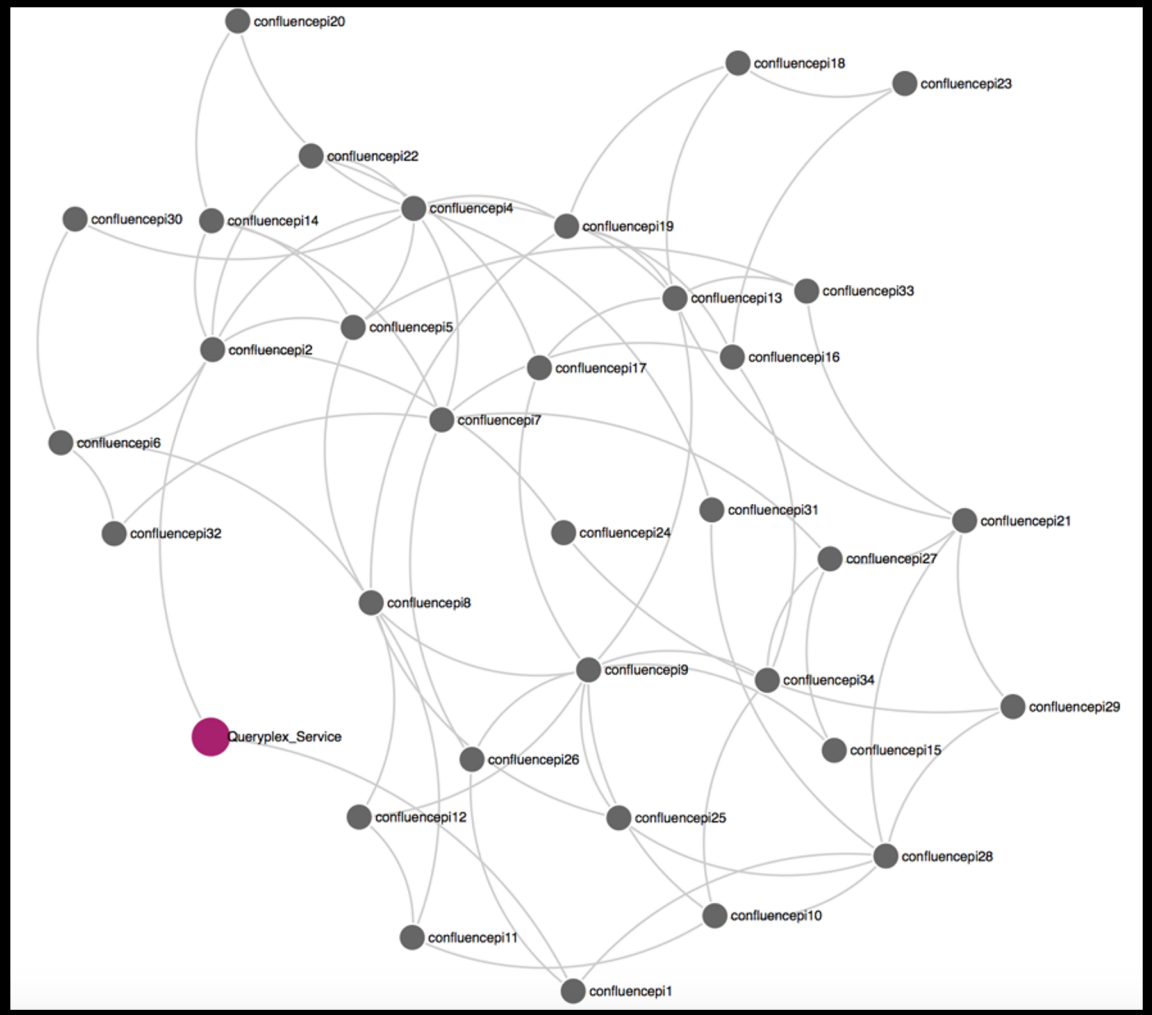
# IBM’s Unique Approach to Data Virtualization

- Parallel processing mesh providing execution performance and scalability:
  - *Quickly deliver analytics results and easily evolve with new data source demands*
- Service Cluster scalability and Enterprise robustness:
  - *Reliability and ability to quickly adapt to increasing business demand*
- Governance Integration and Security:
  - *Controlled, governed and secure access to virtual data sets within the Cloud Pak for Data Platform*
- Common SQL engine, rich set of SQL dialects, application portability
  - *Retain the use of existing applications and tools within the business*
- Richness of automation and discovery of Data:
  - *Enable more self service and increase productivity (while retaining access control)*
- Underlying platform flexibility with Cloud Pak for Data:
  - *Grow or move your Analytics and Virtualization platform with any environment*



# Creating the Constellation

- Small packet of Data Virtualization software deployed on each data node, approximately 50MB.
- Automatic dynamic and resilient organization of data sources.
- Query compiler builds a collaborative query plan, forcing the nodes to collaborate



# Broad support for common data source types

*More to be added in the future.*

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

## Cloud Pak for Data (Nov 2019)

- Db2 family for HDM
- Db2 for iSeries, zSeries
- Db2 for z/OS
- Big SQL
- IIAS, PDA (Netezza)
- Informix
- Derby
- Oracle
- SQL Server
- MySQL
- PostgreSQL
- Apache Hive, HDP Hive
- Cloudera Impala
- Teradata*
- MongoDB
- Hive
- Excel, CSV, Text*
- Sybase
- MariaDB
- Snowflake
- Z Data Sources through IBM DVM Integration
  - VSAM, IMS, CICS, Adabas
- Map-R (Hive)

## In the roadmap pipeline

- BigQuery ( patch in 2019, GA Q1, 20)
- SAP HANA
- SAP BW
- Amazon Redshift
- Salesforce
- SAS
- Interbase
- Apache Drill
- Amazon Dynamo, Aurora
- Generic use of any JDBC access
- CouchDB
- Stream / MQ
- Apache Spark SQL
- Apache Kafka
- Cloudant
- Pivotal Greenplum
- Cassandra

# Data Virtualization Manager for z/OS

# Data Virtualization Manager for z/OS

## More than 35 supported data sources

### IBM Z

ADABAS

VSAM

IMS

IBM DB2

CICS

IDMS

CA IDMS

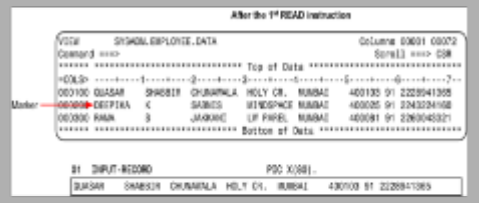
Systems Management Facility SMF

Data for IT Operational Analytics

Syslogs log streams

LOG

Sequential-file



File Systems

zFS

HFS

SFM

Magnetic Tape Virtual Tape

### Non-IBM Z

IBM Db2 Family, Informix

hadoop

BigInsights

ORACLE

HIVE

mongoDB.

Microsoft SQL Server

20+ DRDA data sources

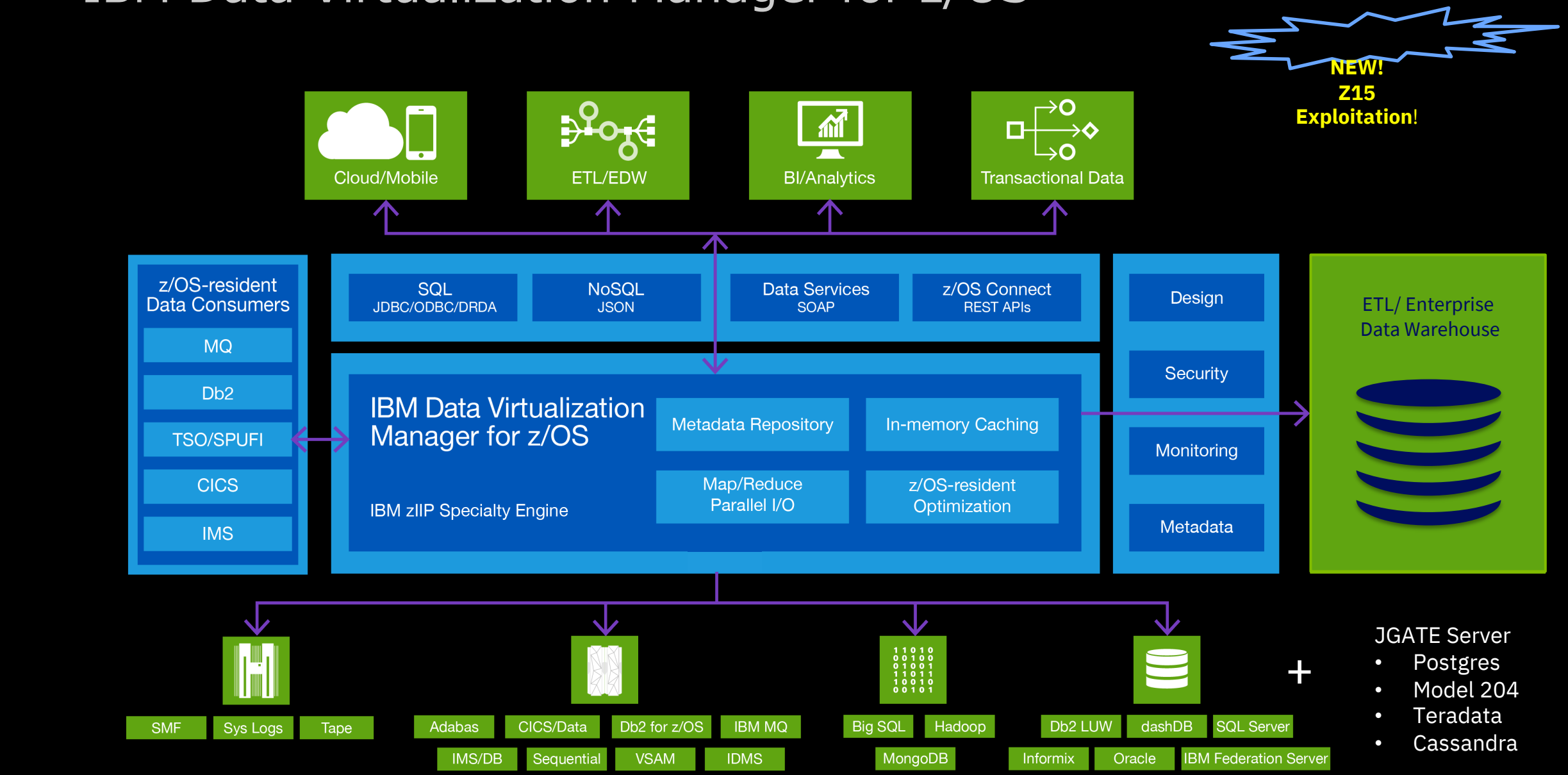
Other DBs

THE Open GROUP

**... exposed to data consumers**

**in SQL (JDBC/ODBC/DRDA) and NoSQL (JSON) formats**

# IBM Data Virtualization Manager for z/OS



# DVM for z/os  Use Cases

- **Expand Db2 applications beyond just Db2**
  - New features making Db2 as the Datahub for non-Db2 data sources
  - ”Any Db2 application to any data”
  - Large Insurance company

- **Providing z data for testing scenarios**
  - DVM enabled customer’s dev tool to access IMS
  - Multi-national Telecomm company

✓ **Creating a virtual data lake**
  - System of record data stayed on z reducing ETL
  - Inventory information located across continents is virtualized with DVM
  - Multi-national Automotive company

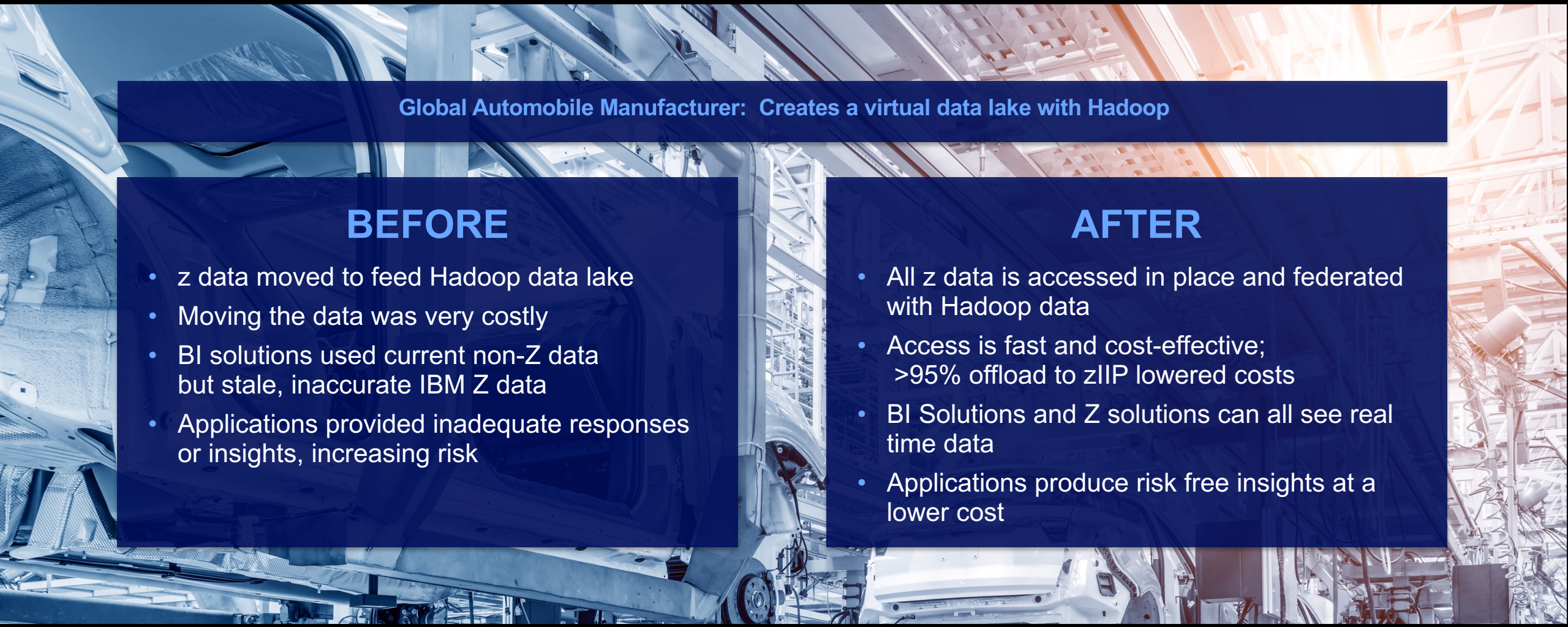
✓ **Modernization with DVM and z/os Connect**
  - DVM used to provide Rest access to z/os data (VSAM, IMS, etc) as well as Db2 z
  - DVM used to access IMS data directly as backup to unstable data warehouse on distributed platform
  - Global Airline

✓ **Shrinking app dev cycle with DVM and ETL**
  - No win situation with ETL (customer politics and deep investment)
  - Large Financial

- **Improving data access to aircraft maintenance records**
  - 4 hr ETL process  to Oracle ODS reduced
  - DVM showed no significant increase in resource utilization on z
  - PoC showed accessing IMS data in place as a good option  over data movement
  - Global  Airline

- **IBM’s Classic Federation Upsell**
  - 5x the performance in PoC (Bank)
  - Classic Fed no longer being enhanced
  - Migration plan being developed by Rocket - - available soon

# Make Data Simple & Accessible



**Global Automobile Manufacturer: Creates a virtual data lake with Hadoop**

## BEFORE

- z data moved to feed Hadoop data lake
- Moving the data was very costly
- BI solutions used current non-Z data but stale, inaccurate IBM Z data
- Applications provided inadequate responses or insights, increasing risk

## AFTER

- All z data is accessed in place and federated with Hadoop data
- Access is fast and cost-effective; >95% offload to zIIP lowered costs
- BI Solutions and Z solutions can all see real time data
- Applications produce risk free insights at a lower cost

# Cloud Pak for Data DV combined with DVM for z/os

## Cloud Pak for Data DV and DVM integration

- Extends the CPD DV service to the mainframe using DVM for z/OS
- Provides seamless interface and access to ALL Z data
- Auto-discovers z data, cataloging, (Gather/Manage, Understand, Govern)
- Leverages real-time persisted data for analytics and ML using z data
- Enables Developers and Business Users lacking mainframe skills
- Reduces costs and complexity of ETL when working with z data
- DVM connectivity with Integrated DRDA Facility, Db2 UDTF support (User Defined Table Function)
- ALL DVM for z/os functions (read/write, API support, data sources support, exploitation of hardware (z15!) are available – no changes to DVM

## Products for Z :

- DVM for z/os (**5698-DVM**) – business as usual
- Cloud pak for Data zPPA: **D1YH6LL** (Enterprise Edition) or **D1ZXLLL** (Native Edition)

## Client already has DVM?

- There is no change to their license. No additional cost..
- Action: Sell Cloudpak for Data

