

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

1,4

Organize

2,3,4,6,7

Analyze

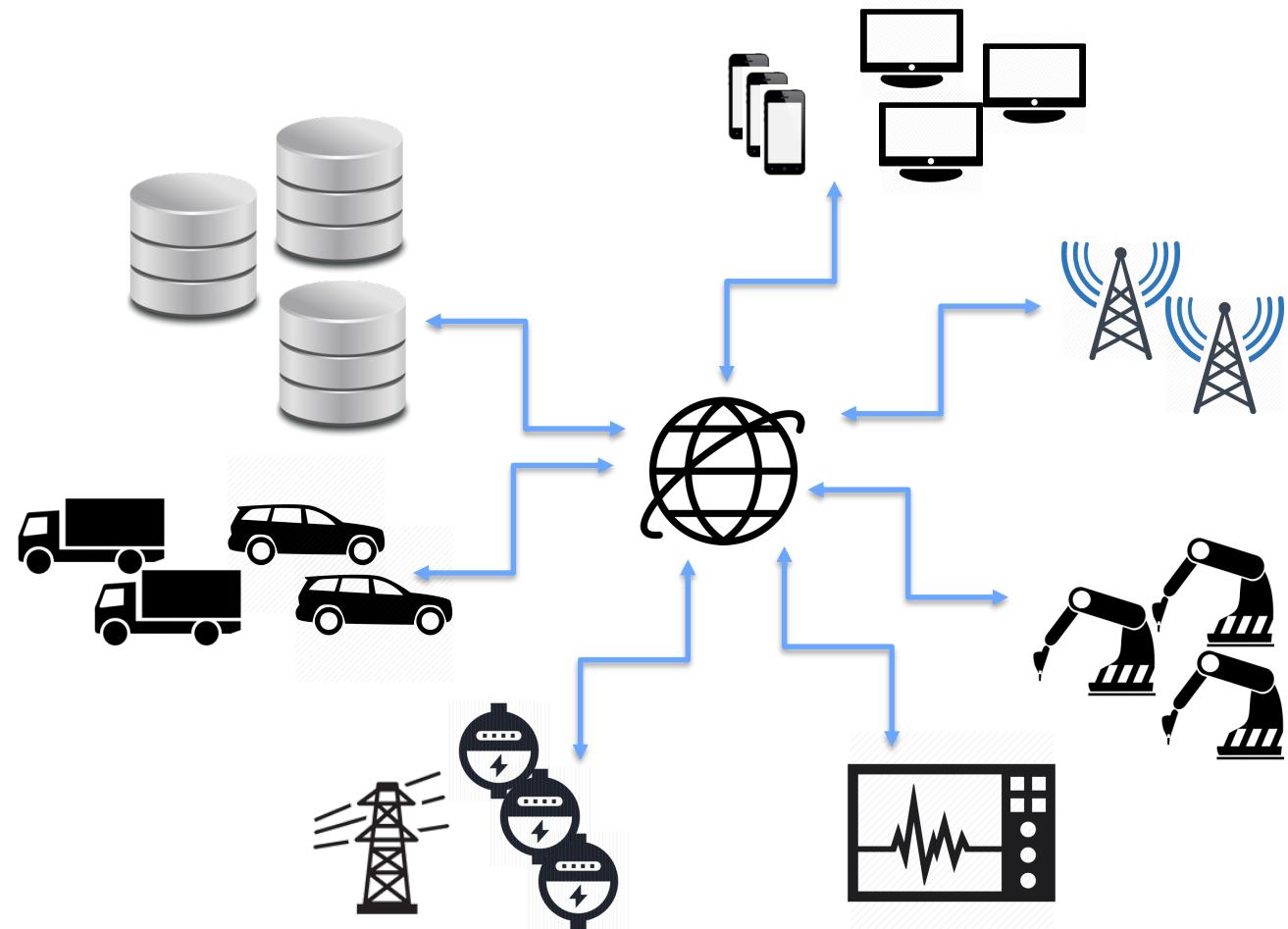
2,3,4,5

Data is Everywhere

Number of Sources and volume rapidly increasing

More and more heterogeneous

Highly distributed – internal and external

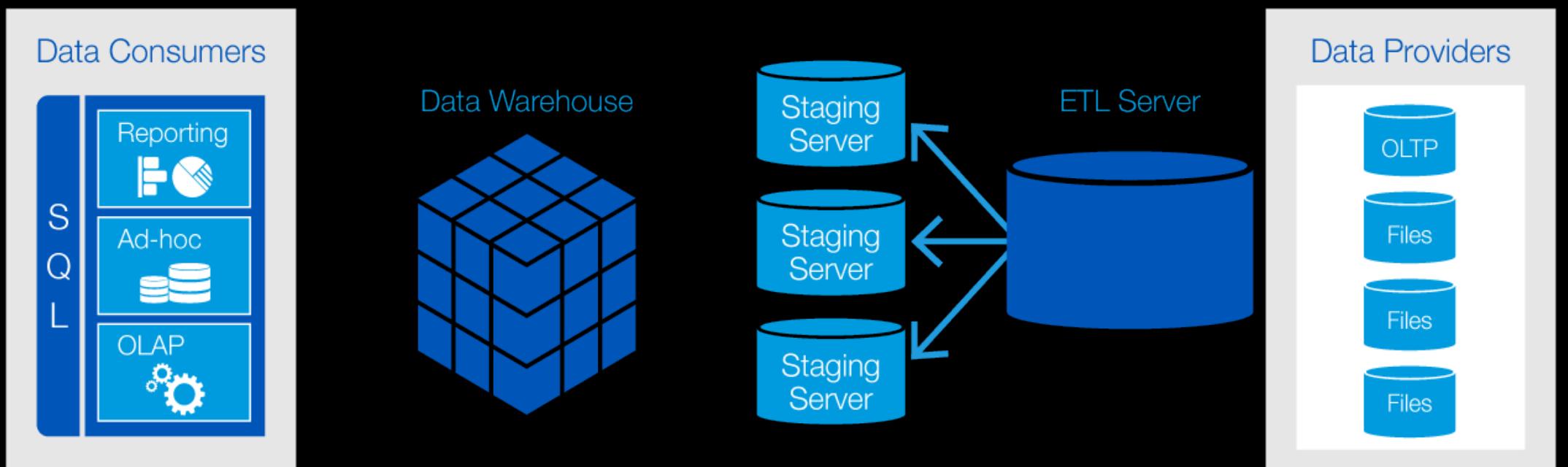


Traditional Data Integration

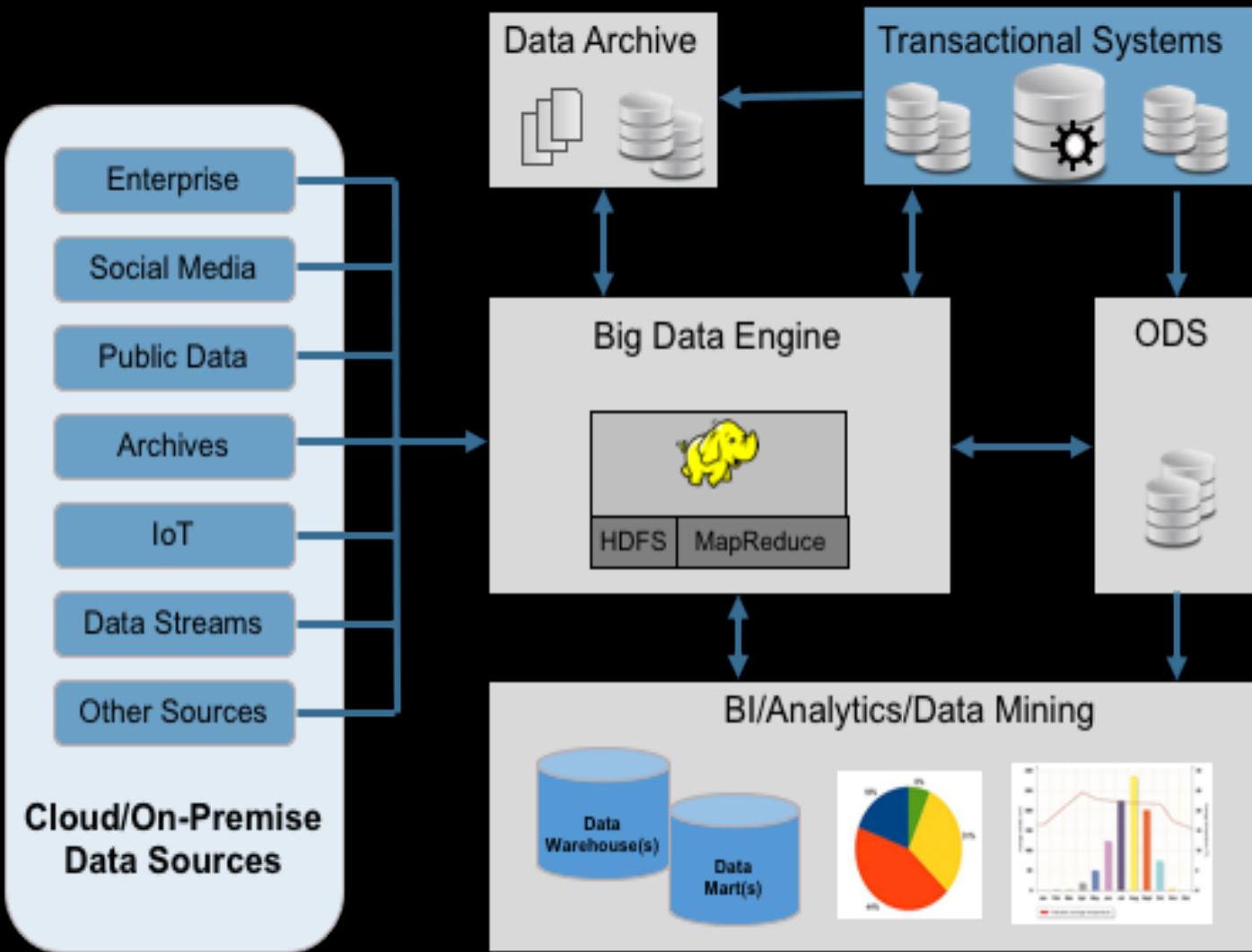
Not Viable to continue to Move or Copy *all* Data
(using extract, transform, load ETL)



- 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?

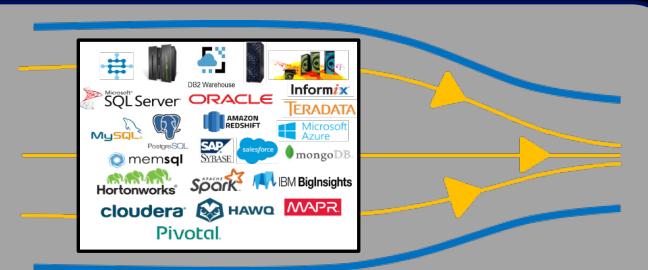


Data virtualization key Use Cases

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!

Scale! 1 or 1,000 at once

- More than 10x better for several important use cases

Schema discovery and folding

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

Rich application capabilities

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

Secure!

