

26<sup>th</sup> – 28<sup>th</sup> April 2022

# Informatica Cloud Data Integration Bootcamp

Global Technical Alliances (GTA) and PTS Team

# Agenda – Day2

**9:30 AM–11:30 AM BST | 10:30 AM CEST – 12:30 PM CEST | 2:00 PM–4:00 PM IST**

**1** Advanced Pushdown

**2** Cloud Data Quality

**3** Cloud Data Governance Integration

**4** Enterprise Data Catalog Integration

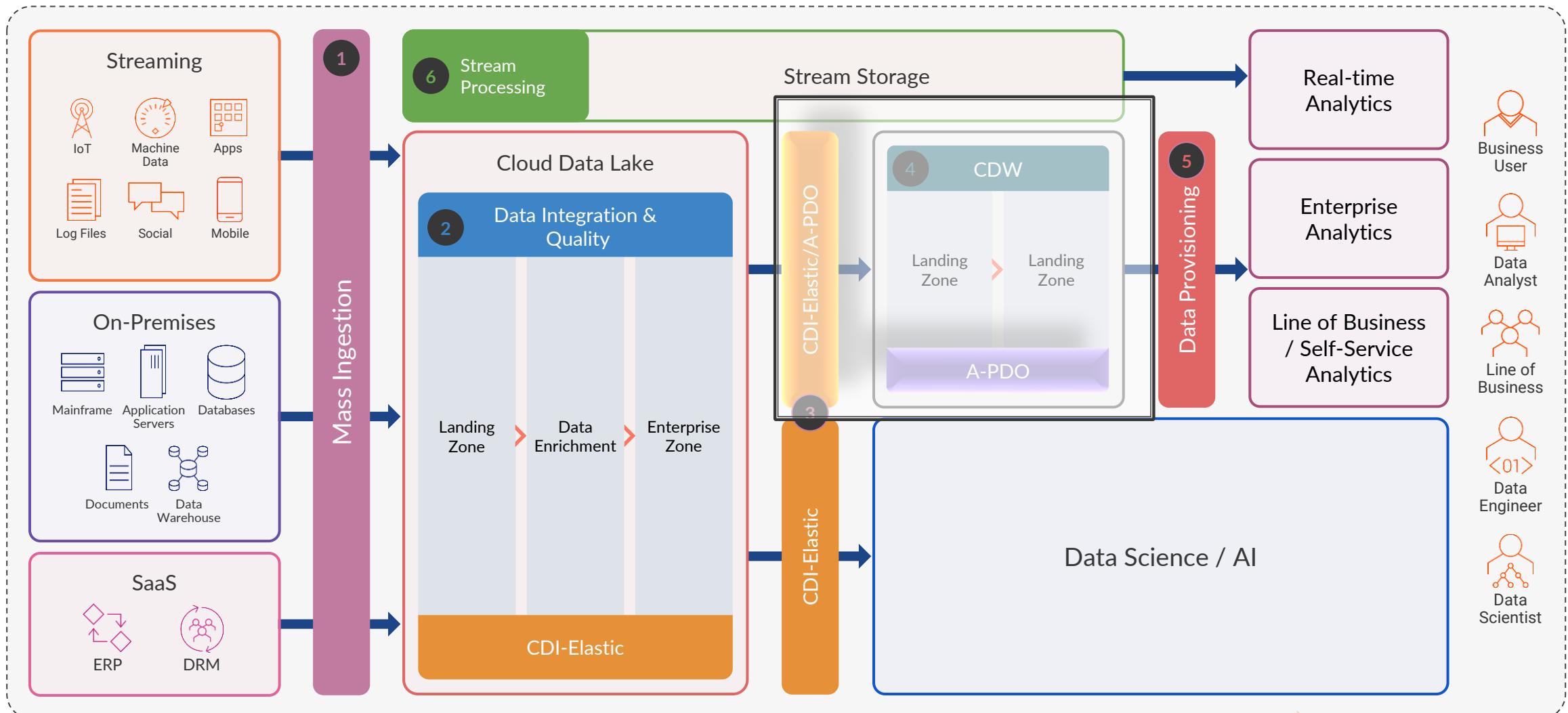
**5** Cloud Integration Hub (CIH)

**6** Cloud Test Data Management



# Advanced Pushdown Optimization(APDO)

# Informatica Data Warehouse and Datalake Architecture



# Cloud Data Management Trends

“By 2022, public cloud services will be essential for 90% of data and analytics innovation” – Gartner