# IBM Data Virtualization Manager– Key Enhancements

| | |
|---|---|
| Integrated DRDA Facility | *Allows joining data from multiple data sources even when the data sources have different catalogs or are located across sysplexes. Example is joining data across IMS servers and VSAM catalogs that is spread out across sysplexes*.<br><br>Technology: A DVM Server can now connect to and access virtual data from one or many DVM servers. Data can be shared between DVM servers without being on the same Sysplex. |
| Db2 UDTF support (User Defined Table Function) | *Allows any application to access any DVM virtualized data using DB2 via SQL. Ex: DB2 app accesses and updates VSAM records using SQL.*<br><br>Technology: DVM supports the creation of a Db2 UDTF on any virtualized object defined to a DVM server within any Db2 subsystem connected to DVM. Once the UDTF has been created any connection/application to Db2 can easily access this data using Db2's SQl engine. |
| JAVA Gateway | *Any JAVA application can access data without needing a JDBC driver to be purchased for the data source. Customers are using this to access* ***Teradata, Postgres, Model 204 and Cassandra data***.<br>Technology: A DRDA application server can be placed on any system to access data via a JDBC driver and make it accessible to any application. Avoids dealing with OEM DB to drive DRDA support. |

# **Cloud Pak for Data –** Self-serve ready

Foundational “out of the box” multicloud data & AI services

**Open, Extensible Platform**


App Developers & Analytics Ops


Business Partners

**Personalized, Collaborative Platform**

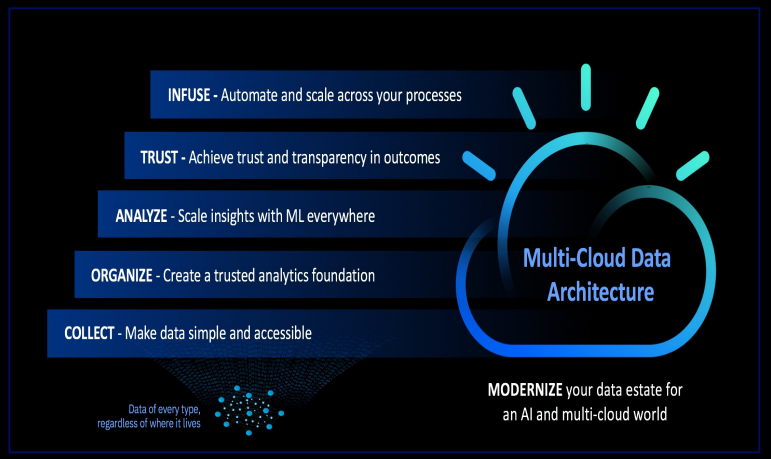

Data Engineers


Data Stewards


Data Scientists


Business Users

**The Ladder to AI**




**MODERNIZE** your data estate for AI in a multi-cloud world

APIs Integrated User Experience

Extensible : “*add-ons*”, accelerators and Solutions

***Modular***: - *provision services & scale out when needed*

**Collect & Connect**

- Data virtualization
- Provision SQL & NOSQL Databases
- Warehouses & Marts
- Event Ingestion & Streaming Analytics
- Distributed compute – Apache Spark

**Organize & Integrate**

- Discovery & search
- Data transformation
- Data catalogs, quality & Curation
- Business glossary
- Policies, rules & privacy

**Analyze & Infuse**

- Data Science & Visualization
- Dashboards & reporting
- AUTO-AI, ML deployments & operations
- AI Trust and Transparency - Explainability & Bias detection
- AI services –Chat, NLU

**Core Services**

- Logging
- Monitoring
- Metering
- Storage Volumes
- Auditing
- Security
- Identity Access Mgmt.
- Docker Registry , Helm

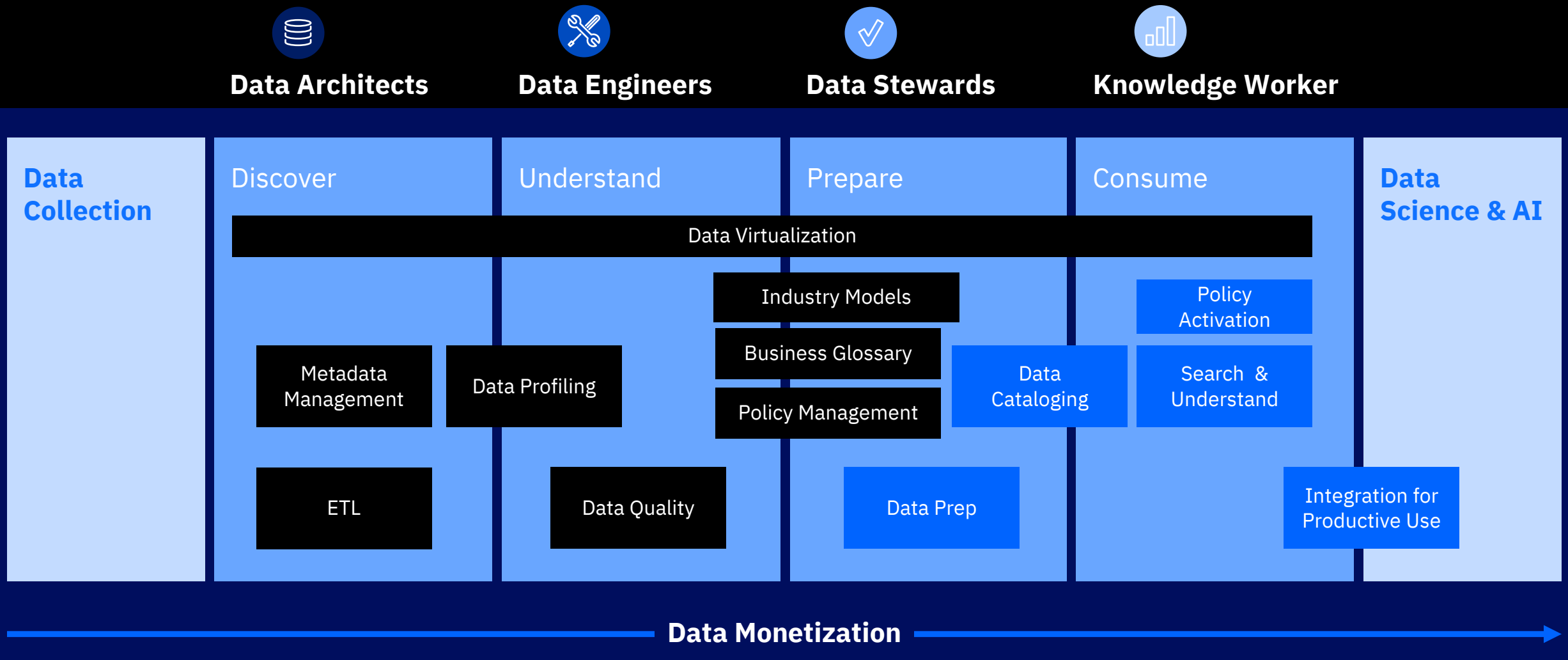
RED HAT OPENSHIFT

IBM Cloud aws Azure Google Cloud openstack

**Hyperconverged** System

# Supporting the continuum from data collection to consumption

## Enabling velocity, scalability, and traceability



# Data Governance

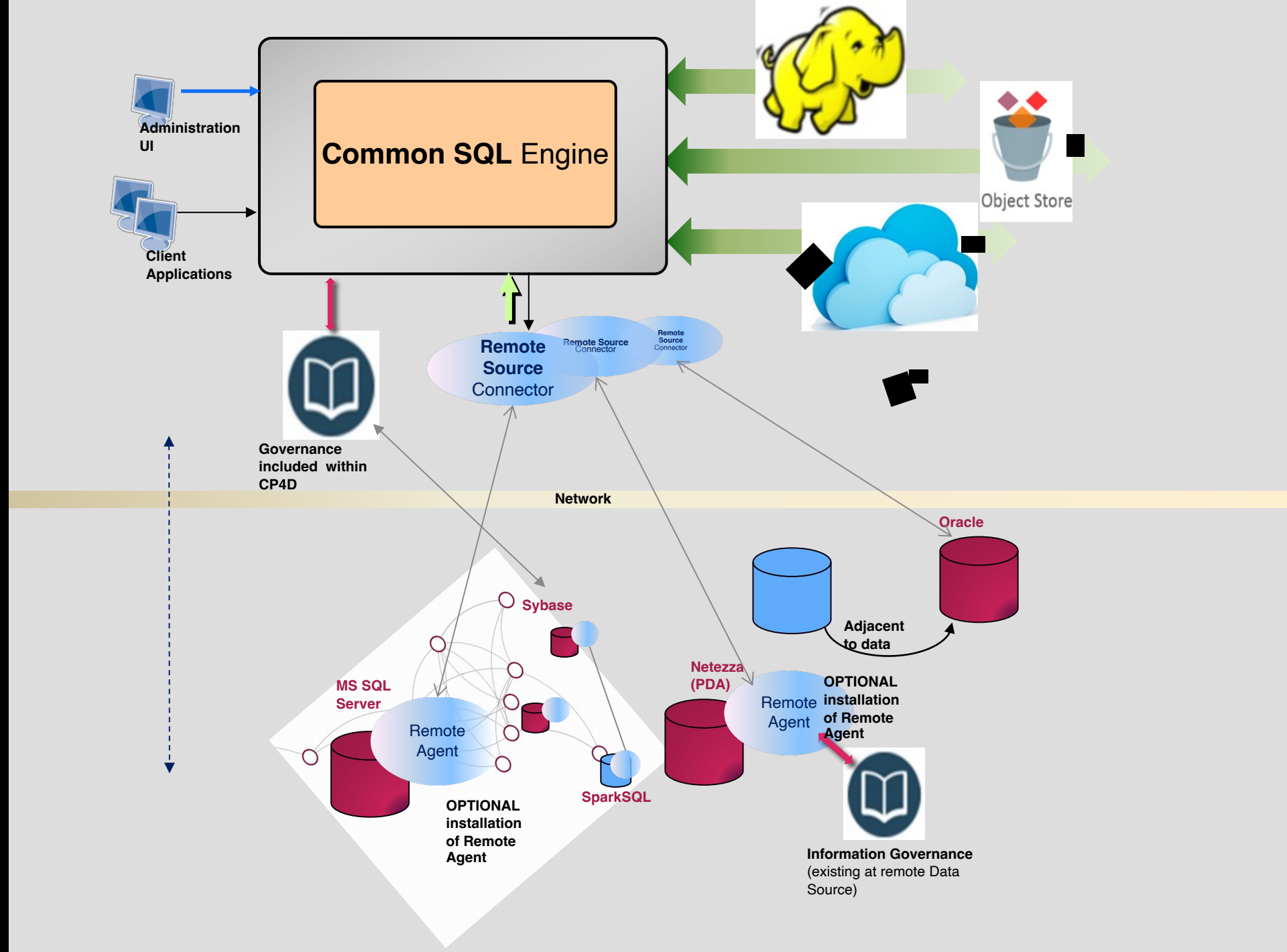
**Classifications**: Each data asset has a classification that describes the sensitivity of the data. Classifications are provided by Watson Knowledge Catalog.

**Data classes**: For relational data sets, one data class is assigned to each column during profiling. Data classes are provided by Watson Knowledge Catalog or can be shared from Information Governance Catalog.

**Business terms**: You create business terms in the **Business Glossary** tool to define business concepts in a standard way for your enterprise. Business terms can also be shared from Information Governance Catalog

**Policies and rules**: You control access to data by creating policies and rules in the **Policy Manager** tool. Within rules, you can include classifications, data classes, business terms, or tags to identify the data to control.

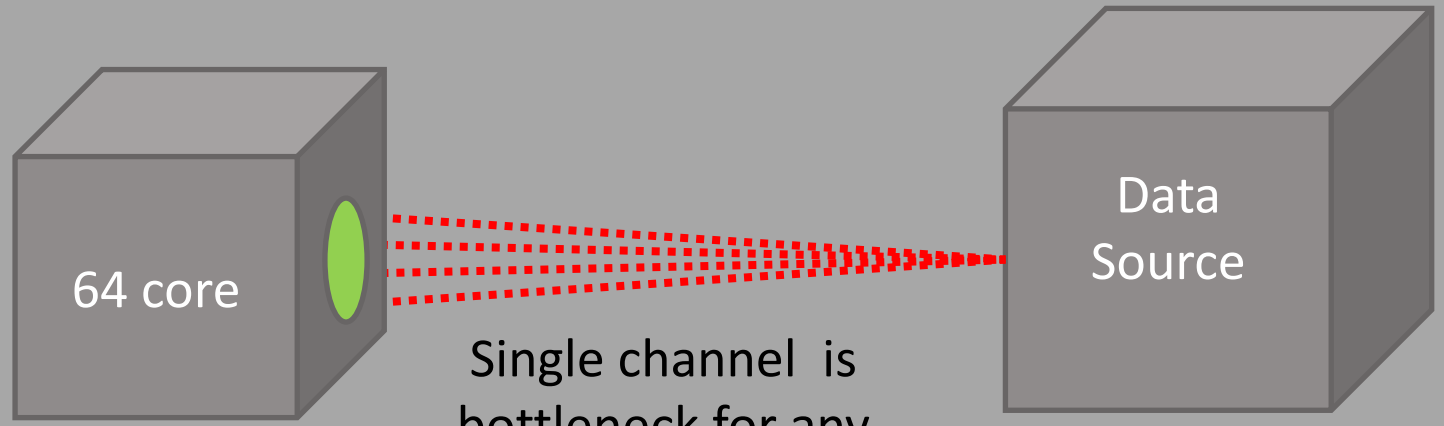
# Architecture



*Bringing parallelism to each data source.*

Traditional processing queries the data source (on right) then processes large results on a single thread (executor) on left.

Solution: Data Virtualization queries data source (on right). Merges results in many **parallel threads** on left by leveraging parallel read streams.

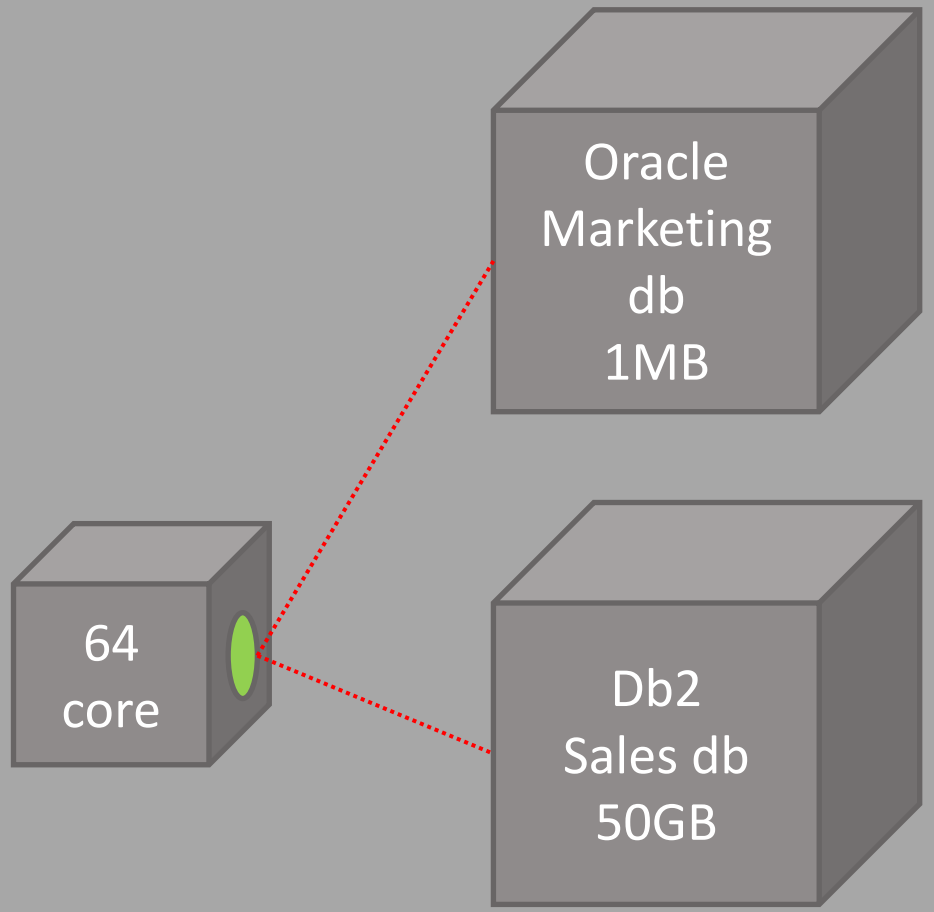



Single channel is bottleneck for any large result set. Data Virtualization resolves this through parallel read from sources.

