Digital Transformation with Data Fabric – Watson Query/ Data Virtualization

1. Overviews
2. Technical Details
3. Typical Use Case Patterns

Piotr Mierzejewski
Director Db2 & BigData
IBM Data and AI

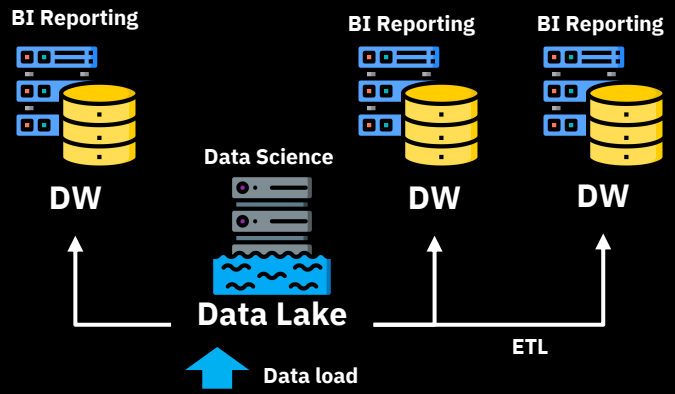
Goal: Be able to expose any or all of organization’s critical data assets no matter the location, platform, technology, or schema.

Query 2 or 2,000 data systems with a single query as One

The Vision of IBM Watson Query

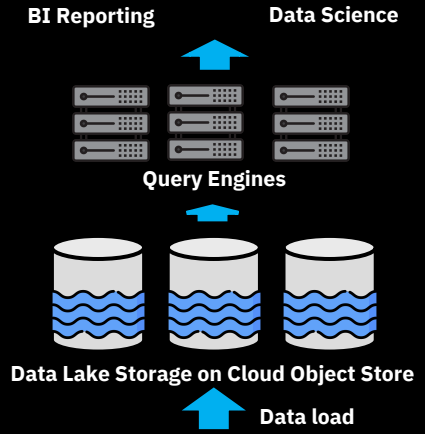
- Ultimately, the goal of Watson Query is to provide the data management experience in Cloud Pak for Data
- Specifically, with a focus on Big Data and how it has evolved in recent years

Classic Big Data Architecture



- Requires many complex data pipelines
- Difficult to scale and maintain
- Duplicates of data
- Isolated BI and Data Science environments

Modern Big Data Architecture



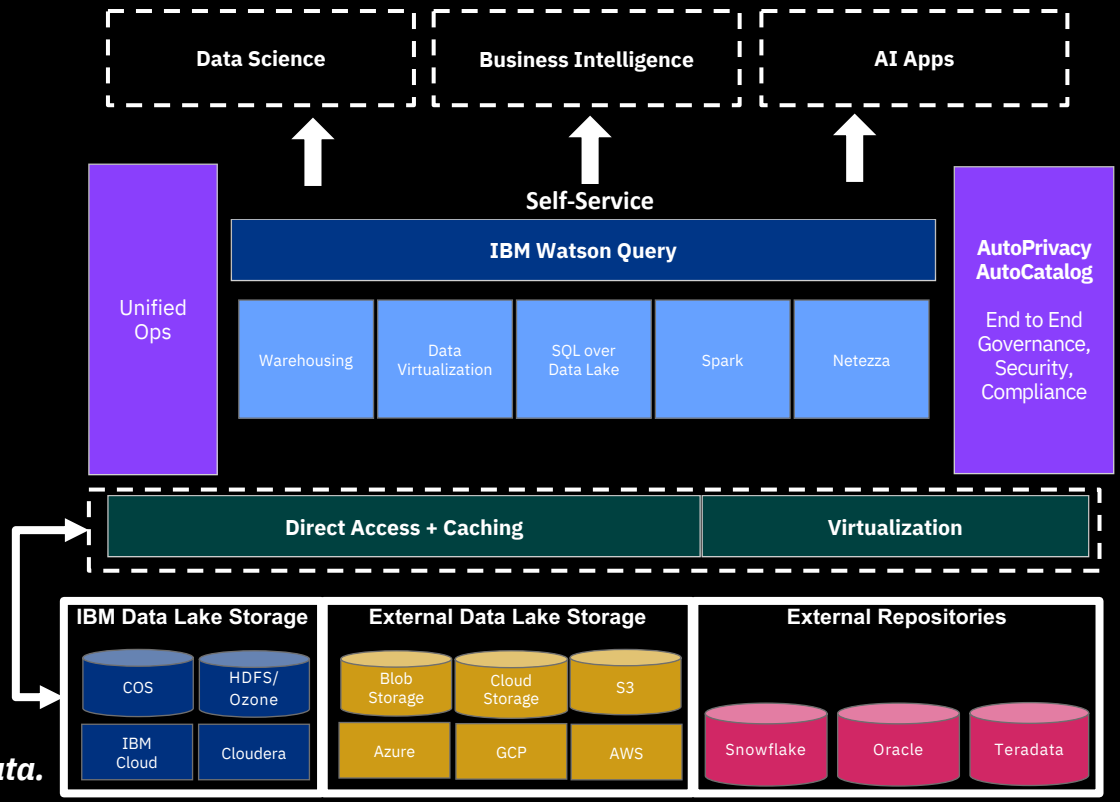
- Scale compute and storage separately
- Handles unstructured and structured data
- Minimal data movement and pipeline
- Single environment for BI and Data Science

IBM Watson Query

A universal query layer that automates how you access, update and unify data across any source or type (clouds, warehouses, lakes, etc.) without the need for data movement or replication. With intelligent performance optimizations, peta-byte scale and visual query building experiences, it abstracts the complexity of multiple query engines to provide simplified self-service data across an organization

Distributed Query Experience

- **One Query Experience.** Multiple capabilities from virtualization to warehousing
- **End to End Integrated Governance** with AutoCatalog and AutoPrivacy
- **Open Data Formats** to work with data on any cloud and on-prem
- **Petabyte scale** landing, exploration, shaping, governance, and AI
- **Cloud-native** with **consumption pricing**, **instant-on**, and **available** across any zone
- **Fully elastic** with separate compute and storage
- Unified access **without data movement**
- **Hybrid cloud:** Available to deploy on multiple clouds (on prem, private, public)



Any data. Anywhere.

