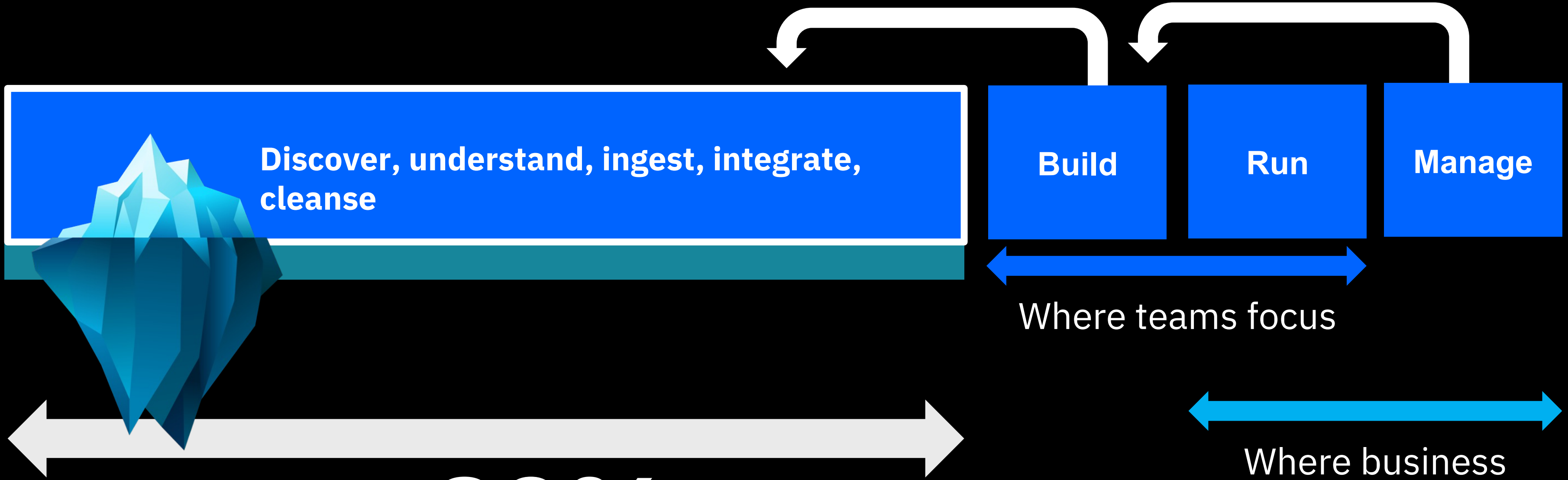
Data Lifecycle

Prepared for ADP – August 2022

Nigel Jones, David Radley, Sepideh Seifzadeh, Lena Woolf

Getting Data to your AI Initiatives is Hard



Discover, understand, ingest, integrate, cleanse

Build

Run

Manage

Where teams focus

Where business impact is created

Where **80%** of time is spent

*Source: Data scientist report, Figure Eight Inc data production

data consumption

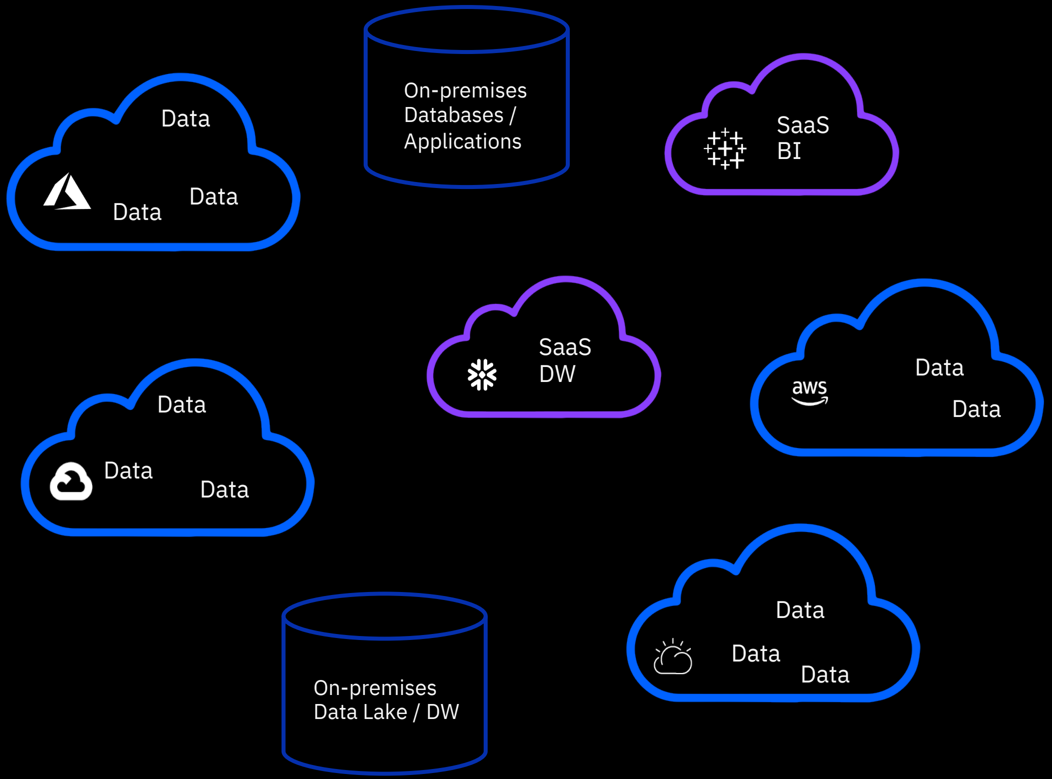
The three myths of cloud modernization

Myth #1: The cloud will simplify my landscape

Myth #2: The cloud will eliminate data silos

Myth #3: The cloud “takes care” of governance / compliance

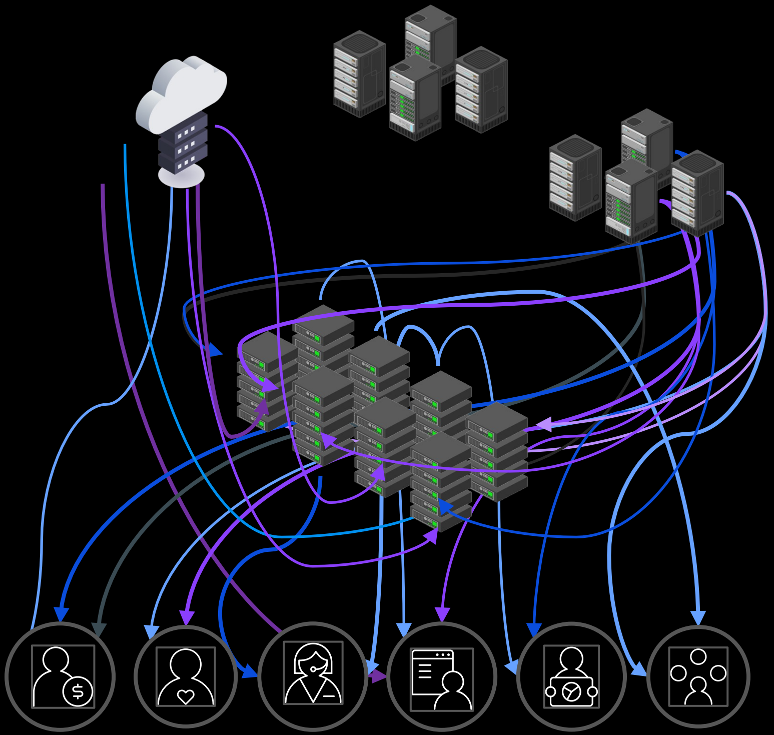
The landscape will not be less complex with cloud



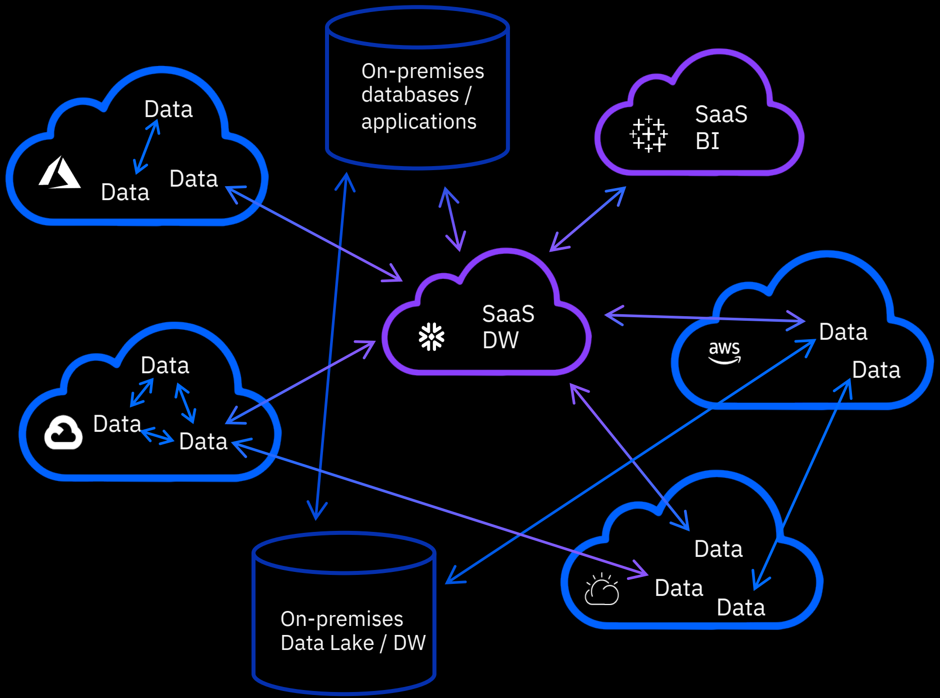
1. The move to cloud is a move to multiple cloud platforms
2. Many on-premises systems will still be around for years to come
3. Point SaaS solutions will continue to expand the footprint of tools and applications