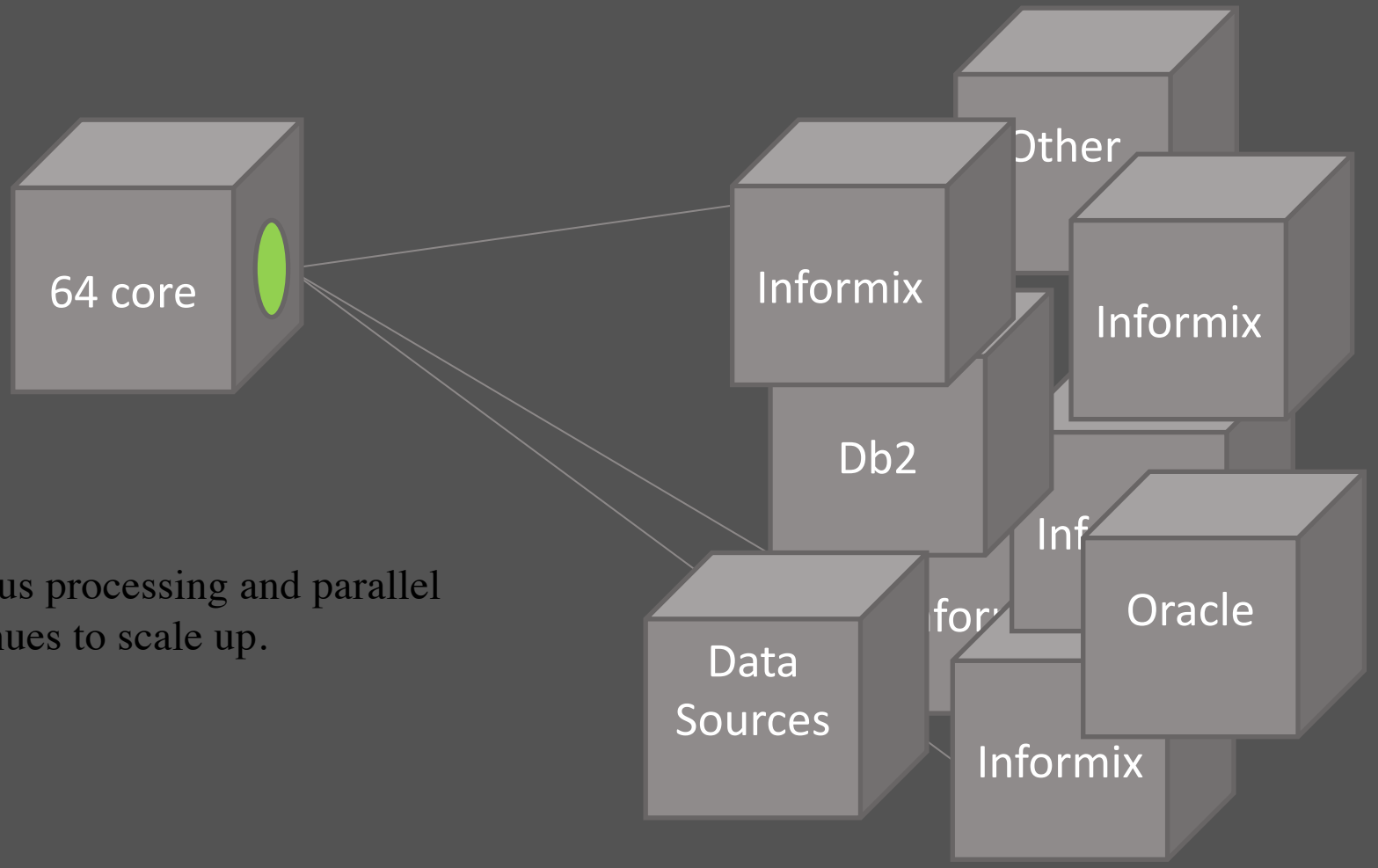
# Efficient Cross Source Joins

**Problem**: Joining data between multiple databases is a common pattern, but brutally slow. With traditional processing, data from both tables is shipped over the network to the server, where the join is processed.

**Solution**: Data Virtualization uses the **early filtering** techniques to dramatically reduce transmission costs. Data from the smaller table (inner) is used to build a small filter query that is applied to the larger (outer) table before data is transmitted. In many cases this is 85% effective at filtering data that will be removed by the join before the join is processed, leading commonly to a ~10x acceleration.



# 20 databases? Traditional processing gets worse as the number of databases increases



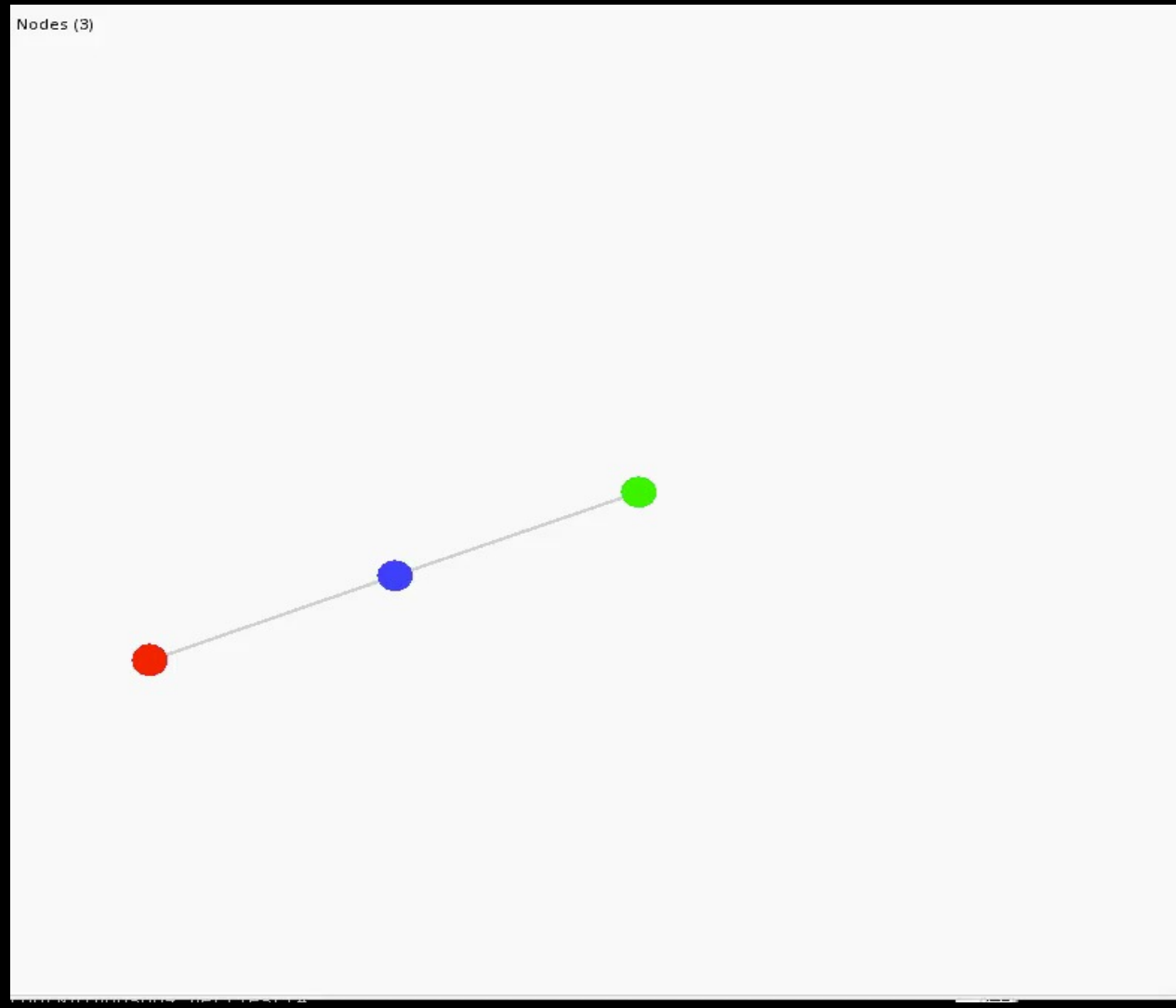

Simultaneous processing and parallel fetch continues to scale up.

# 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.

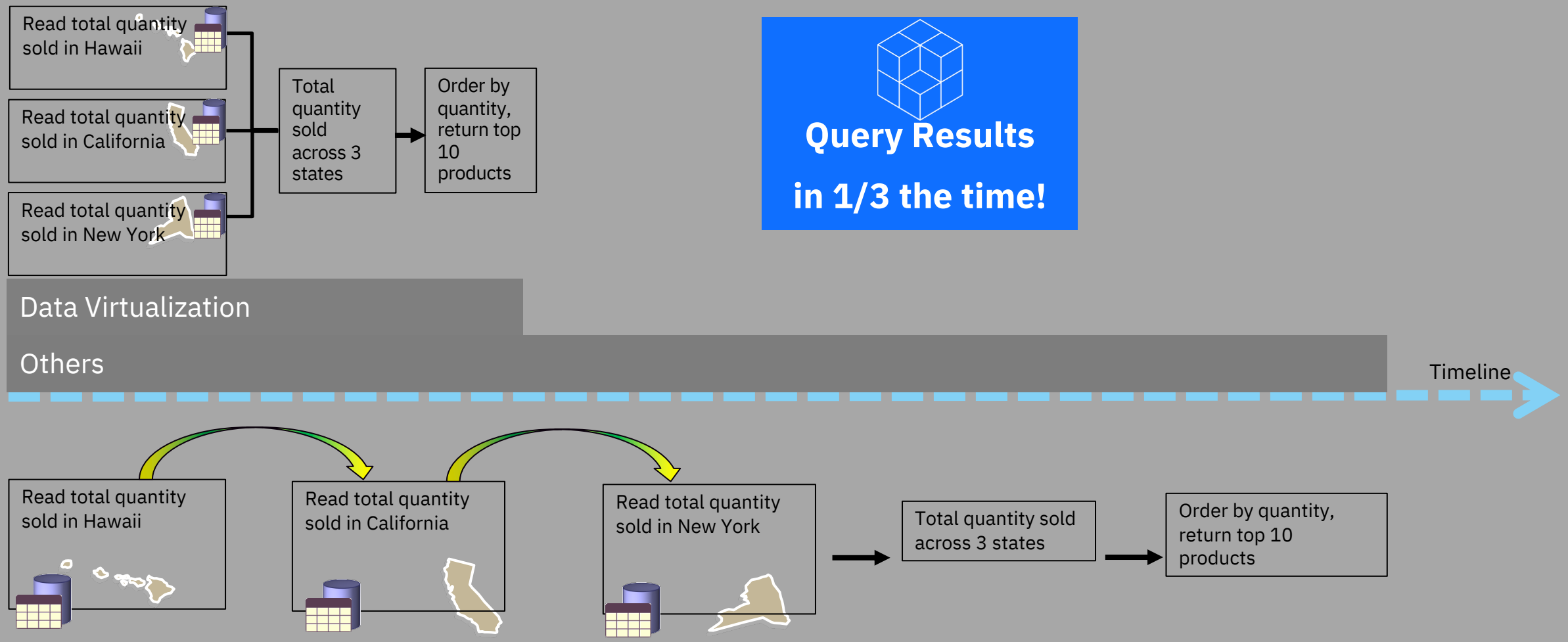
- Latency aware connection between nodes
  - Which nodes connect to which others?
  - Fastest reply strategy

- Diameter of the constellation (i.e. the number of hops between the two furthest nodes) grows logarithmically. Small diameter is ideal for communications.



# Smart Query Processing

IBM Data Virtualization reduces query time by using Parallel Processing, Pushdown Optimization and Connection Pooling



# DV Engine Generated SQL for Distribute Aggregation

Original

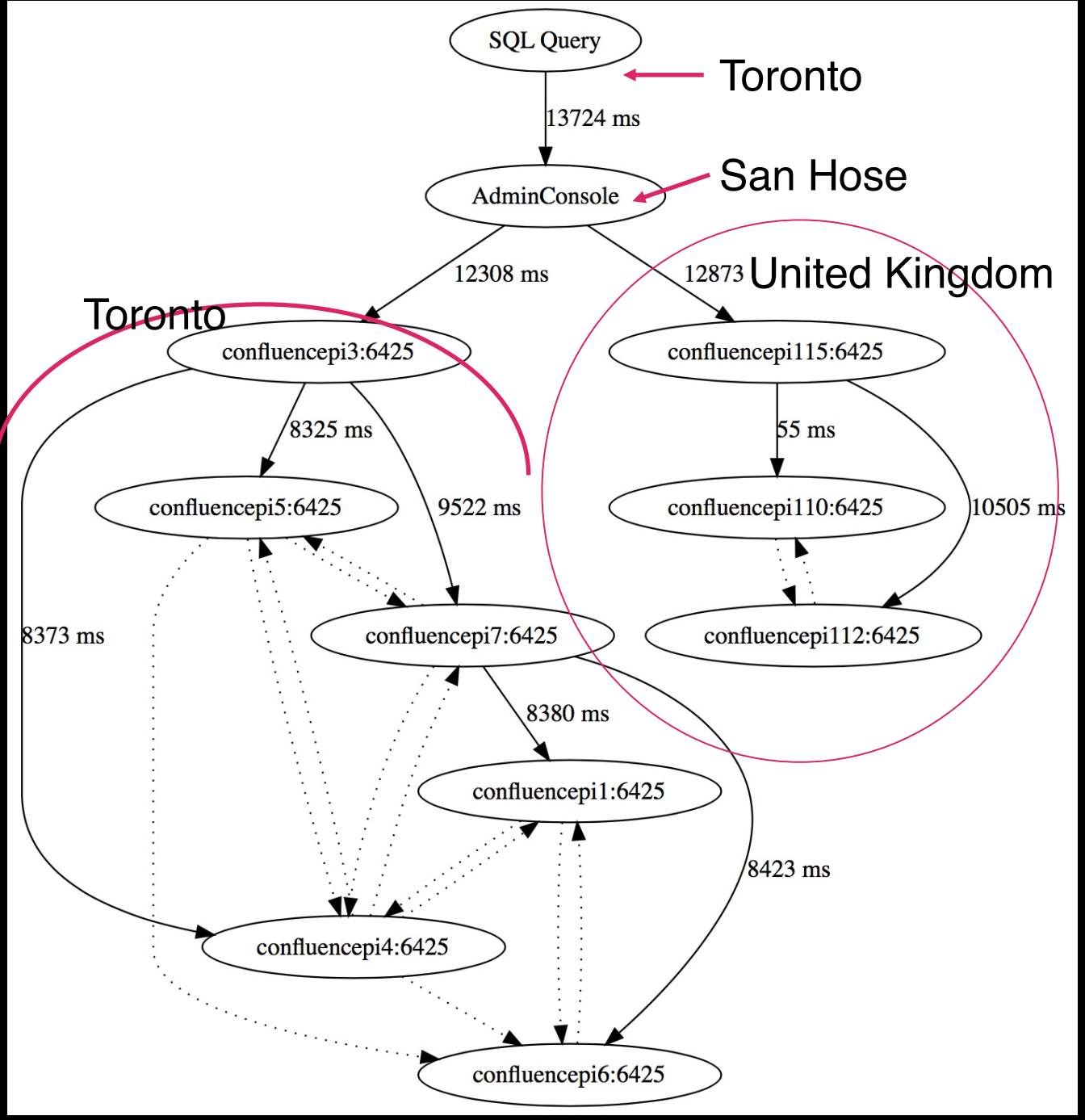
```
SELECT COUNT(PROMOVALUE2) FROM PROMOTION
```

Remote SQL

```
SELECT SUM( A0.C0)
  FROM (
    SELECT A1.C0 C0
      FROM new com.ibm.db2j.GaianQuery(
            'SELECT COUNT( A2."PROMOVALUE2") C0
               FROM new com.ibm.db2j.GaianTable(
                      ''PROMOTION'',
                      ''SOURCELIST=(MYSQL10000:"POPS_node1", MYSQL10001:"POPS_node2",
                                    MYSQL10002:"POPS_node3", MYSQL10003:"POPS_node4",
                                    MYSQL10004:"POPS_node5")'',
                      '''"PROMOKEY" INTEGER, "PROMOTYPE" INTEGER, "PROMODESC" CHAR(30),
                        "PROMOVALUE" DECIMAL(5, 2), "PROMOVALUE2" DECIMAL(5, 2),"PROMO_COST" DECIMAL(9, 2)''
             ) A2',
            'GAIAN_EXTENDER=[pAgg]','',
            'C0 DECIMAL(5, 2)') A1
       ) A0
```

# Query Processing in the Constellation

- Fixed execution within the constellation is impossible because of the highly dynamic nature of the network.
- Each node instead simultaneously sends the relevant portions of the query to both the connected data source to it's peers in the network.
- Combines and process the results as they are received.
- Duplicate results are avoided by a given node only returning results to the first peer that requested them.
- Implicitly results in balanced processing of the query through the constellation.