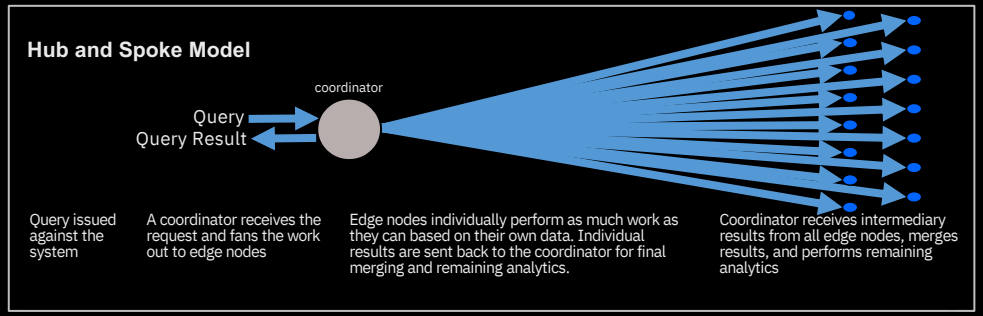
Key Architectural Differentiation

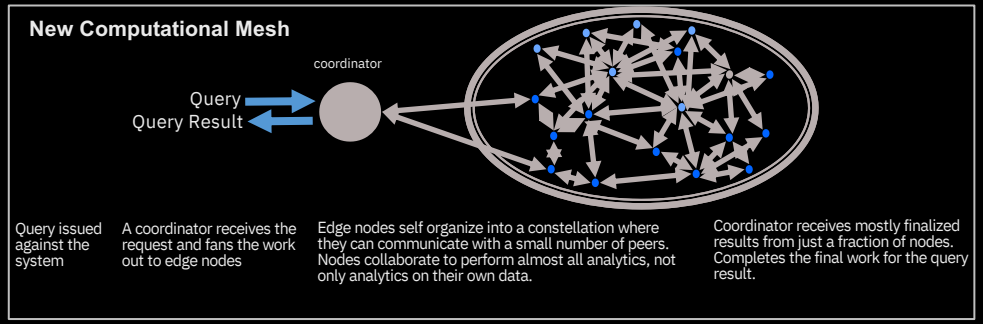
Hub and spoke execution models:

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

IBM is first to market with a parallel processing model:

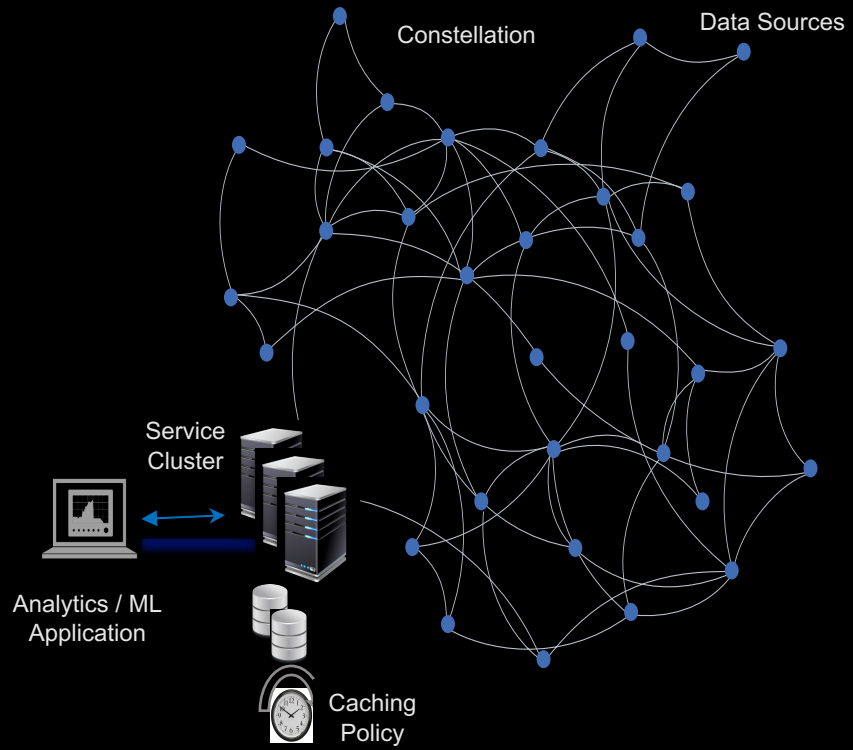
- Theoretically unlimited scalability
- Ease of addition/removal of sources
- Execution pushed down into the constellation mesh





Mesh: Remote Connectors and Data Discovery

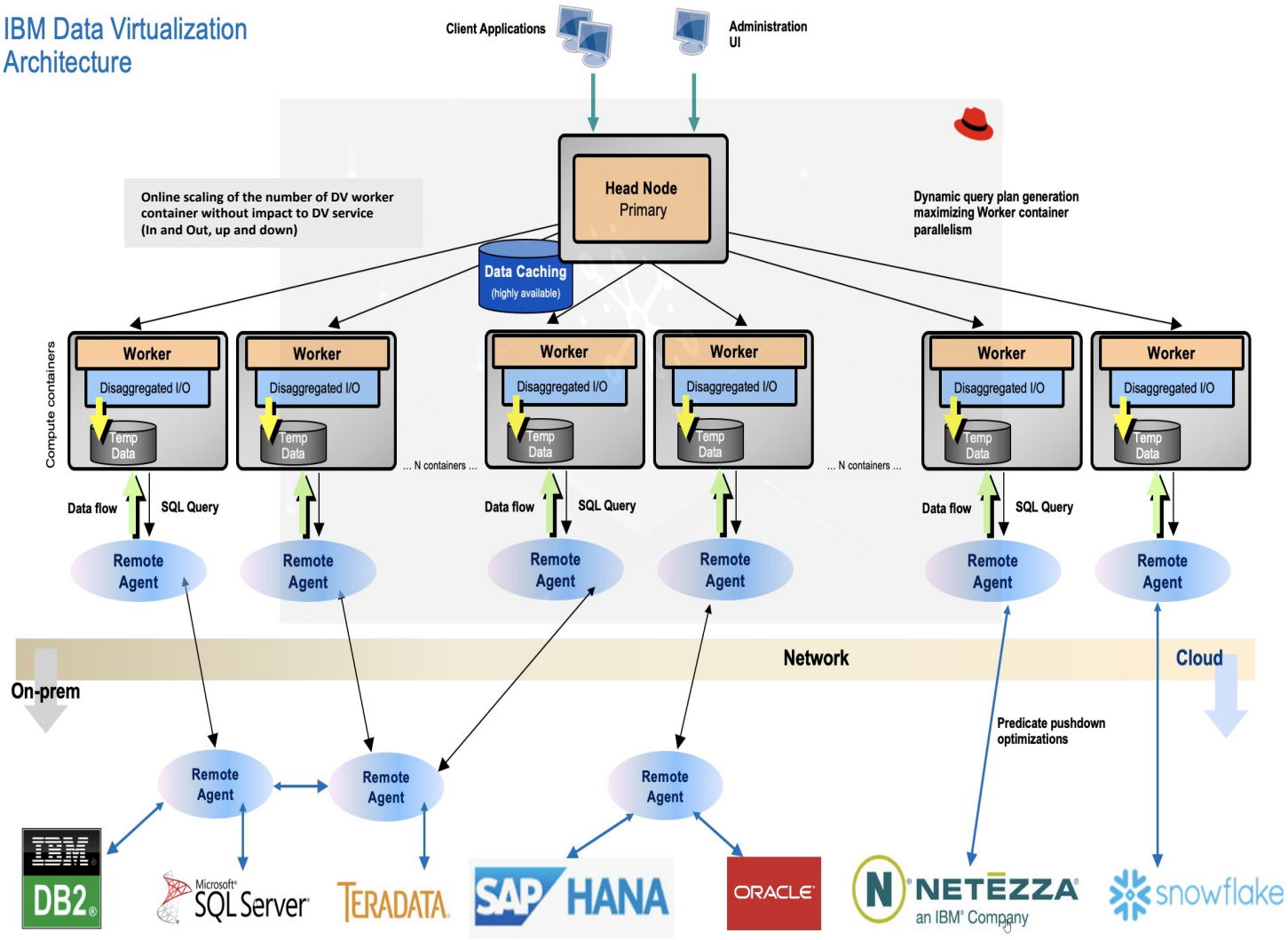
- Parallel processing mesh providing execution performance and scalability:
  - *Quickly deliver analytics results and easily evolve with new data source demands*
- Provides resilient connections between data sources:
  - *Reliability and ability to quickly adapt to increasing business demand*
- Scales 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*