**Cloud adoption**



**Emergence of ELT**



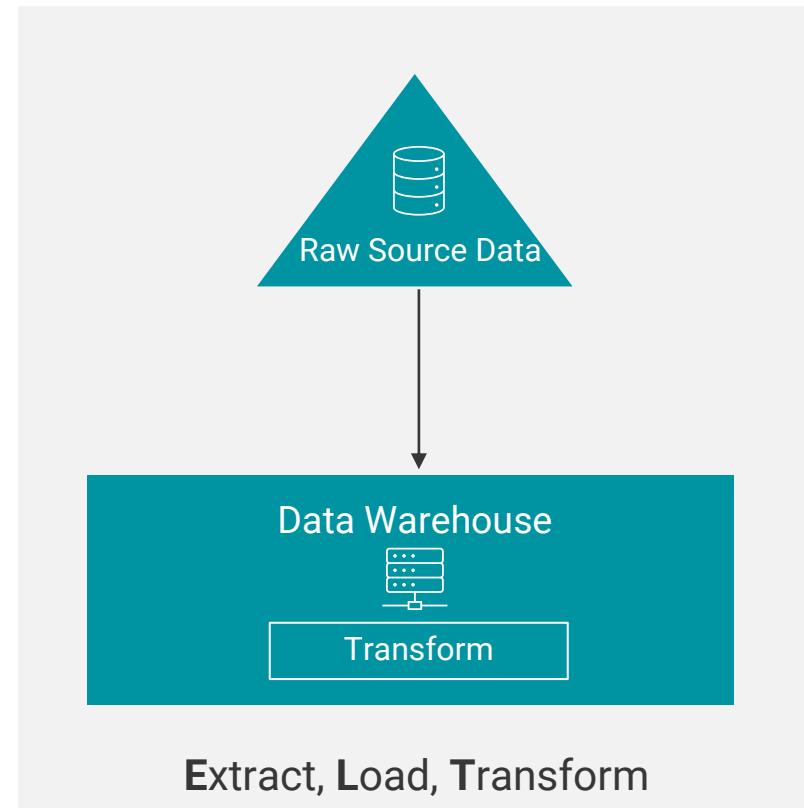
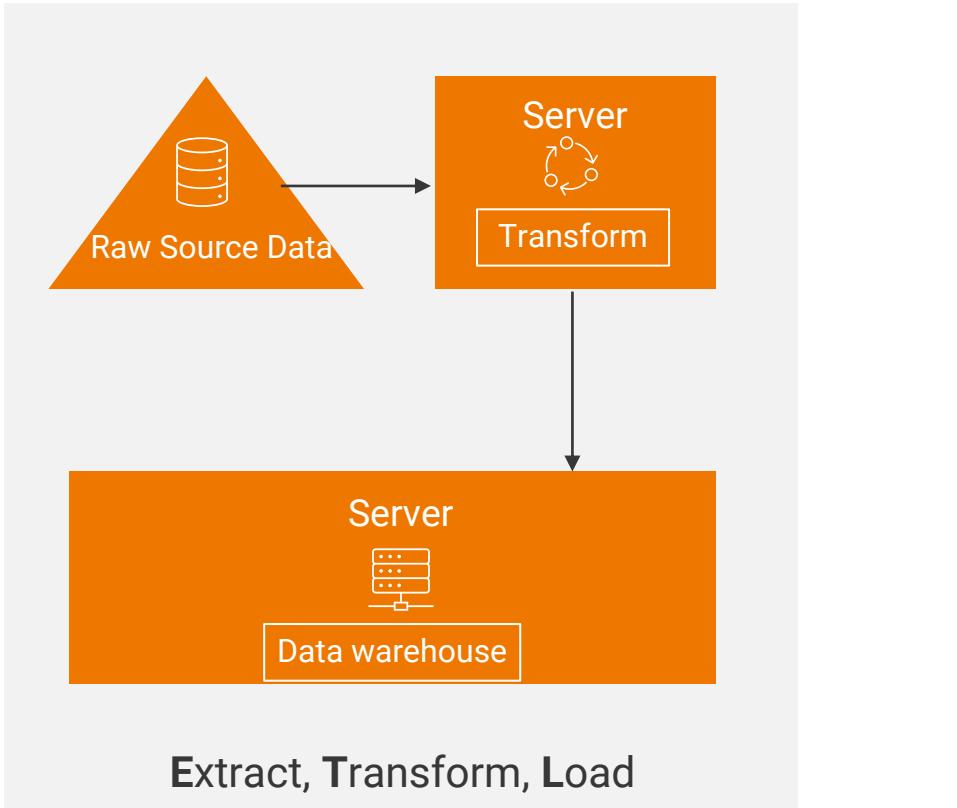
**Cloud Data Lake, Data warehouse, Lakehouse**

## Research Circle Survey on data management strategy from Gartner -

- By 2022, public cloud services will be essential for 90% of data and analytics innovation.
- By 2023, cloud-based AI will increase 5x from 2019 making AI one of the top workload categories in the cloud.
- Out of 200 survey respondents, more than one-third (38%) are already using cloud data warehouses (CDWs). Long term, 43% expected to have all of their data in the cloud, with the remainder planning to pursue hybrid models that leverage both cloud and on-premises data warehouses.
- While the use of CDWs is already widespread, only 16% currently use data lakes. More than half (56%) plan to use data lakes in the future, and another 26% are considering doing so.

# ETL vs ELT

- ETL and ELT differs from each other when it comes to where data processing occurs



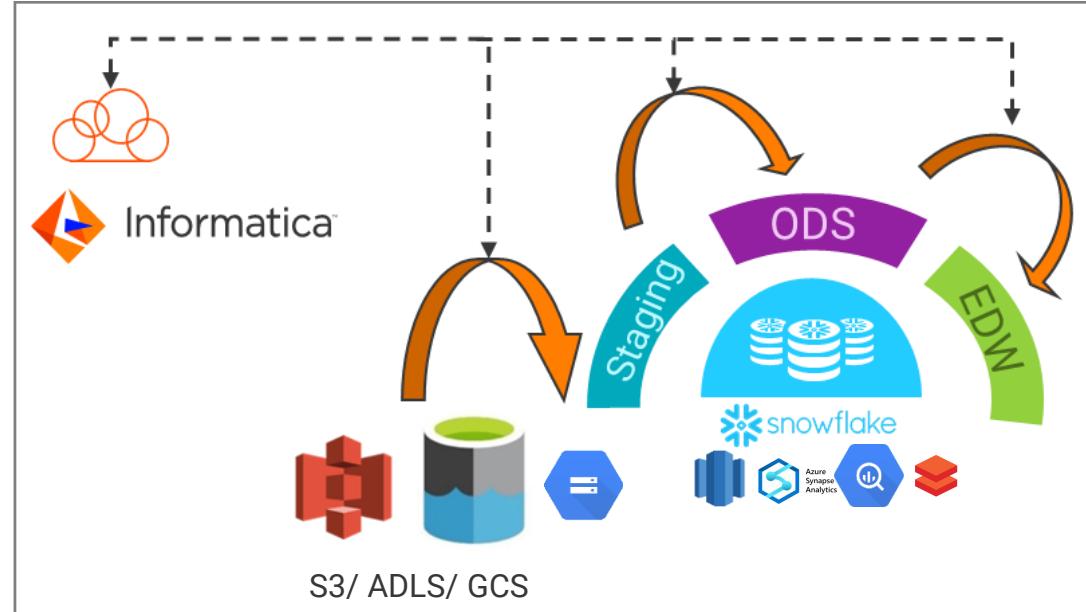
# Advanced Pushdown Optimization(APDO)