Data silos will not go away with cloud

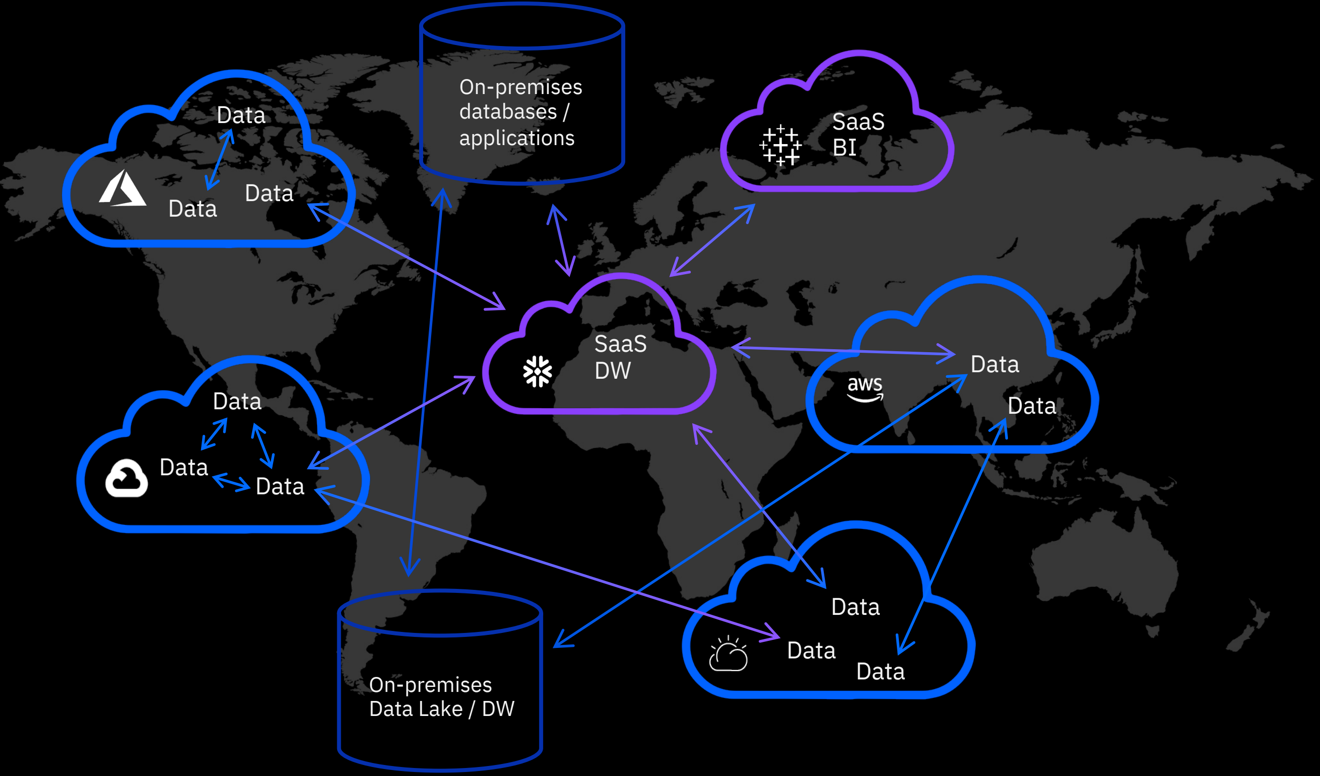
Past



Future



Cloud will not ”take care” of global data governance / compliance

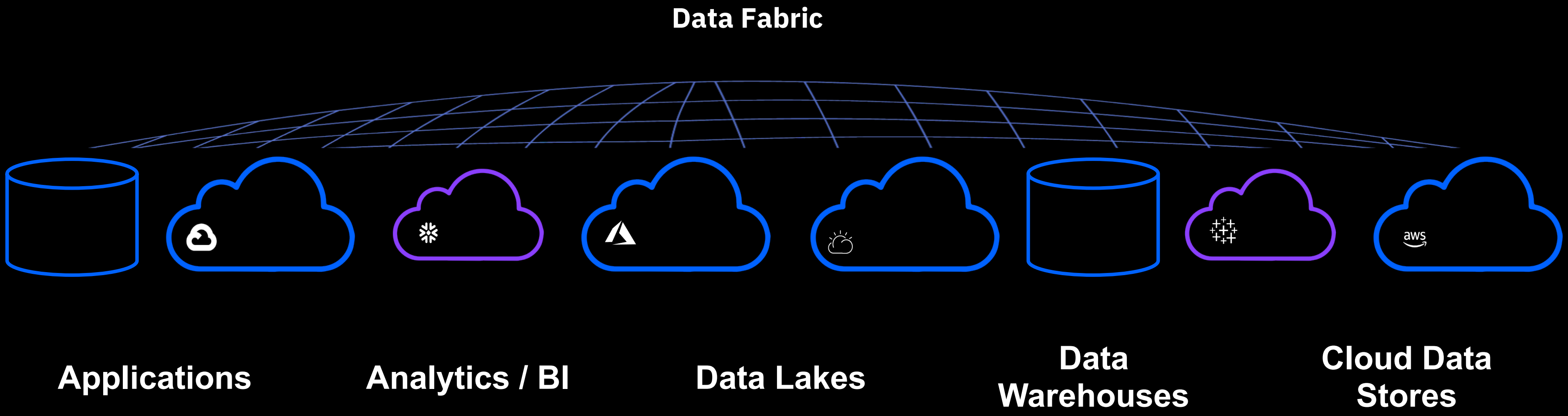


Data
Data
Data
On-premises databases / applications
SaaS BI
SaaS DW
Data
Data
Data
aws
Data
Data
Data
Data
Data
On-premises Data Lake / DW

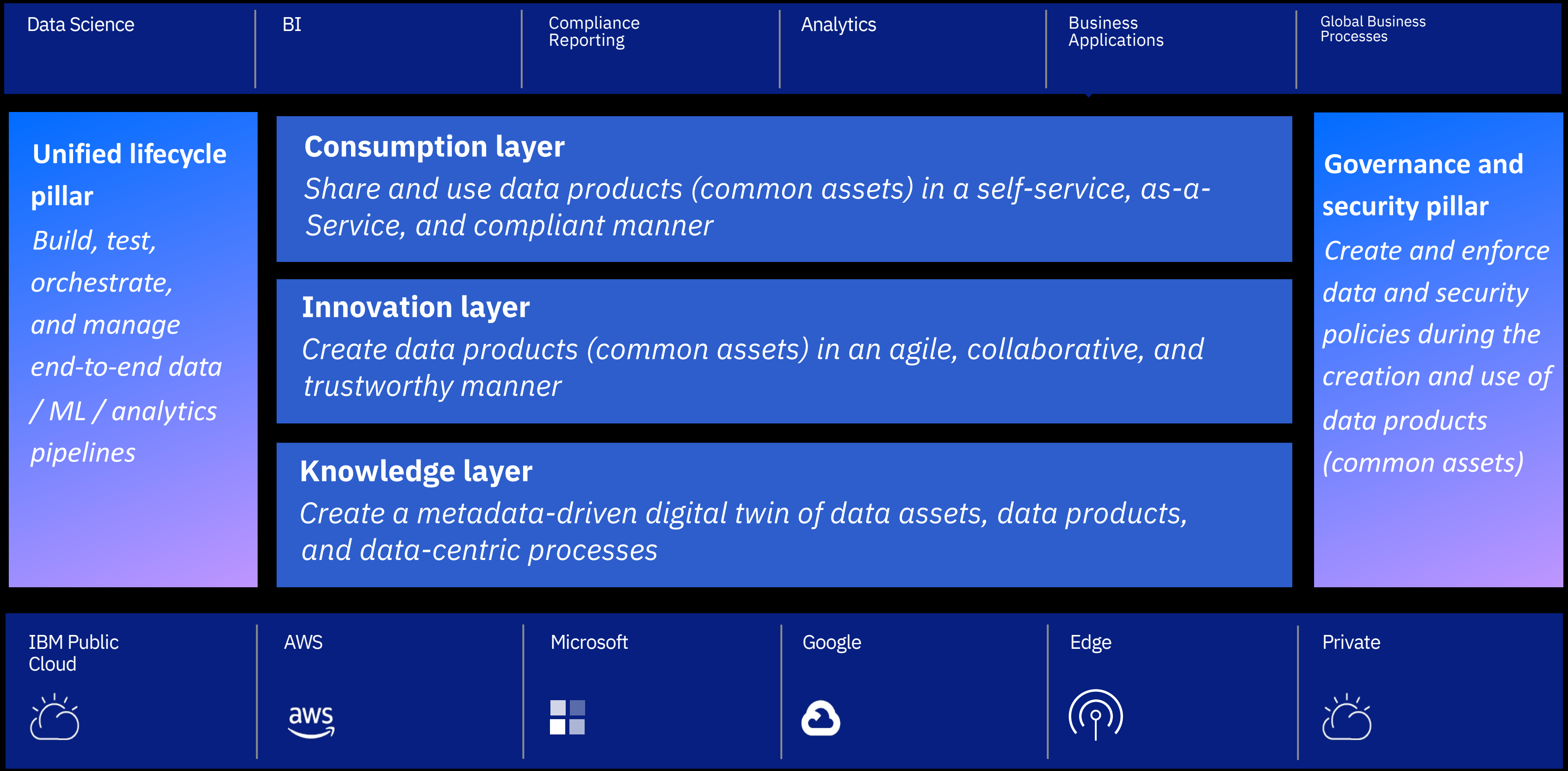
*Global data governance and compliance will require a **hybrid multi-cloud** solution*

The Data Fabric enables a hybrid multi cloud data architecture

The Data Fabric intelligently and automatically connects the right data, at the right time, to the right people, with appropriate governance.



IBM Offering (Data Fabric) conceptual vision



Data Science

BI

Compliance Reporting

Analytics

Business Applications

Global Business Processes

Unified lifecycle pillar

Build, test, orchestrate, and manage end-to-end data / ML / analytics pipelines

Consumption layer

Share and use data products (common assets) in a self-service, as-a-Service, and compliant manner

Innovation layer

Create data products (common assets) in an agile, collaborative, and trustworthy manner

Knowledge layer

Create a metadata-driven digital twin of data assets, data products, and data-centric processes

Governance and security pillar

Create and enforce data and security policies during the creation and use of data products (common assets)

IBM Public Cloud

AWS

Microsoft

Google

Edge

Private

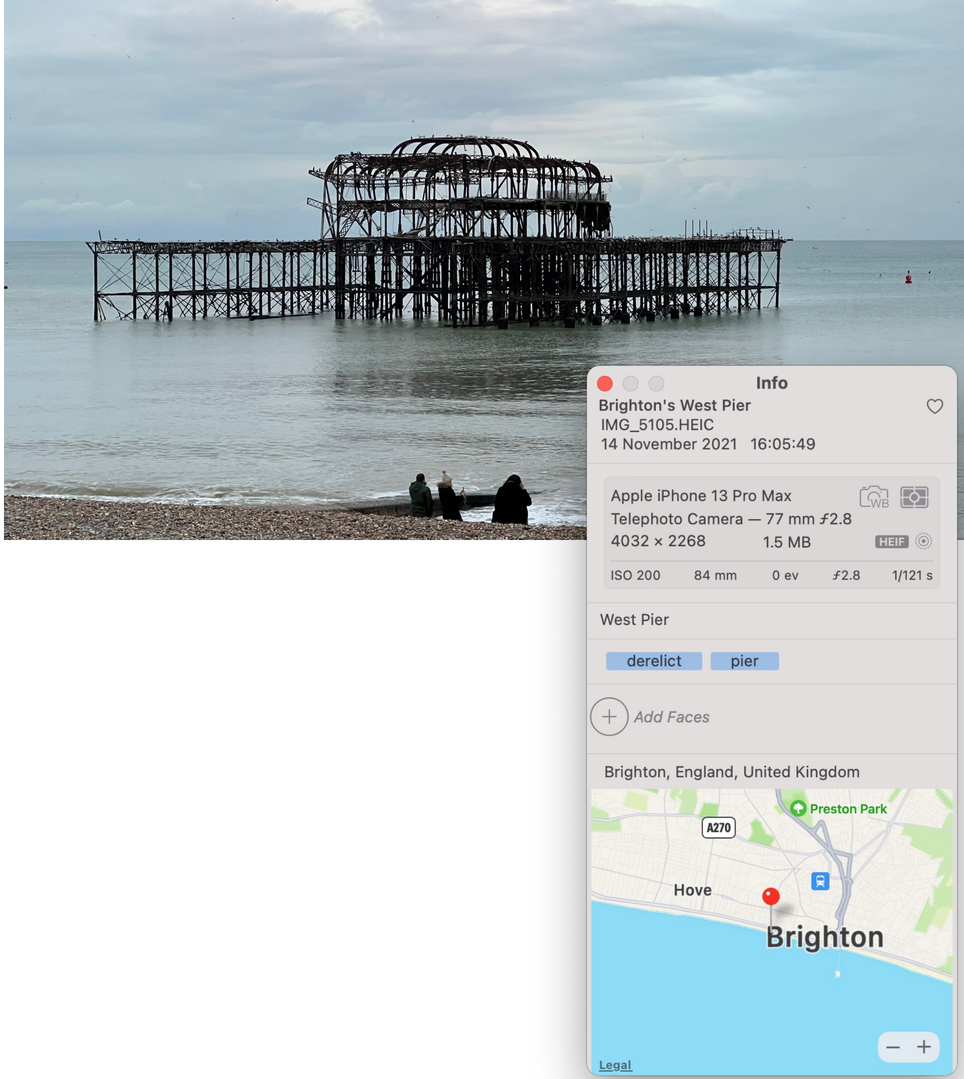
Why do we need metadata?