Data Virtualization architecture provides multiple layers that can be tuned according to the workload.

- Multi-worker support in the cluster scales both horizontally and vertically to handle high concurrency and data volumes.
- DV remote agents scale to support higher number of connected data sources and multi-source queries.
- Optimizer dynamically chooses the best number of worker to use for portions of the query based on data sizes and workload

IBM Data Virtualization Architecture



Broad support for common data source types

More to be added in the future.

Cloud Pak for Data

DV 1H 2022, DV aaS MVP on 1H, 2022

- Db2 family for HDM
- Db2 for iSeries, zSeries
- Db2 for z/OS
- Big SQL
- IIAS, PDA (Netezza)
- Informix
- Amazon RDS (Oracle, SQLServer)
- Denodo (Data Virtualization as a source)
- 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)
- BigQuery
- SAP S4/HANA & BW (JDBC / OData connection)
- Amazon Redshift
- Salesforce
- Greenplum
- Apache Spark SQL
- COS / S3

In the roadmap pipeline

- TM1 (REST API)
- Apache Kafka
- Cassandra
- SAS
- JSON
- REST API
- TIBCO (Data Virtualization as a source)
- Interbase
- Apache Drill
- Amazon Dynamo DB, Aurora DB
- CouchDB
- Stream / MQ
- Cloudant
- DataBricks Delta Lake

*Driver not distributed
@Not available in DV aaS

Key Features

Collaborative Compute Model

Remote Processing and Discovery

Governance Integration

Caching

Schema Folding

Compute Scalability

Data and Result Caching

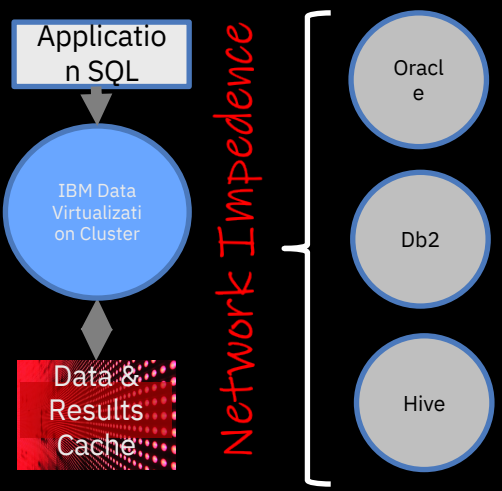
Automated cache recommendations to further optimize query performance-