## Advanced pushdown

Converts and processes data pipelines to native ecosystem commands and SQL queries for faster processing at lower cost while ensuring data stays within the ecosystem

### Features

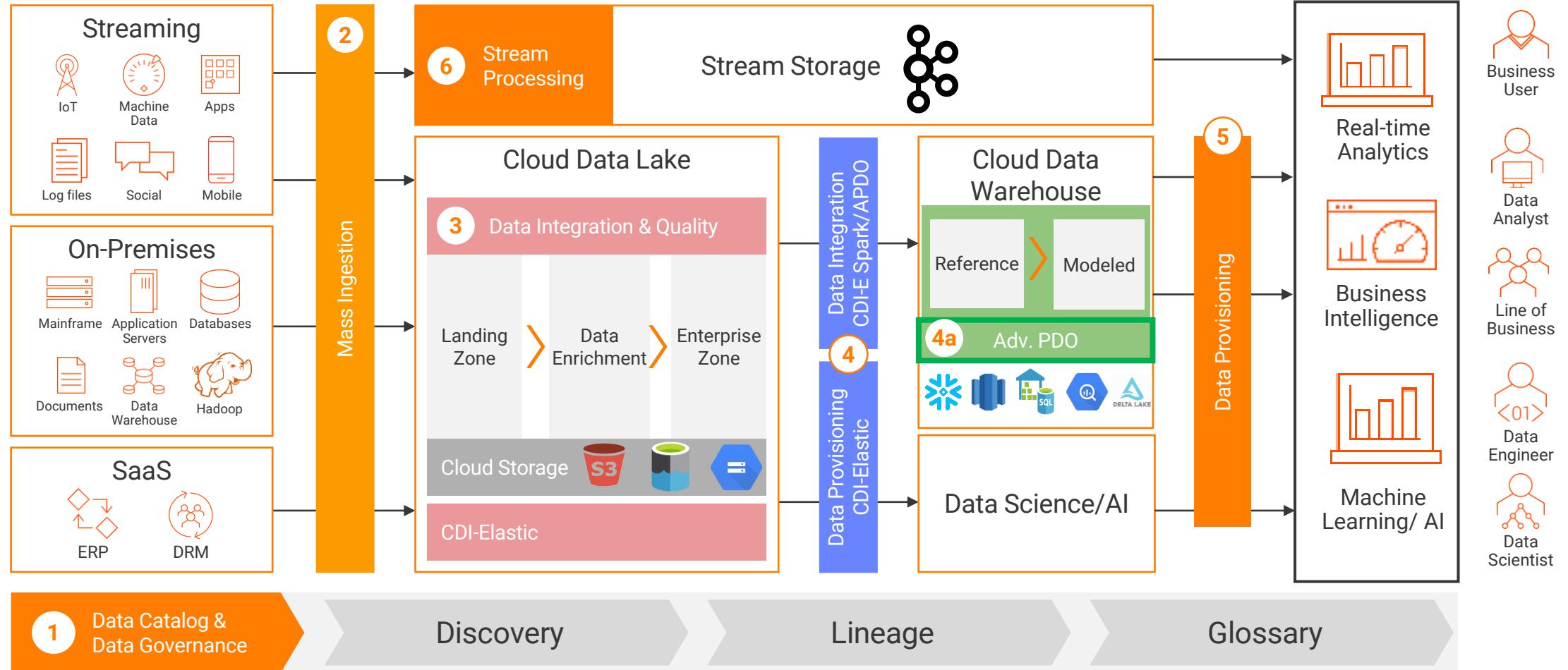
- Data pipeline logic gets translated into Cloud ecosystem based native SQL (**SQL Based PDO**) or native ecosystem API/ commands (**Ecosystem PDO**) based on the Data integration pattern
- Support for **Full, Source, Partial** PDO
- Broadest array of connectors and support for all major ecosystems (CDL/CDWs)
- Ecosystem agnostic
- Simple drop-down option in GUI with no need to learn proprietary commands



Enable faster processing with zero data egress charges through advanced pushdown optimization



# Informatica Reference Architecture for Data Management



# ODBC PDO vs APDO

ODBC Based Pushdown Optimization	Advanced Pushdown Optimization
Developed 15+ years ago. No further plans to expand transformation/function support	Specifically designed to support CDW/CDL patterns. Major expansion plan for transformations/function support, more features in roadmap.
An ODBC connection needs to be created and used in mappings	Advanced Pushdown Optimization is a native connector feature. No separate ODBC connection required
Classical CDW patterns only	Multiple patterns within CDW, CDL, including classical CDW
Requires Secure Agent	All-cloud: Works on Informatica Runtime, Informatica Advanced Serverless (supports secure agent as well)
Supports only ODBC connection features	Existing connector features are supported (example: any advanced authentication options)
No separate license required	Enabled with IPU based model. For Non-IPU, requires separate license.