- Metadata enables data to be used outside of the application that created it.
  - Analytics and decision making
  - New business applications
  - Reporting and compliance
- Metadata describes the format and content of data allowing people to judge which data set to use for a new project
  - Structure
  - Meaning
  - Origin
  - Valid values and quality
  - Usage and ownership
  - Regulations and classifications that apply
  - <more>
- Metadata describes the business context and classification of data allowing automated governance processes to operate.

What is metadata?

- Where was it taken?
- When was it taken?
- What device was used?
- What settings were used?
- Photographers labels, title

-> Metadata adds context



Info

Brighton's West Pier
IMG_5105.HEIC
14 November 2021 16:05:49

Apple iPhone 13 Pro Max
Telephoto Camera — 77 mm ƒ2.8
4032 × 2268 1.5 MB HEIF

ISO 200 84 mm 0 ev ƒ2.8 1/121 s

West Pier

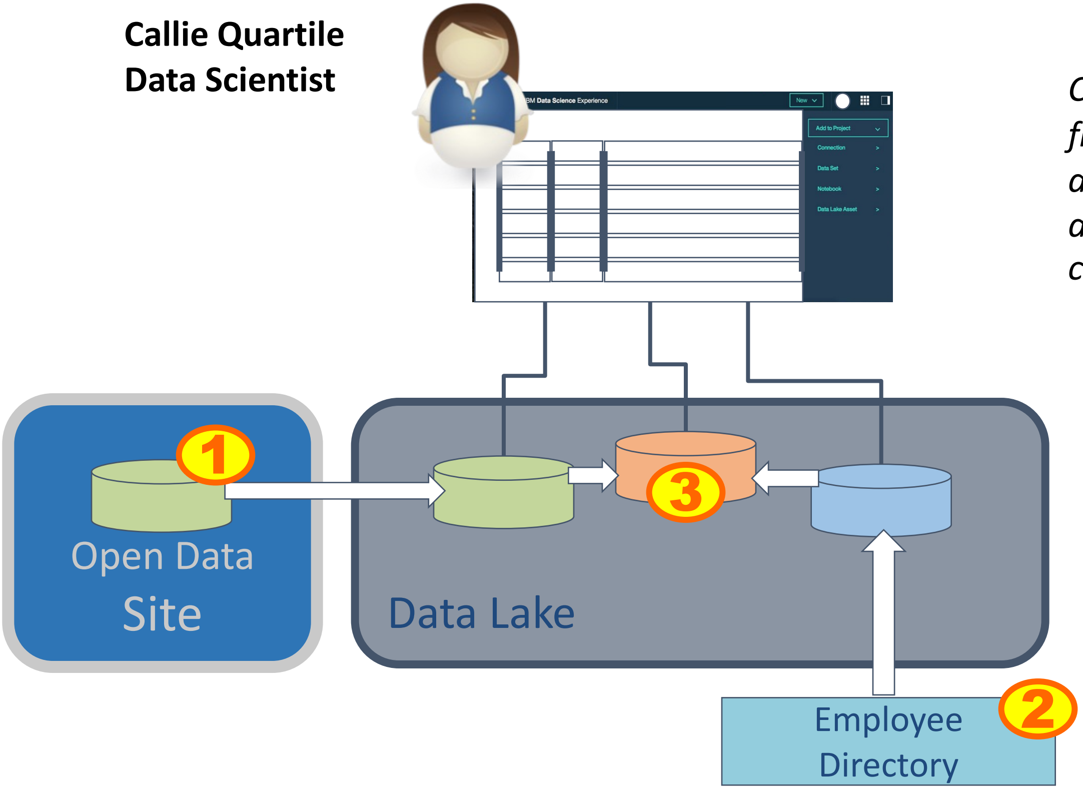
derelict pier

Add Faces

Brighton, England, United Kingdom

The perils of reusing data …

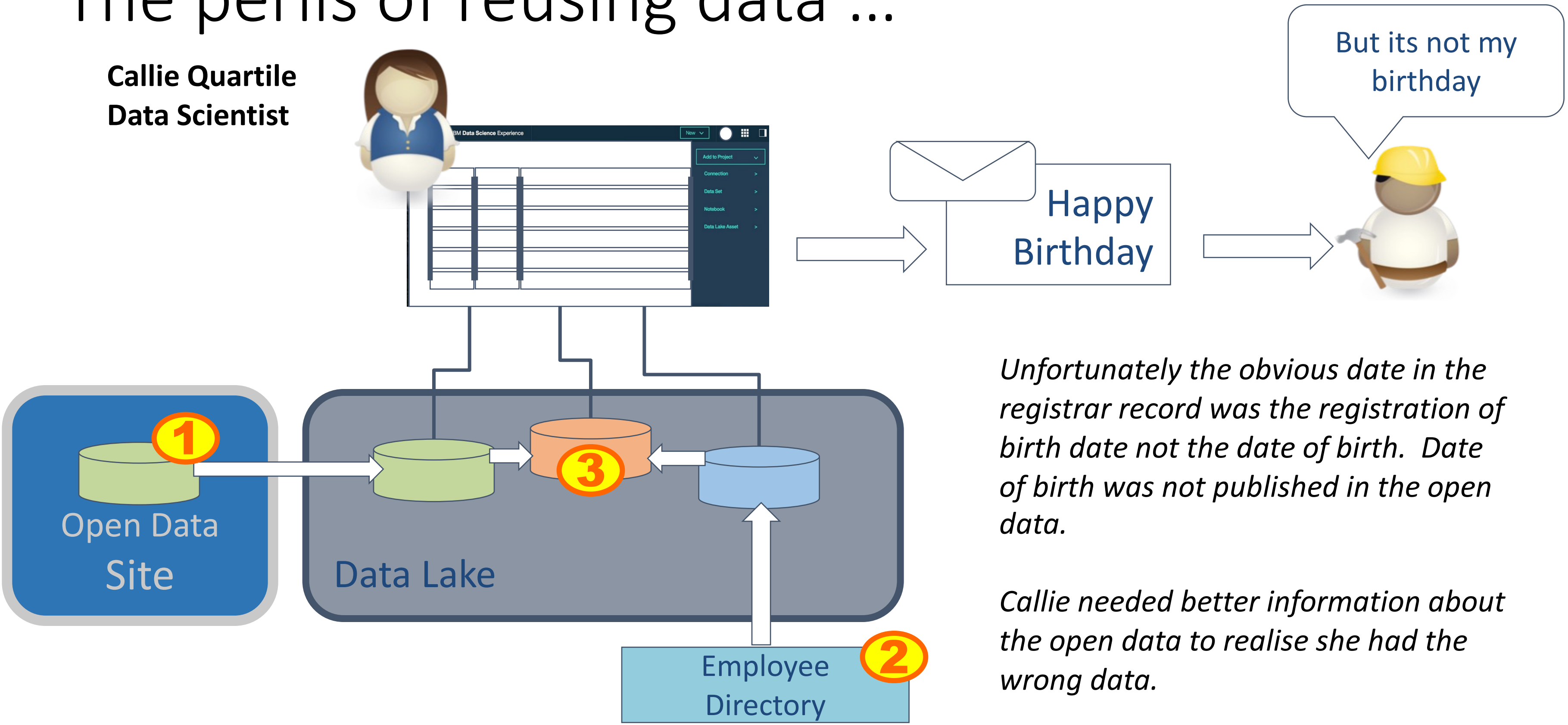
Callie Quartile
Data Scientist



Callie Quartile uses (1) open data from the local government registrar and (2) data from the employee directory to (3) create a birthday card service for the company.

The perils of reusing data …

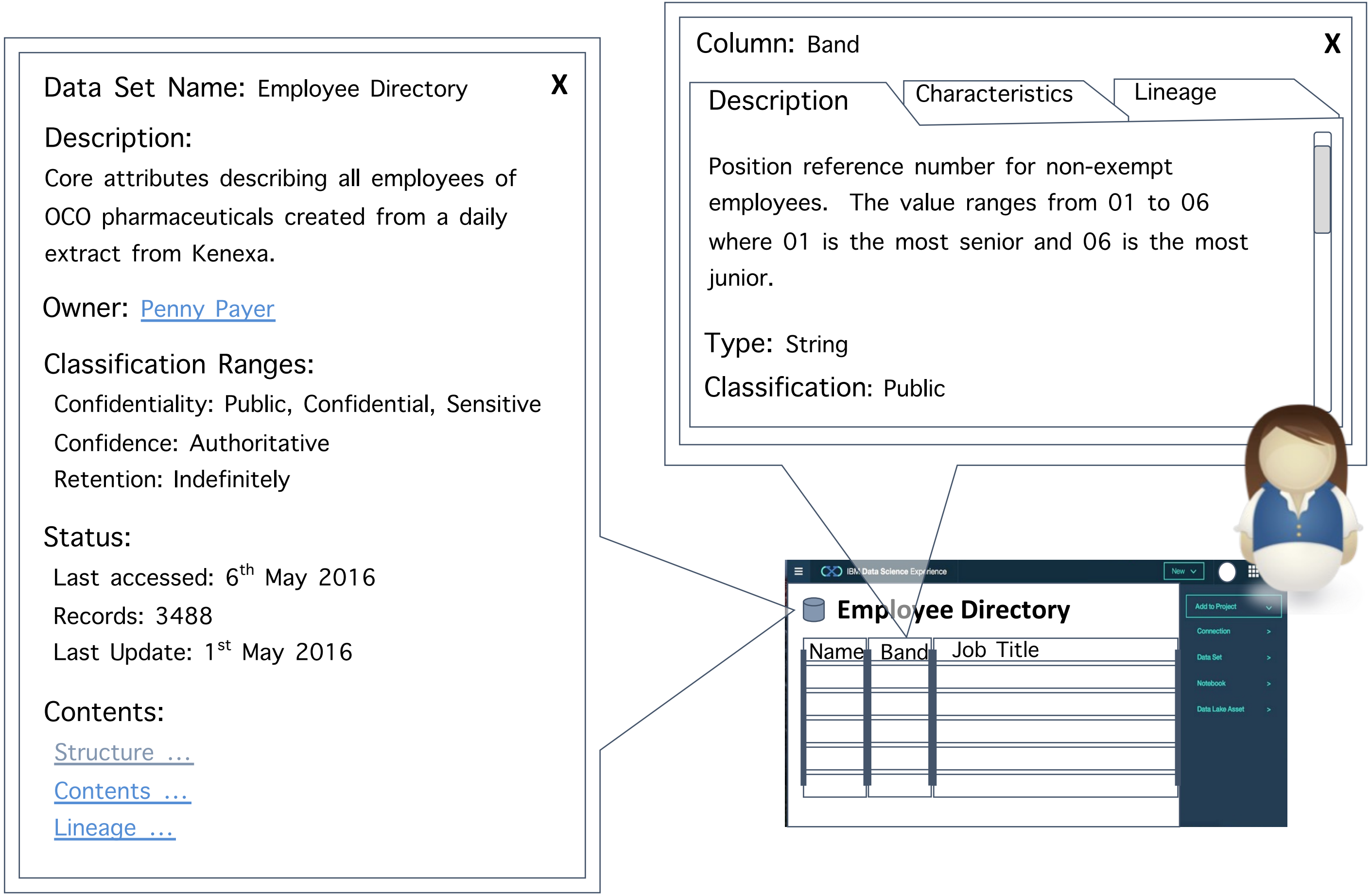
Callie Quartile
Data Scientist



Unfortunately the obvious date in the registrar record was the registration of birth date not the date of birth. Date of birth was not published in the open data.