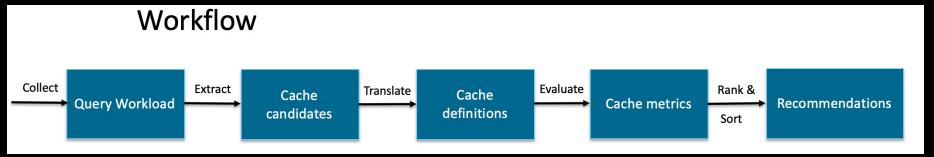
Powerful

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

Under the hood

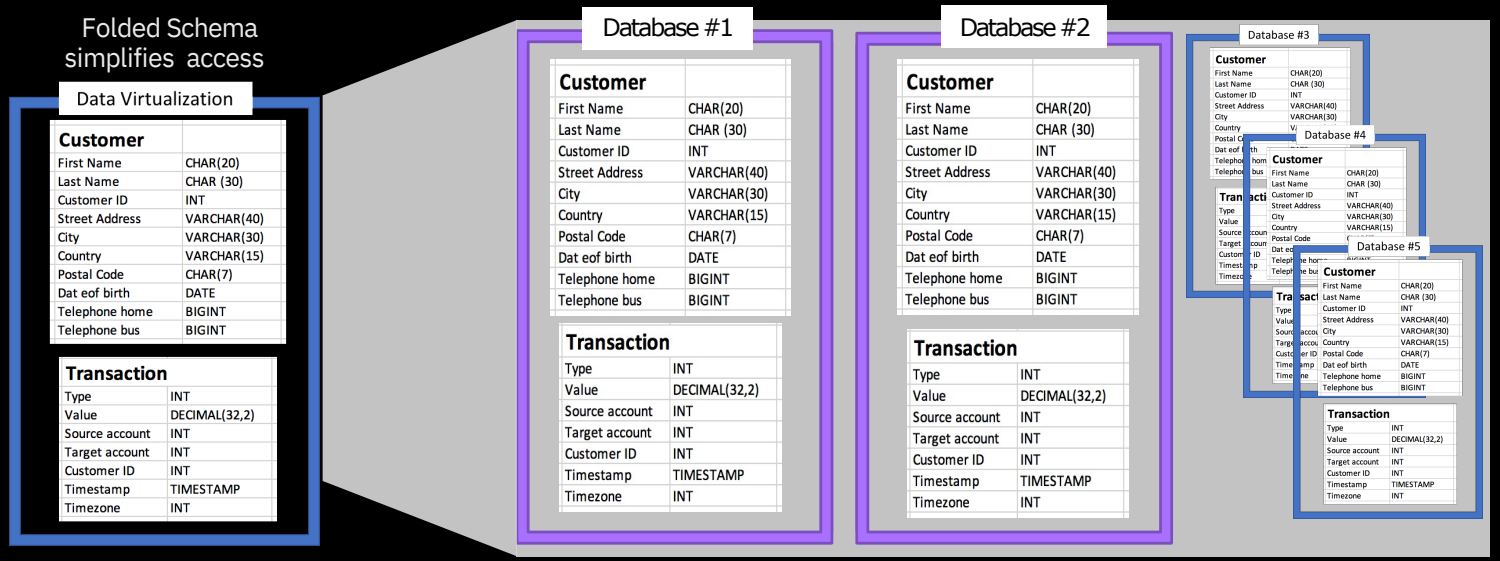
- Advanced query compiler determines whether to use cached data and results for part or all of a query result (MQT & ML based)
- Automation for implementing cache based on your query history





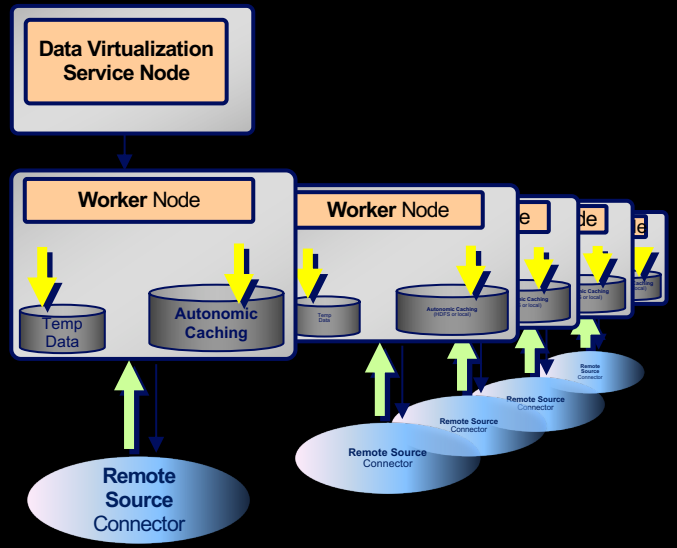
Schema Folding – Simplify Your Data

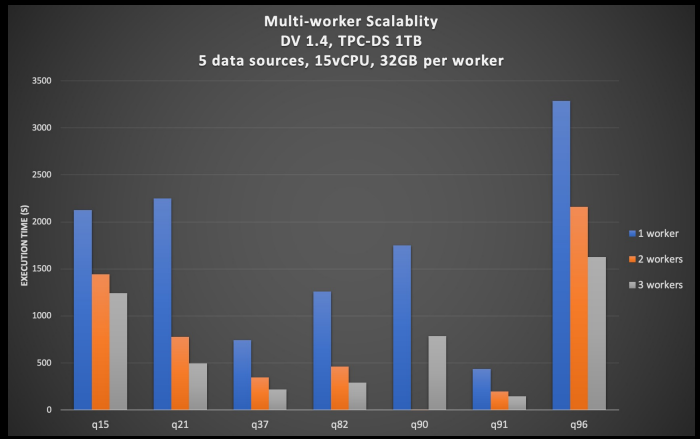
- Common or similar schemas appear in multiple databases.
  - E.g. branch database for a bank or retailer.



Service Compute Scalability

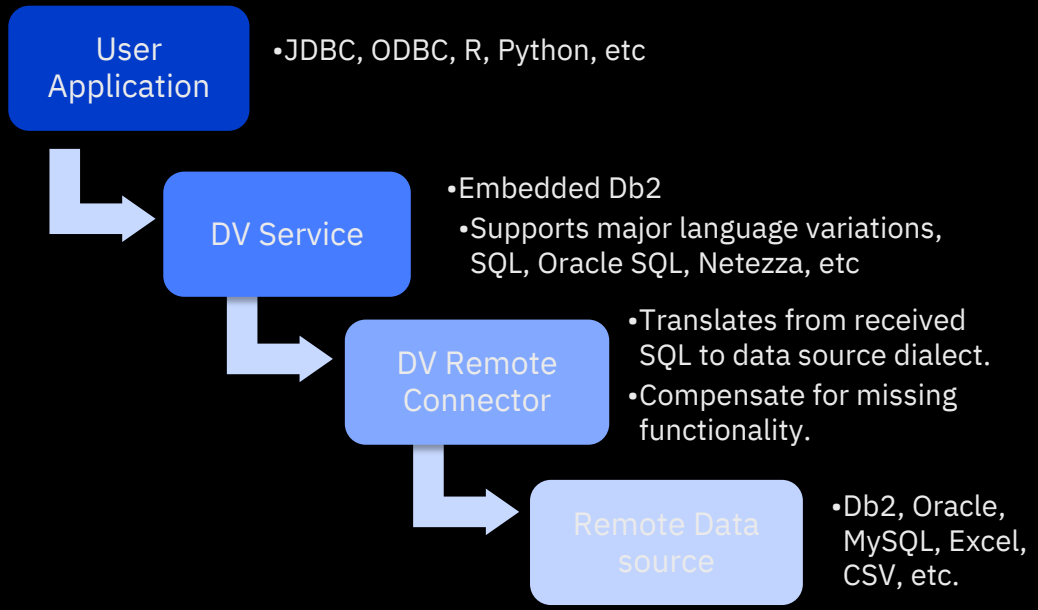
- Supporting multiple worker nodes to improve query performance for large workloads and data sets.
- Parallelized data fetch from data sources.
- When using remote connectors, additional advantages for network parallelism as data can flow on separate network routes to reach the cluster.





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.



Security

Overall approach to Governance Integration and Security.

- Controlled, governed and secure access to virtual data sets within the Cloud Pak for Data Platform

Governance

- Provides data classification (business term) and policy control for virtualized data from any source.
- Data provenance and lineage to show how each source contributes to the analytics.