# Use Case1

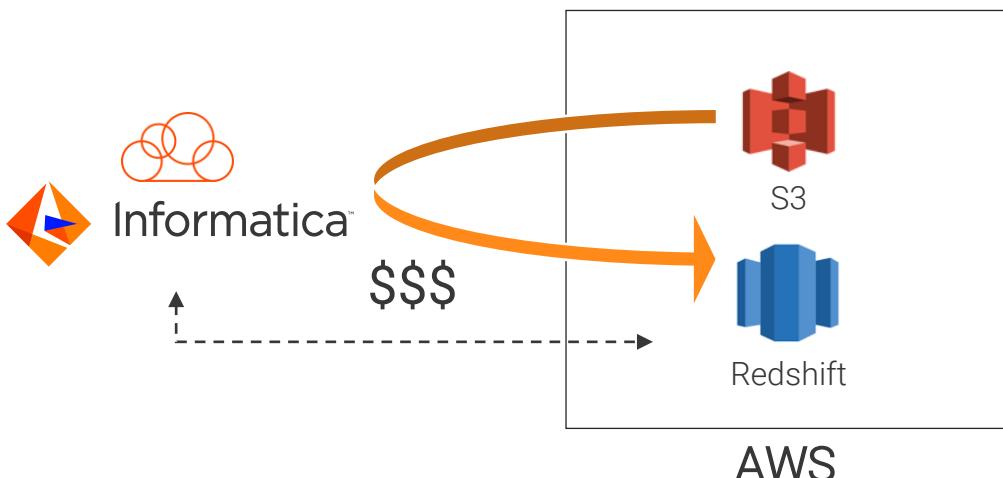
## Ecosystem Pushdown

# Use Case 1: Ecosystem Pushdown



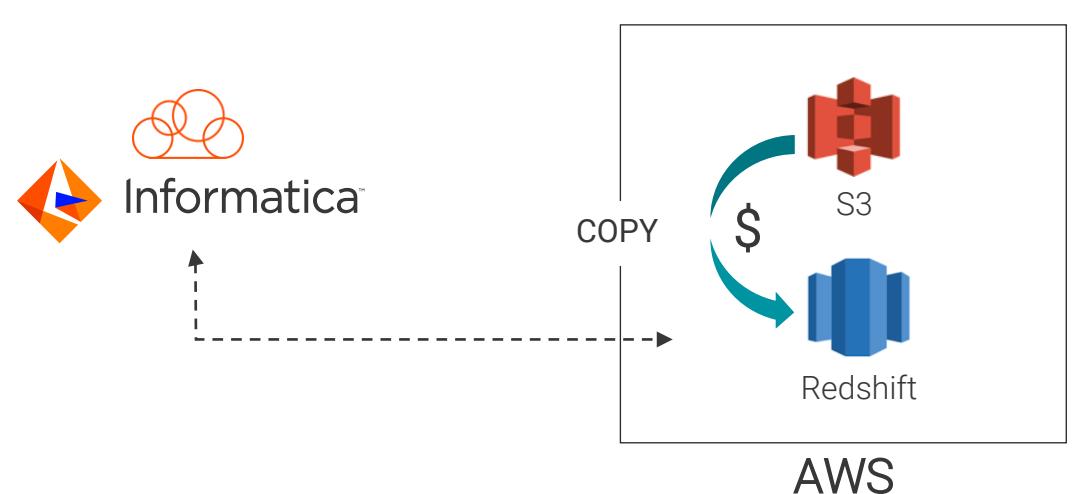
Ecosystem Pushdown transfers data from cloud data lake to data warehouse using native ecosystem commands

Without Pushdown



Loading data from Data lake to Data warehouse using  
Informatica engine

With Pushdown



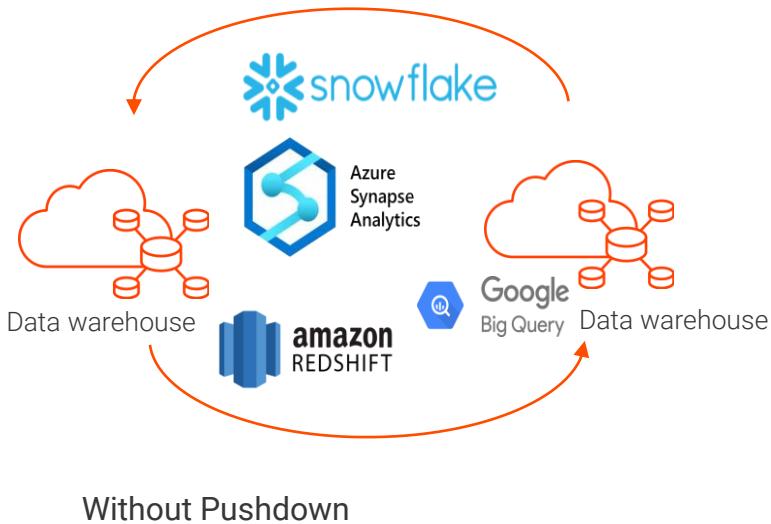
Loading data from Data lake to Data warehouse using AWS  
commands



# Use Case2

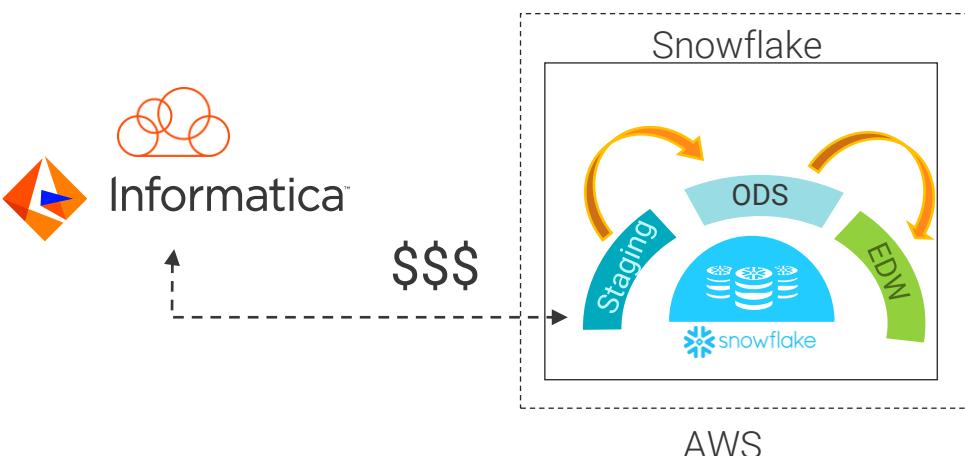
## Datawarehouse Pushdown

# Use Case 2: Data warehouse Pushdown



Data warehouse pushdown uses SQL queries to move data from staging area to ODS and ODS to EDW within a data warehouse