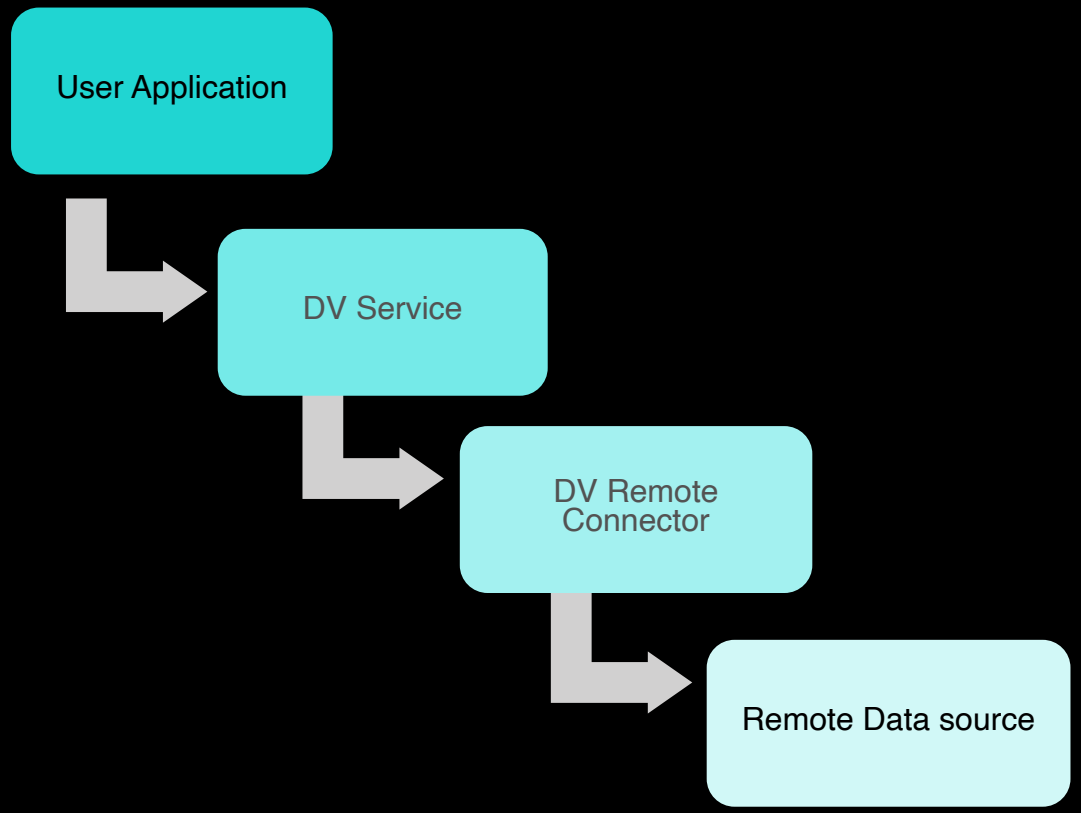


# Language Translation in Data Virtualization

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.

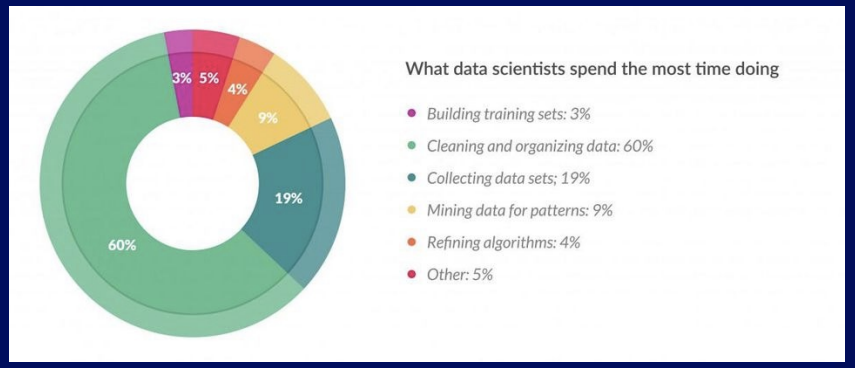
Multiple levels of translation as we move from the applications through the constellation down to the data source.



# The Big Problem!

**80%**

of data is either inaccessible, un-governed, untrusted, unanalyzed




Hence, we need to ‘organize’ / ‘catalog’ and build an Information Architecture (IA)

# The AI Ladder

A prescriptive approach to accelerating the journey to AI

AI

**INFUSE** – Operationalize AI with trust and transparency

**ANALYZE** - Scale insights with AI everywhere

**ORGANIZE** - Create a trusted analytics foundation

**COLLECT** - Make data simple and accessible

Data of every type,
regardless of where it lives

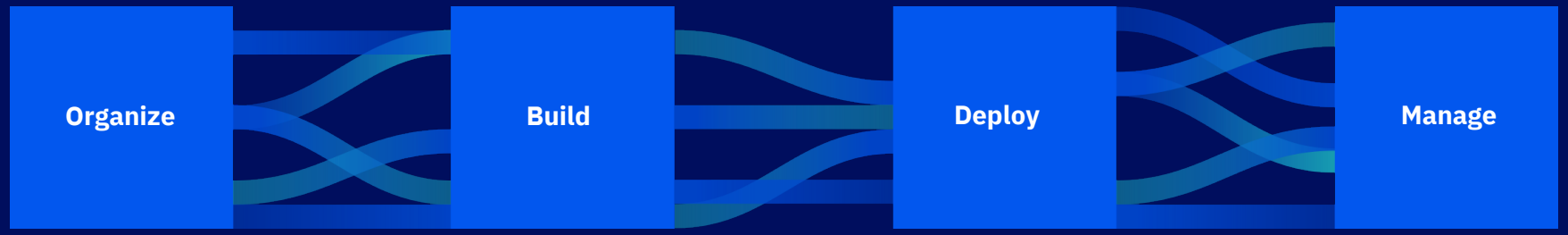
**MODERNIZE**
your data estate for an
AI and multicloud world

**There is no AI without an IA** (information architecture)

# 4 critical steps to operationalizing AI

## Business-Ready data

1. What AI teams Want to Focus on




2. Where 80% of their time spent on

3. Where organizations see business results

| Discover | Integrate | Govern | Data Profiling | Curate | Organize | Asset Quality |
|---|---|---|---|---|---|---|
| • Intelligent Search & Explore | • Data Connectors (50+) | • Policy Manager & Rules | • Auto Profiling (60+ Classifiers) | • Data Refinery + Operations | • Cataloging | • Review & Understand |

# IBM’s AI Portfolio

Everything you need for Enterprise AI, on any cloud

The Ladder to AI

Interact with Pre-built AI Services

**Watson Application Services**

| Organize | Build | Deploy | Manage |
|---|---|---|---|
| **Watson Knowledge Catalog** | **Watson Studio** | **Watson Machine Learning** | **Watson OpenScale** |

OS AI Open Source Frameworks

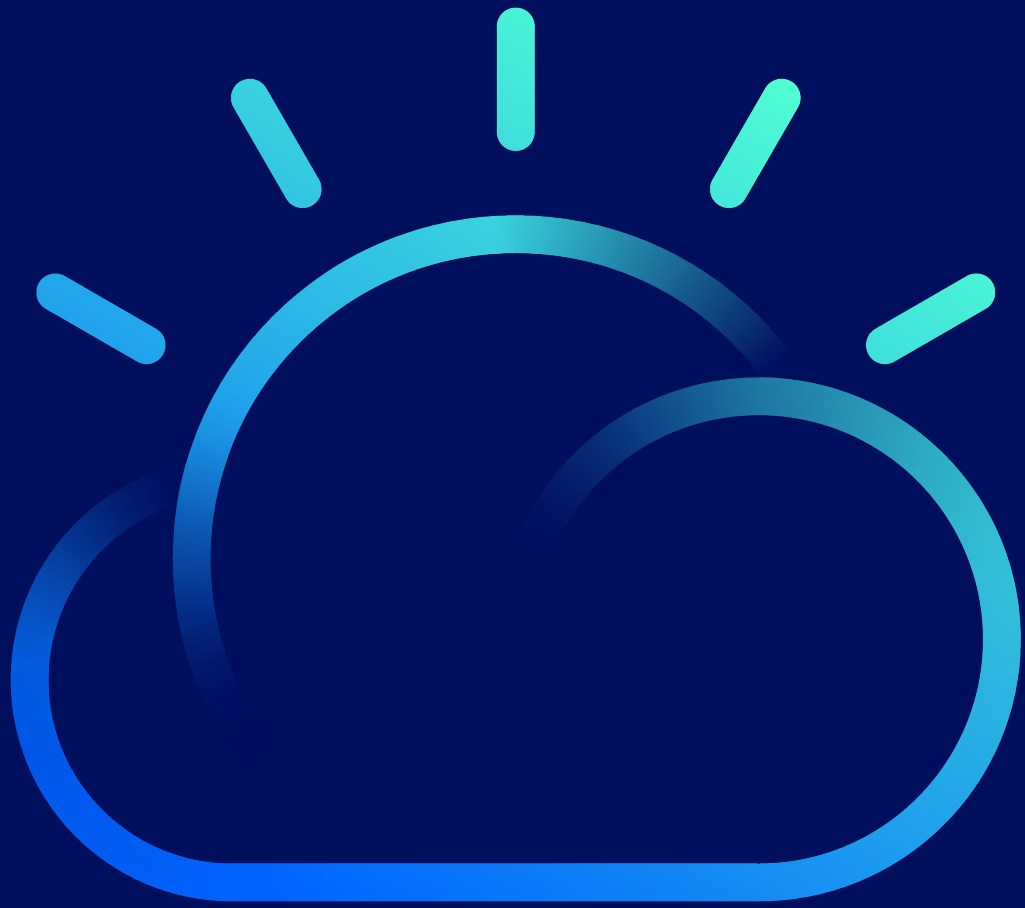
Unify on a Multicloud Data Platform

**IBM Cloud Private for Data**

# IBM Watson Knowledge Catalog

*An intelligent, integrated  enterprise data catalog and data governance platform*

- Auto Profiling
- Quality and Anonymization
- Data Governance
- Tags & Business terms
- Policy & Rules
- Column Classifiers (160+)
- Search and Explore

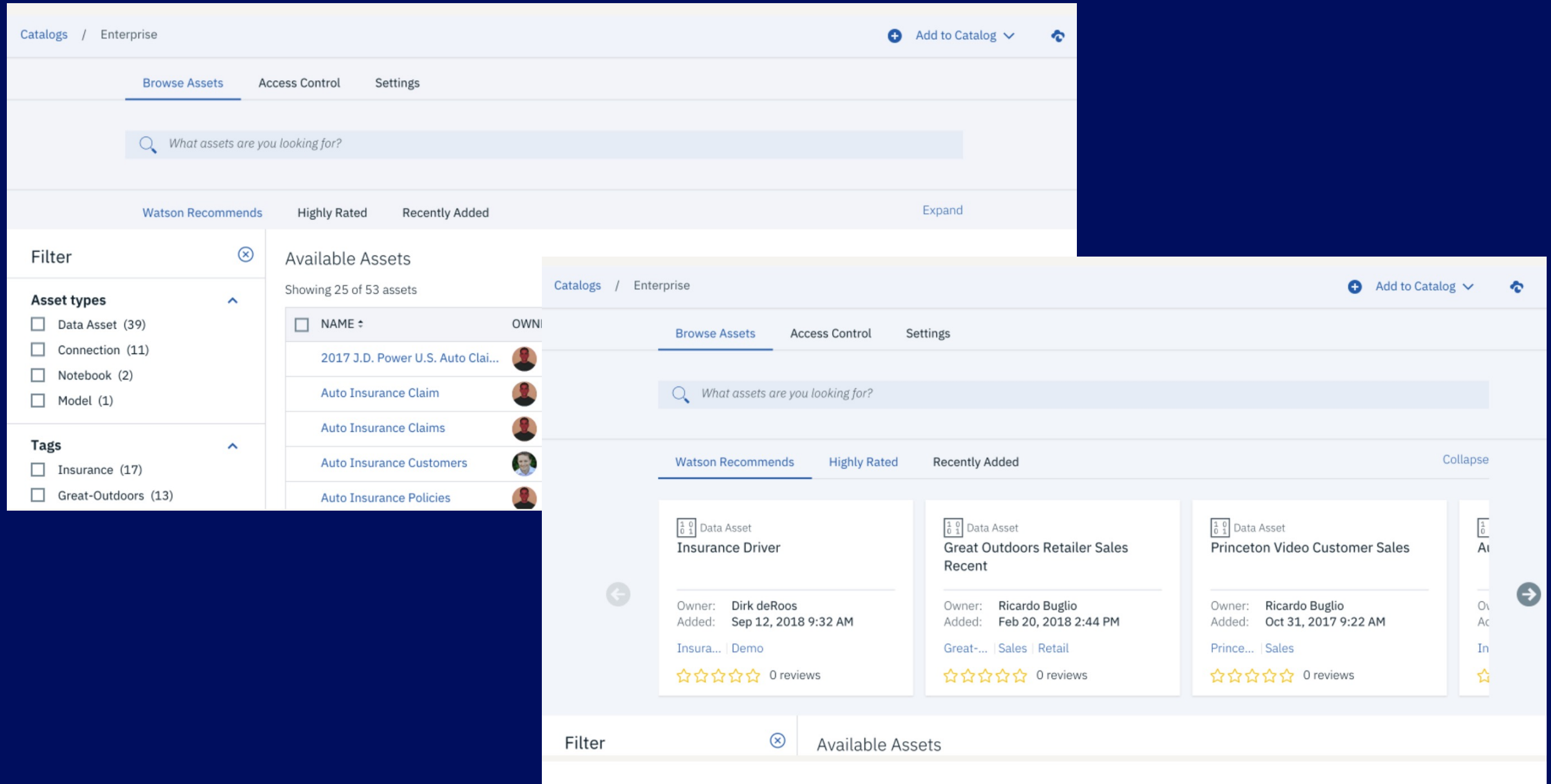


**Users**

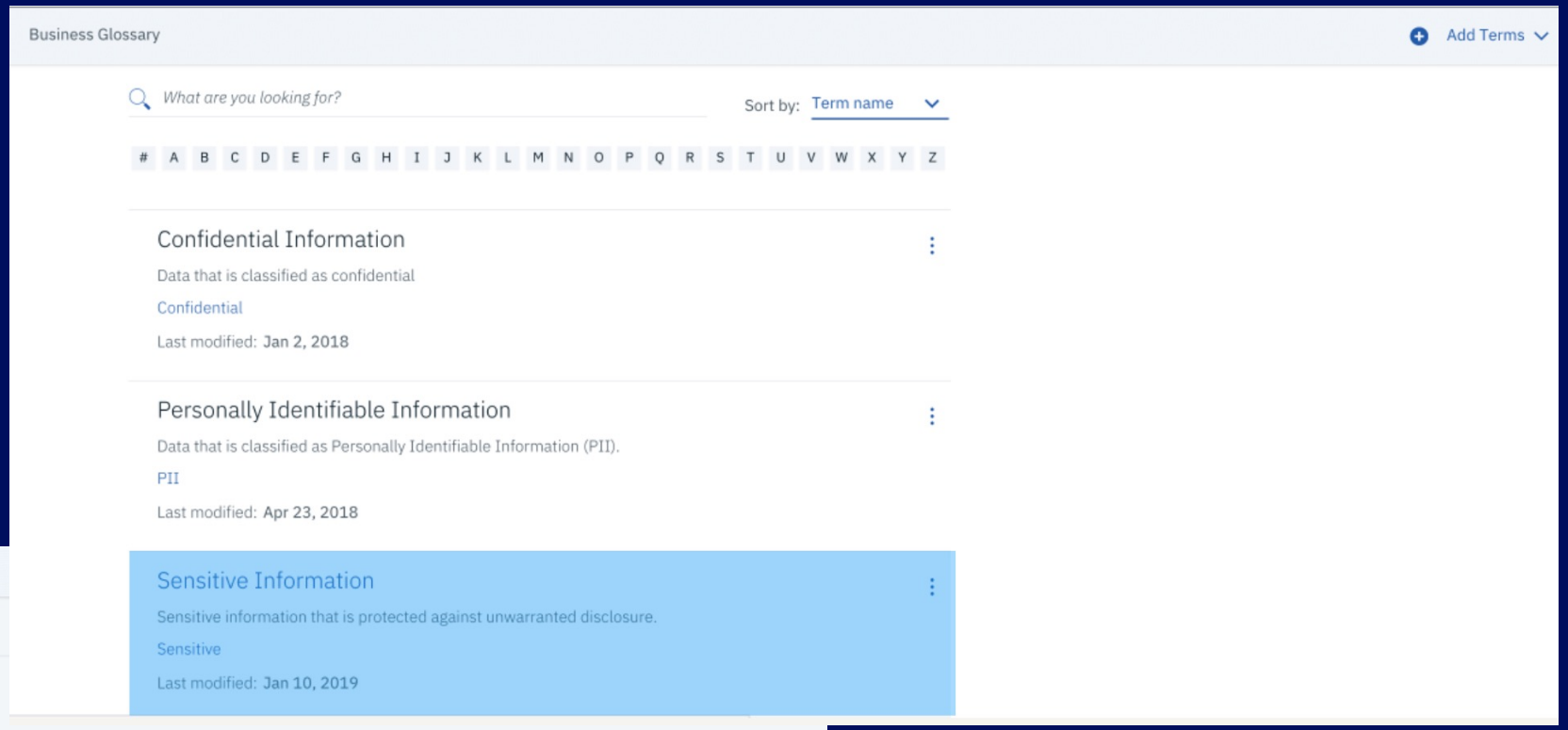
- CDO
- Data Quality Analysts
- Data Stewards/Architects
- Business Analysts
- Data Engineer
- Data Scientists