Security

- Data lives in the original source with only results ever being transmitted.
- Fully encrypted communications between all components of the solution.
- Strong authentication and permission management.
- Policy enforcement and masking capabilities.
- User Group Support *

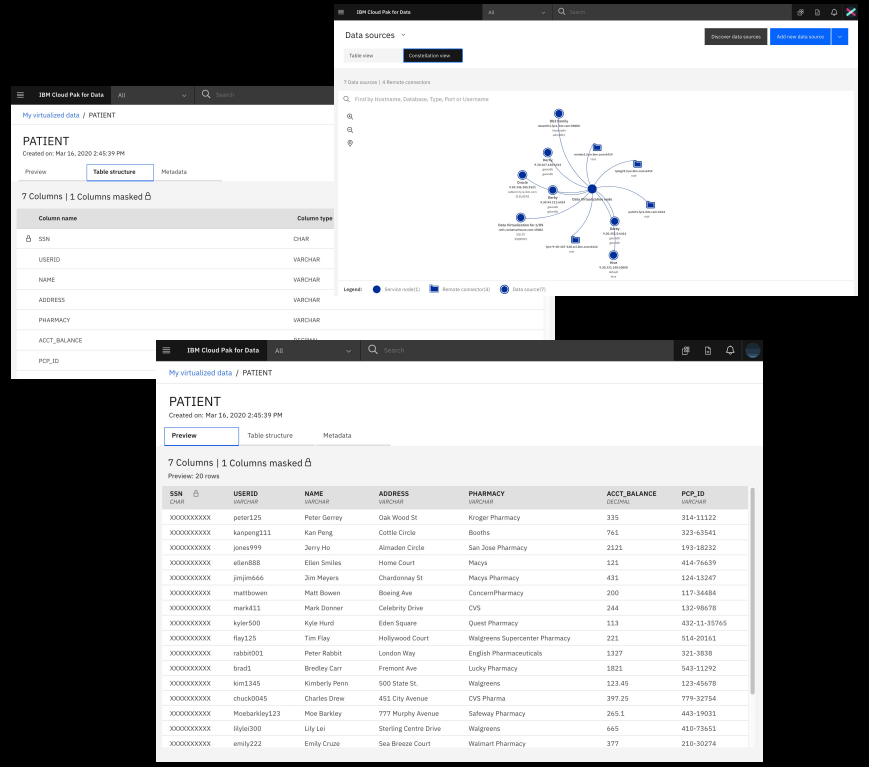
Policies, via a Runtime Catalog

- Policies can capture corporate data mandates and can have sub-policies
- Additional rules can be associated with policies.
- A data protection rule can be added to the policy so that the policy to protect sensitive data can be enforced.

Data Virtualization Today
Governance Policy and rules apply across

Governance in real-time for all virtual objects

- Govern the Data fabric using policies, data protection rules, or deny policy across all Virtual objects
- Apply masking rules to your virtual objects in real-time
- With deep integration policies and rules result in same behavior when accessed within, or outside of Cloud Pak for Data platform
- Uniformity of enforcement for all virtual objects , irrespective of the Data Sources data type or structure



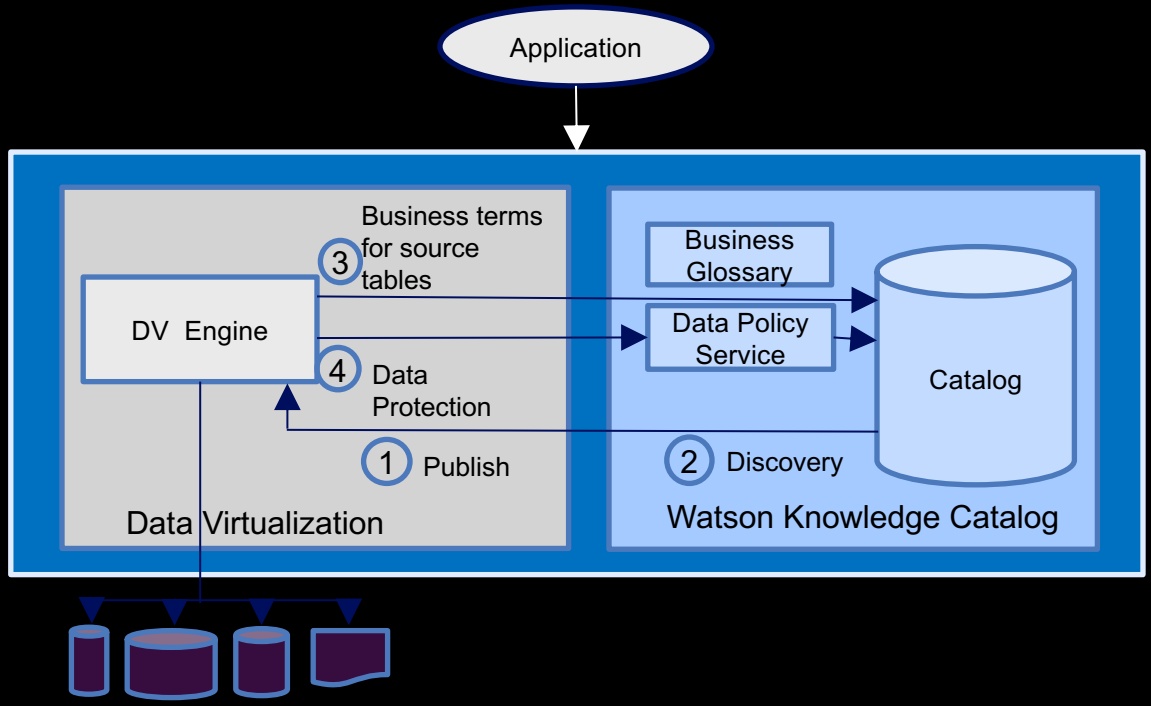
Integrated Governance

I. DV Publishes remote assets to WKC.

II. WKC Discovery performs classification, scoring, term association

III. DV retrieves Business Terms for the source tables to give the users a common understanding of the data.

IV. DV obtains information about the policies it needs to enforce upon data access.



What problem are we solving? What's the opportunity?