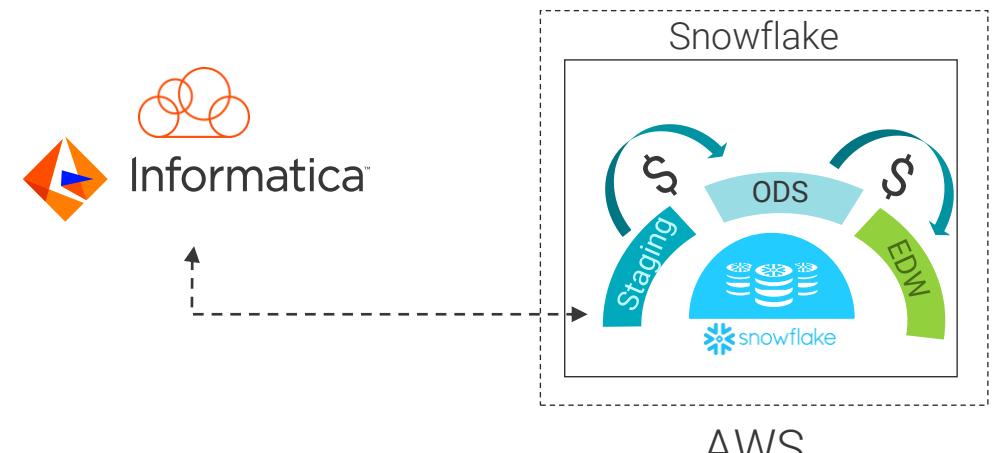
Without Pushdown



Loading data from staging to ODS in Snowflake using Informatica engine

© Informatica. Proprietary and Confidential.