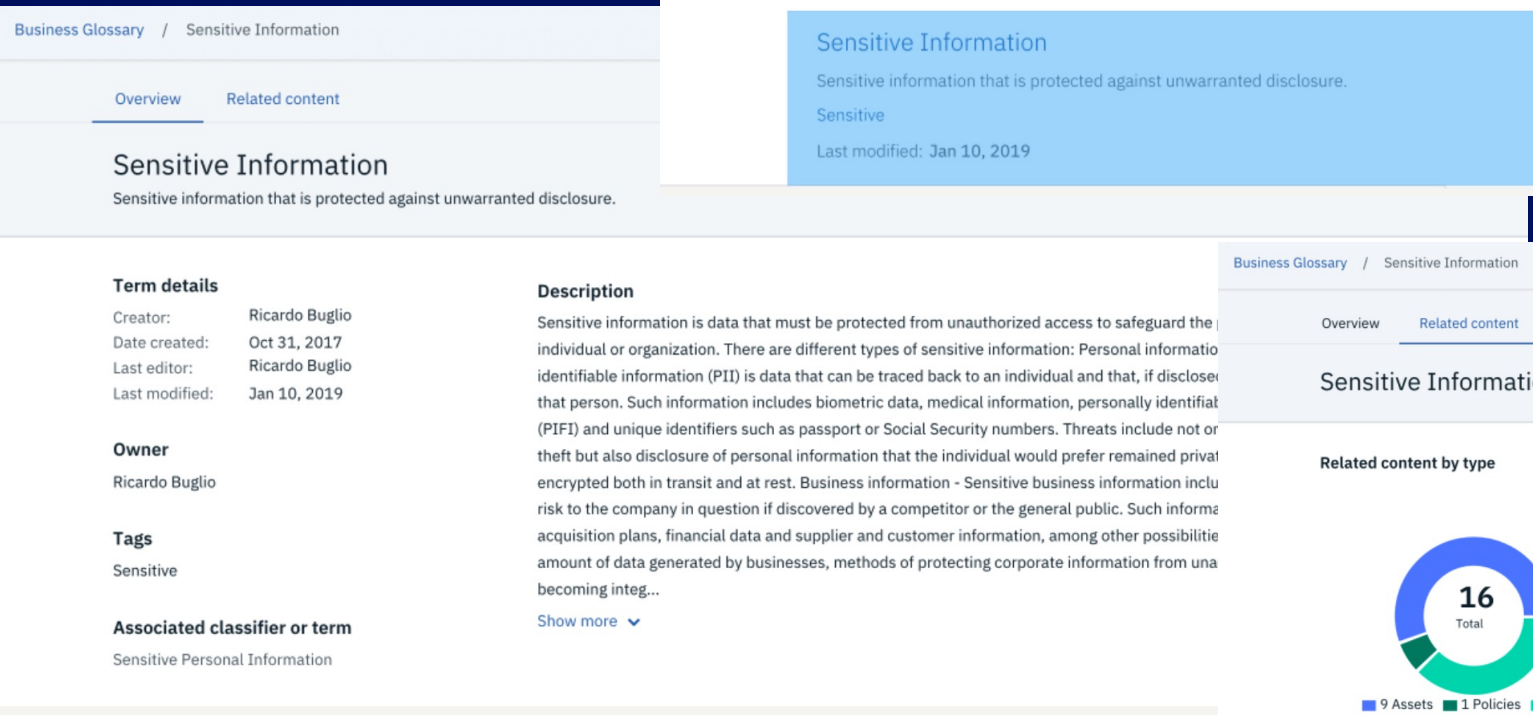
IBM **Cloud**

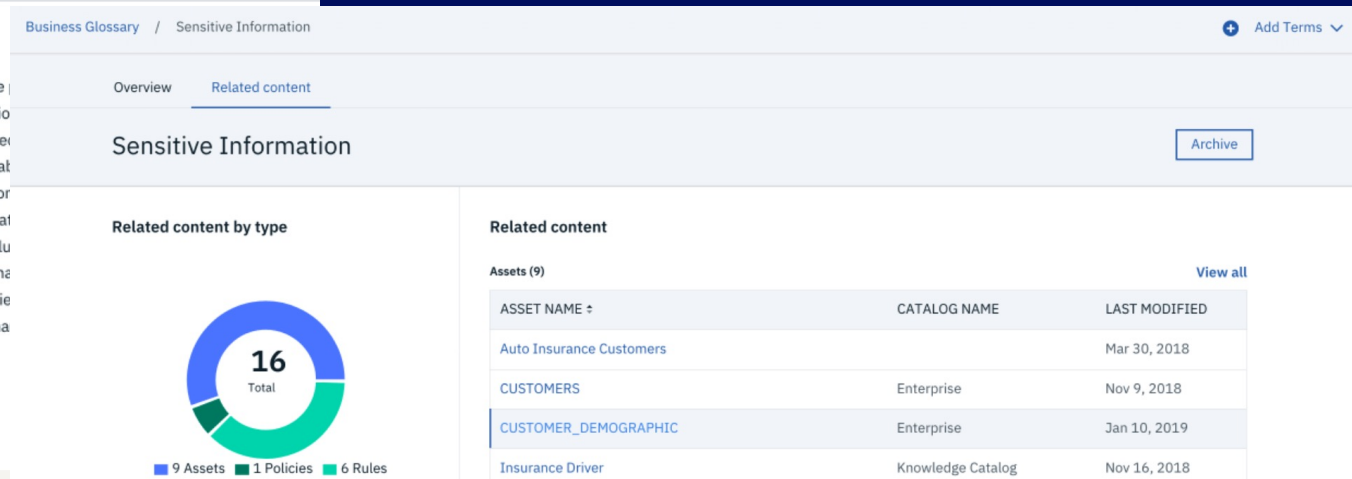
# Watson Knowledge Catalog:  Search and Explore



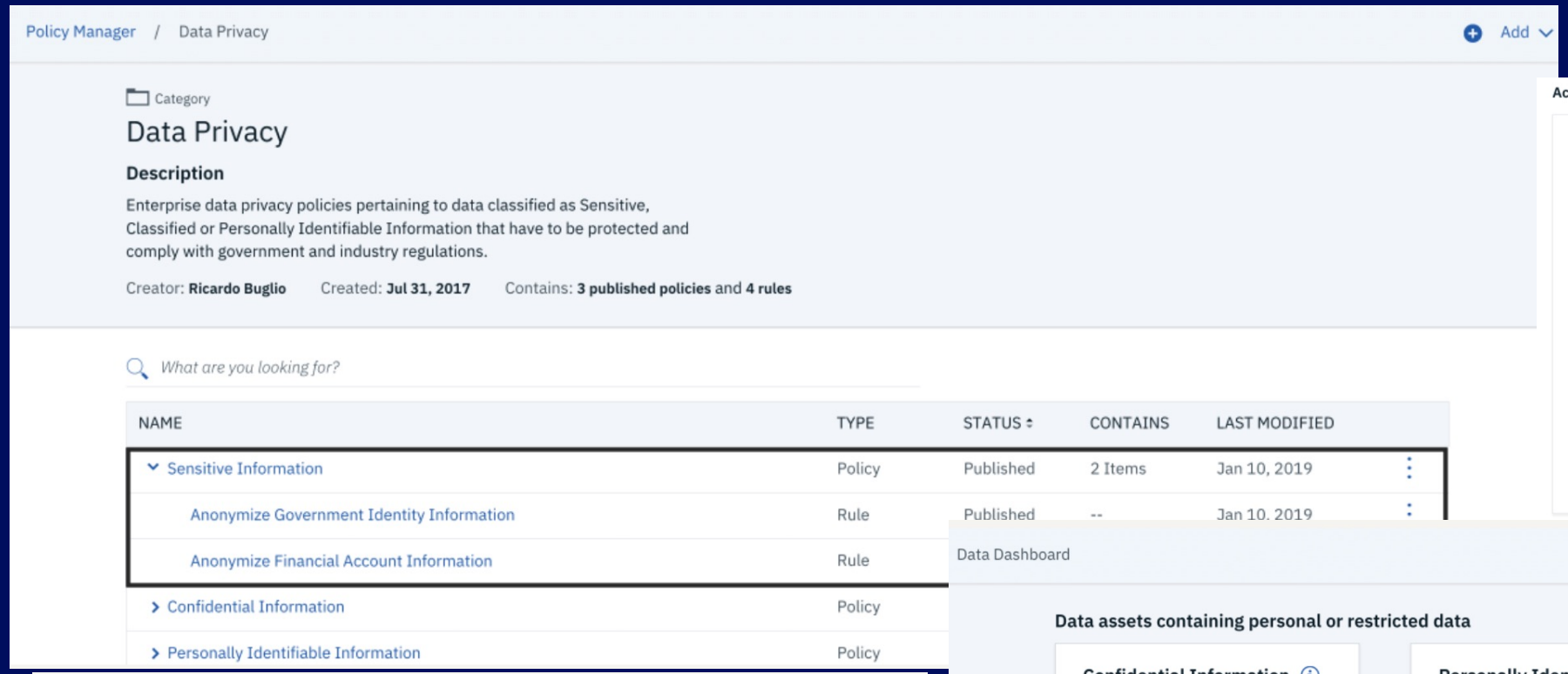
# Watson Knowledge Catalog:  Glossary Terms