Callie needed better information about the open data to realise she had the wrong data.

Metadata should bring as much information about the data sets to Callie's data science as is known collectively by the organization.



Data Set Name: Employee Directory X

Description:

Core attributes describing all employees of OCO pharmaceuticals created from a daily extract from Kenexa.

Owner: Penny Payer

Classification Ranges:

Confidentiality: Public, Confidential, Sensitive

Confidence: Authoritative

Retention: Indefinitely

Status:

Last accessed: 6th May 2016

Records: 3488

Last Update: 1st May 2016

Contents:

Structure …

Contents …

Lineage …

Column: Band X

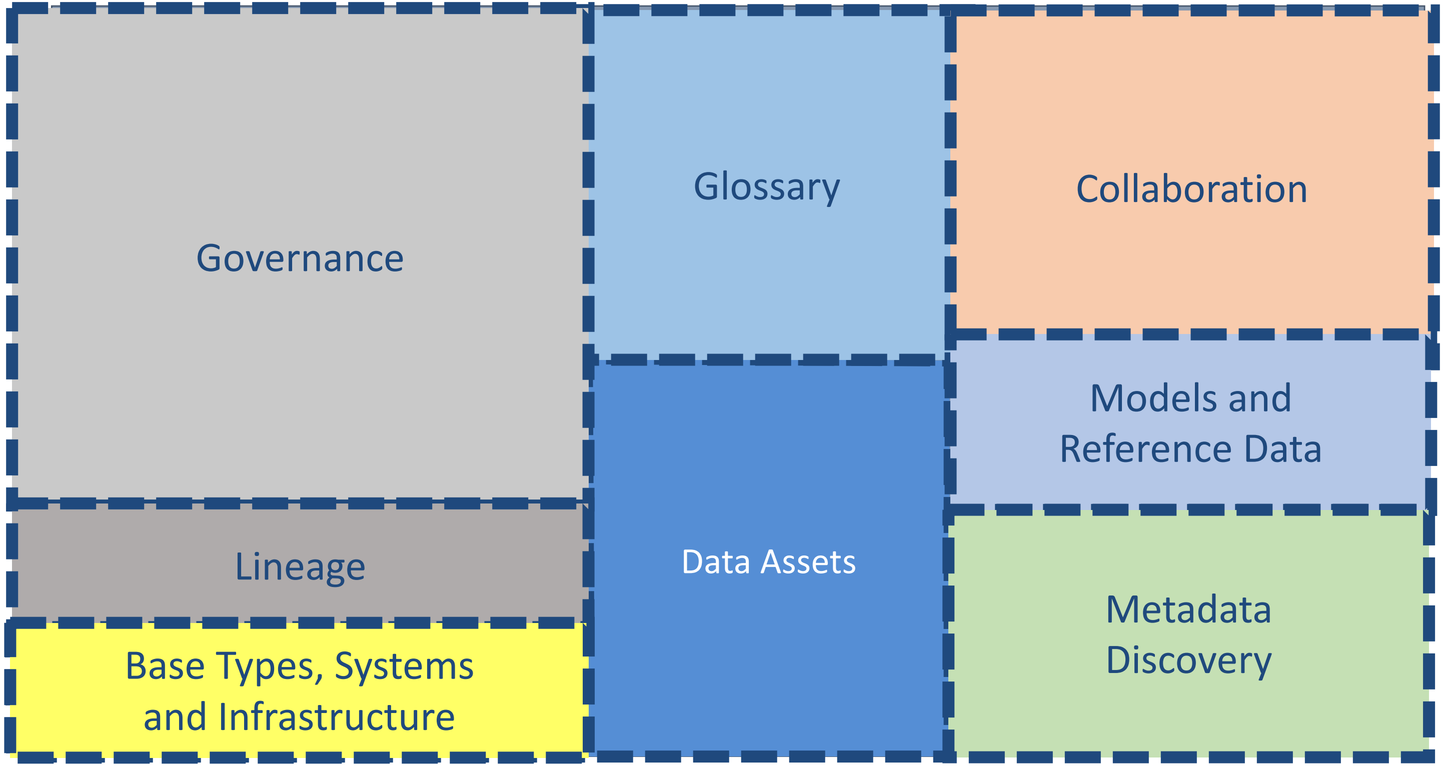
Description | Characteristics | Lineage

Position reference number for non-exempt employees. The value ranges from 01 to 06 where 01 is the most senior and 06 is the most junior.

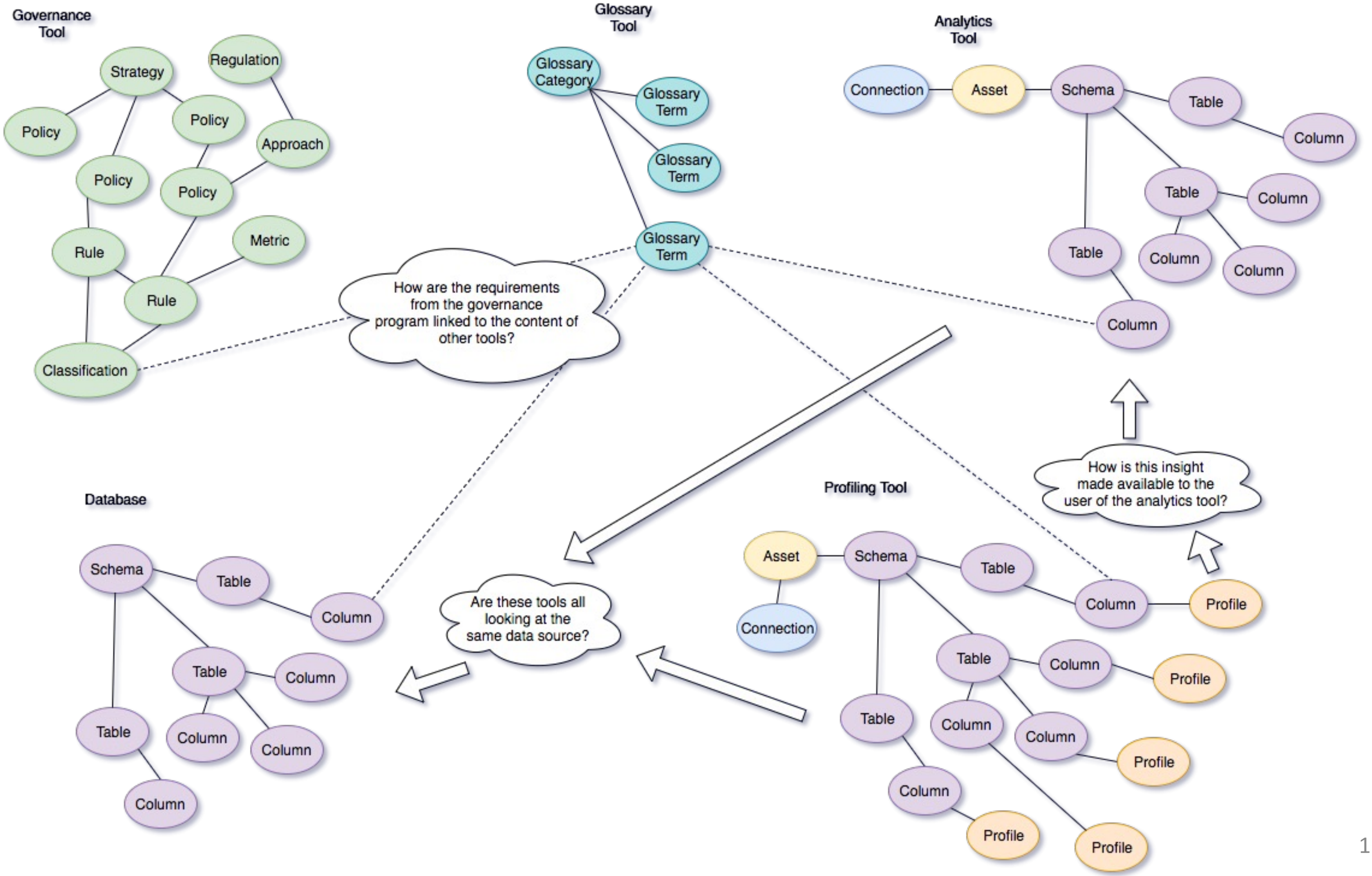
Type: String

Classification: Public

Scope of metadata for a data driven organization



Metadata linkage

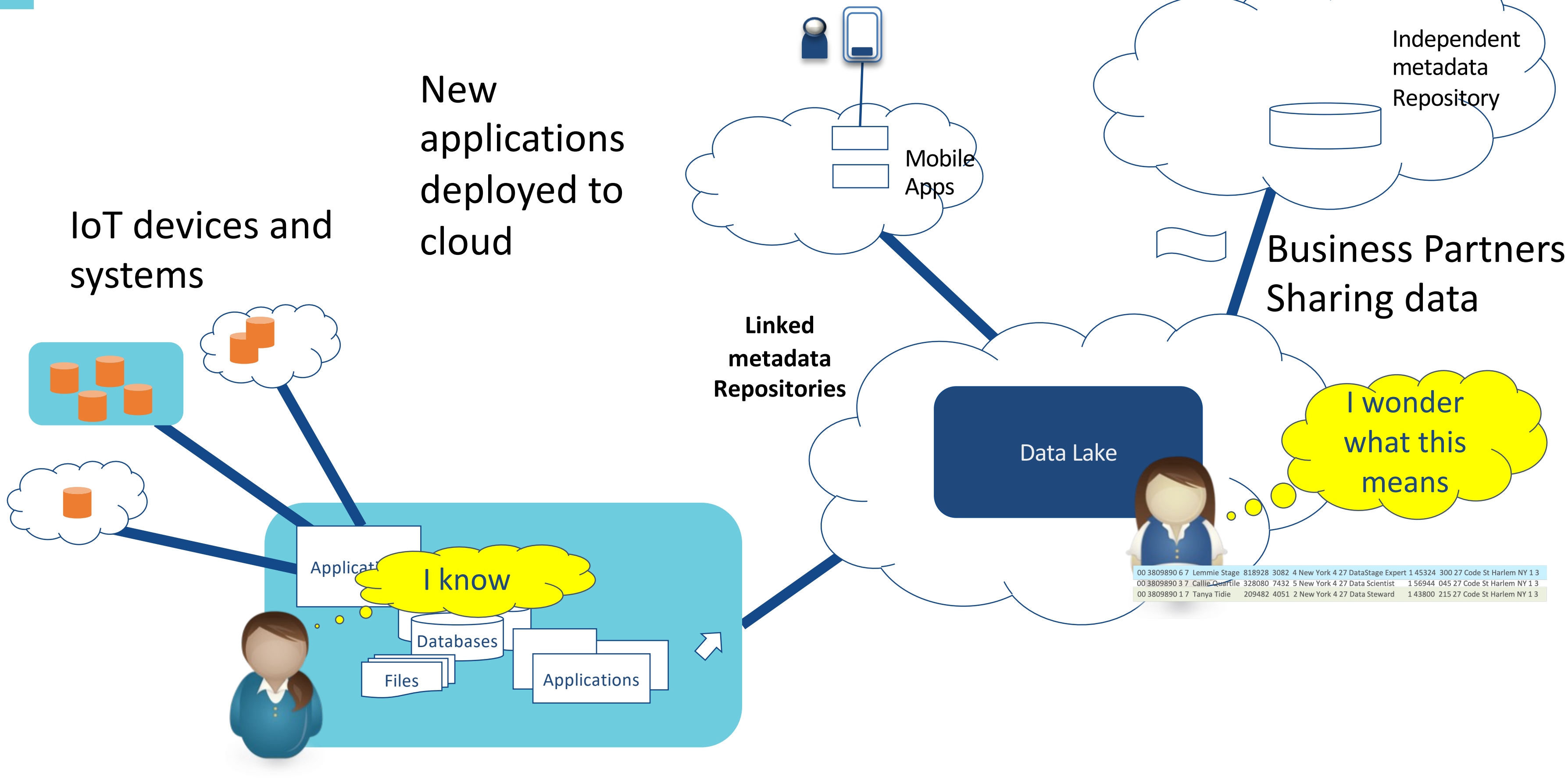


How are the requirements from the governance program linked to the content of other tools?