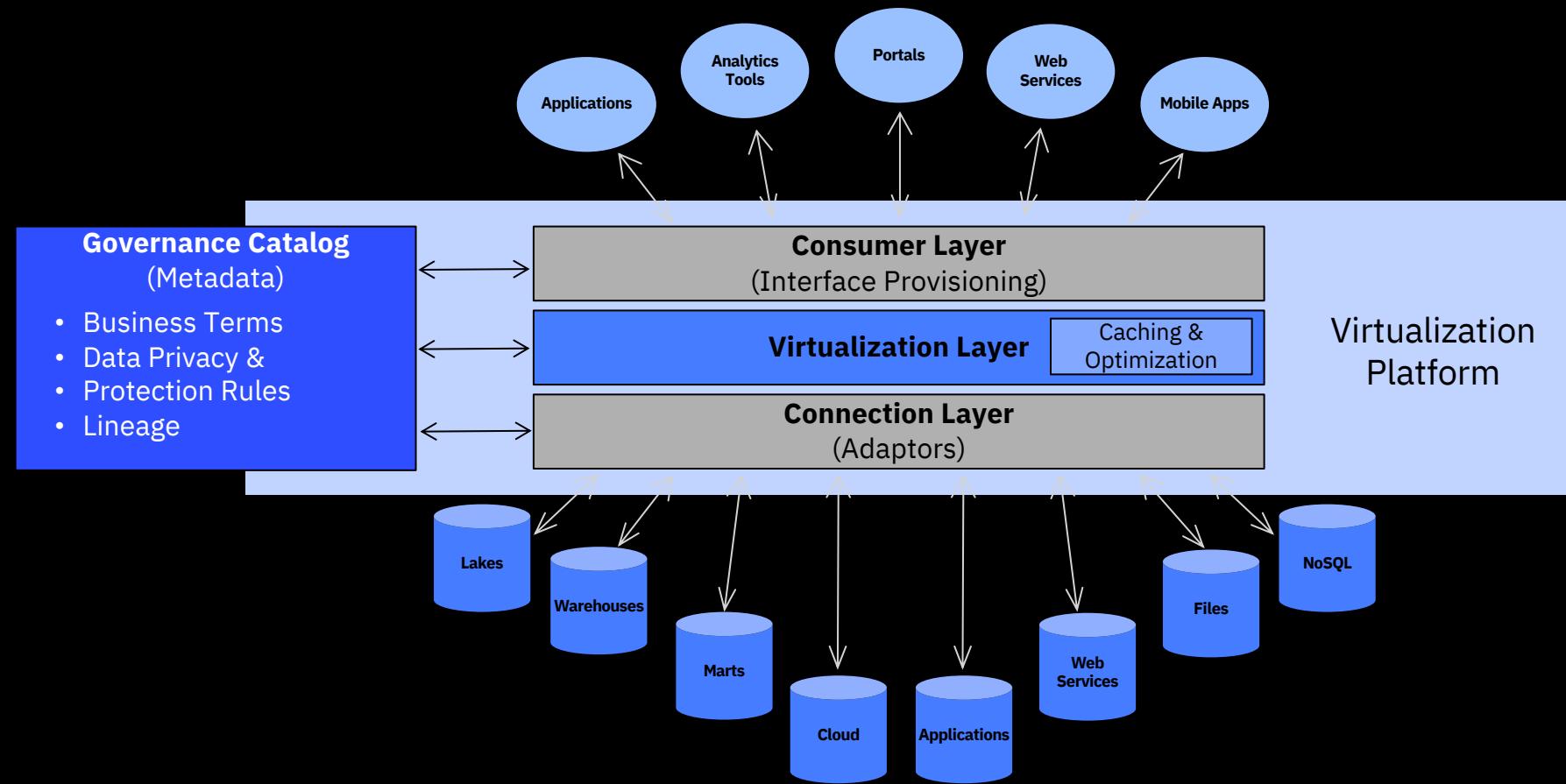
- Strict access controls
- Fully encrypted communications

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

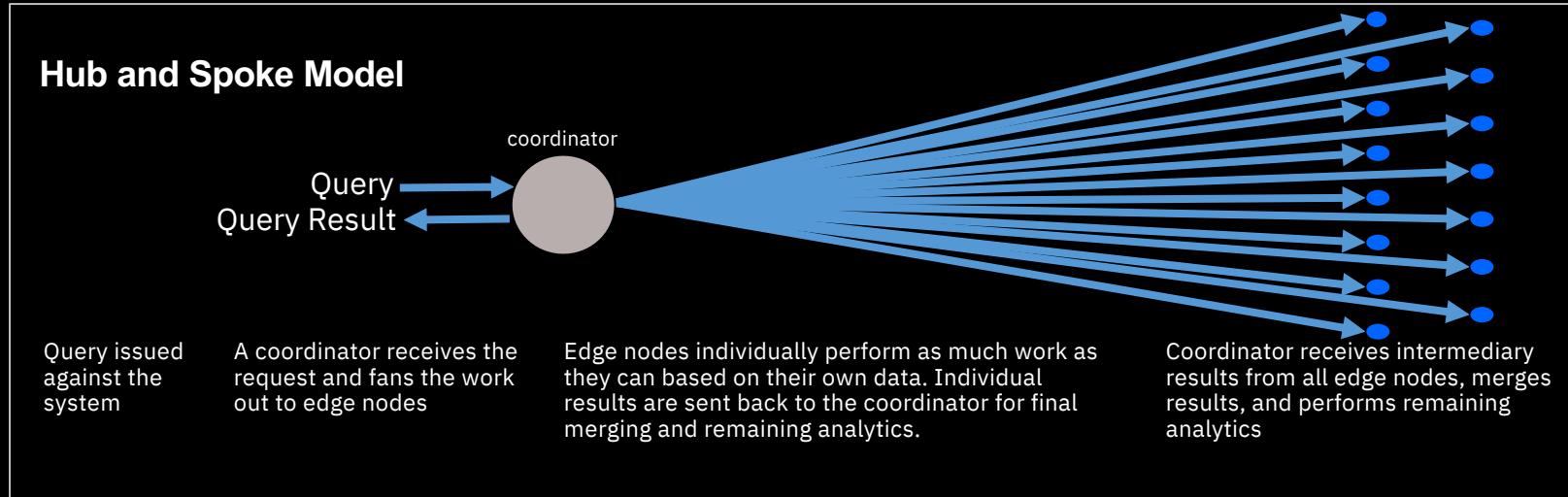
The ability to view, access, manipulate and analyze data without the need to know or understand its physical format or location, and without having to move or copy it.



Key Architectural Differentiation

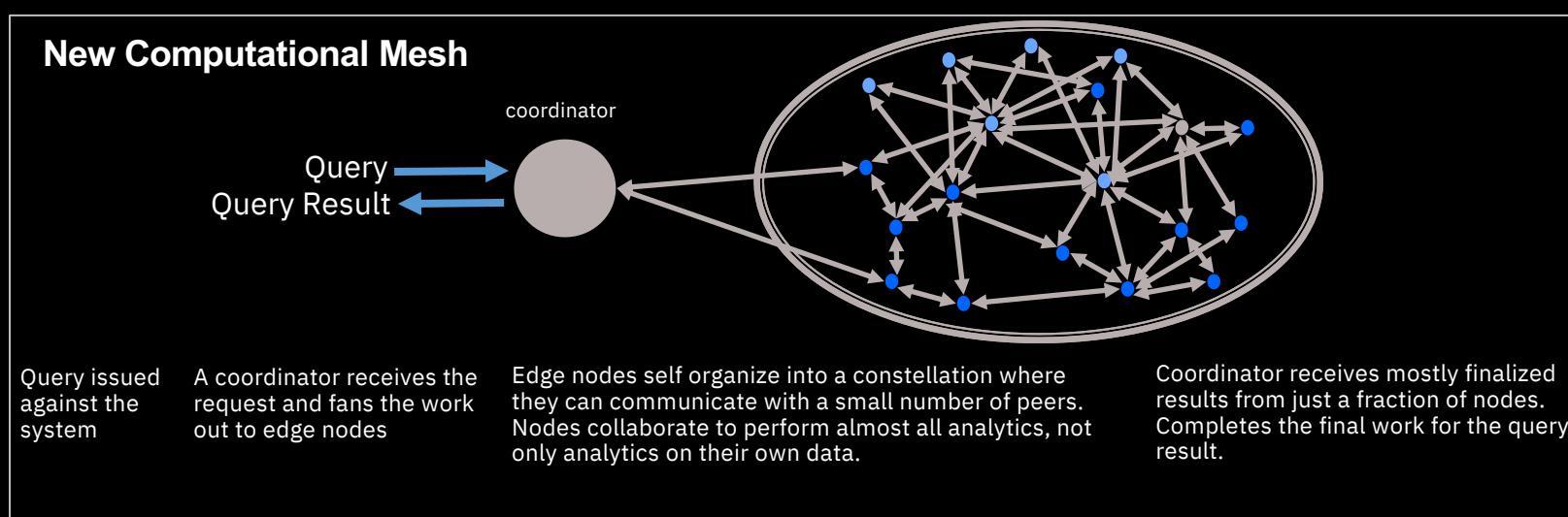
Hub and spoke execution models:

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



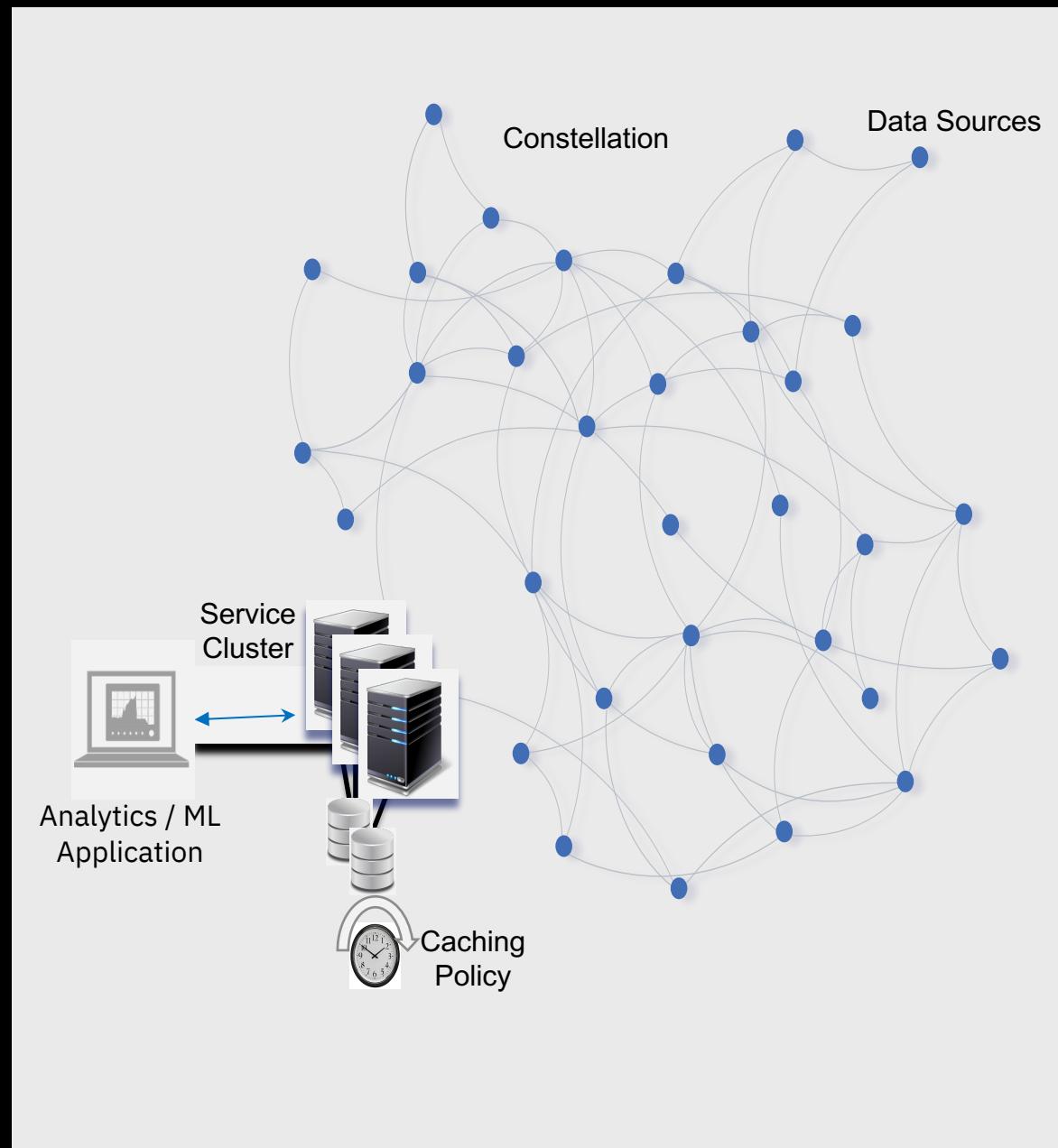
IBM is first to market with a parallel processing model:

- 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



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

Broad support for common data source types

*More to be added
in the future.*

Data Virtualization Manager for z/OS

Data Virtualization Manager for z/OS

More than 35 supported data sources

IBM Z

ADABAS



VSAM



DB2



Syslogs
log streams



Sequential-file



File Systems

zFS

HFS

SFM

Magnetic Tape
Virtual Tape



Systems Management Facility
SMF
Data for IT Operational Analytics

CA IDMS

Non-IBM Z

IBM Db2 Family, Informix



ORACLE



20+ DRDA data sources

THE
Open
GROUP

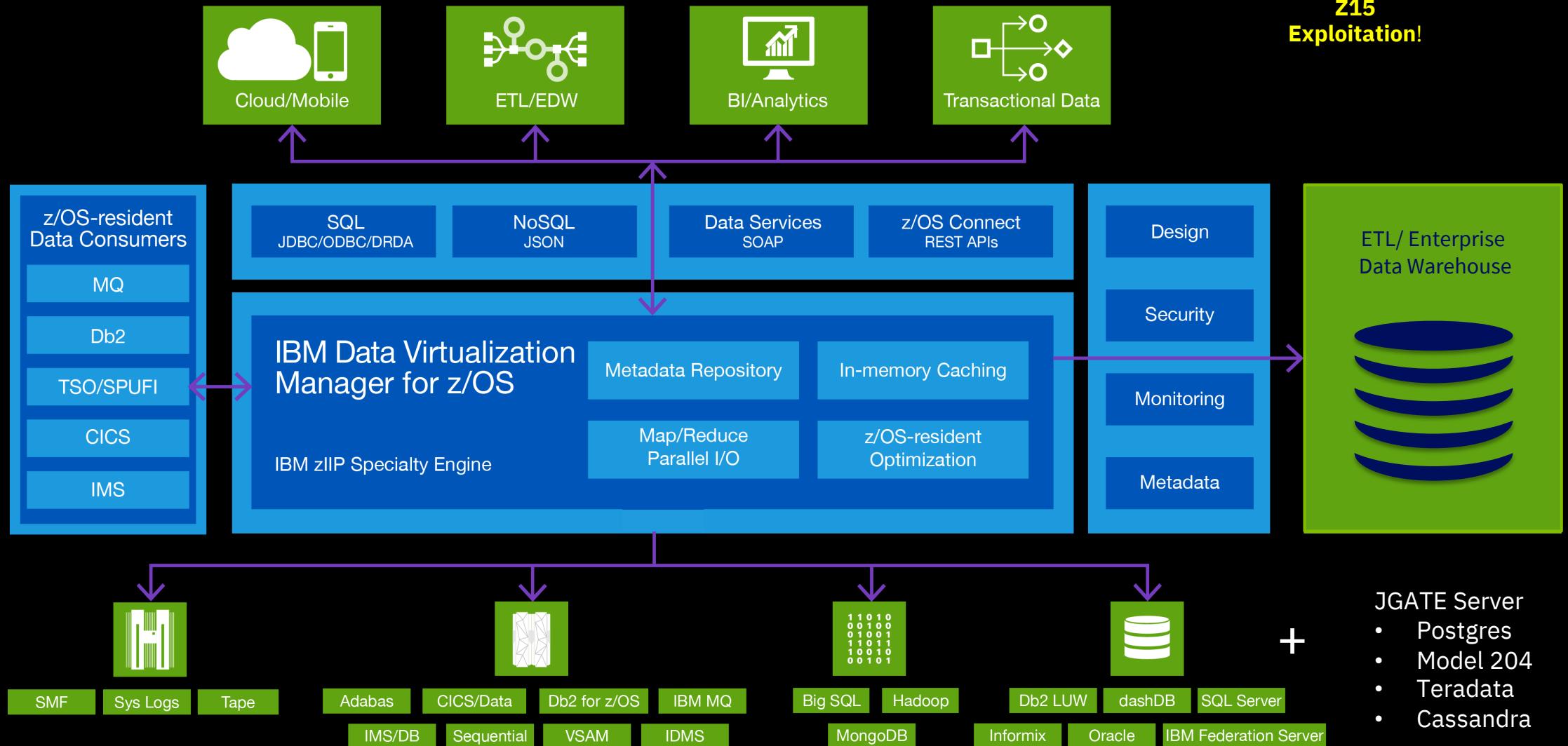
Other DBs



... exposed to data consumers
in SQL (JDBC/ODBC/DRDA) and NoSQL (JSON) formats

IBM Data Virtualization Manager for z/OS

NEW!
Z15
Exploitation!



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	<p><i>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.</i></p> <p>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.</p>
Db2 UDTF support (User Defined Table Function)	<p><i>Allows any application to access any DVM virtualized data using DB2 via SQL. Ex: DB2 app accesses and updates VSAM records using SQL.</i></p> <p>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.</p>
JAVA Gateway	<p><i>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.</i></p> <p>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.</p>

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



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



IBM Cloud



Azure

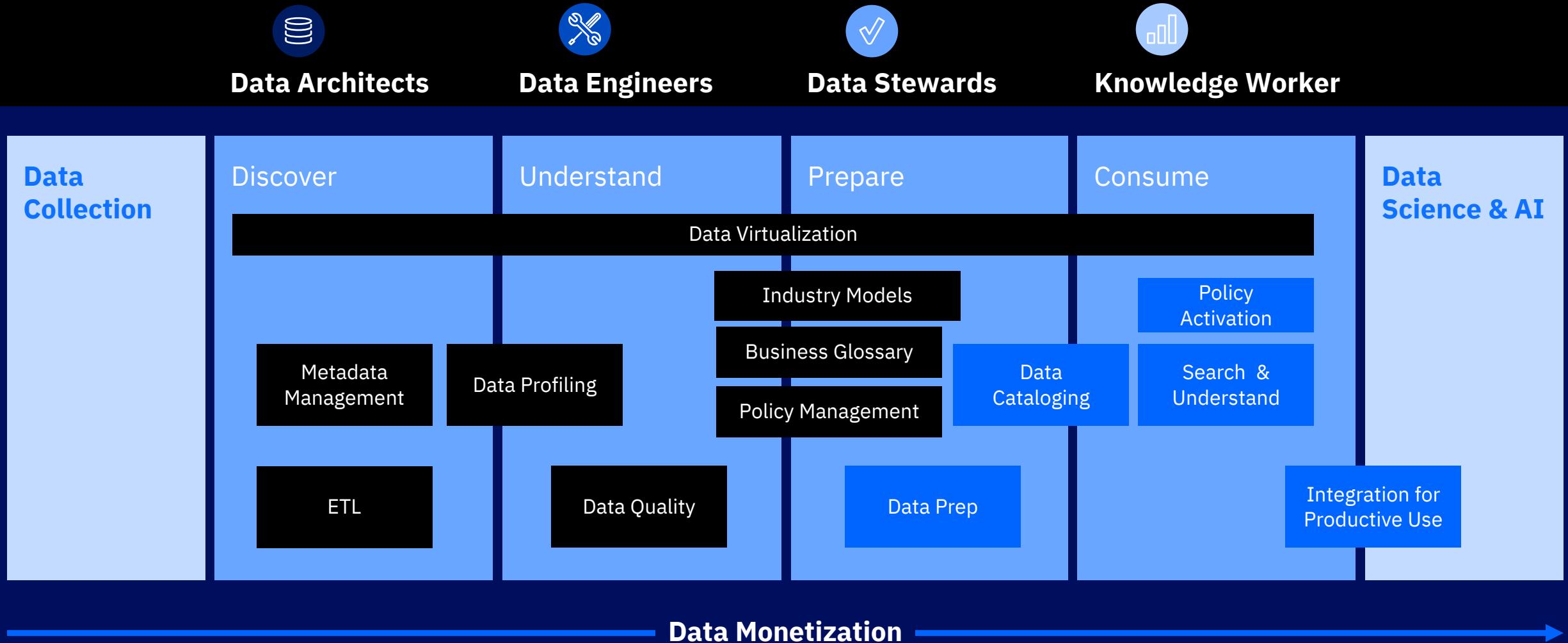
Google Cloud



Hyperconverged
System

Supporting the continuum from data collection to consumption

Enabling velocity, scalability, and traceability



Data Governance

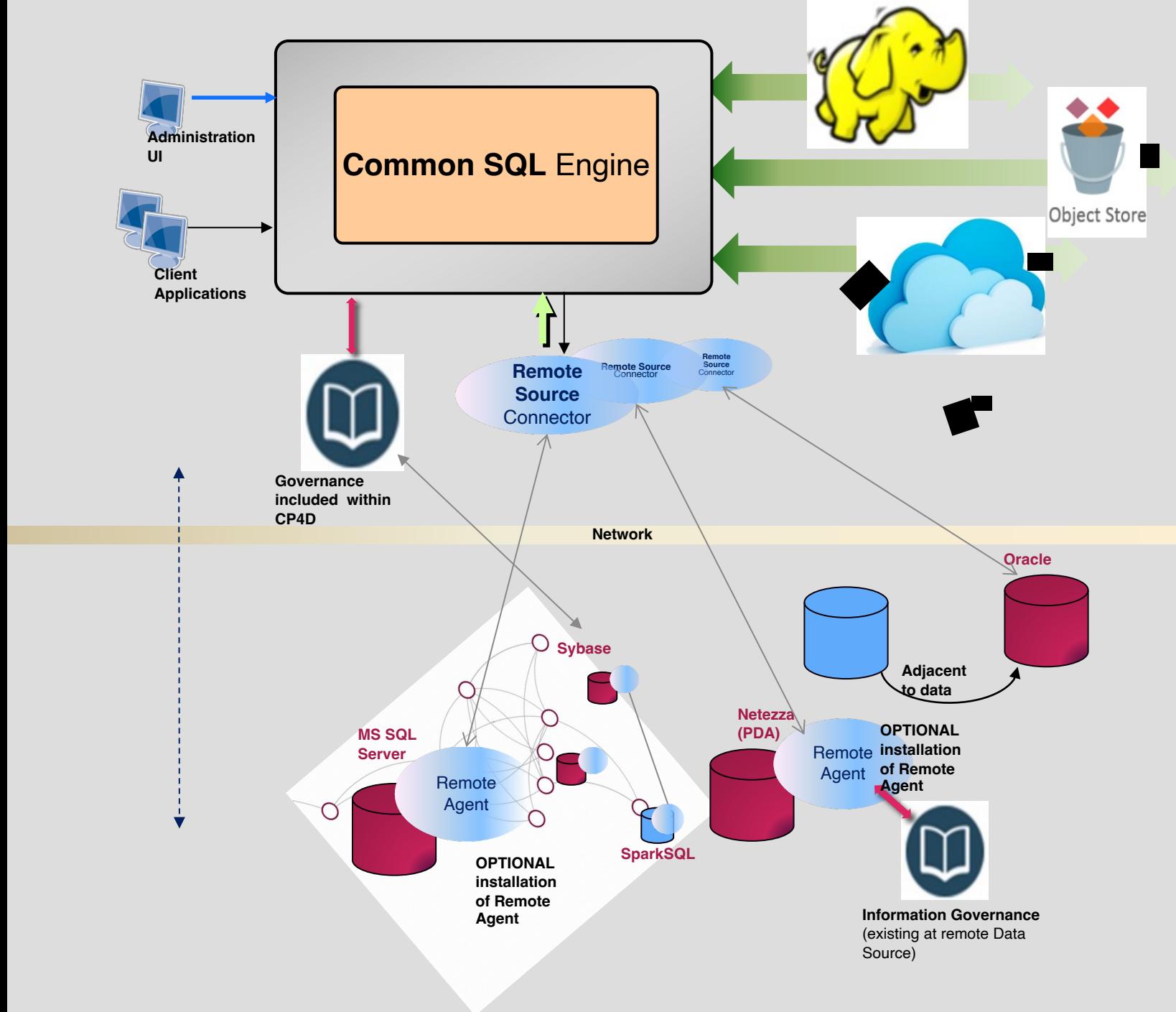
Classifications: Each data asset has a classification that describes the sensitivity of the data. Classifications are provided by Watson Knowledge Catalog.