Exposing Hybrid Data Quickly via Data Fabric (& Key Attributes)

What? Hybrid data fabric for *making existing platforms flexible*

- Use data without knowing format/location, moving or copying

Why? Constant change and need for automated governance

- Ease of query / access to query across sources
- Hybrid: Connect and Access Across Sources
- Scalable and Fast: 1 or 1,000s!
- Distributed: Regional Storage/Governance & Parallel Access
- Governed: Policies applied during search, query, usage

Benefits:

- Time/Effort to Access: from Months to Minutes
- Use of Data: Effective/Proper, in-context
- Time to Value/Innovation: 80% of time spent understanding data
- Tech / Data Modernization: technology modernization/take-out, isolating data impacts (e.g. Data Lakes)

Steps to increase Time-to-Result, with Reduced Effort

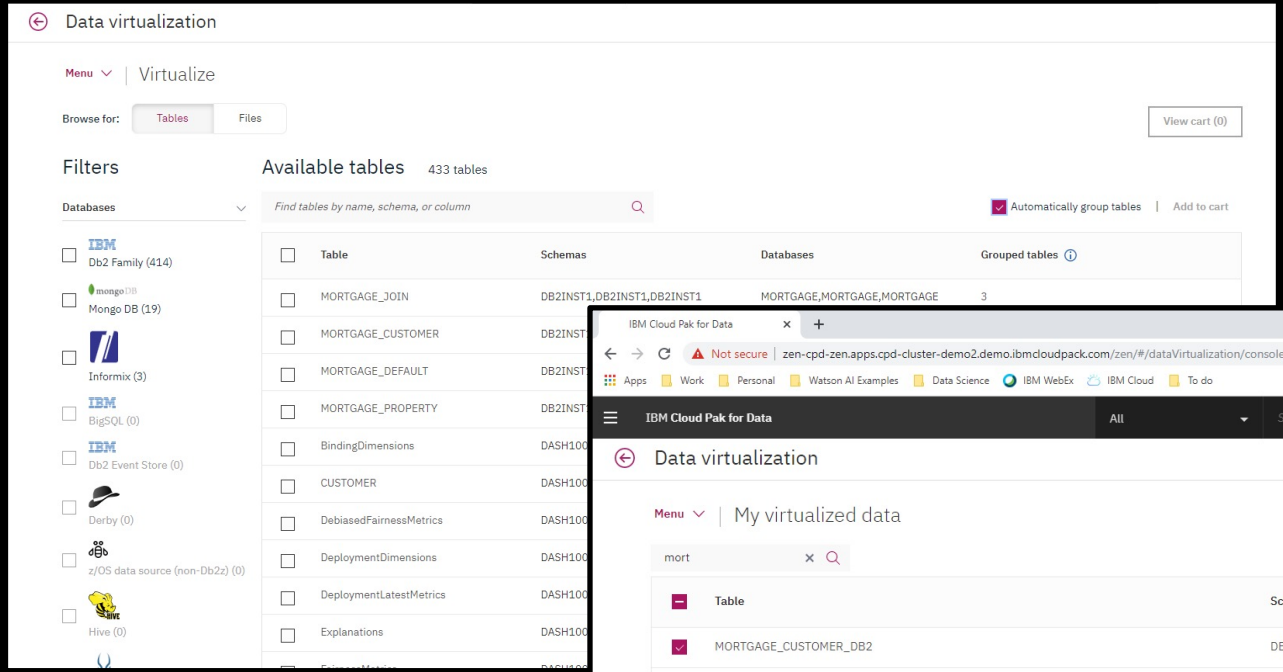
1. Connect to Existing Data Assets (DBs, etc.) or Search Cataloged Data Assets

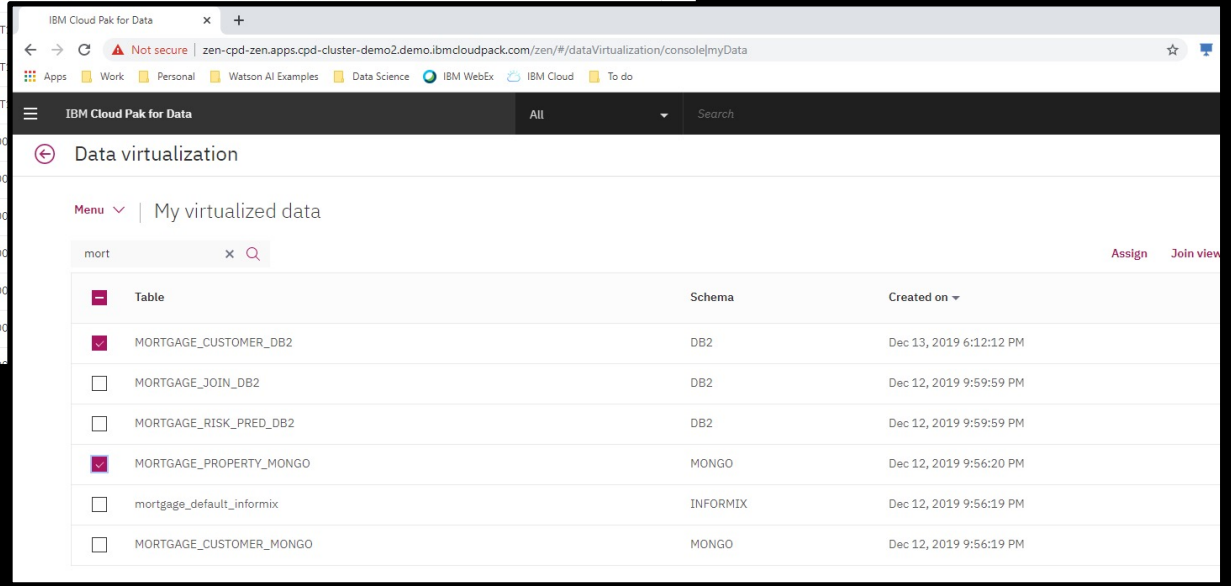
2. If the User/SME doesn’t already know the ”design”, …
   - Discover and Profile the Quality of Data Assets
   - Discover the De Facto “Design” between Hybrid Sources (PK->FK Key Relationship Analysis)

3. Create a new Virtual Object, Querying Across Data Sources
   - Based on already-known de-facto “design” PK->FK Key Relationships