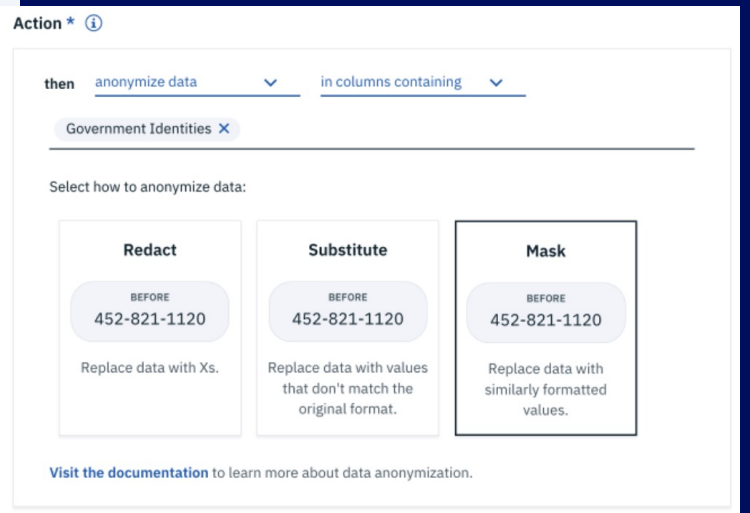


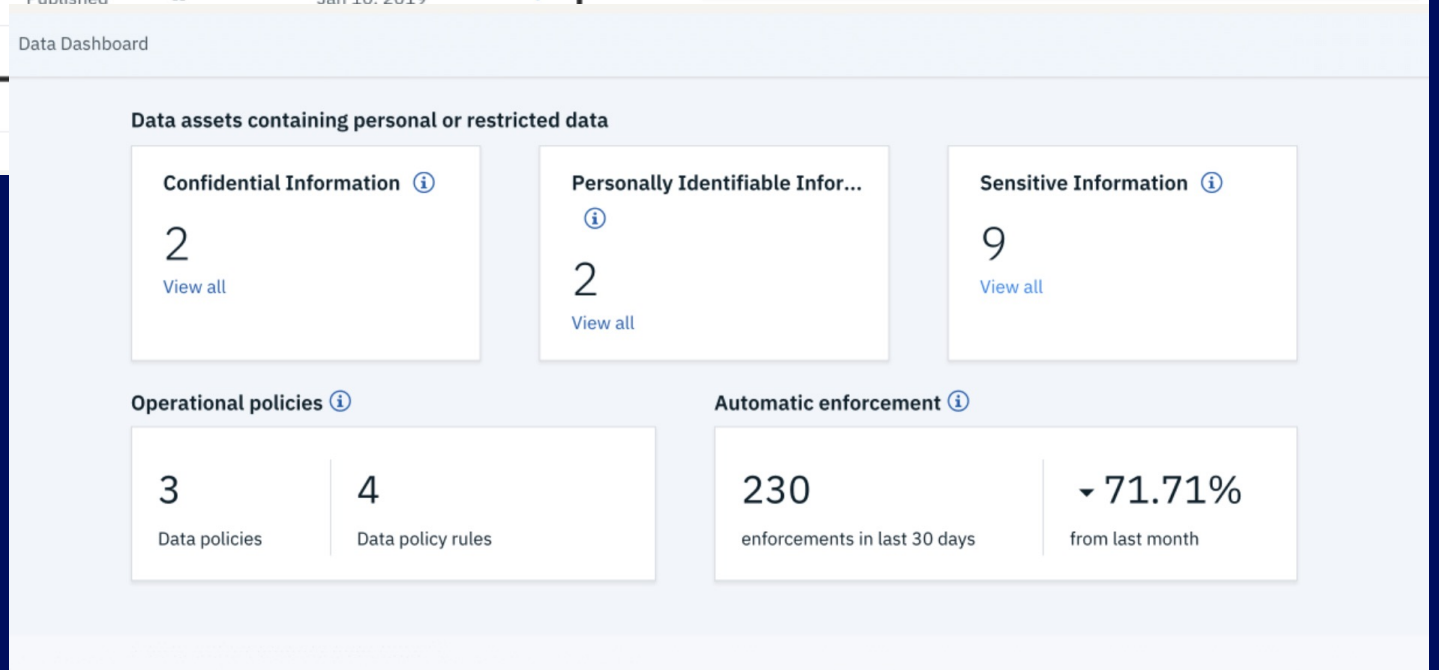



# Watson Knowledge Catalog:  Data Rules and Policies

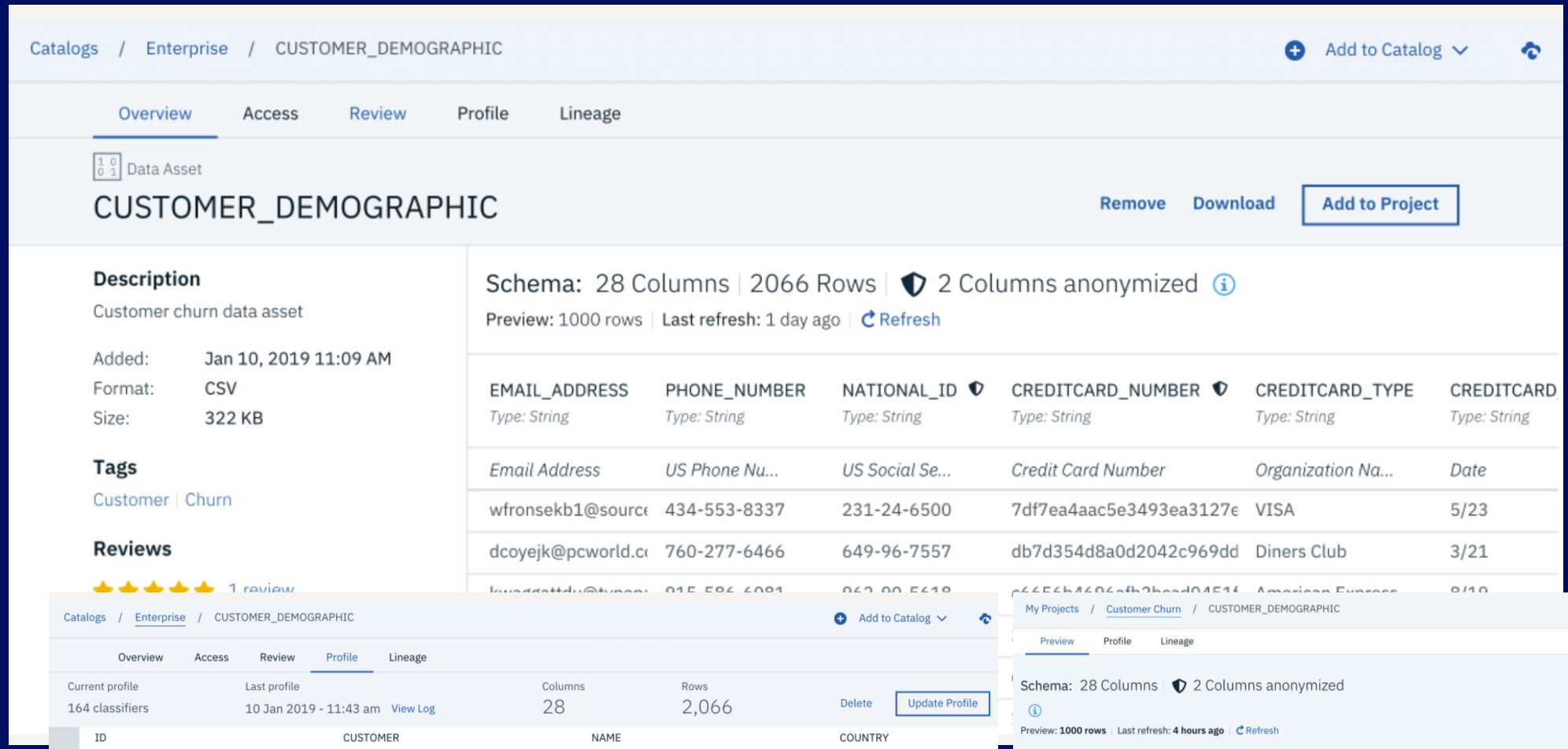


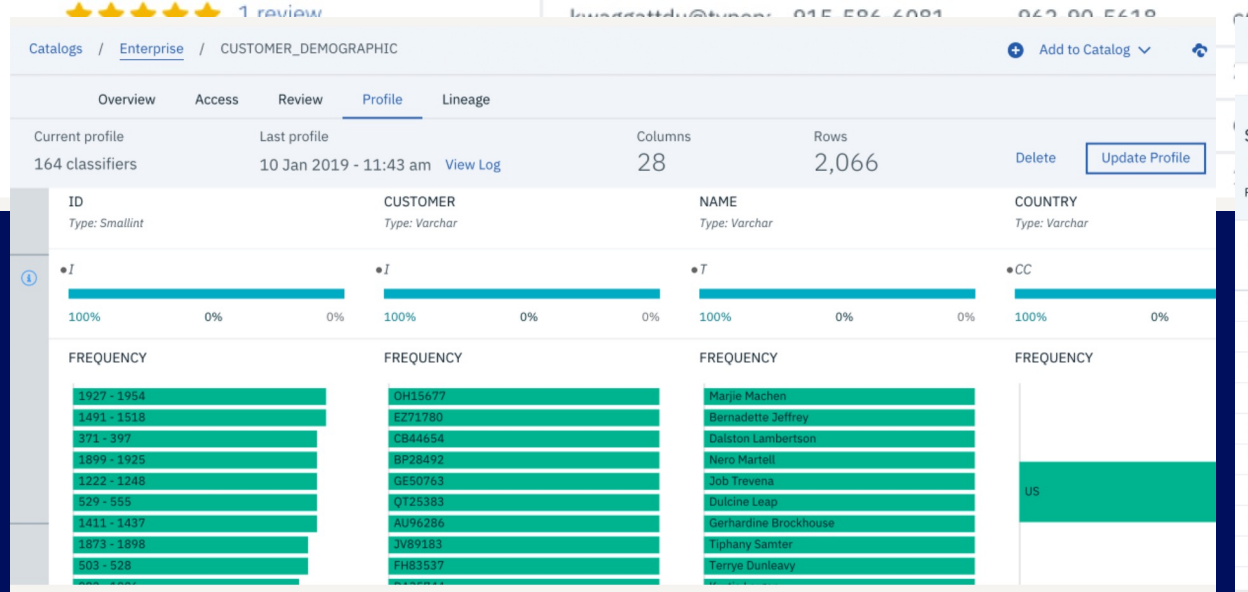



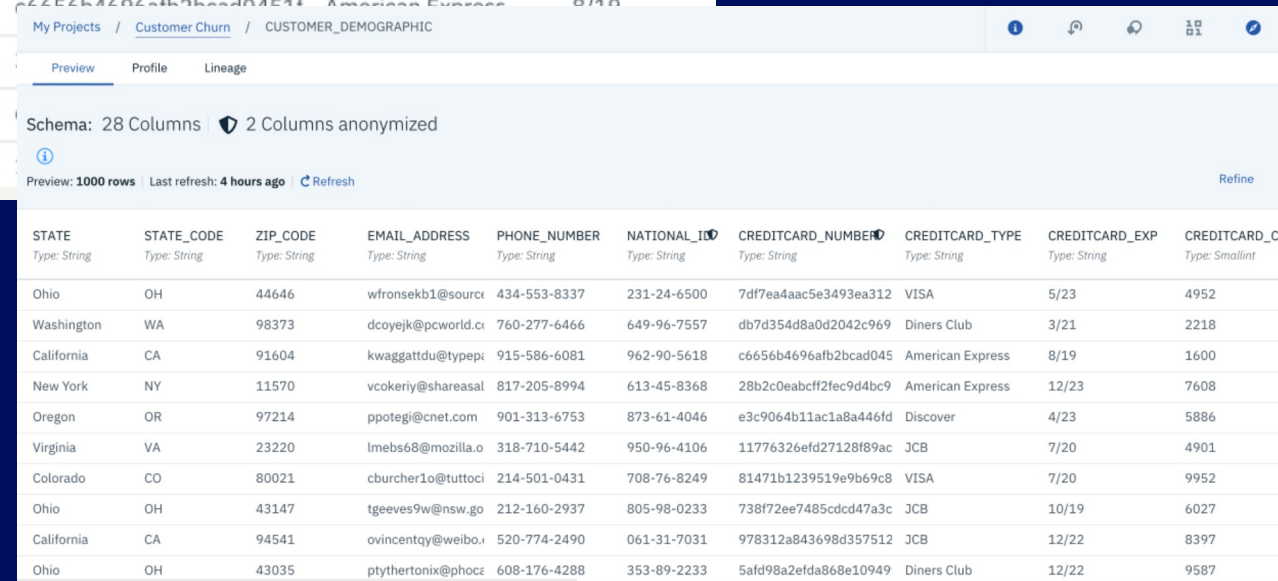




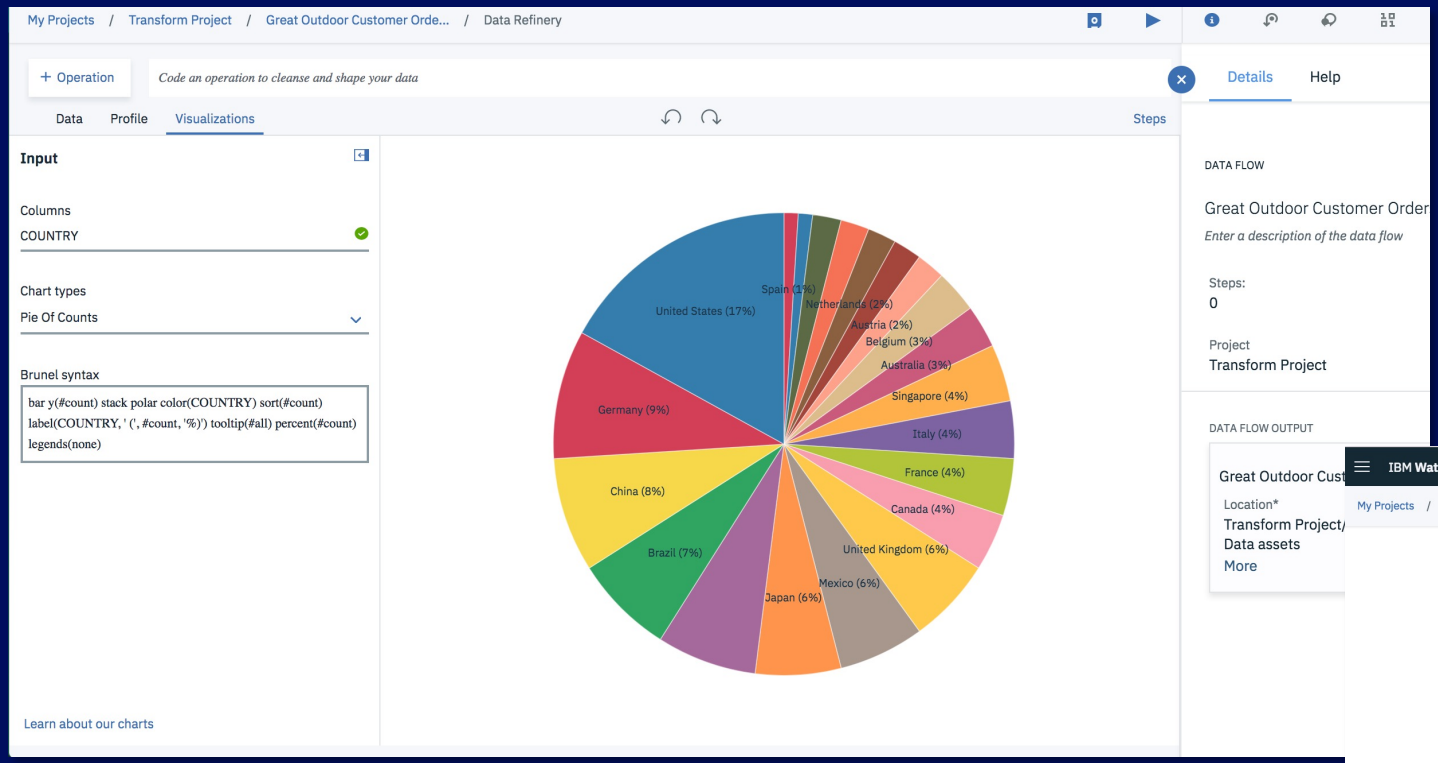
# Watson Knowledge Catalog: Review and Understand

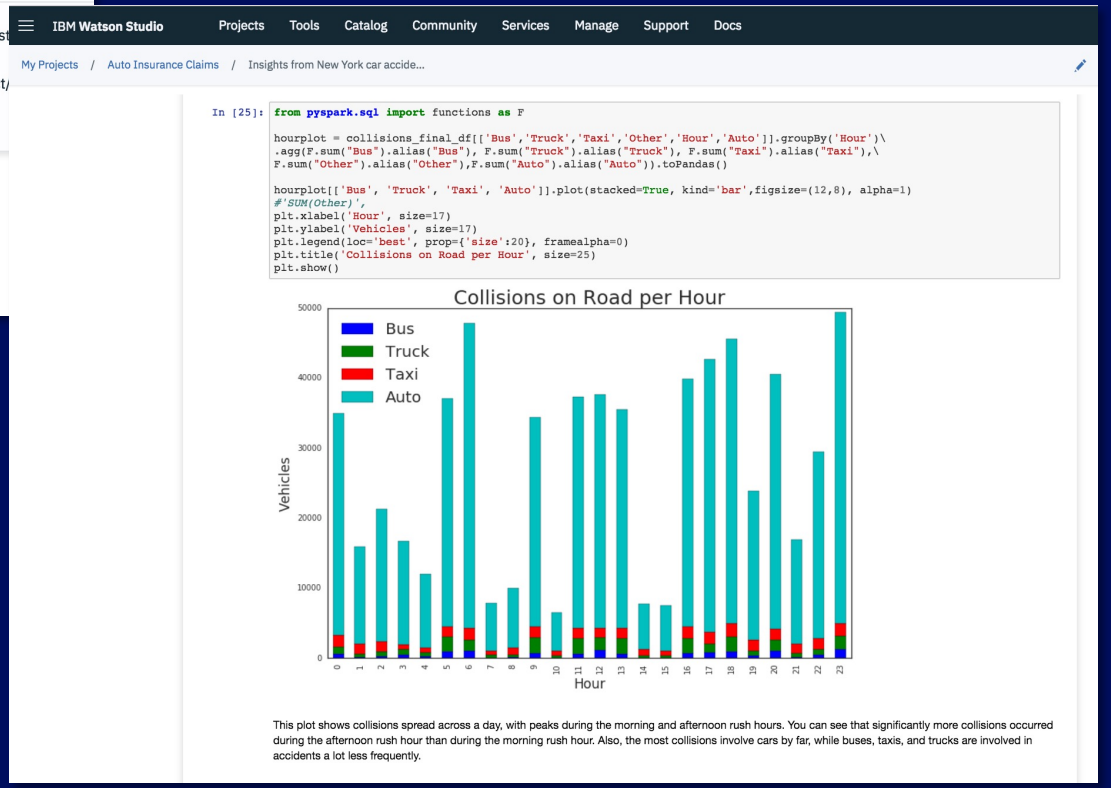


# Watson Knowledge Catalog:  Data Preparation & Data Science tools



# Business Problems Solved by WKC

**Problem #1: Inability to find the right data at the right time for reporting to the business and for creating models**
**Persona:** Data Scientist or Business Analyst
**Solution:** Knowledge Catalog indexes and classifies data assets which can be be quickly found by Intelligent search algorithms that can guide your data scientists and BAs to find the best data for their purpose. Once found they can move/share the data into their AI / Analytics projects.
**Value** :Improve the findability and use of data assets for business analysis or model development

**Problem #2: Inability to easily inventory data assets, document it and apply governance rules to protect data**
**Persona:** Data Steward
**Solution:** Knowledge Catalog provides an easy way to build and index all assets across your business functions & then enforce access and anonymization using policies & rules.
**Value:** Intelligent data curation helps curate data faster and governed and protected data is made available to the business users to trust and use it with confidence.

**Problem #3: Information locked away in department silos**
**Persona:** Data Scientist or Business Analyst
**Solution:** Knowledge Catalog makes it easy for even unstructured knowledge to be shared and accessible across the enterprise. ‘Auto Profiling’ and ‘data connections’ to 60+ sources makes it easy to catalog and organize. Cataloging of data creates a ‘shop for data’ online shopping experience experience.
**Value:** Find and use data through self-service access to the data

**Problem #4: Inability to deliver data efficiently and reliably to the business**
**Persona:** Data Engineer
**Solution**: Knowledge Catalog provides self-service access to data. The business users no longer need to wait on IT to get the data they need and the Data Engineers can focus on the business critical tasks essential for business growth. CDO organization is directly in charge of the organization’s data mine.
**Value:** Quickly deliver data as and when new data sources are added to the data pipeline.

**Problem #5: Business and Software teams not in sync in a project**
**Persona:** Application Developer / Business Analyst
**Solution:** IBM Knowledge Catalog provides an intelligent business terms and tags that connects business with software / IT orgs so, that all are on the same page.
**Value:** Commonality between business and software teams.

# Key Benefits

## Augmented Data Management

Automate the collection, curation and understanding of your data. Detect business entities, quality, bias and validity of you data to build an automatic level trust in your data.

***Automate the mundane***

## Data Prep & AI-powered Findability

Utilize automatic suggestions on how data can be best prepared for use. Intelligent suggestions to guide users to the best data for faster, smarter analytics and AI.

***Use the data to get the data***

## Policy Activation for Data Governance

Autonomous enforcement of data and AI governance policies, providing automatic decisions to mask and protect data for a particular purpose.

***Activate your governance program***

## AI Governance

Support your adoption and roll out of trusted AI. Ensure that your models remain compliant for quality and bias.