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.

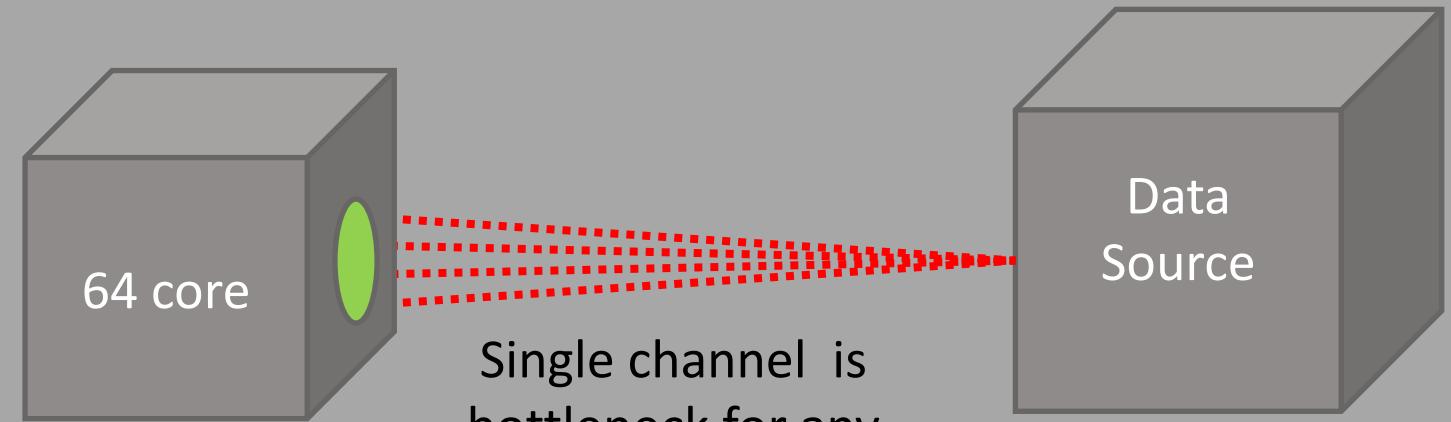
Architecture



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.

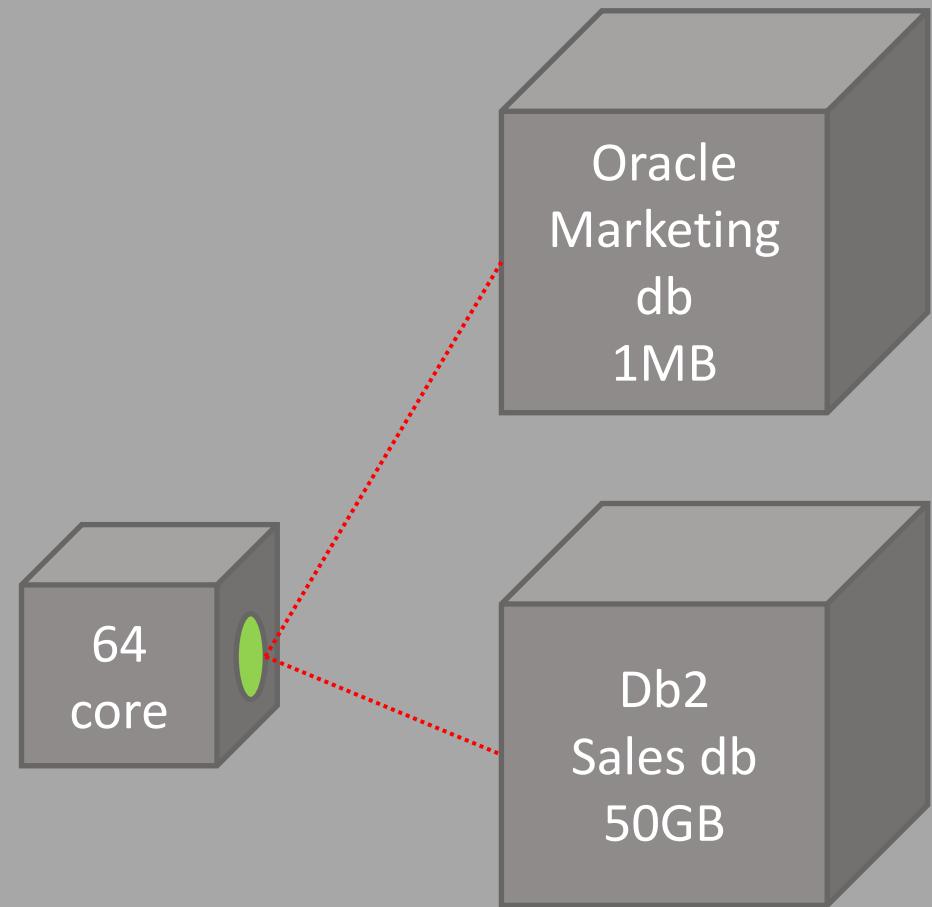


Single channel is bottleneck for any large result set. Data Virtualization resolves this through parallel read from sources.

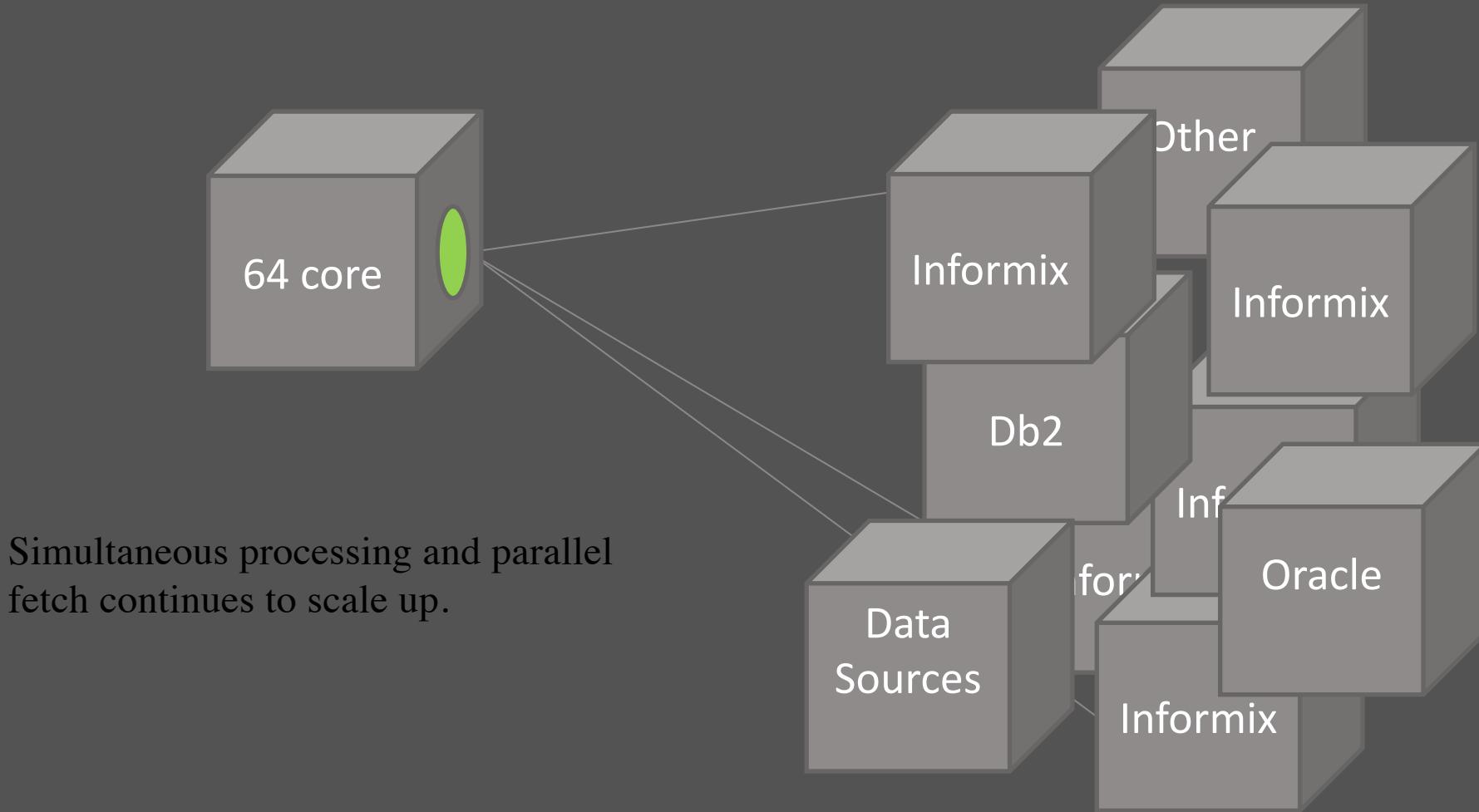
Efficient Cross Source Joins

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.



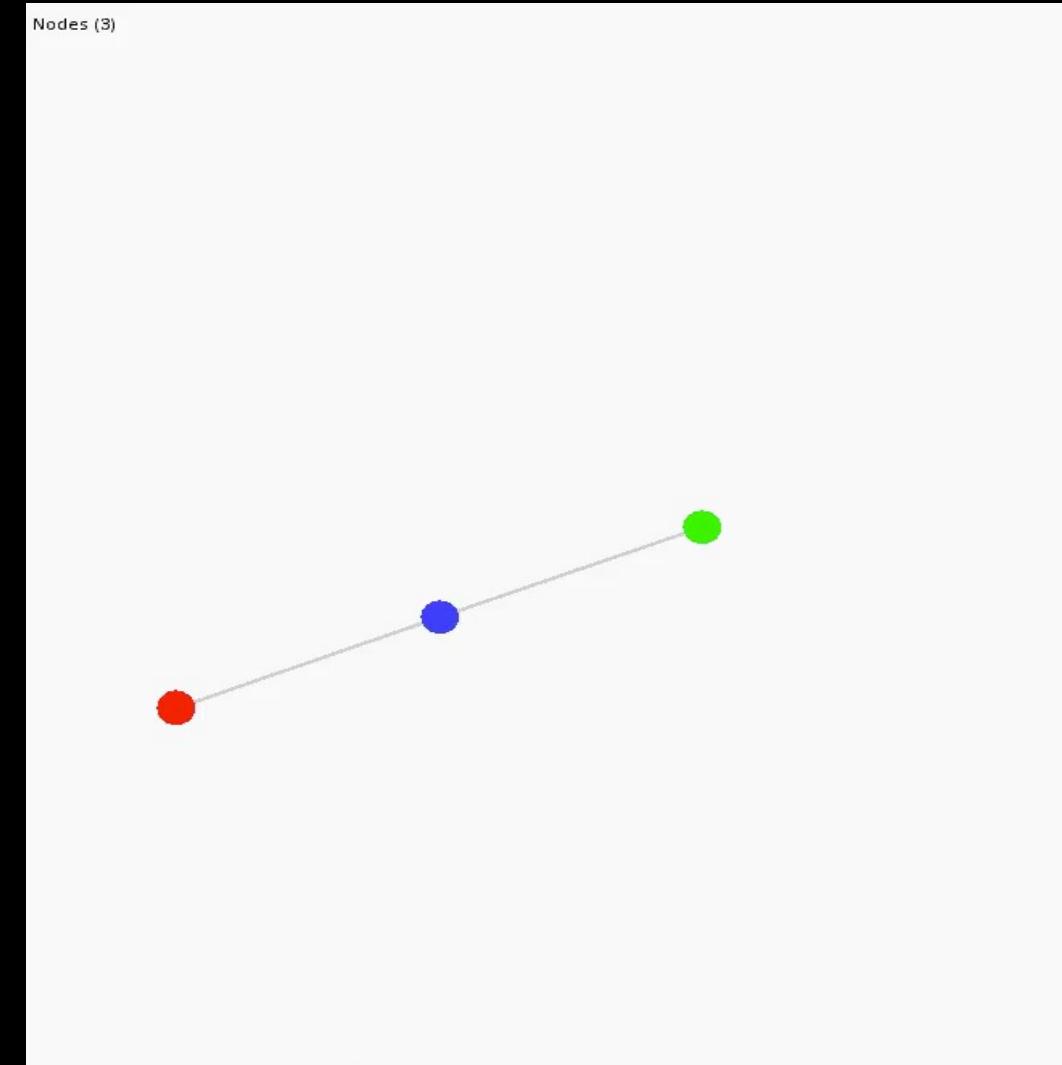
20 databases? Traditional processing gets worse as the number of databases increases