With Pushdown



Loading data from staging to ODS in Snowflake using Informatica engine

**Informatica**

# APDO Benefits



## Zero data egress cost

No data egress cost since data does not move out of your underlying cloud infra



## Faster than traditional ETL

Up to 20x faster than traditional ETL



## Ecosystem agnostic

Not hardwired with a specific DW vendor  
Vendor switching is not difficult

## Easy switching between runtime options

Enable pushdown at runtime with a click



## Extensive connector support

Support all major cloud ecosystems and connectivity

## No code experience

You don't need a qualified coder to operate INFA pushdown

# DEMO





# Cloud Data Quality

# Informatica Intelligent Cloud Services



# Enabling **Any User** with Self-Service Data Quality

Powerful User Interactions

Business Focused



Architect



IT Specialist



SaaS /EDW  
Owner



Data  
Scientist



Citizen  
Integrators



Data  
Analyst



Business  
Leader

Scalable

Enterprise-ready

Simple

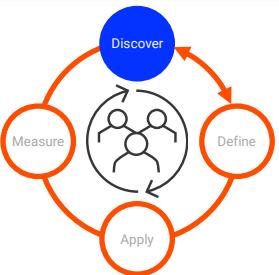
# Data Quality Methodology



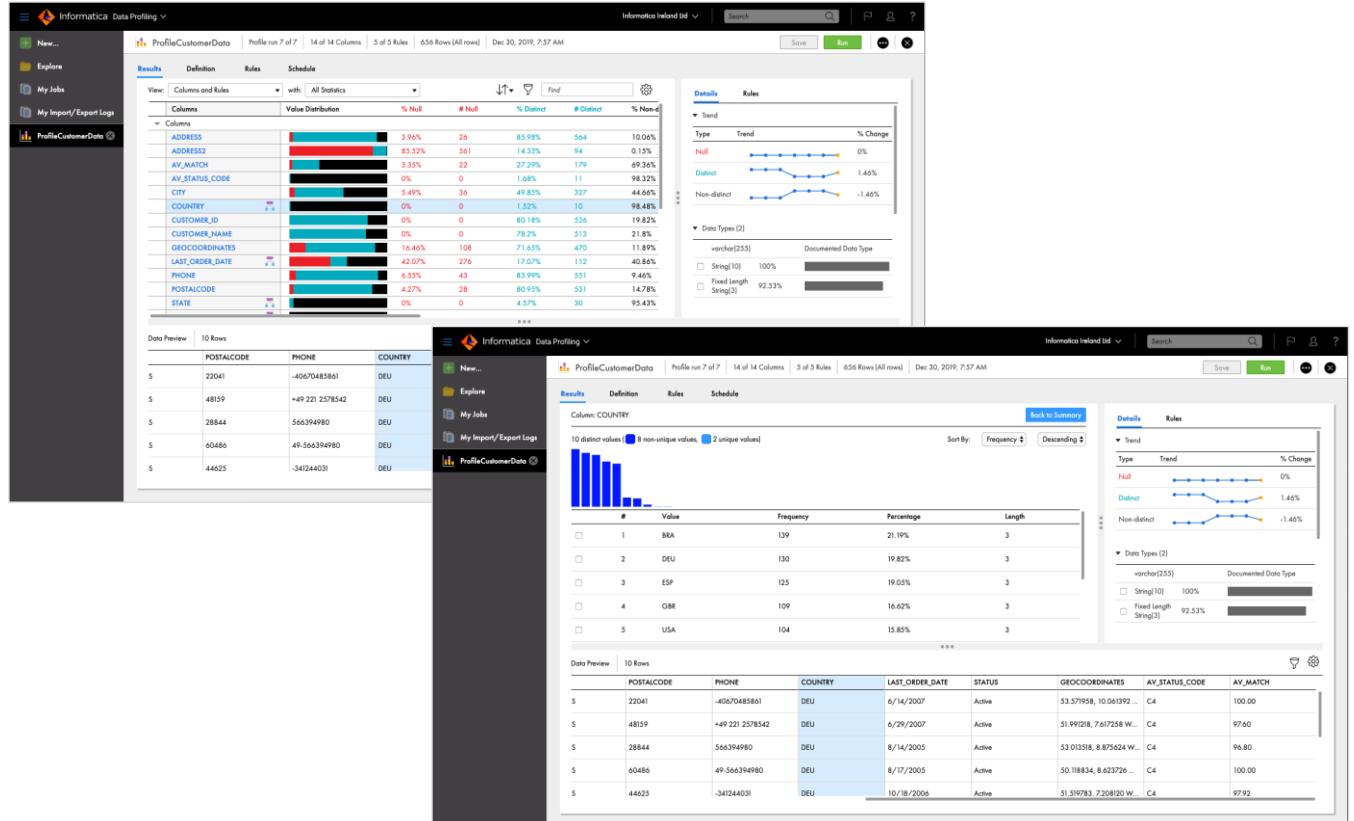
Informatica®

# Discover

## Enterprise Discovery and Profiling



Profile data to examine its structure and context using out-of-the-box templates



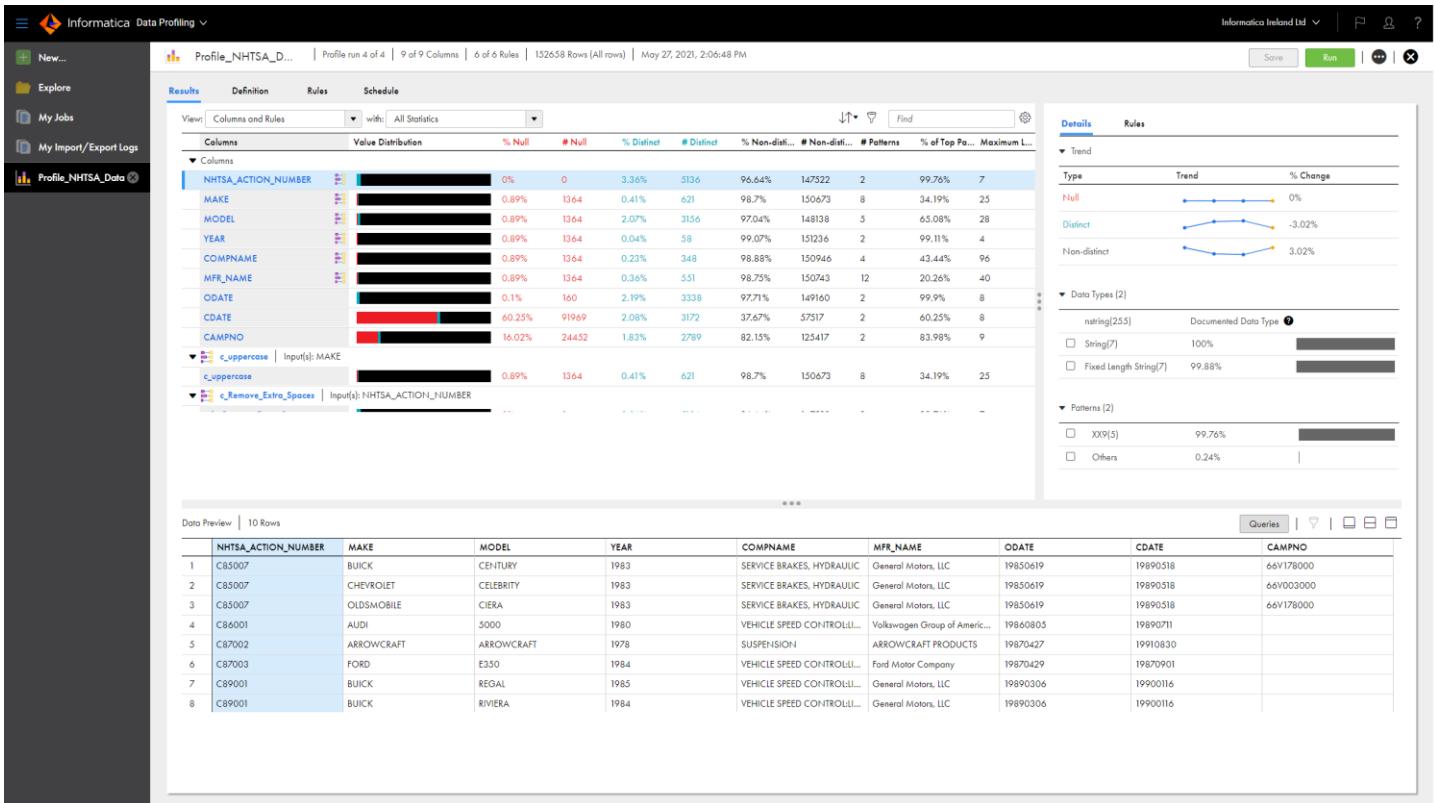
Drill down to see details and filter on results

Compare profile runs to identify trends over time

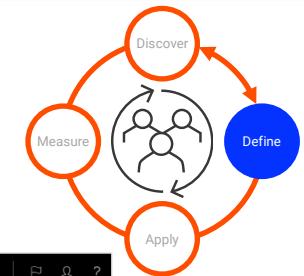
# Cloud Data Profiling

## Get to Know Your Data

- Identify potential data issues
- Track aggregated changes on data over time
  - Provide statistics on data
  - Identify uniqueness or repeating values (value frequencies)
  - Identify patterns and formats
- Can profile Data Quality Assets
  - What-if analyses to determine rule fit