How is this insight made available to the user of the analytics tool?

Are these tools all looking at the same data source?

Knowledge is separated in space and time



IoT devices and systems

New applications deployed to cloud

Mobile Apps

Independent metadata Repository

Business Partners Sharing data

Linked metadata Repositories

Data Lake

I wonder what this means

Applicat

I know

Databases

Files

Applications

00 3809890 6 7	Lemmie Stage	818928	3082	4 New York 4 27 DataStage Expert	1 45324	300 27 Code St Harlem NY 1 3
00 3809890 3 7	Callie Quartile	328080	7432	5 New York 4 27 Data Scientist	1 56944	045 27 Code St Harlem NY 1 3
00 3809890 1 7	Tanya Tidie	209482	4051	2 New York 4 27 Data Steward	1 43800	215 27 Code St Harlem NY 1 3

Different personas need different services

Callie Quartile
Data Scientist

Find data
Understand data
Manage analytics models

Build data strategy
Define governance program
Monitor progress

Jules Keeper
Chief Data Officer

Different personas need different services

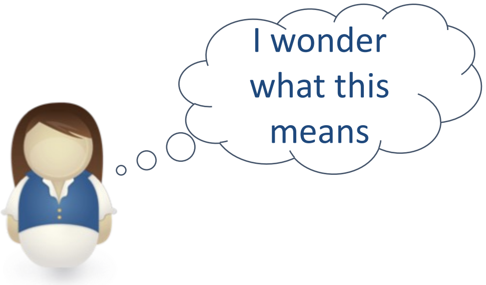
Tanya Tidie
Clinical Trials Administrator

Maintain accurate patient records
Catalog clinical trials data
Demonstrate good data management practices

Understand risks to organization
Set up protection
Monitor for suspicious activity

Ivor Padlock
Chief Security Officer

Curation



```
00 3809890 6 7  Lemmie Stage  818928  3082  4 New York 4 27 DataStage Expert 1 45324  300 27 Code St Harlem NY 1 3
00 3809890 3 7  Callie Quartile  328080  7432  5 New York 4 27 Data Scientist      1 56944  045 27 Code St Harlem NY 1 3
00 3809890 1 7  Tanya Tidie      209482  4051  2 New York 4 27 Data Steward       1 43800  215 27 Code St Harlem NY 1 3
```

Scared to share

Faith Broker
Business Team

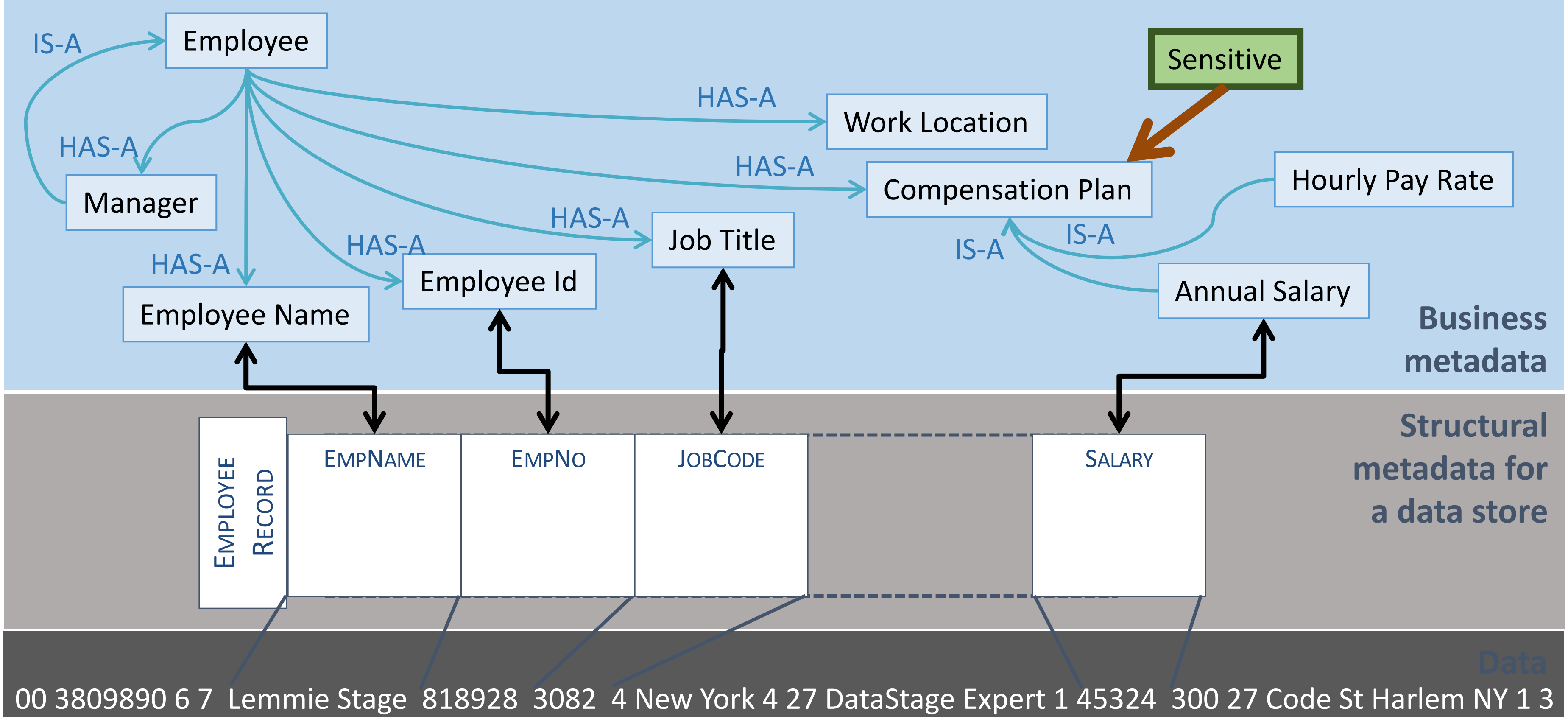
Faith Broker has been doing some simple analysis on the HR data of the company. She wants to share this data with Callie Quartile to do some detailed work. However, she does not want Callie to see the sensitive personal information in the record.

```
00 3809890 6 7  Lemmie Stage  818928  3082  4 New York 4 27 DataStage Expert 1 45324  300 27 Code St Harlem NY 1 3
00 3809890 3 7  Callie Quartile  328080  7432  5 New York 4 27 Data Scientist      1 56944  045 27 Code St Harlem NY 1 3
00 3809890 1 7  Tanya Tidie       209482  4051  2 New York 4 27 Data Steward       1 43800  215 27 Code St Harlem NY 1 3
```

```
00 3809890 6 7  Lemmie Stage  818928  3082  4 New York 4 27 DataStage Expert 1 XXXXX  XXX 27 Code St Harlem NY 1 3
00 3809890 3 7  Callie Quartile  328080  7432  5 New York 4 27 Data Scientist      1 XXXXX  XXX 27 Code St Harlem NY 1 3
00 3809890 1 7  Tanya Tidie       209482  4051  2 New York 4 27 Data Steward       1 XXXXX  XXX 27 Code St Harlem NY 1 3
```

Callie Quartile
Data Scientist

Using glossary function for semantic processing



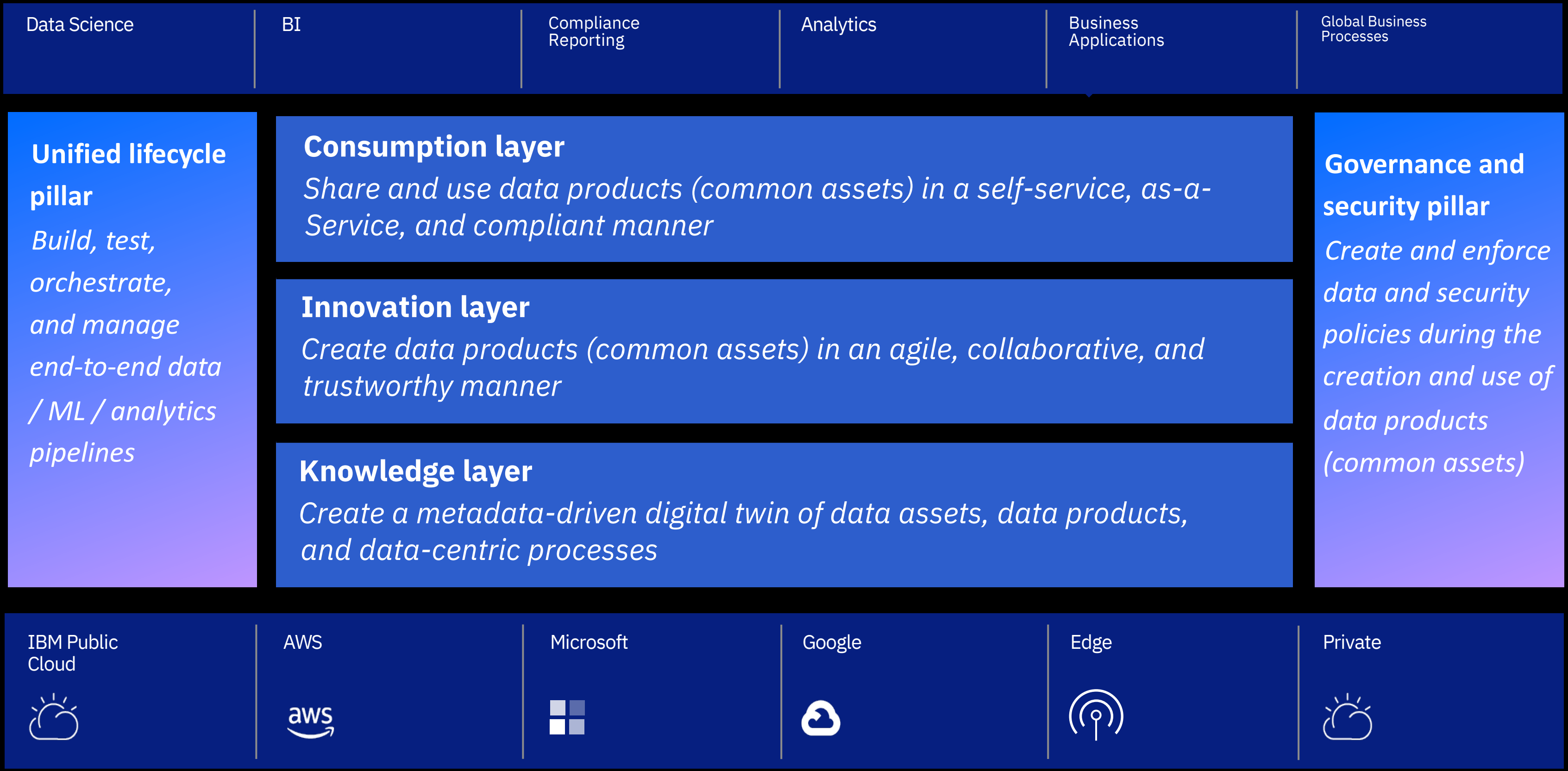
Data needs to work harder …

- Regulations and a need to operate a coherent, connected business made it necessary to extract data from original application, combine it and use it in new contexts.
- Data is now like a tortoise without its protected shell.
- The infrastructure and people that support this data need to recreate the protected shell for their data.