Real System test

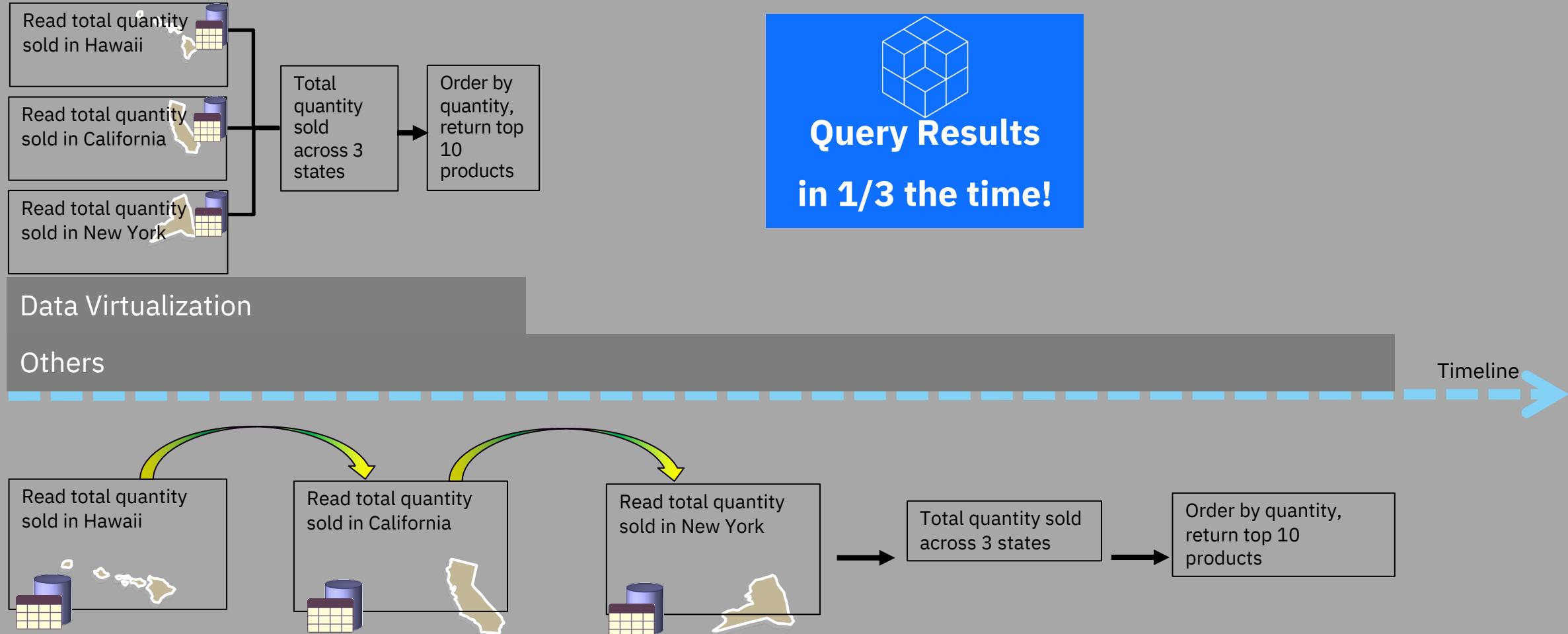
Growing a Constellation

- 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',
```

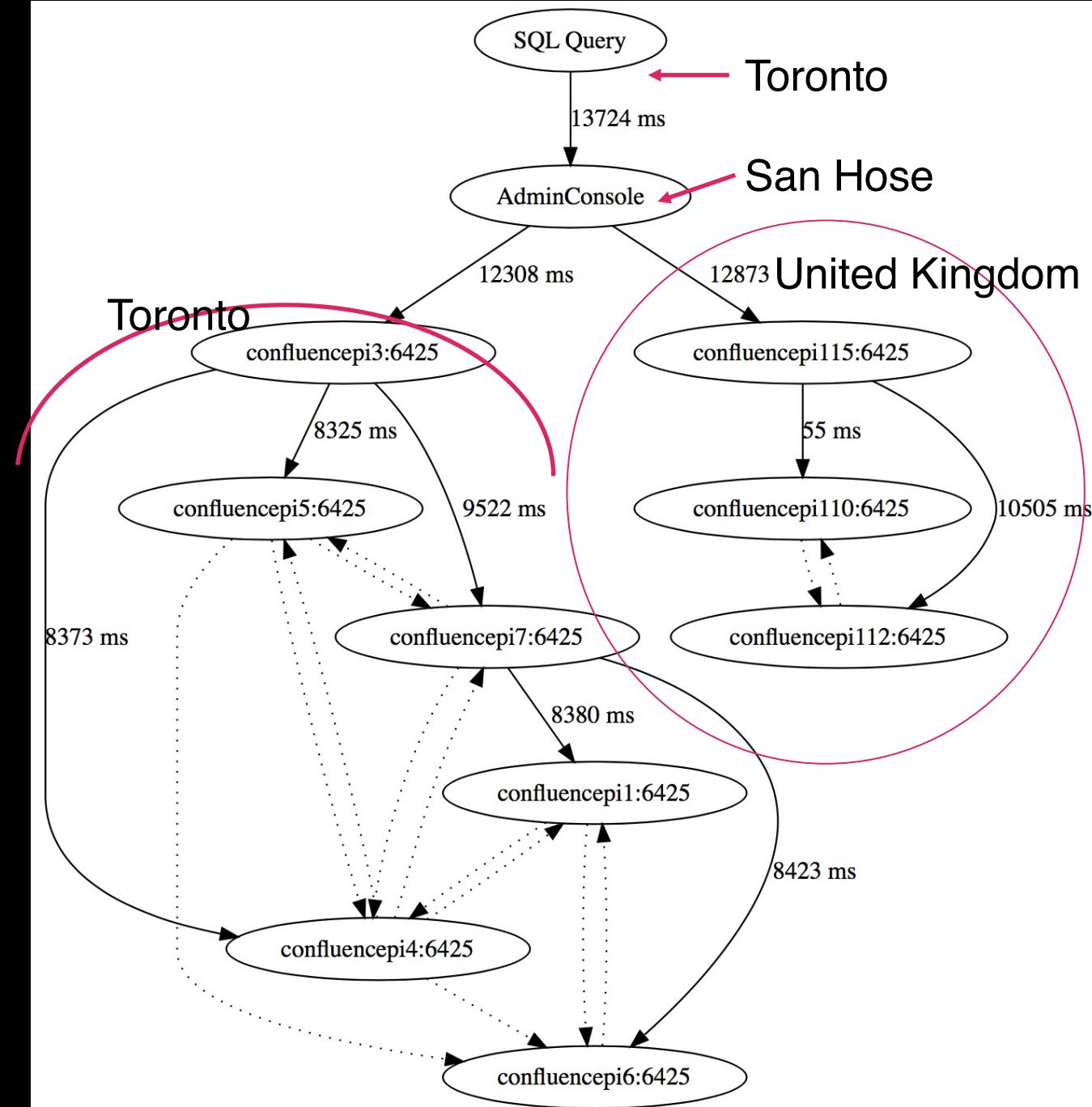
```
' GAIAN_EXTENDER=[pAgg] ', '' ,
```

```
' 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.

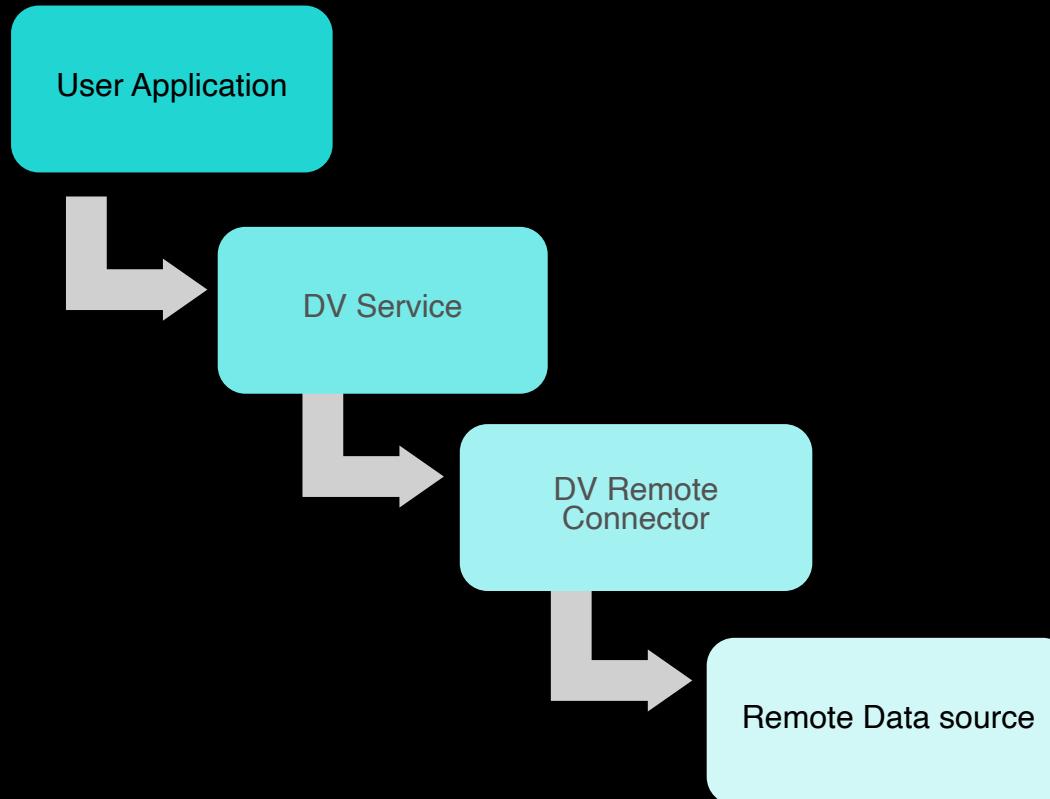


Language Translation in Data Virtualization

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.

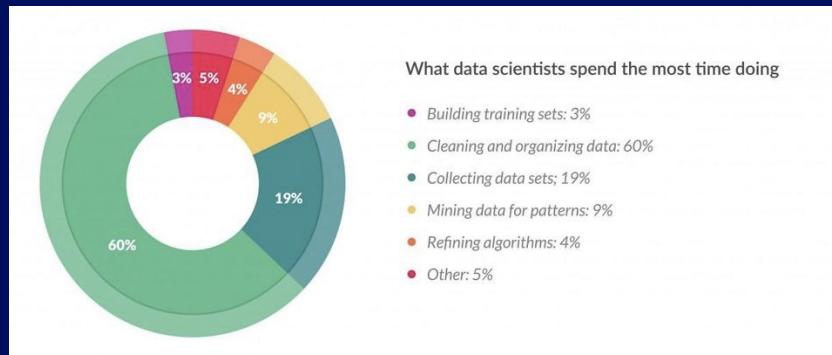
Multiple levels of translation as we move from the applications through the constellation down to the data source.