***Govern AI across your enterprise***

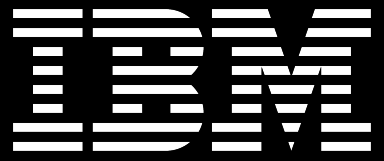
# 3 Things you can do now to learn more

1. Visit: https://www.ibm.com/cloud/watson-knowledge-catalog

2. Garage Materials to learn quickly:
   https://www.ibm.com/cloud/garage/dte/tutorial/fuel-data-innovation-and-ai-ibmr-watson-knowledge-catalog
   https://www.ibm.com/cloud/garage/dte/producttour/ibm-watson-knowledge-catalog-guided-demo-explore-active-policy-management

3. Try WKC for free:
   https://dataplatform.cloud.ibm.com/registration/stepone?apps=data_catalog&context=wdp

# Q&A



VISION

Public

Private

On-Prem

# IBM Watson Knowledge Catalog

## One Unified Asset Catalog

One single, seamless catalog experience that serves different personas on public, private and on-prem

All Governance and Integration capabilities (UGI) converged with WKC

**Claire** - Chief Data Officer

**Dominik** - Data Steward

**Quan** - Data Quality Analyst

**Daniel** - Data Engineer

**Emmet** - System Admin

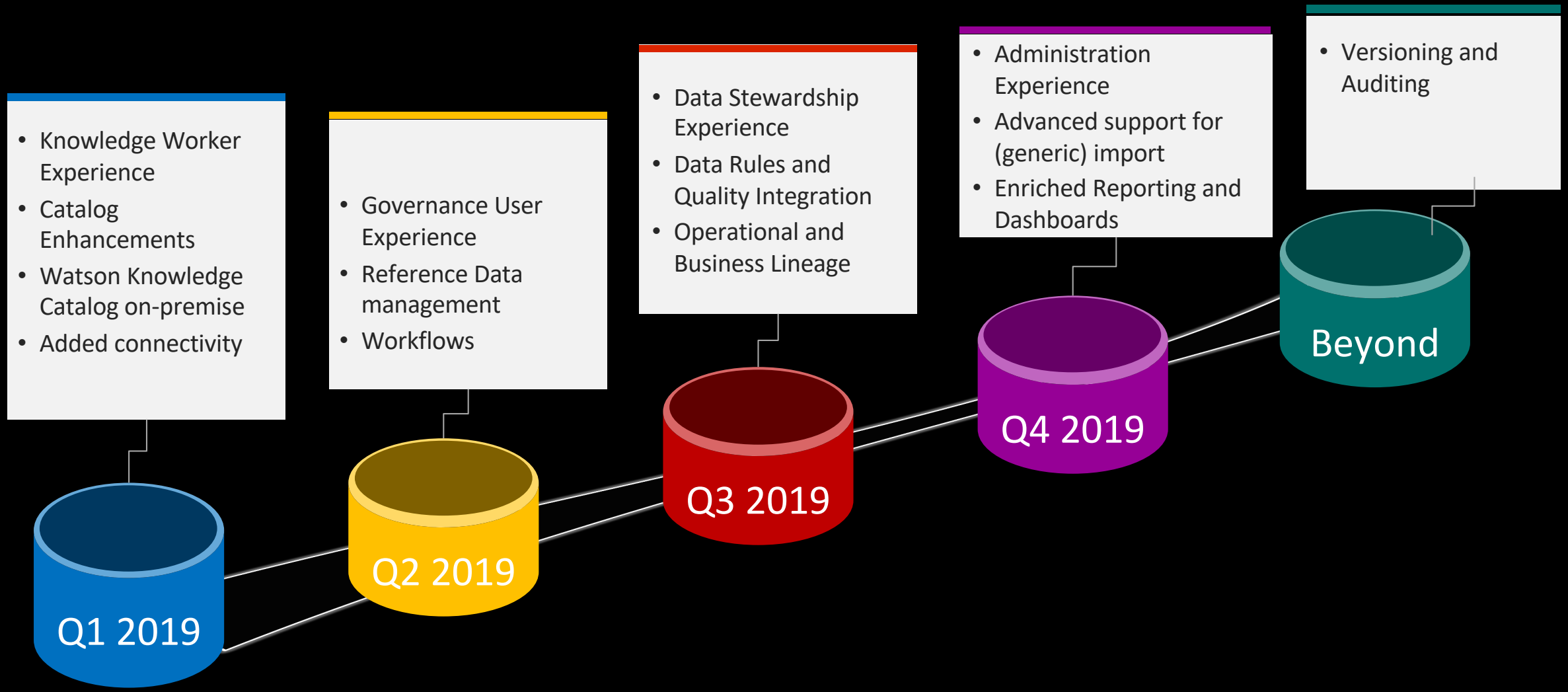
**Betty** - Business Analyst

**Chris** - Data Scientist

# IBM Watson Knowledge Catalog

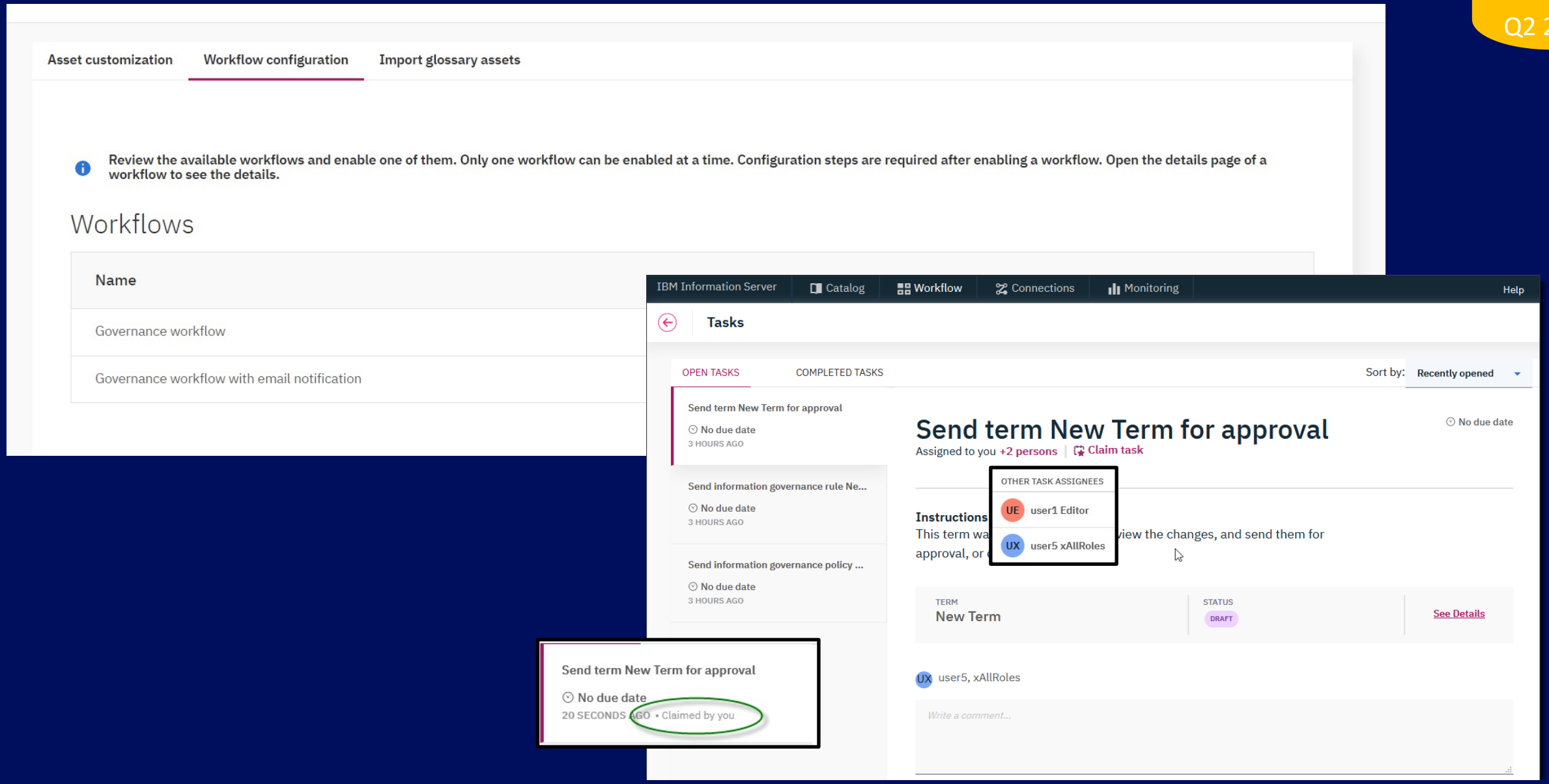
**One unified catalog**

2019 Roadmap and Strategic Vision



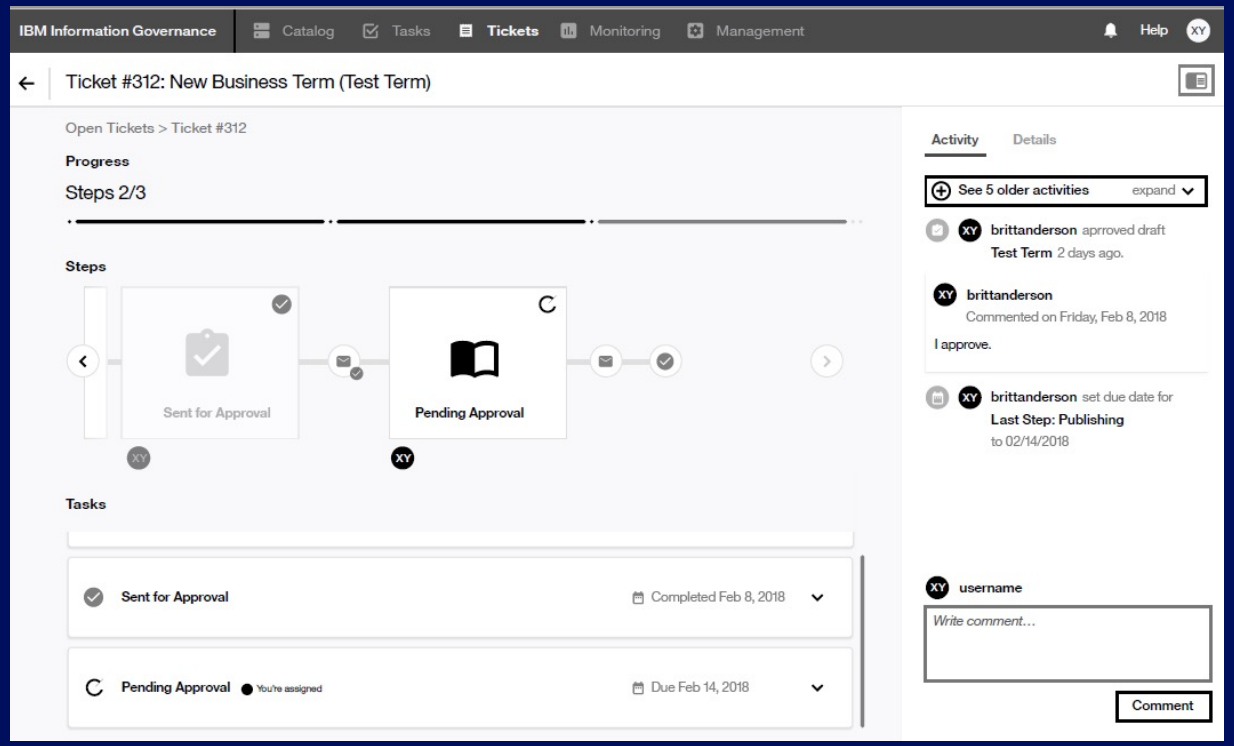
# Advanced Workflow

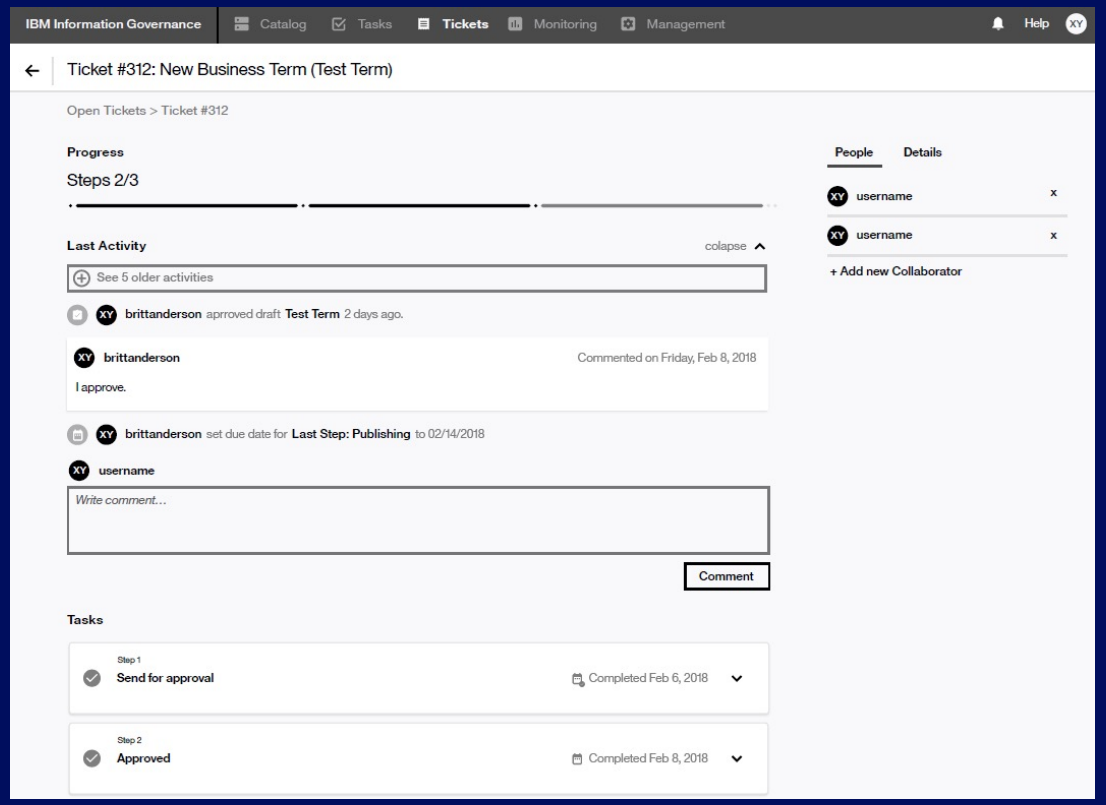
Q2 2019



# Advanced Workflow

Q2 2019



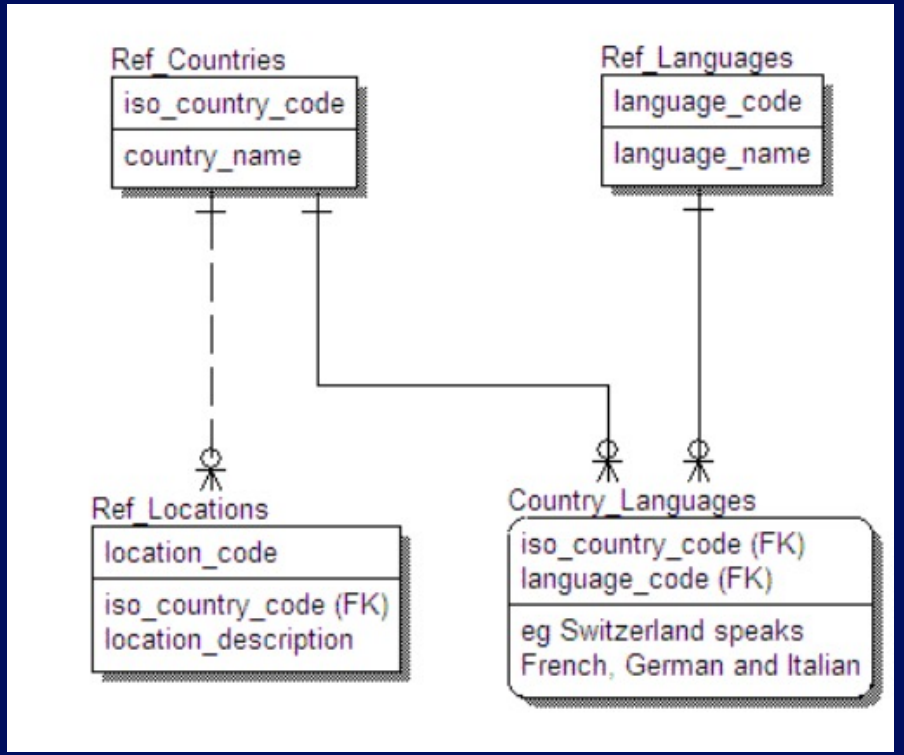
# Reference Data


Support the ability to capture, manage and exploit *Reference Data* across our Unified Governance & Integration portfolio and extend value across other IBM offering such as MDM or Guardiam. Reference Data serves the purpose of recording and making known the permissible values of an Entity or Concept, and further aligning such values to enrich, develop and identify Glossary, Policy or Classification objects or as support for Data Integration projects.