What is at stake?

- Value when you use it effectively
  - New business opportunities
  - Cross-sell/up-sell
  - Operational efficiencies and agility (including compliance)
  - Used across multiple business processes
- Cost/Risk if you abuse data
  - Data breaches, not following privacy policies
  - Regulatory compliance issue, loss of reputation, etc.
- Cost/Risk if you lose it (availability/backup)
  - Business outage
  - Compliance issues
- Cost/Risk if you confuse it (data quality)
  - Bad business decisions
  - Customer satisfaction problems
  - More regulatory compliance issues
  - Breaking of contractual obligations

IBM Offering (Data Fabric) conceptual vision



Data Science

BI

Compliance Reporting

Analytics

Business Applications

Global Business Processes

Unified lifecycle pillar

Build, test, orchestrate, and manage end-to-end data / ML / analytics pipelines

Consumption layer

Share and use data products (common assets) in a self-service, as-a-Service, and compliant manner

Innovation layer

Create data products (common assets) in an agile, collaborative, and trustworthy manner

Knowledge layer

Create a metadata-driven digital twin of data assets, data products, and data-centric processes

Governance and security pillar

Create and enforce data and security policies during the creation and use of data products (common assets)

IBM Public Cloud

AWS

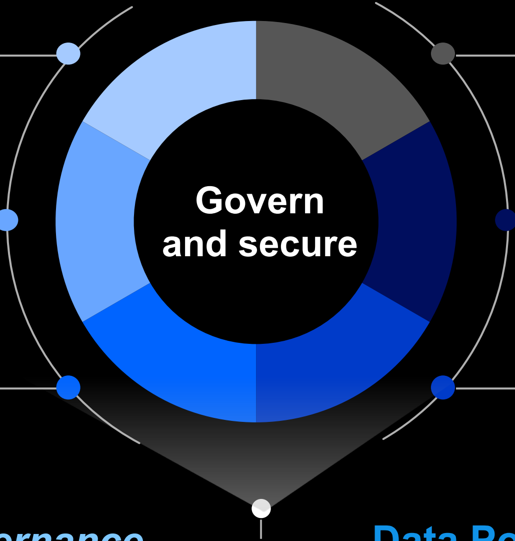
Microsoft

Google

Edge

Private

Governance and security pillar



Know your data

Have confidence in the quality and the source it originates from with a full understanding of content and usability

Trust Your Data

Have confidence in the quality and the source it originates from with a full understanding of content and usability

Protect Your Data

Access the data you need without the risk of regulatory compliance violations

Data Governance

Establish a data governance foundation of well understood business glossary of metadata, and governance policies and rules

Data Quality and Lineage

Provide easy access to data with automatic discovery, quality analysis, profiling, classification and business term assignment

Data Privacy

Autonomous enforcement of data and AI governance policies, providing automatic decisions to mask and protect data

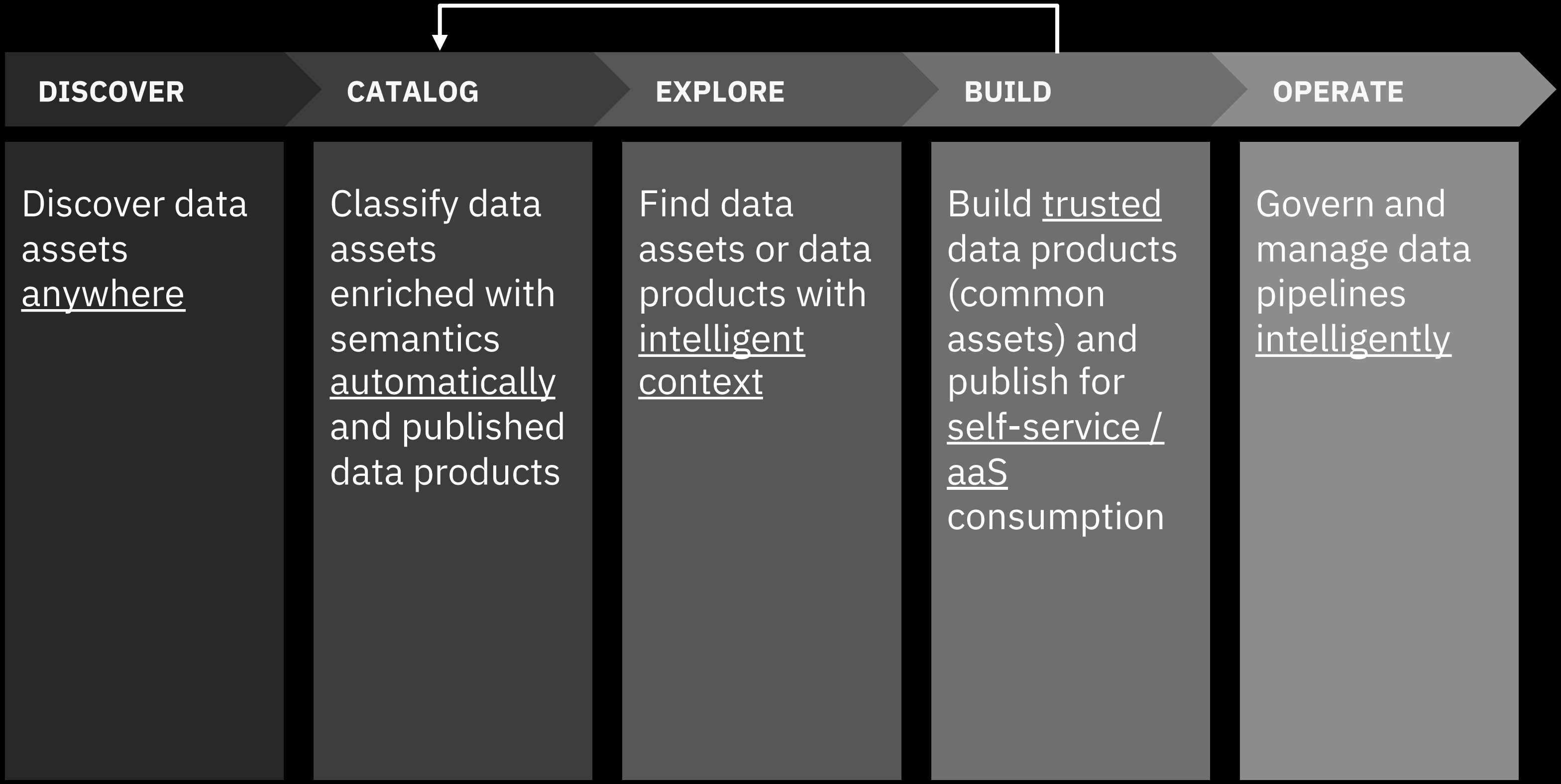
Design governance

Adopt a governance and security by design approach to ensure unified compliance.

Data Policies

Create and enforce data policies at both a local and global level

Lifecycle pillar



DISCOVER

Discover data assets anywhere

CATALOG

Classify data assets enriched with semantics automatically and published data products

EXPLORE

Find data assets or data products with intelligent context

BUILD

Build trusted data products (common assets) and publish for self-service / aaS consumption

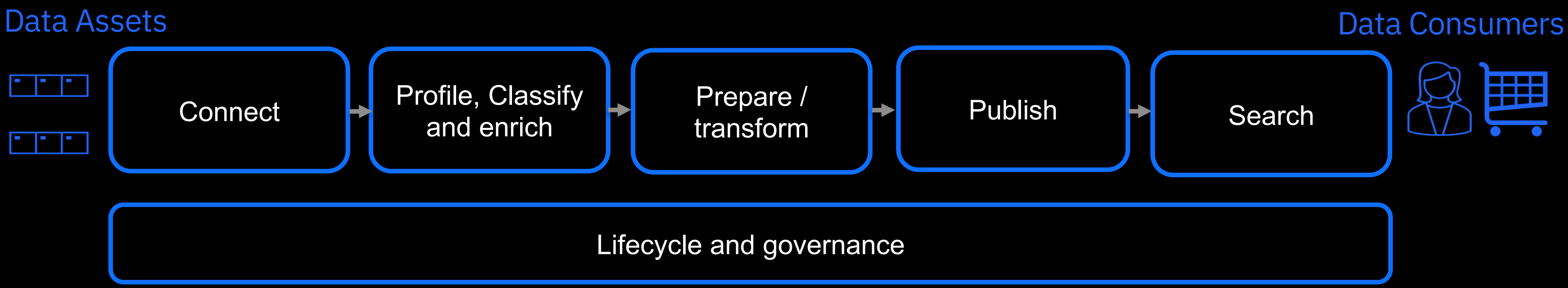
OPERATE

Govern and manage data pipelines intelligently

Leverage the knowledge layer to augment the end-to-end lifecycle

	DISCOVER	CATALOG	EXPLORE	BUILD	OPERATE
Metadata	Capture of ALL data fabric metadata across operational sources, Data Lakes, Data Warehouses, etc. (e.g., technical, business, operational, social, etc.) such as popular queries, logs from ETL jobs, lineage events, schema changes, data quality changes, etc.				
+ ***Insights***	Leverage AI / ML to discover patterns / correlations from the metadata and operationalize this insight into the data fabric				
↓ ***Activation*** ***(Examples)***	Automatically recommend changes to downstream pipelines from schema drift	Automated tagging based on patterns learned from manual stewardship	Automated suggestions based on popular queries or relationships in the data	Automated suggestions of next best data source to integrate	Automated data pipeline troubleshooting based on target SLAs

Data Fabric enables a “trusted factory” approach for innovation



Connect

Establish connectivity to physical data sources.

Profile, Classify and enrich

Assess the quality of data assets. Classify data assets, assign data policies and rules, and enrich with semantics.

Prepare / transform

Engineer data assets into trusted data products.

Publish

Publish data products.

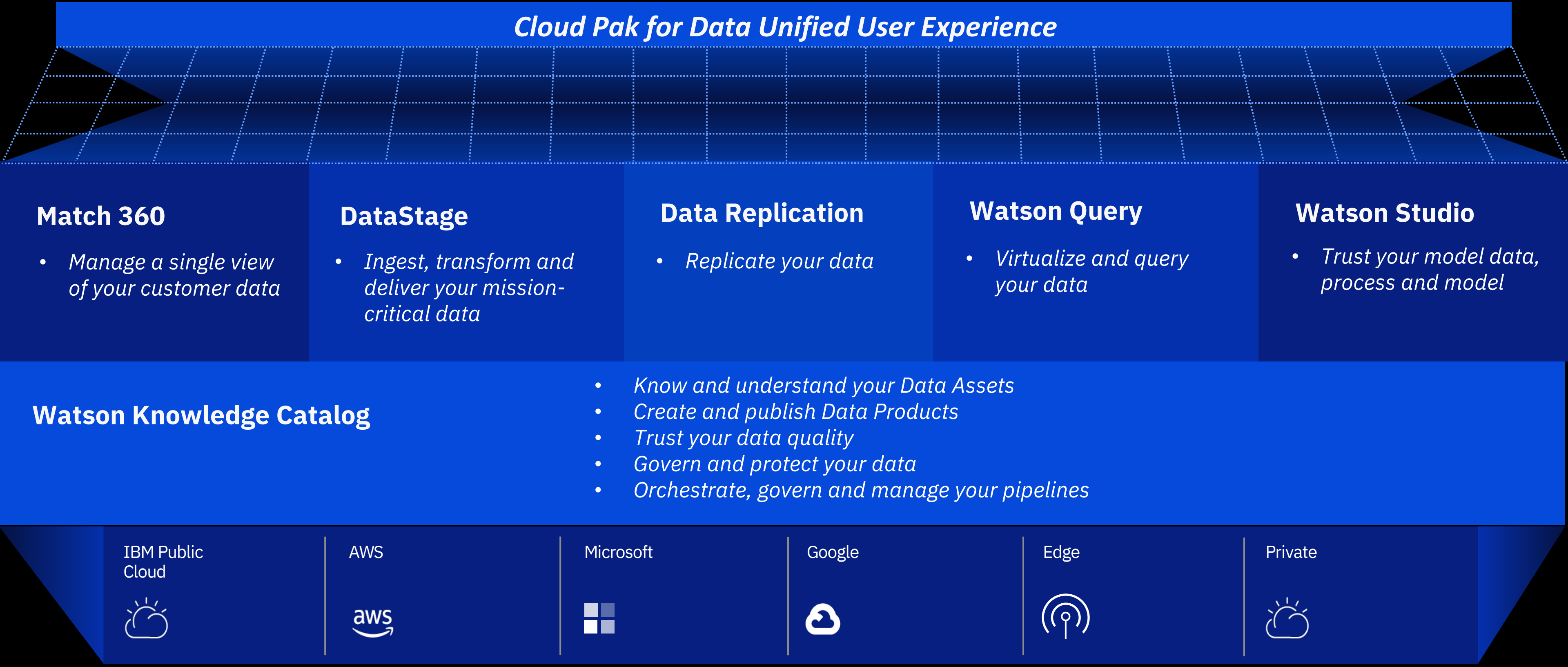
Search

Search and find data products using Natural Language.