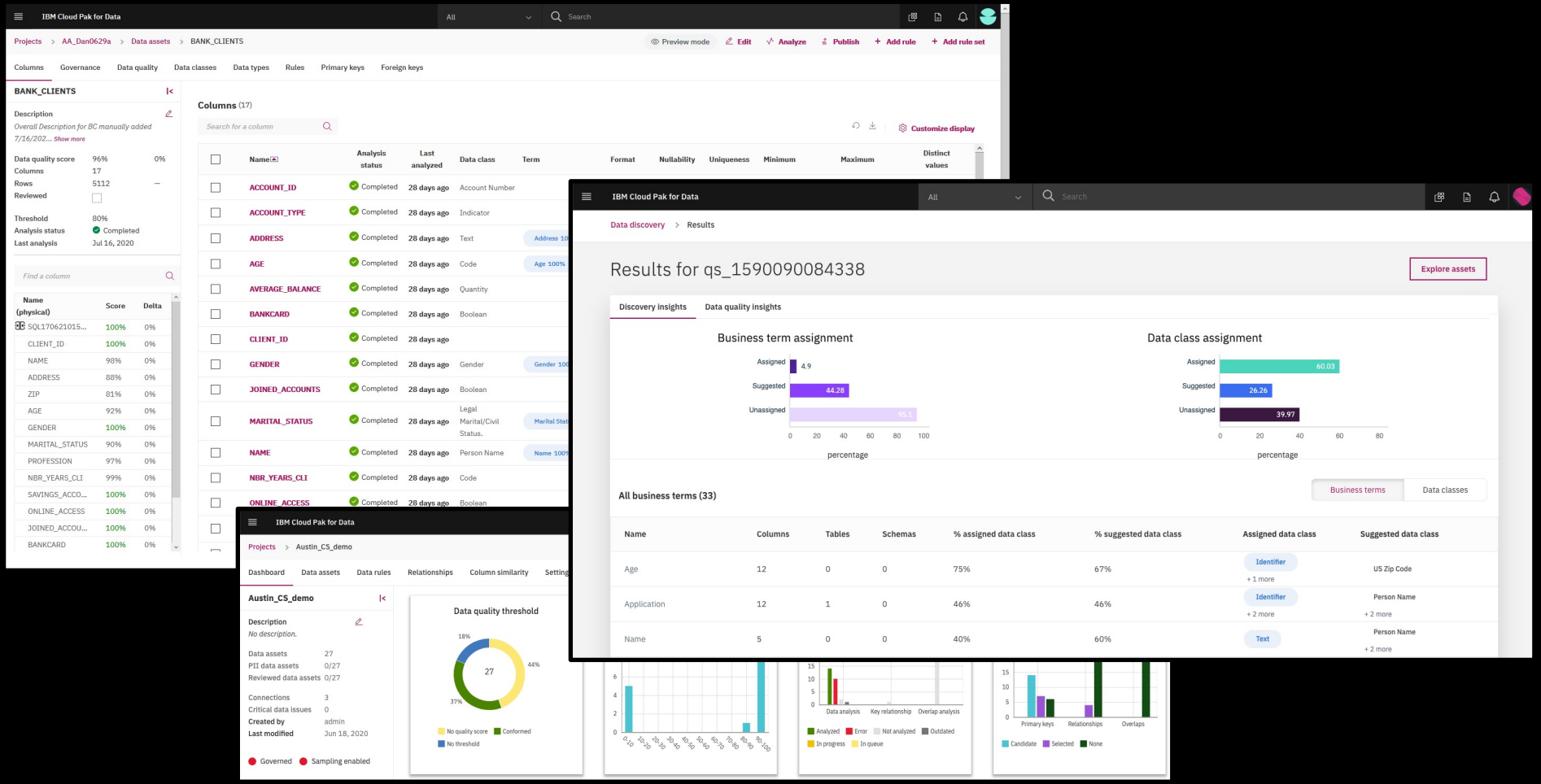
4. Make Virtual Data Available to Others (Users, Platforms, Applications)