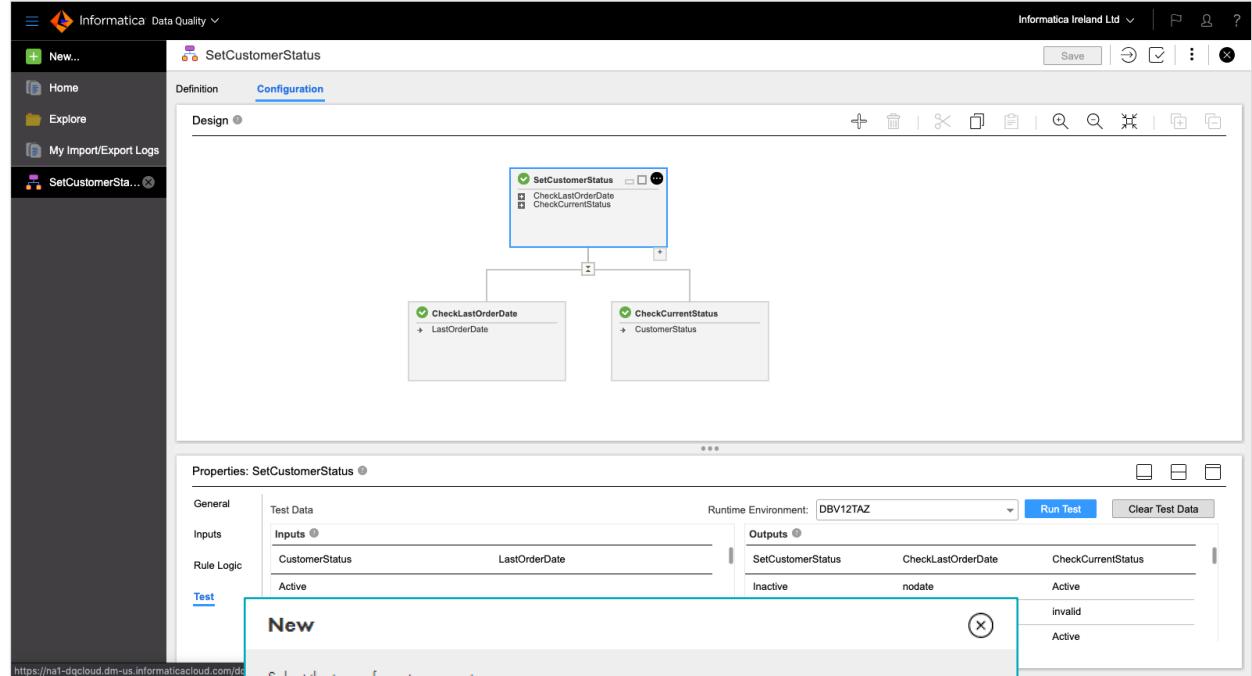
# Define Business Rule Definition



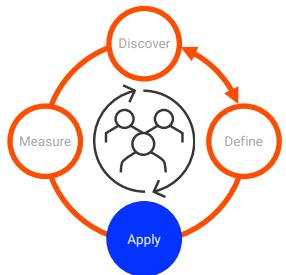
Empower the business to lead data quality initiatives

Reduce project cycles

Enable IT to focus on strategic projects



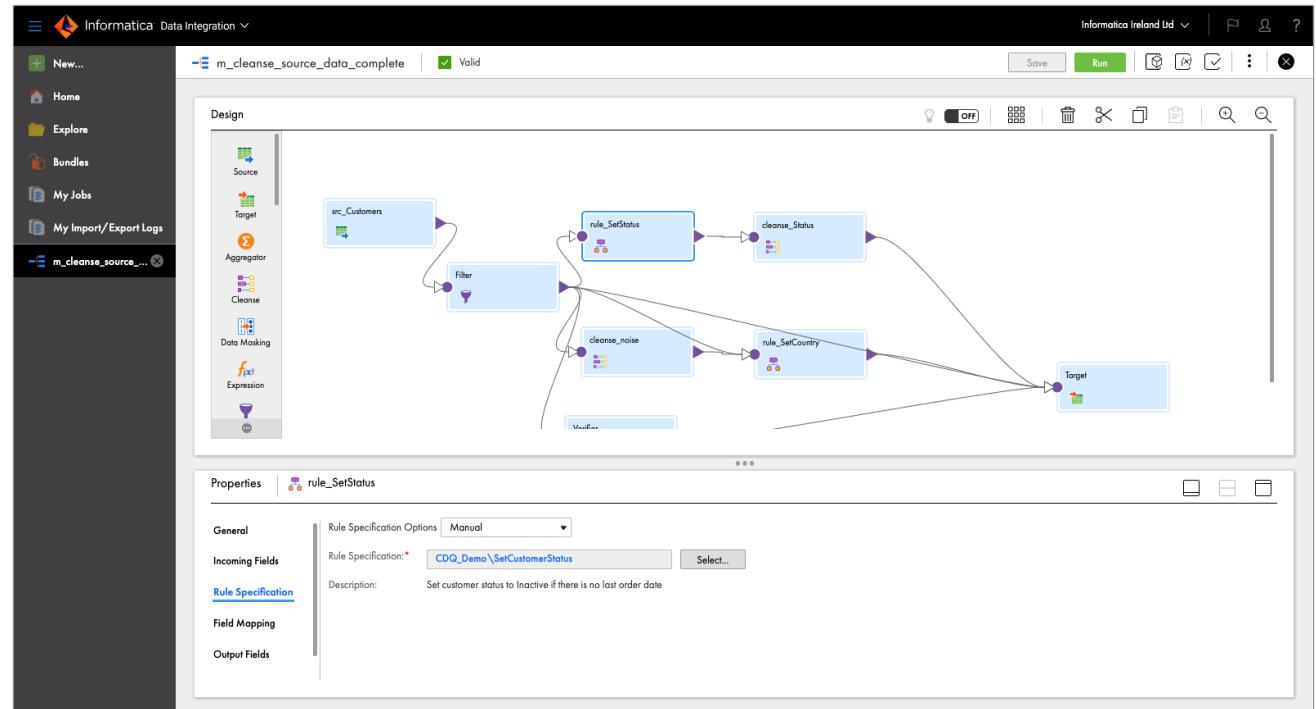
# Apply Centralized Re-usable Rules



Consistently apply data quality rules across the enterprise in support of data governance