Lifecycle and governance

Implement DataOps principles throughout the lifecycle and enforce governance end-to-end.

IBM Data Fabric overview of capabilities



Cloud Pak for Data Unified User Experience

Match 360

- Manage a single view of your customer data

DataStage

- Ingest, transform and deliver your mission-critical data

Data Replication

- Replicate your data

Watson Query

- Virtualize and query your data

Watson Studio

- Trust your model data, process and model

Watson Knowledge Catalog

- Know and understand your Data Assets
- Create and publish Data Products
- Trust your data quality
- Govern and protect your data
- Orchestrate, govern and manage your pipelines

IBM Public Cloud

AWS

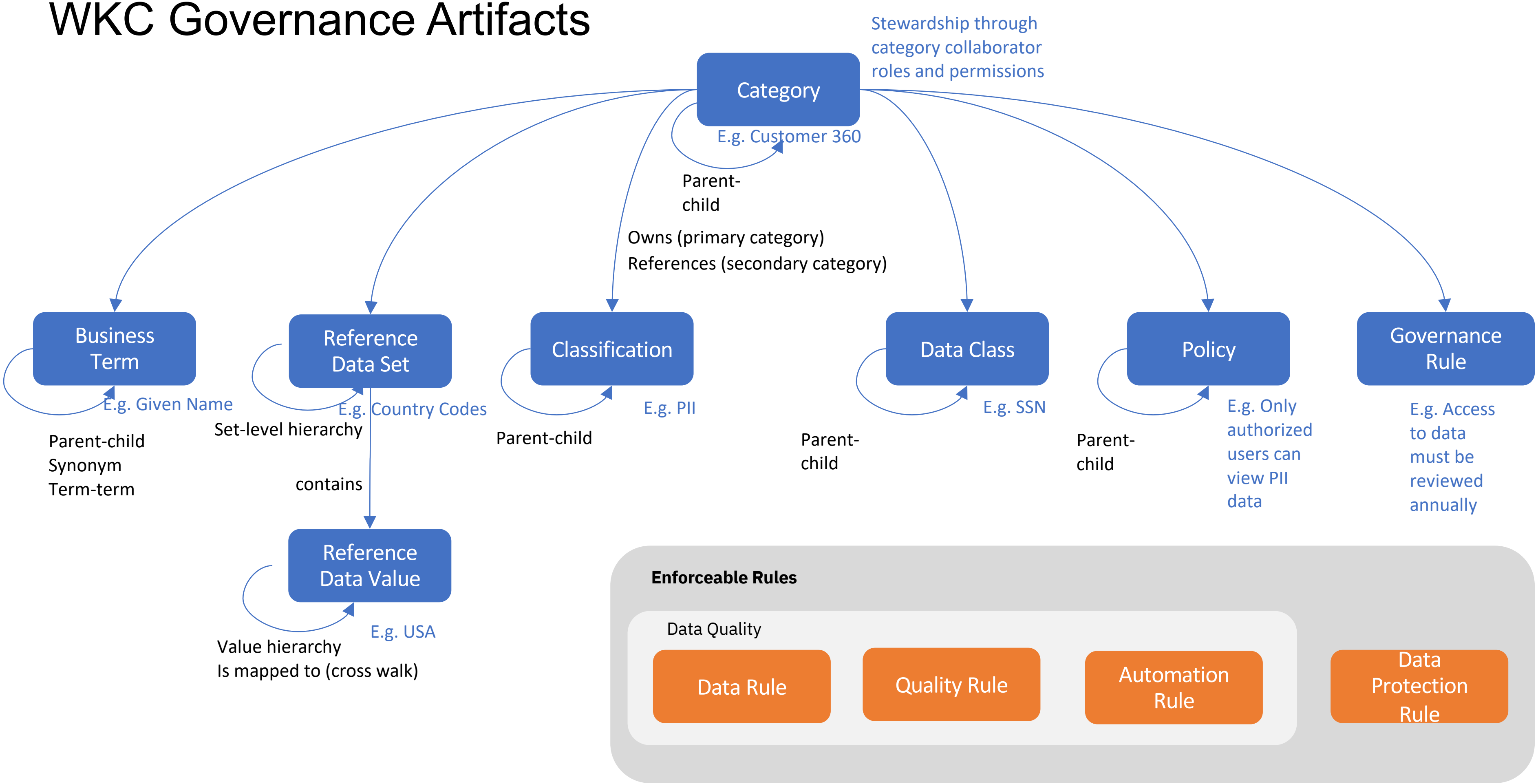
Microsoft

Google

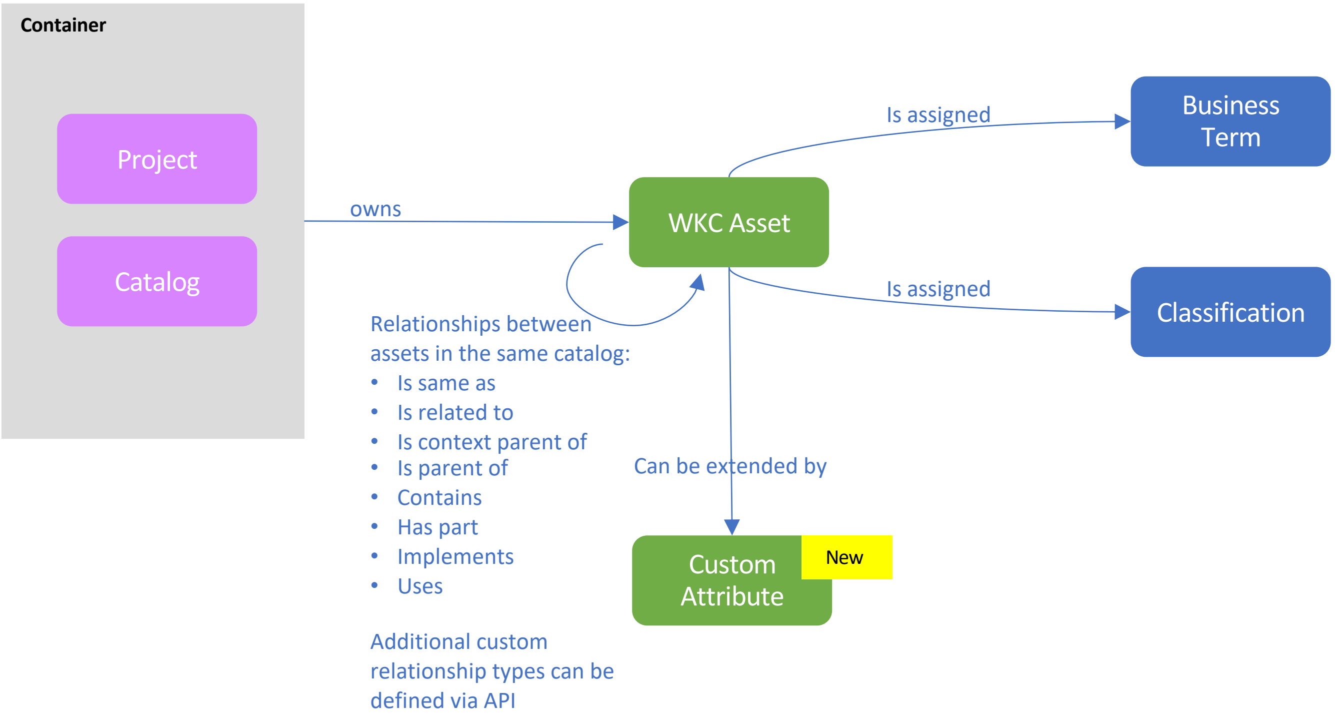
Edge

Private

WKC Governance Artifacts



WKC Assets and Relationships



Container

Project

Catalog

owns

WKC Asset

Is assigned

Business Term

Is assigned

Classification

Relationships between assets in the same catalog:

- Is same as
- Is related to
- Is context parent of
- Is parent of
- Contains
- Has part
- Implements
- Uses

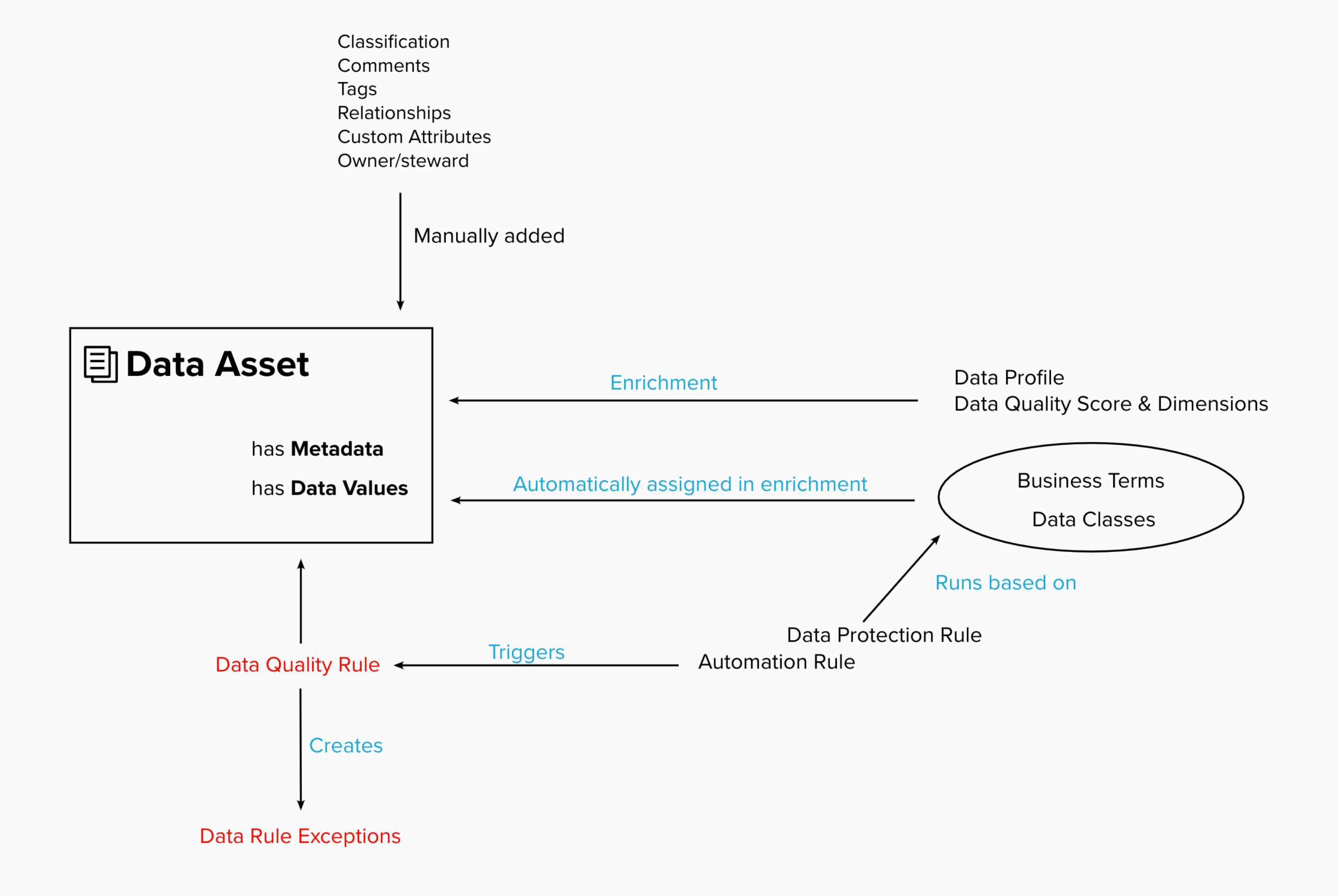
Additional custom relationship types can be defined via API