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

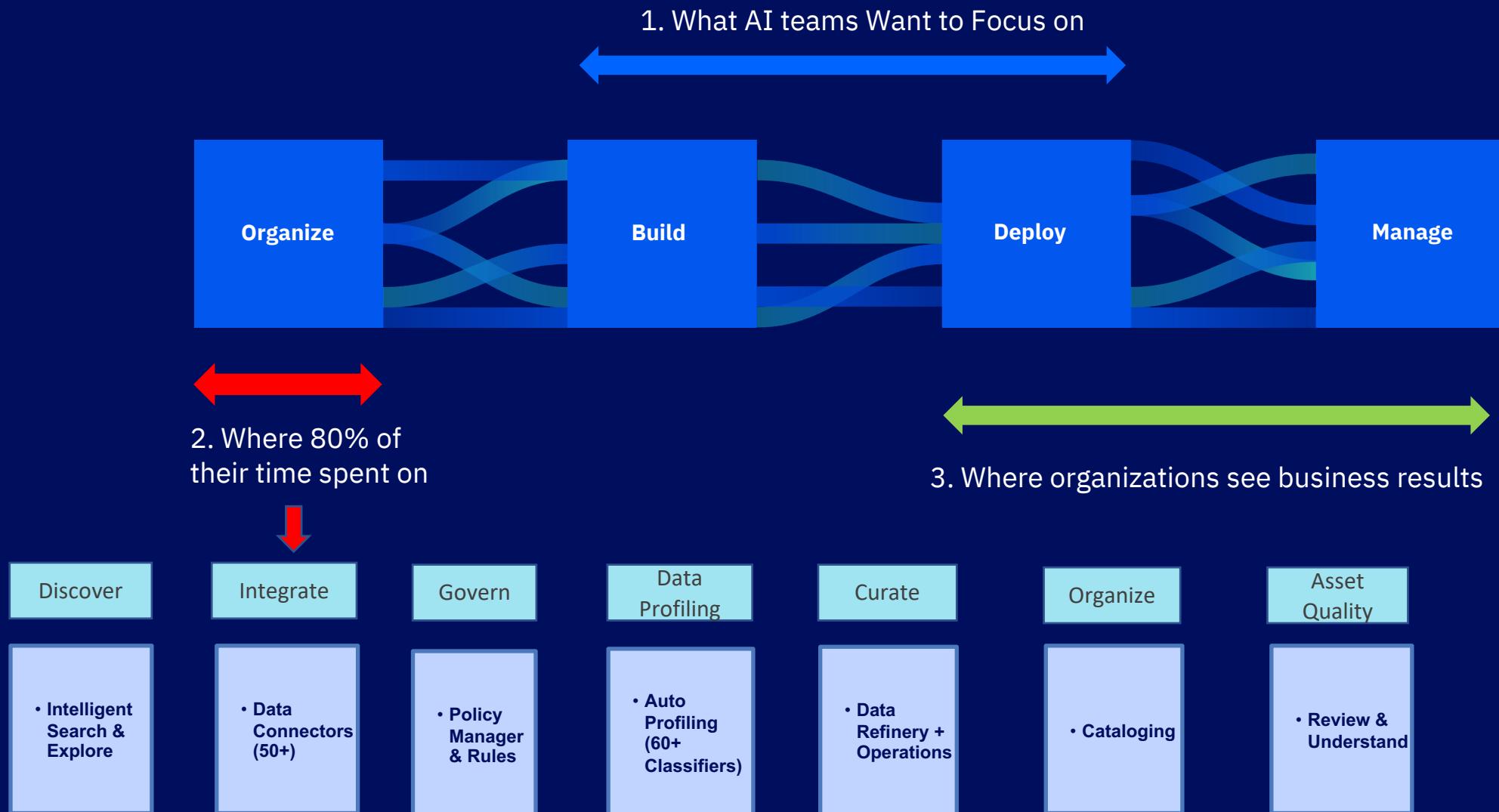


MODERNIZE
your data estate for an
AI and multicloud world

There is no AI without an IA (information architecture)

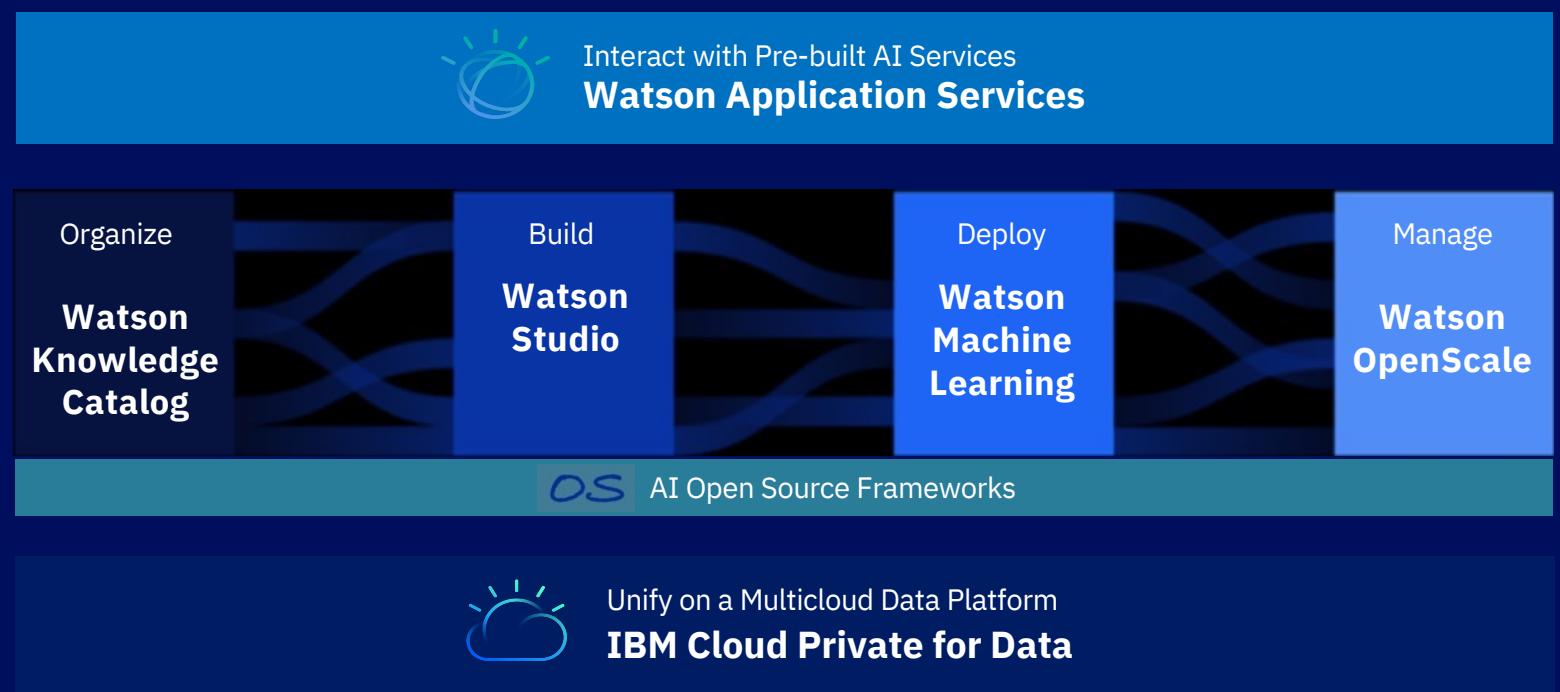
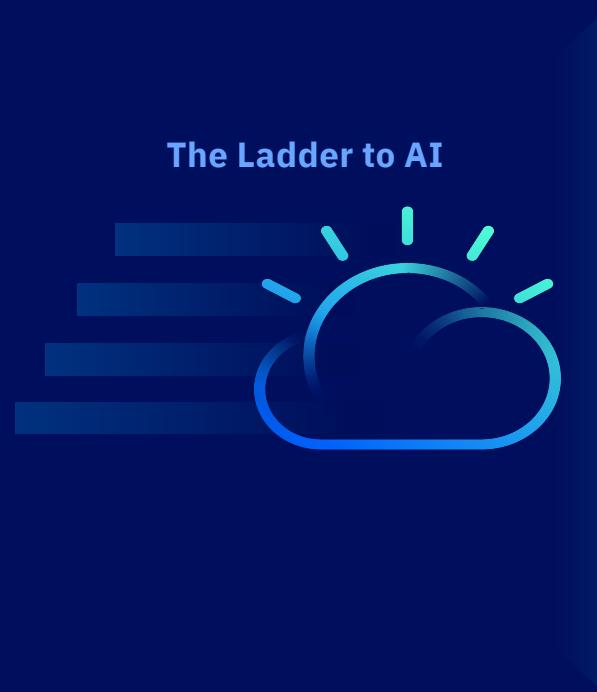
4 critical steps to operationalizing AI

Business-Ready data



IBM's AI Portfolio

Everything you need for Enterprise AI, on any cloud



IBM Watson Knowledge Catalog

An intelligent, integrated enterprise data catalog and data governance platform



IBM Cloud

Watson Knowledge Catalog: Search and Explore

Catalogs / Enterprise

Add to Catalog

Browse Assets Access Control Settings

What assets are you looking for?

Watson Recommends Highly Rated Recently Added Expand

Filter Asset types Tags

Available Assets

Showing 25 of 53 assets

NAME	OWNER
2017 J.D. Power U.S. Auto Clai...	
Auto Insurance Claim	
Auto Insurance Claims	
Auto Insurance Customers	
Auto Insurance Policies	

Browse Assets Access Control Settings

What assets are you looking for?

Watson Recommends Highly Rated Recently Added

Filter Available Assets

Owner: Dirk deRoos
Added: Sep 12, 2018 9:32 AM
Insurance | Demo
★★★★★ 0 reviews

Owner: Ricardo Buglio
Added: Feb 20, 2018 2:44 PM
Great-... | Sales | Retail
★★★★★ 0 reviews

Owner: Ricardo Buglio
Added: Oct 31, 2017 9:22 AM
Prince... | Sales
★★★★★ 0 reviews

Watson Knowledge Catalog: Glossary Terms

Business Glossary

What are you looking for? Sort by: Term name

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Confidential Information
Data that is classified as confidential
Confidential
Last modified: Jan 2, 2018

Personally Identifiable Information
Data that is classified as Personally Identifiable Information (PII).
PII
Last modified: Apr 23, 2018

Sensitive Information
Sensitive information that is protected against unwarranted disclosure.
Sensitive
Last modified: Jan 10, 2019

Business Glossary / Sensitive Information

Overview Related content

Sensitive Information
Sensitive information that is protected against unwarranted disclosure.

Term details
Creator: Ricardo Buglio
Date created: Oct 31, 2017
Last editor: Ricardo Buglio
Last modified: Jan 10, 2019

Owner
Ricardo Buglio

Tags
Sensitive

Associated classifier or term
Sensitive Personal Information

Description

Sensitive information is data that must be protected from unauthorized access to safeguard the individual or organization. There are different types of sensitive information: Personal identifiable information (PII) is data that can be traced back to an individual and that, if disclosed, could compromise their privacy. Such information includes biometric data, medical information, personally identifiable (PII) and unique identifiers such as passport or Social Security numbers. Threats include not only theft but also disclosure of personal information that the individual would prefer remained private. Business information - Sensitive business information includes financial information, acquisition plans, financial data and supplier and customer information, among other possibilities. The amount of data generated by businesses, methods of protecting corporate information from unauthorized access and the consequences of a data breach are becoming integral parts of business strategy.

Show more ▾



Business Glossary / Sensitive Information

Overview Related content

Sensitive Information

Related content by type

Related content

Assets (9) View all

ASSET NAME	CATALOG NAME	LAST MODIFIED
Auto Insurance Customers		Mar 30, 2018
CUSTOMERS	Enterprise	Nov 9, 2018
CUSTOMER_DEMOGRAPHIC	Enterprise	Jan 10, 2019
Insurance Driver	Knowledge Catalog	Nov 16, 2018

Watson Knowledge Catalog: Data Rules and Policies

Policy Manager / Data Privacy

+ Add ▾

Category

Data Privacy

Description

Enterprise data privacy policies pertaining to data classified as Sensitive, Classified or Personally Identifiable Information that have to be protected and comply with government and industry regulations.

Creator: Ricardo Buglio Created: Jul 31, 2017 Contains: 3 published policies and 4 rules

What are you looking for?