1. Connect to Existing Data Assets (DBs, etc.) or Search Cataloged Data Assets

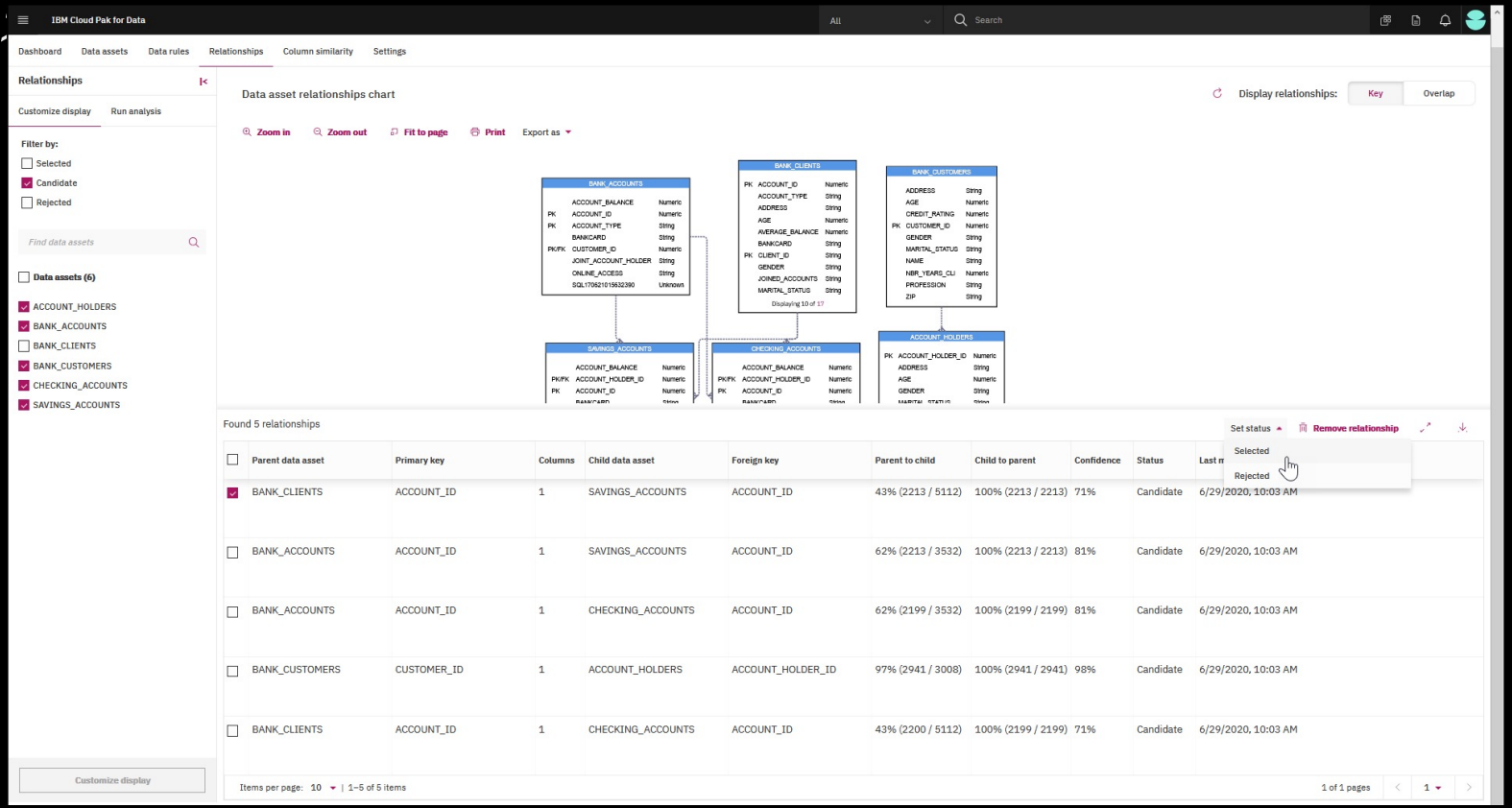




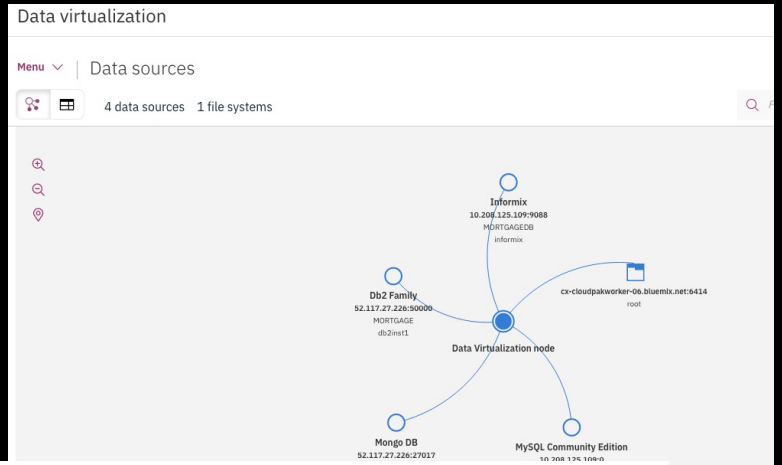
If needed, Discover and Profile the Quality of Data Assets

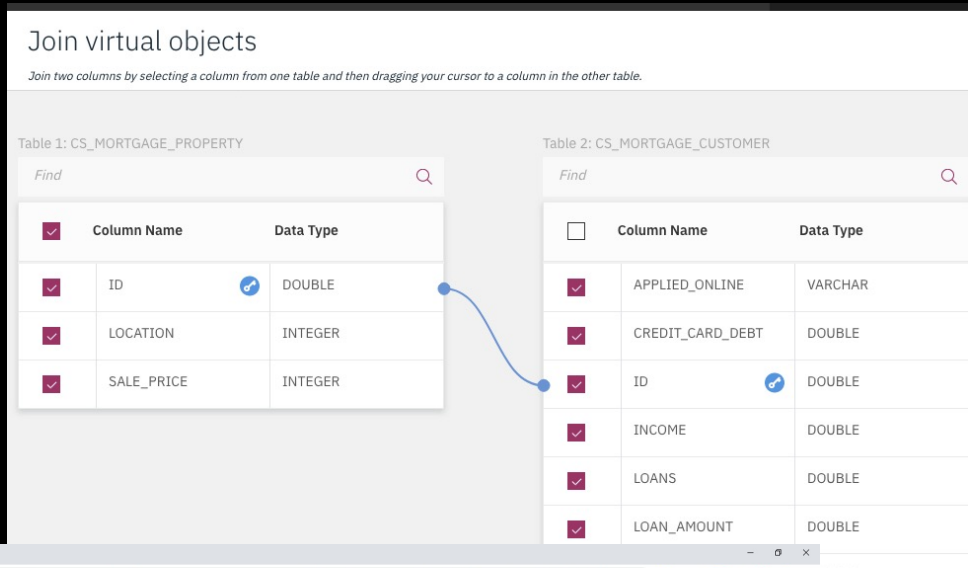


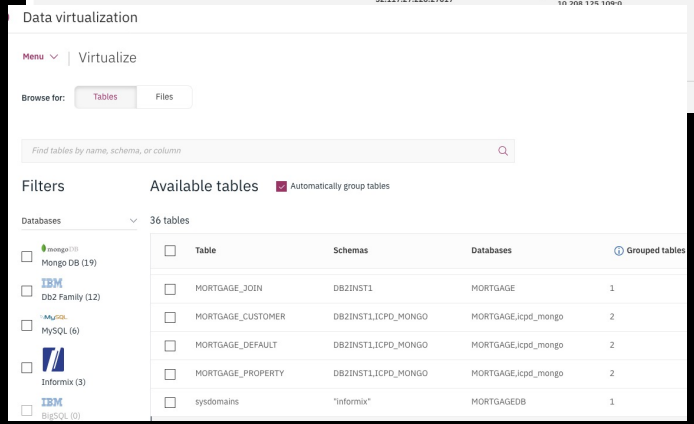
If needed, Discover the De Facto “Design” between Hybrid Sources *(PK-)*

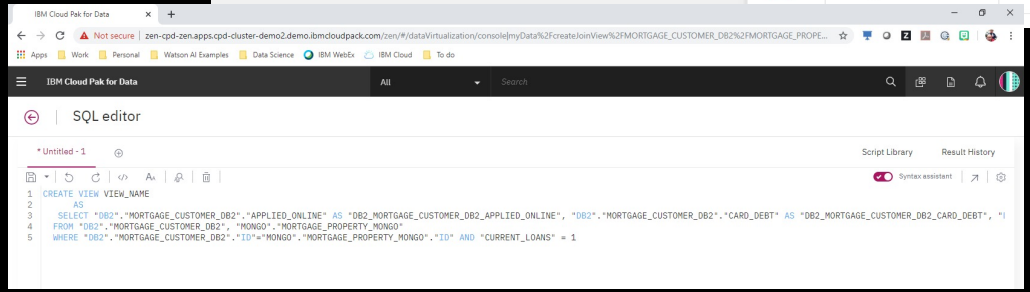


3. Create a new Virtual Object, Querying Across Data Sources (based on de-facto “design” PK->FK Key Relationships)









4. Make Virtual Data Available to Others (Users, Platforms, Applications)