Can be extended by

Custom Attribute

New

WKC Conceptual Model



Classification
Comments
Tags
Relationships
Custom Attributes
Owner/steward

Manually added

Data Asset

has **Metadata**

has **Data Values**

Enrichment

Data Profile
Data Quality Score & Dimensions

Automatically assigned in enrichment

Business Terms

Data Classes

Runs based on

Data Protection Rule

Automation Rule

Triggers

Data Quality Rule

Creates

Data Rule Exceptions

What is Data Virtualization and Watson Query?

- Data virtualization is a capability that enables users to integrate and query data in real-time without movement.
- Watson Query is the name of the service on IBM Cloud that provides users with data virtualization capabilities, while heavily integrating with other Cloud Pak for Data services

Data Virtualization Capabilities

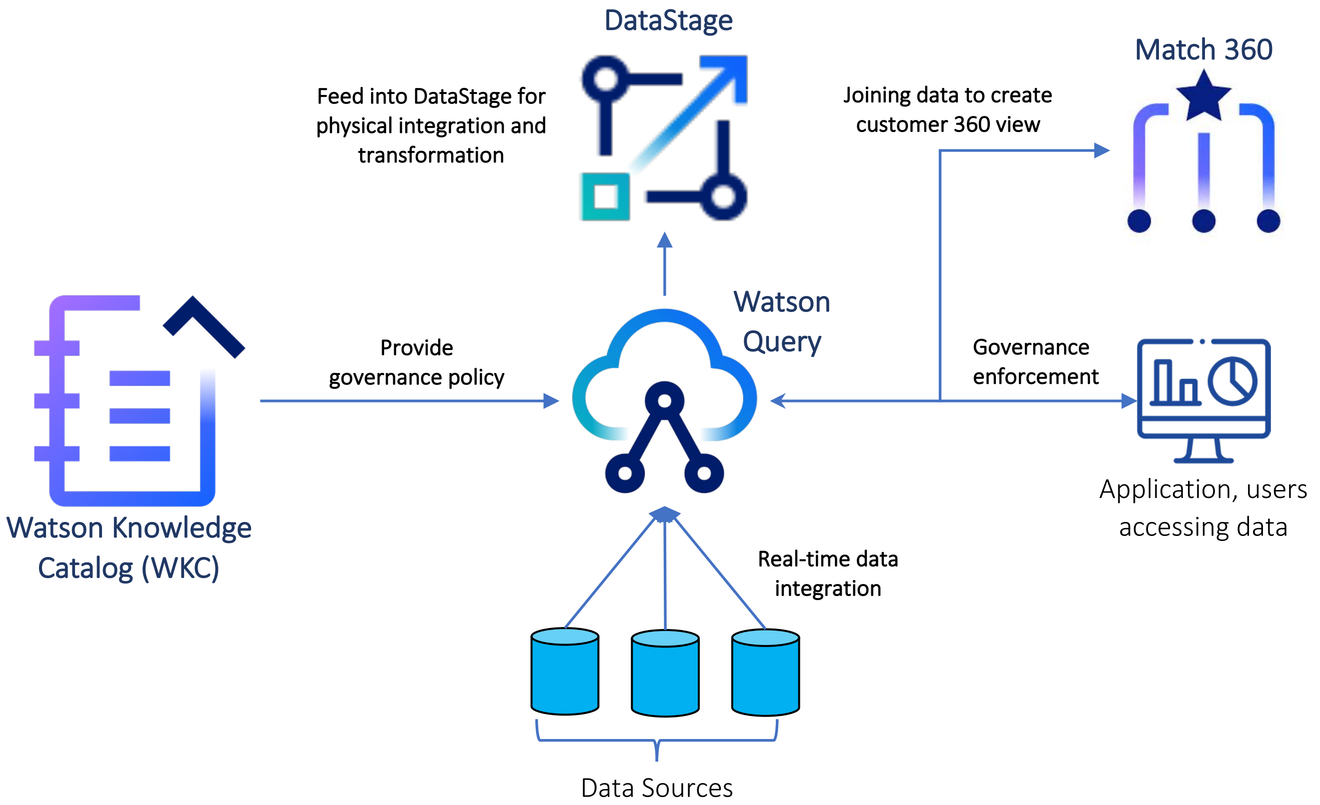
- Connect, access, and govern any data without the need for data movement - Access structured and unstructured disparate data on demand, without need for ETL or creation of copies.
- Abstract complexity from data consumers - A virtual semantic layer across all your data sources allows users to quickly connect to, join, and analyze data from multiple sources without needing to understand back-end database technologies.
- Create virtual views over multiple data sources - No need for transformations; create a virtual view of required data that can be shared within your organization.
- Governance integration and security - Controlled, governed and secure access to virtual data sets through native integration with Watson Knowledge Catalog.

Watson Query Experience

- **One query experience** over multiple data sources, types, and form factors.
- **Integrated governance** with Watson Knowledge Catalog to provide governed data access.
- **Open data formats** to work with data on any cloud and on-premises
- Real-time data integration **without data movement**.
- **Intelligent cache recommendation** accelerating query performance with minimal user input.

Watson Query and Data Fabric

Watson Query provides data virtualization to support data access and governance enforcement for Data Fabric



DataStage

Feed into DataStage for physical integration and transformation

Match 360

Joining data to create customer 360 view

Watson Knowledge Catalog (WKC)

Provide governance policy

Watson Query

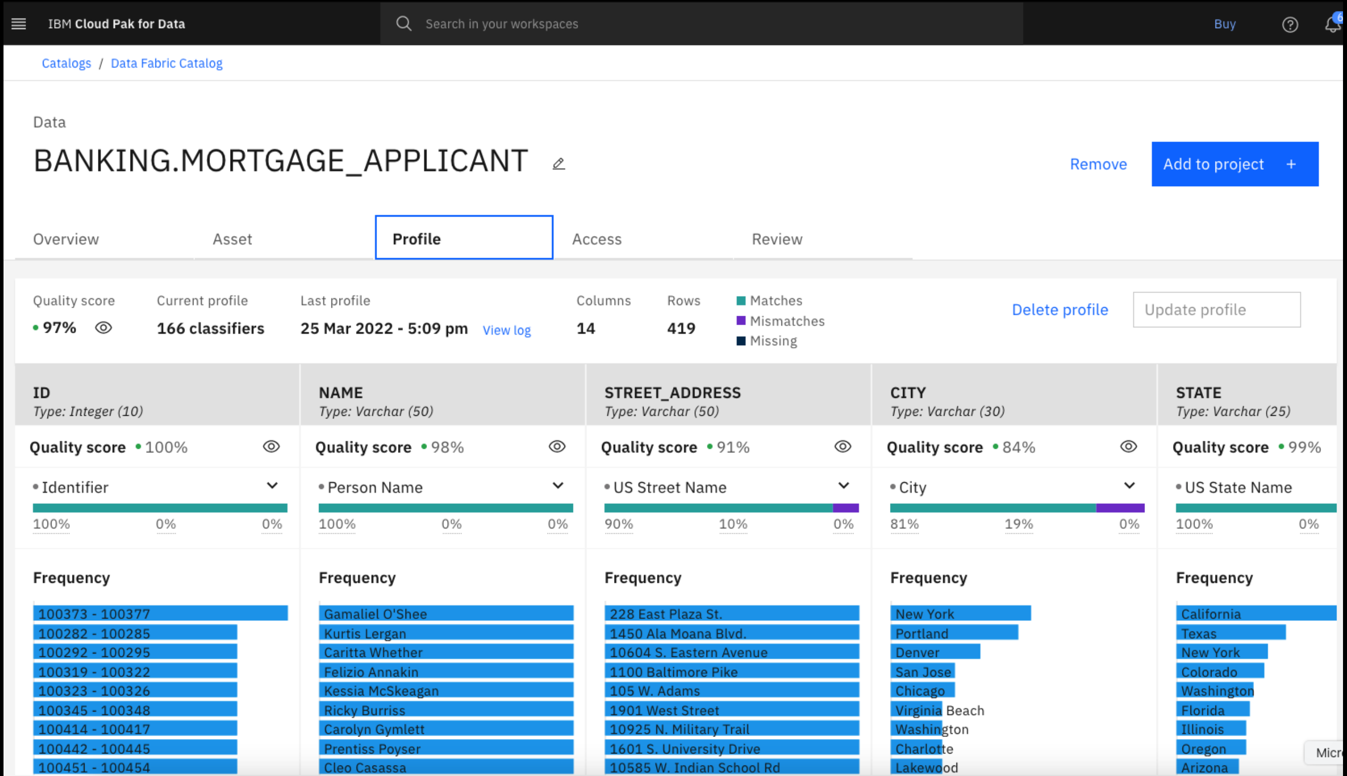
Governance enforcement

Application, users accessing data

Real-time data integration

Data Sources

Demo



Data

BANKING.MORTGAGE_APPLICANT

Quality score	Current profile	Last profile	Columns	Rows
• 97%	166 classifiers	25 Mar 2022 - 5:09 pm	14	419

■ Matches
■ Mismatches
■ Missing

ID	NAME	STREET_ADDRESS	CITY	STATE
Type: Integer (10)	Type: Varchar (50)	Type: Varchar (50)	Type: Varchar (30)	Type: Varchar (25)
Quality score • 100%	Quality score • 98%	Quality score • 91%	Quality score • 84%	Quality score • 99%
• Identifier	• Person Name	• US Street Name	• City	• US State Name
100%	100%	90%	81%	100%
0%	0%	10%	19%	0%
0%	0%	0%	0%	
Frequency	Frequency	Frequency	Frequency	Frequency
100373 - 100377	Gamaliel O'Shee	228 East Plaza St.	New York	California
100282 - 100285	Kurtis Lergan	1450 Ala Moana Blvd.	Portland	Texas
100292 - 100295	Caritta Whether	10604 S. Eastern Avenue	Denver	New York
100319 - 100322	Felizio Annakin	1100 Baltimore Pike	San Jose	Colorado
100323 - 100326	Kessia McSkeagan	105 W. Adams	Chicago	Washington
100345 - 100348	Ricky Burriss	1901 West Street	Virginia Beach	Florida
100414 - 100417	Carolyn Gymlett	10925 N. Military Trail	Washington	Illinois
100442 - 100445	Prentiss Poyser	1601 S. University Drive	Charlotte	Oregon
100451 - 100454	Cleo Casassa	10585 W. Indian School Rd	Lakewood	Arizona