NAME	TYPE	STATUS	CONTAINS	LAST MODIFIED	⋮
▼ Sensitive Information	Policy	Published	2 Items	Jan 10, 2019	⋮
Anonymize Government Identity Information	Rule	Published	--	Jan 10, 2019	⋮
Anonymize Financial Account Information	Rule				Data Dashboard
➤ Confidential Information	Policy				
➤ Personally Identifiable Information	Policy				

Data Dashboard

Policy enforcements over time ⓘ

Dec 16, 2018 - Jan 16, 2019



Action * ⓘ

then anonymize data

in columns containing

Government Identities ✖

Select how to anonymize data:

Redact

BEFORE
452-821-1120

Replace data with Xs.

Substitute

BEFORE
452-821-1120

Replace data with values that don't match the original format.

Mask

BEFORE
452-821-1120

Replace data with similarly formatted values.

[Visit the documentation](#) to learn more about data anonymization.

Data assets containing personal or restricted data

Confidential Information ⓘ

2

[View all](#)

Personally Identifiable Infor... ⓘ

2

[View all](#)

Sensitive Information ⓘ

9

[View all](#)

Operational policies ⓘ

3

Data policies

4

Data policy rules

Automatic enforcement ⓘ

230

enforcements in last 30 days

▼ 71.71%

from last month

Watson Knowledge Catalog: Review and Understand

Catalogs / Enterprise / CUSTOMER_DEMOGRAPHIC

+ Add to Catalog



Overview

Access

Review

Profile

Lineage

Data Asset

CUSTOMER_DEMOGRAPHIC

Remove Download

Add to Project

Description

Customer churn data asset

Added: Jan 10, 2019 11:09 AM

Format: CSV

Size: 322 KB

Tags

Customer | Churn

Reviews

1 review

Schema: 28 Columns | 2066 Rows | 2 Columns anonymized

Preview: 1000 rows | Last refresh: 1 day ago |

EMAIL_ADDRESS	PHONE_NUMBER	NATIONAL_ID	CREDITCARD_NUMBER	CREDITCARD_TYPE	CREDITCARD_EXP
Type: String	Type: String	Type: String	Type: String	Type: String	Type: String
Email Address	US Phone Nu...	US Social Se...	Credit Card Number	Organization Na...	Date
wfronsek1@source...	434-553-8337	231-24-6500	7df7ea4aac5e3493ea3127e	VISA	5/23
dcoyejk@pcworld.co...	760-277-6466	649-96-7557	db7d354d8a0d2042c969dd	Diners Club	3/21

Catalogs / Enterprise / CUSTOMER_DEMOGRAPHIC

+ Add to Catalog



Overview

Access

Review

Profile

Lineage

Current profile

Last profile

Columns 28

Rows 2,066

Delete

Update Profile

ID

CUSTOMER

NAME

COUNTRY

Type: Smallint

Type: Varchar

Type: Varchar

Type: Varchar

I

I

T

CC

FREQUENCY

FREQUENCY

FREQUENCY

FREQUENCY

1927 - 1954

OH15677

EZ71780

CB44654

1491 - 1518

BP28492

GE50763

QT25383

371 - 397

BP28492

Job Trenera

Dalston Lamberton

1899 - 1925

GE50763

Nero Martell

Bernadette Jeffrey

1222 - 1248

AU96286

Dulcine Leap

Marie Machen

529 - 555

JV89183

Tiphany Samter

Gerhardine Brockhouse

1411 - 1437

FH83537

Terrye Dunleavy

Job Trenera

1873 - 1898

503 - 528

1927 - 1954

Ohio

503 - 528

OH15677

Ohio

Ohio

503 - 528

OH15677

Ohio

Ohio

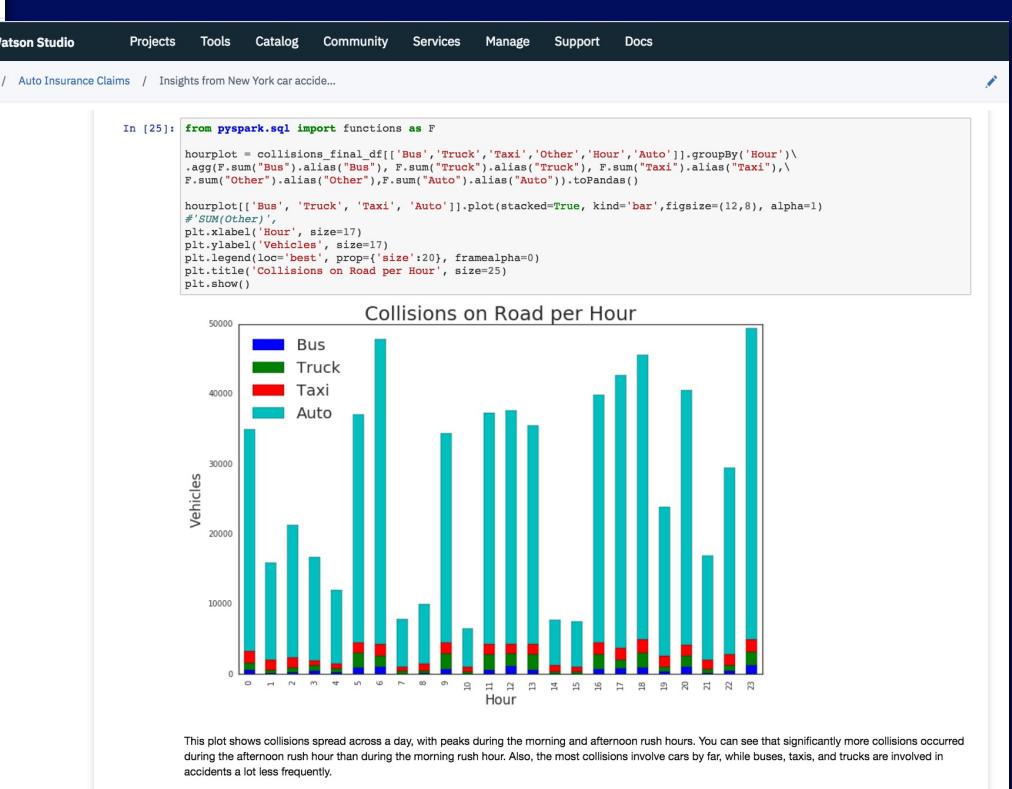
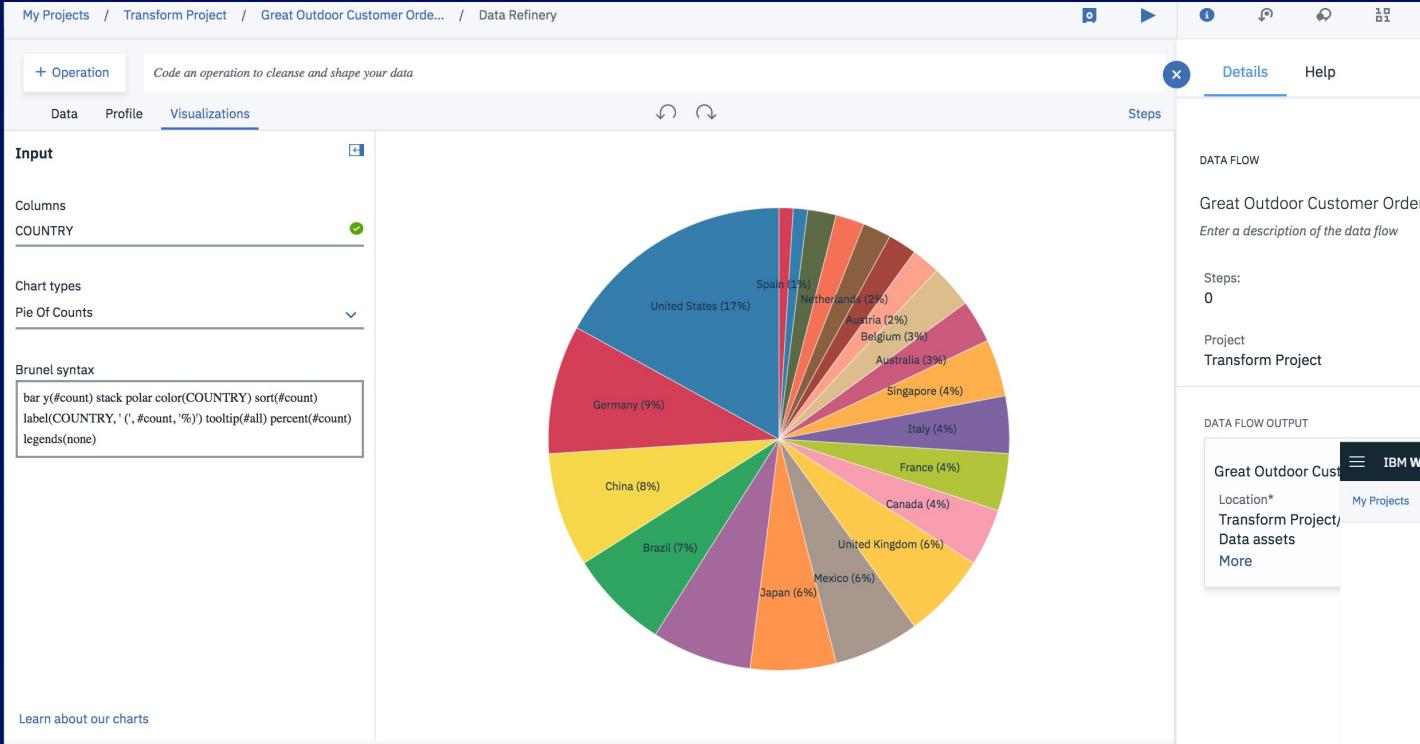
Schema: 28 Columns | 2 Columns anonymized

Preview: 1000 rows | Last refresh: 4 hours ago |

Refine

STATE	STATE_CODE	ZIP_CODE	EMAIL_ADDRESS	PHONE_NUMBER	NATIONAL_ID	CREDITCARD_NUMBER	CREDITCARD_TYPE	CREDITCARD_EXP	CREDITCARD_C
Ohio	OH	44646	wfronsek1@source...	434-553-8337	231-24-6500	7df7ea4aac5e3493ea3127e	VISA	5/23	4952
Washington	WA	98373	dcoyejk@pcworld.co...	760-277-6466	649-96-7557	db7d354d8a0d2042c969dd	Diners Club	3/21	2218
California	CA	91604	kwaggatdu@typepe...	915-586-6081	962-90-5618	c6656b4696afb2bcd045	American Express	8/19	1600
New York	NY	11570	vcokerky@shareasal...	817-205-8994	613-45-8368	2802c0eabcf2fec94dbc9	American Express	12/23	7608
Oregon	OR	97214	ppotegi@cnet.com	901-313-6753	873-61-4046	e3c9064b11ac1a8a446fd	Discover	4/23	5886
Virginia	VA	23220	lmebs68@mozilla.o...	318-710-5442	950-96-4106	11776326fed27128f89ac	JCB	7/20	4901
Colorado	CO	80021	cburcher1o@tuttoci...	214-501-0431	708-76-8249	81471b1239519e9b69c8	VISA	7/20	9952
Ohio	OH	43147	tgeeves9w@nsw.go...	212-160-2937	805-98-0233	738f72ee7485cdcd47a3c	JCB	10/19	6027
California	CA	94541	ovincentqy@weibo.i...	520-774-2490	061-31-7031	978312a843698d357512	JCB	12/22	8397
Ohio	OH	43035	ptythertonix@phoc...	608-176-4288	353-89-2233	5af98a2efda868e10949	Diners Club	12/22	9587

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 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.

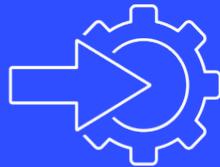
Problem #5: Business and Software teams not in sync in a project

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 your data to build an automatic level trust in your data.



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.



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.