**Reference Data Set:** Set of Reference Data Values that define the permissible values to be used within other constructs, e.g. Database Columns

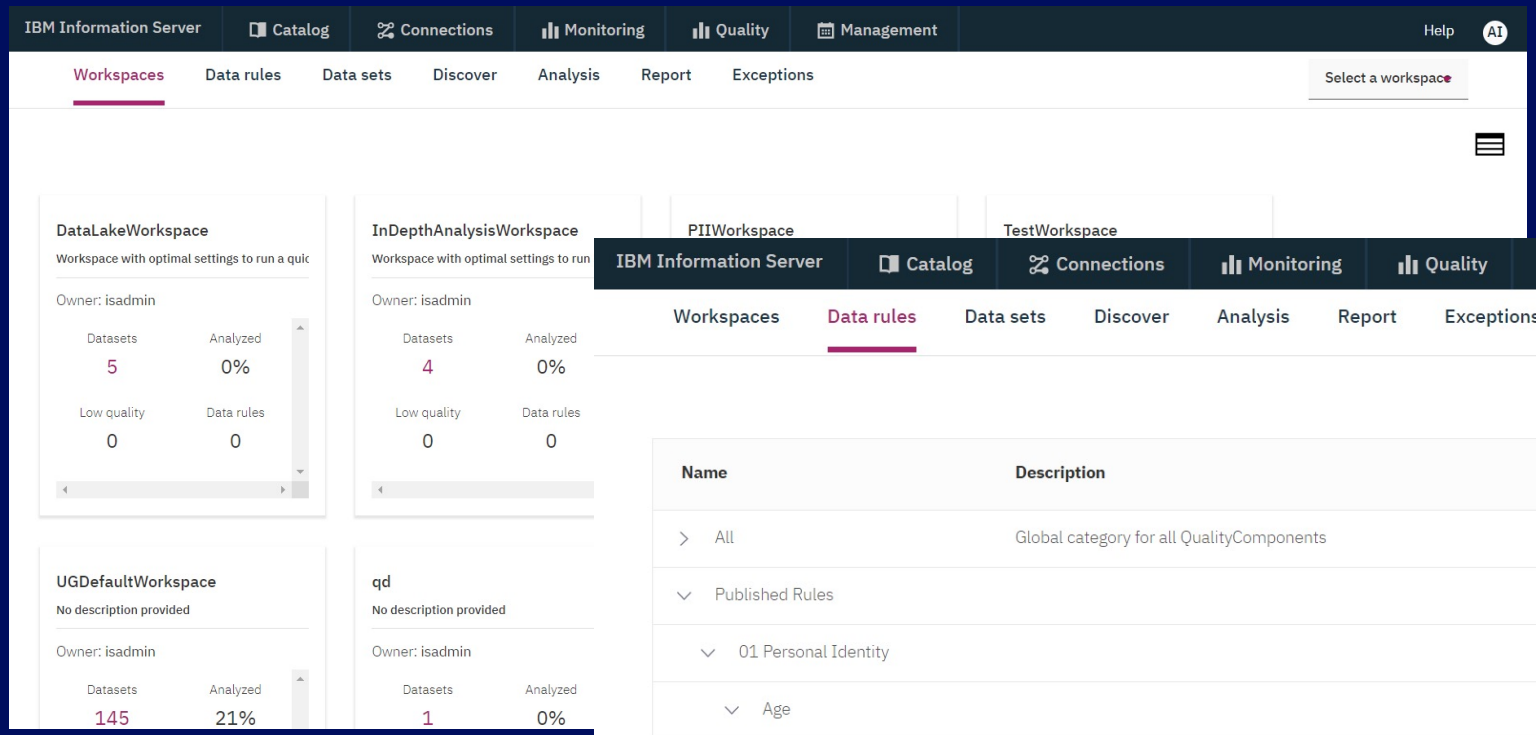
**Reference Data Value:** Individual permissible value contained within a Reference Data Set

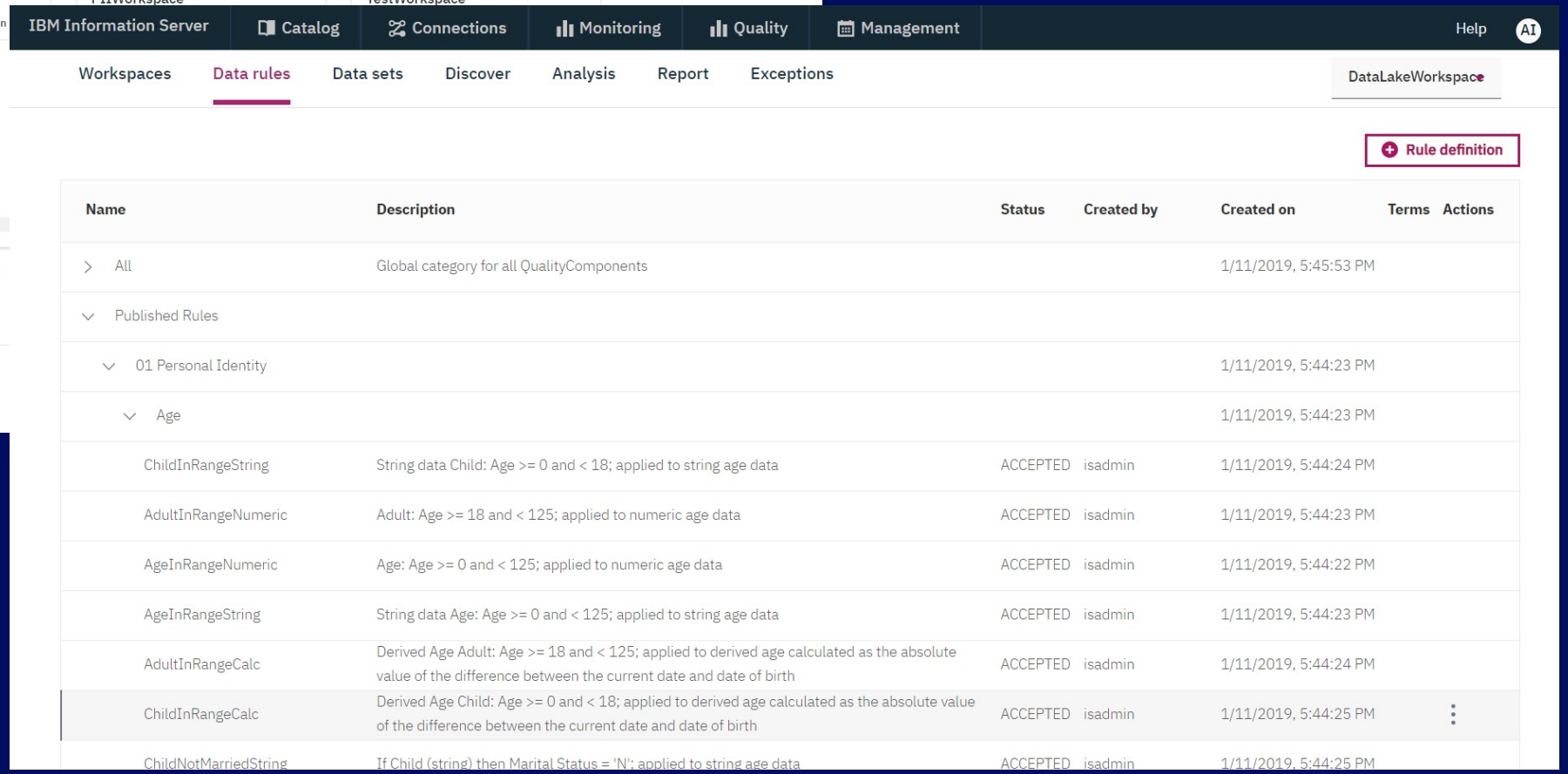
**Reference Data Mapping:** Mapping between individual Reference Data Values belonging to different Reference Data Sets, e.g. mapping between Country Code Value of *DE* and Country Code Name of *Germany* within different Reference Data Sets



# Data Rule Definitions

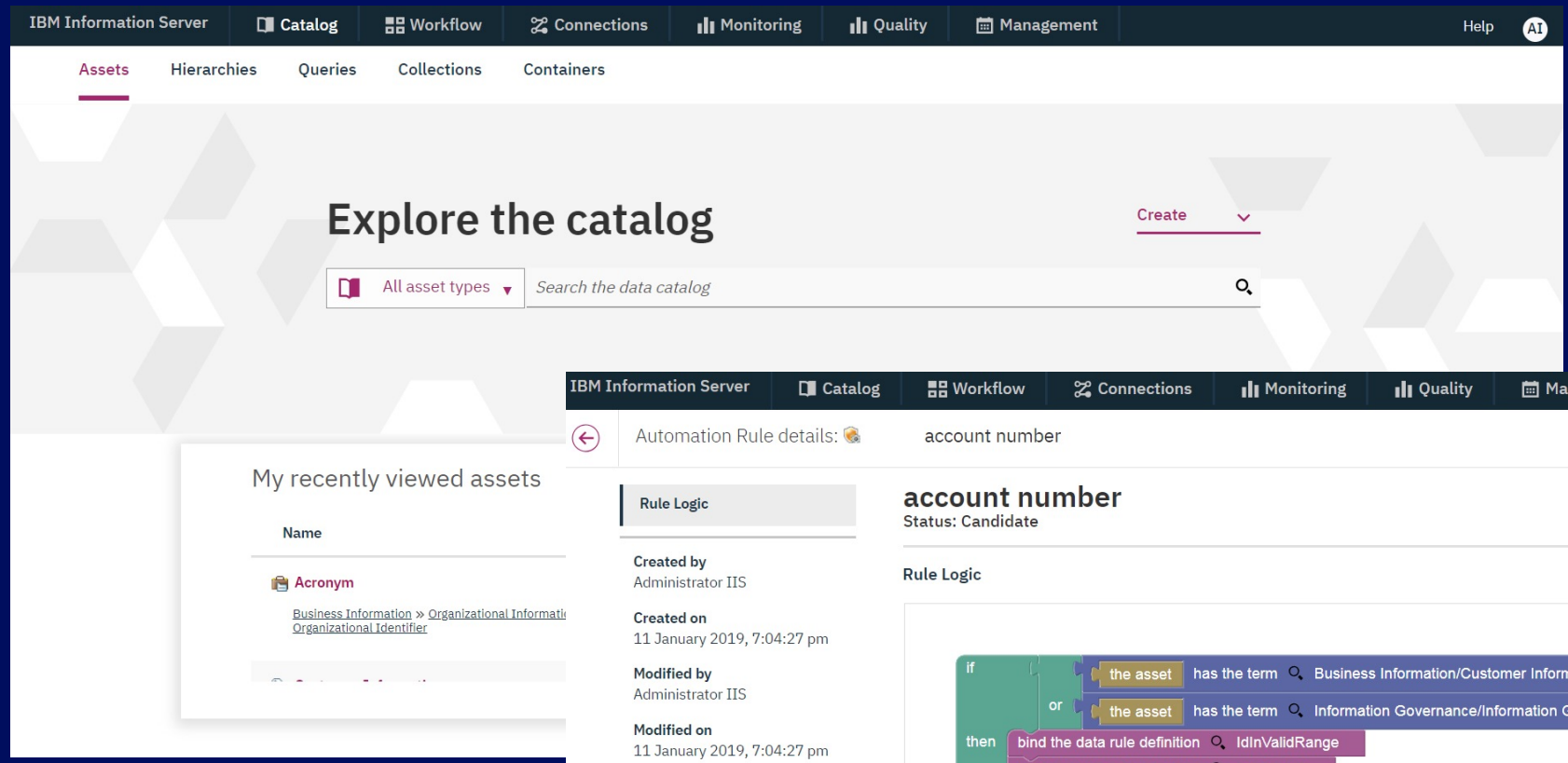
Q3 2019

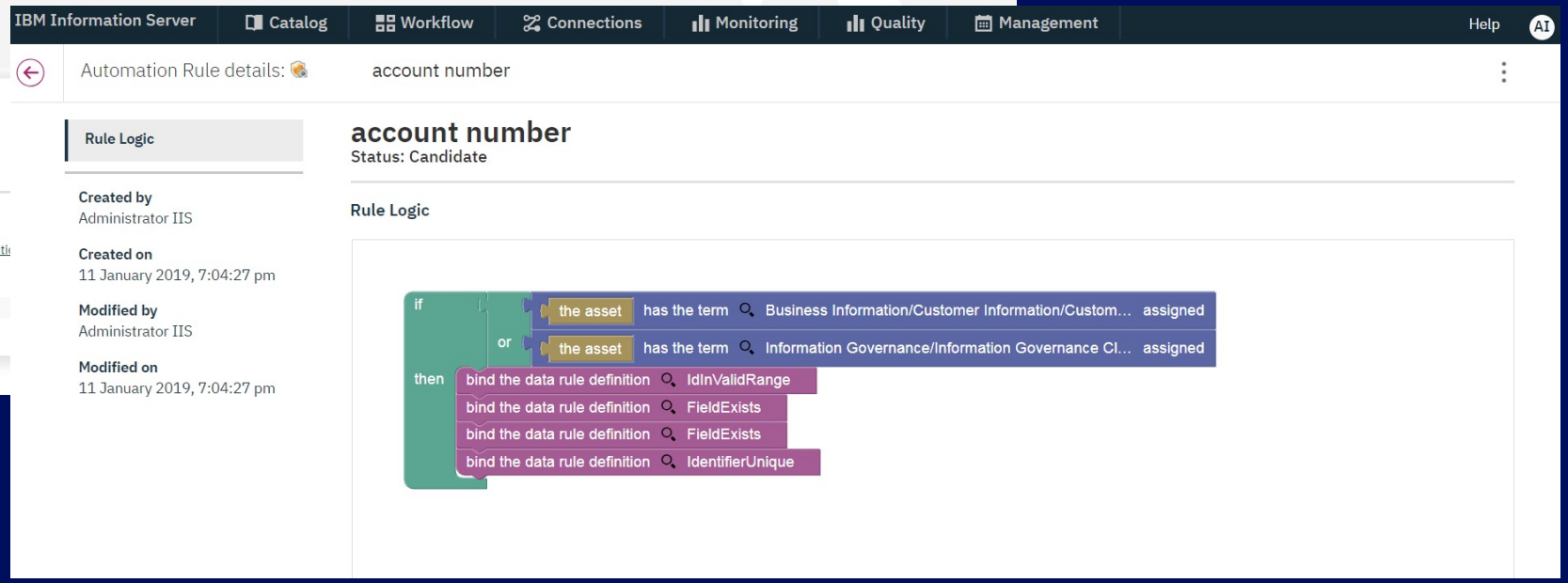


# Workspace and Analysis

Q3 2019



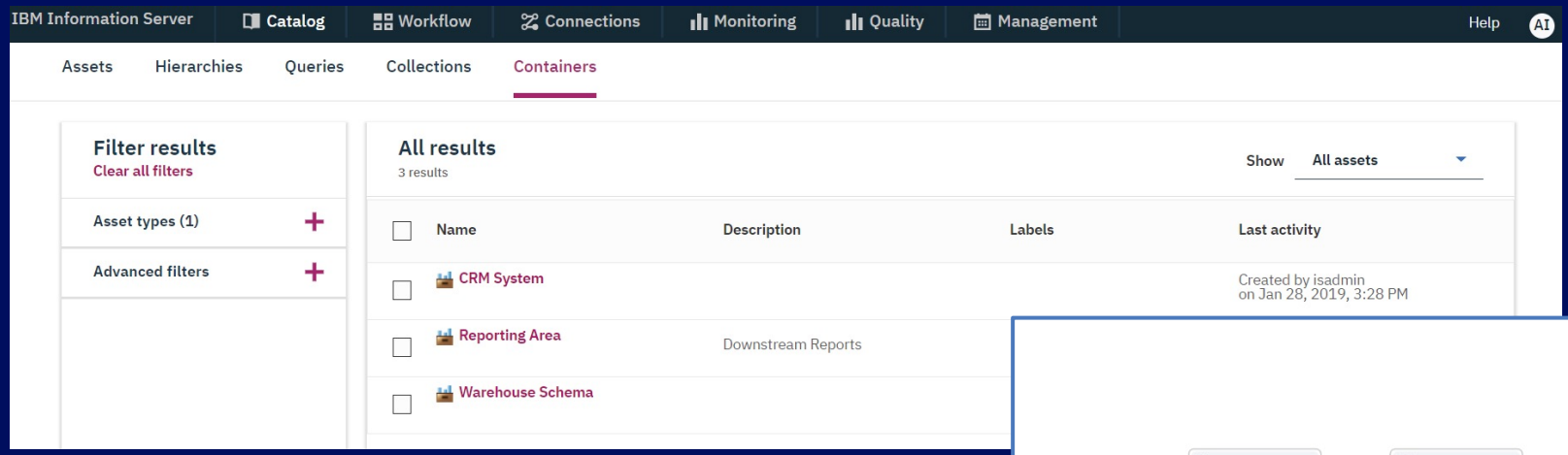



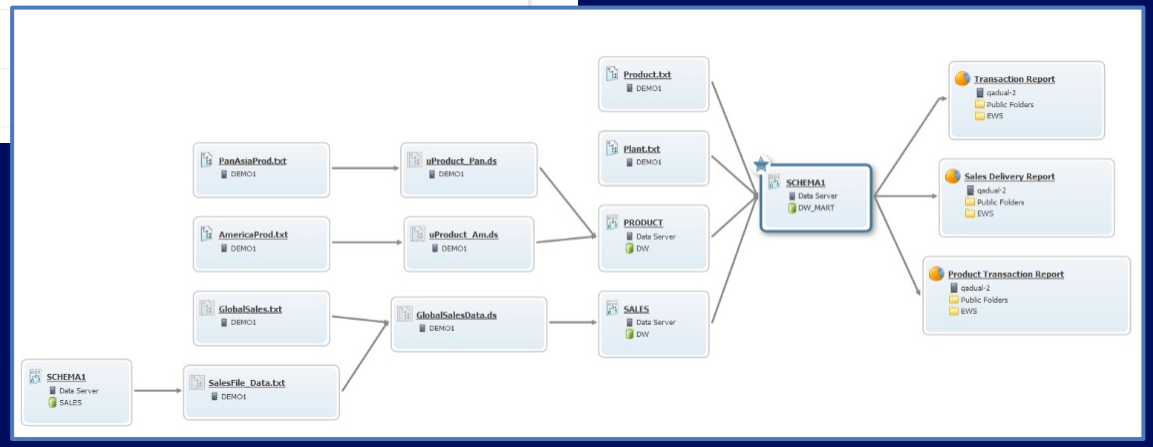
# Business Lineage


**Business Lineage** is not consumable by the business. Business Lineage reflects the data flow and intersection of Data Assets, referencing such Assets according to their technical descriptors and name. Business Lineage is viewed as a filtered Data Lineage report, void of Mapping Documents and ETL Jobs.

**Business Users** want to view Information and the flow of Information across Systems, Projects and Domains, and would like a simplified view of Business Lineage in the language of the Business, and not according to a Development or IT view.

Support the business user audience and their need to display and support Business Lineage in a manner that is understood and consumable.



# Catalog Permission

Q3 2019

Support the ability to restrict the visibility of information according to user and their organizational or business requirement.

Provide a simple administration console, to apply permissions according to existing context or domain information to an individual or group of users.