AI Governance

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

Automate the mundane

Use the data to get the data

Activate your governance program

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:

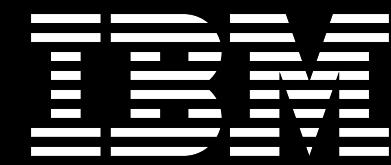
<https://www.ibm.com/cloud/garage/dte/tutorial/fuel-data-innovation-and-ai-ibmr-watson-knowledge-catalog>

<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



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

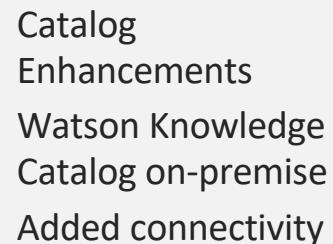
- Knowledge Worker Experience
- Catalog Enhancements
- Watson Knowledge Catalog on-premise
- Added connectivity

- Governance User Experience
- Reference Data management
- Workflows

- Data Stewardship Experience
- Data Rules and Quality Integration
- Operational and Business Lineage

- Administration Experience
- Advanced support for (generic) import
- Enriched Reporting and Dashboards

- Versioning and Auditing



Q3 2019

Q4 2019

Q1 2019

Q2 2019

Beyond

Advanced Workflow

Q2 2019

Asset customization Workflow configuration Import glossary assets

- Review the available workflows and enable one of them. Only one workflow can be enabled at a time. Configuration steps are required after enabling a workflow. Open the details page of a workflow to see the details.

Workflows

Name
Governance workflow
Governance workflow with email notification

IBM Information Server Catalog Workflow Connections Monitoring Help

Tasks

OPEN TASKS COMPLETED TASKS Sort by: Recently opened

Send term New Term for approval
 No due date
3 HOURS AGO

Send information governance rule Ne...
 No due date
3 HOURS AGO

Send information governance policy ...
 No due date
3 HOURS AGO

Send term New Term for approval
Assigned to you +2 persons | Claim task

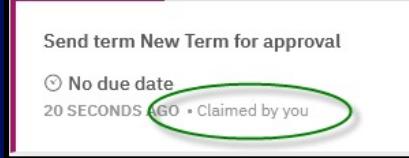
OTHER TASK ASSIGNEES
 UE user1 Editor
 UX user5 xAllRoles

Instructions
This term was created for you to review the changes, and send them for approval, or...

TERM: New Term STATUS: DRAFT See Details

UX user5, xAllRoles

Write a comment...



Advanced Workflow

Q2 2019

IBM Information Governance

Catalog Tasks Tickets Monitoring Management Help XY

Ticket #312: New Business Term (Test Term)

Open Tickets > Ticket #312

Progress
Steps 2/3

Steps

Sent for Approval Pending Approval

Tasks

<input checked="" type="checkbox"/> Sent for Approval	Completed Feb 8, 2018
<input type="checkbox"/> Pending Approval	You're assigned
	Due Feb 14, 2018

Activity Details

+ See 5 older activities expand ▾

XY brittanderson approved draft Test Term 2 days ago.

XY brittanderson Commented on Friday, Feb 8, 2018 I approve.

XY brittanderson set due date for Last Step: Publishing to 02/14/2018

XY username Write comment...

Comment

IBM Information Governance

Catalog Tasks Tickets Monitoring Management Help XY

Ticket #312: New Business Term (Test Term)

Open Tickets > Ticket #312

Progress
Steps 2/3

Last Activity

+ See 5 older activities

XY brittanderson approved draft Test Term 2 days ago.

XY brittanderson Commented on Friday, Feb 8, 2018 I approve.

XY brittanderson set due date for Last Step: Publishing to 02/14/2018

XY username Write comment...

Comment

People Details

XY username X

XY username X

+ Add new Collaborator

Tasks

<input checked="" type="checkbox"/> Step 1 Send for approval	Completed Feb 6, 2018
<input checked="" type="checkbox"/> Step 2 Approved	Completed Feb 8, 2018

Reference Data

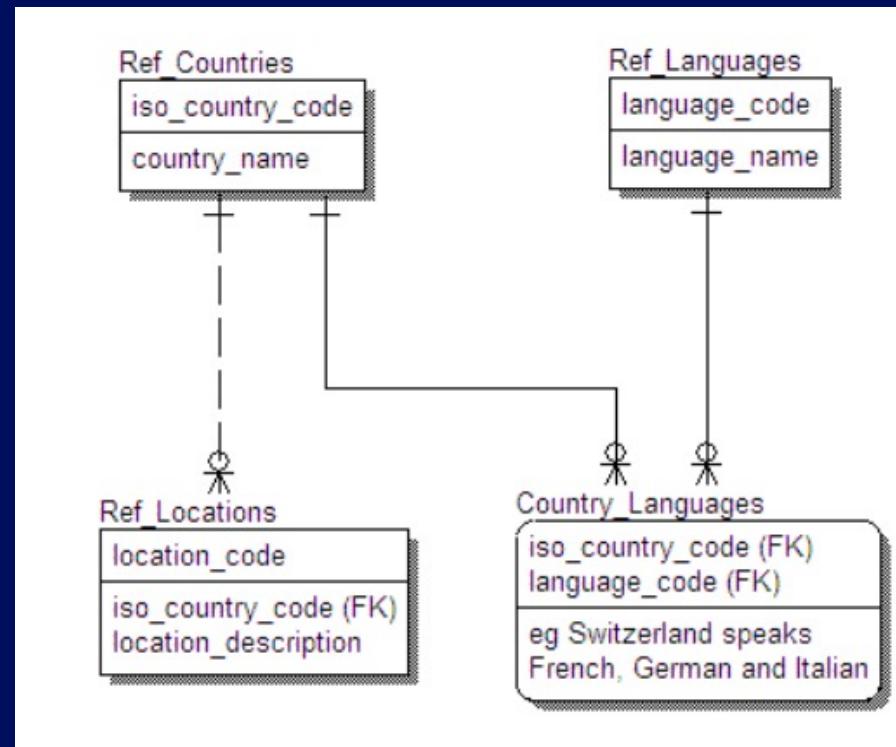
Q2 2019

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



IBM Information Server Catalog Connections Monitoring Quality Management Help AI

Workspaces Data rules Data sets Discover Analysis Report Exceptions Select a workspace

DataLakeWorkspace
Workspace with optimal settings to run a quick analysis.

Owner: isadmin
Datasets: 5 Analyzed: 0%
Low quality: 0 Data rules: 0

InDepthAnalysisWorkspace
Workspace with optimal settings to run a detailed analysis.

Owner: isadmin
Datasets: 4 Analyzed: 0%
Low quality: 0 Data rules: 0

PIIWorkspace
TestWorkspace

IBM Information Server Catalog Connections Monitoring Quality Management

Workspaces Data rules Data sets Discover Analysis Report Exceptions DataLakeWorkspace

Rule definition

Name	Description	Status	Created by	Created on	Terms	Actions
All	Global category for all QualityComponents			1/11/2019, 5:45:53 PM		
Published Rules						
01 Personal Identity				1/11/2019, 5:44:23 PM		
Age				1/11/2019, 5:44:23 PM		
ChildInRangeString	String data Child: Age >= 0 and < 18; applied to string age data	ACCEPTED	isadmin	1/11/2019, 5:44:24 PM		
AdultInRangeNumeric	Adult: Age >= 18 and < 125; applied to numeric age data	ACCEPTED	isadmin	1/11/2019, 5:44:23 PM		
AgeInRangeNumeric	Age: Age >= 0 and < 125; applied to numeric age data	ACCEPTED	isadmin	1/11/2019, 5:44:22 PM		
AgeInRangeString	String data Age: Age >= 0 and < 125; applied to string age data	ACCEPTED	isadmin	1/11/2019, 5:44:23 PM		
AdultInRangeCalc	Derived Age Adult: Age >= 18 and < 125; applied to derived age calculated as the absolute value of the difference between the current date and date of birth	ACCEPTED	isadmin	1/11/2019, 5:44:24 PM		
ChildInRangeCalc	Derived Age Child: Age >= 0 and < 18; applied to derived age calculated as the absolute value of the difference between the current date and date of birth	ACCEPTED	isadmin	1/11/2019, 5:44:25 PM		⋮
ChildNotMarriedString	If Child (string) then Marital Status = 'N'; applied to string age data	ACCEPTED	isadmin	1/11/2019, 5:44:25 PM		

Workspace and Analysis

Q3 2019

The screenshot shows the IBM Information Server Catalog interface. At the top, there's a navigation bar with links for Catalog, Workflow, Connections, Monitoring, Quality, Management, Help, and AI. Below the navigation bar, there are tabs for Assets, Hierarchies, Queries, Collections, and Containers. The main area is titled "Explore the catalog" and features a search bar with "All asset types" dropdown and a "Search the data catalog" input field. A "Create" button with a dropdown arrow is also present. On the left, a sidebar titled "My recently viewed assets" lists an "Acronym" asset with details: Created by Administrator IIS, Created on 11 January 2019, 7:04:27 pm, Modified by Administrator IIS, and Modified on 11 January 2019, 7:04:27 pm. The main content area displays the "account number" asset details. It includes a "Rule Logic" section with the following logic:

```
if [the asset] has the term Business Information/Customer Information/Custom... assigned  
or [the asset] has the term Information Governance/Information Governance Cl... assigned  
then bind the data rule definition IdInValidRange  
bind the data rule definition FieldExists  
bind the data rule definition FieldExists  
bind the data rule definition IdentifierUnique
```

Business Lineage

Q3 2019

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.

IBM Information Server

Catalog Workflow Connections Monitoring Quality Management Help AI

Assets Hierarchies Queries Collections Containers

All results
3 results

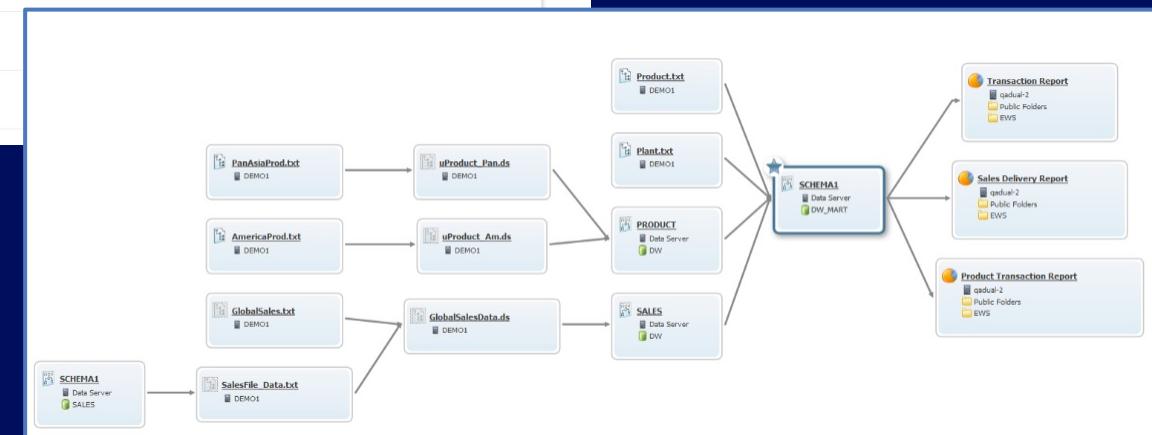
Show All assets

Filter results
Clear all filters

Asset types (1) +

Advanced filters +

Name	Description	Labels	Last activity
CRM System			Created by isadmin on Jan 28, 2019, 3:28 PM
Reporting Area	Downstream Reports		
Warehouse Schema			



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.

The screenshot shows the 'Catalog Permissions' page in the IBM Infosphere Information Governance Catalog - Administration. The left sidebar lists categories: Address Example, andrewwtest, Business Information (Calendar Information, Customer Information), Location Information, Organizational Information (Payment Card Information), Personal Information (Transaction Information), Category test, Industry Models Personal Infor, Information Governance (kk_cat, mateo category, Test Demo Category). The main panel shows a list of users under 'CATEGORY: CUSTOMER INFORMATION'. The 'Only selected users and groups have access.' option is selected. The list includes:

User	Role	User Name
Administrator IIS	Administrator	isadmin
Arron La	Author	arron
Asset Assigner	Assigner	AssetAssigner
Asset Author	Author	AssetAuthor
Basic User	User	BasicUser
biz_steward_1	Steward	biz_steward_1