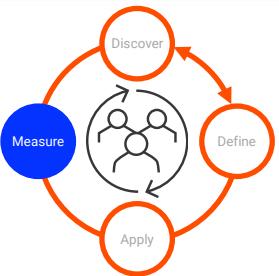
Reduce cost through re-use of centrally managed data quality rules

Streamline the resolution of data quality issues



# Manage and Monitor

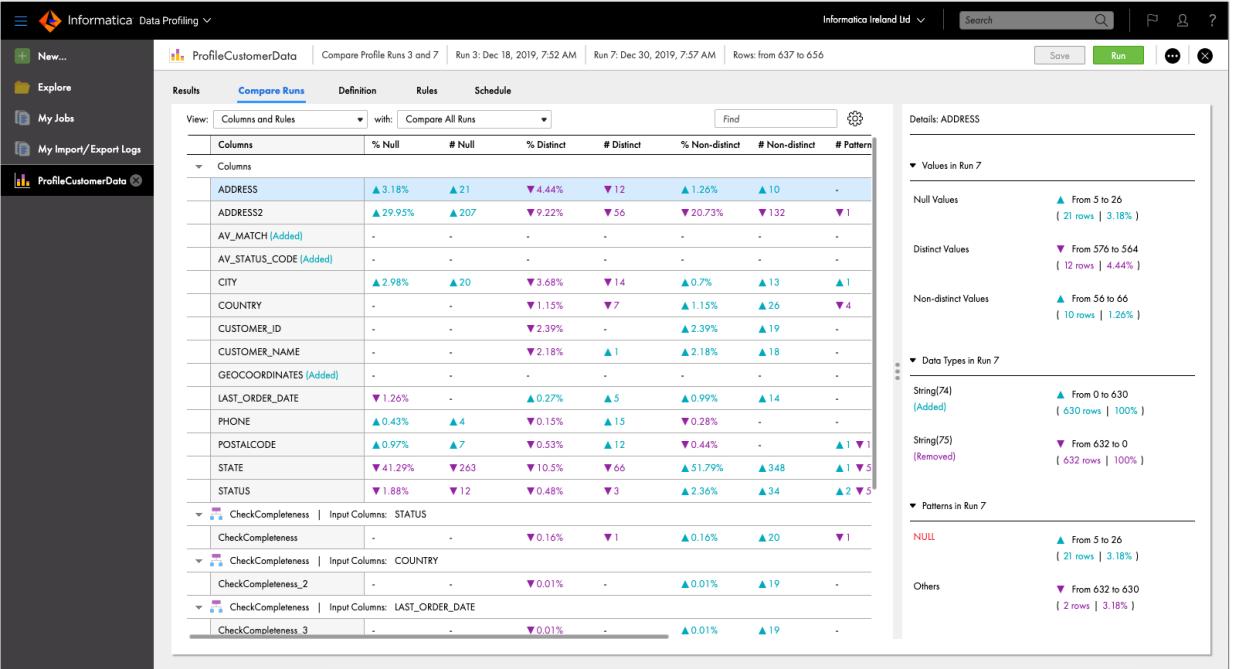
## Provide Continuous Insight



Align data quality and data governance efforts

Track data quality improvements over time

Enable IT to focus on strategic projects



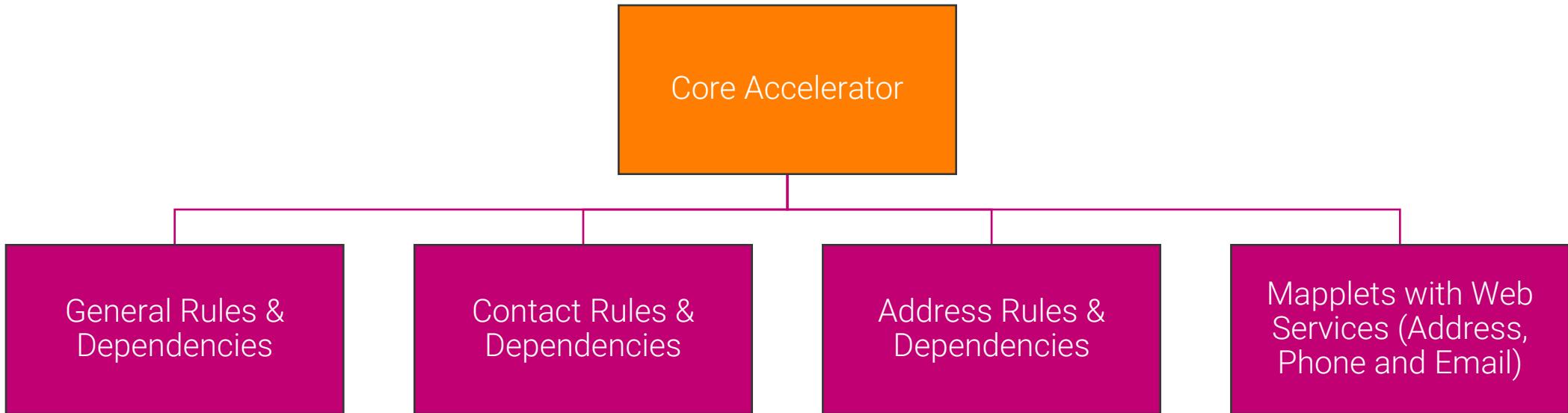


# Data Quality Bundles

Pre-built Data Quality content available  
to accelerate adoption

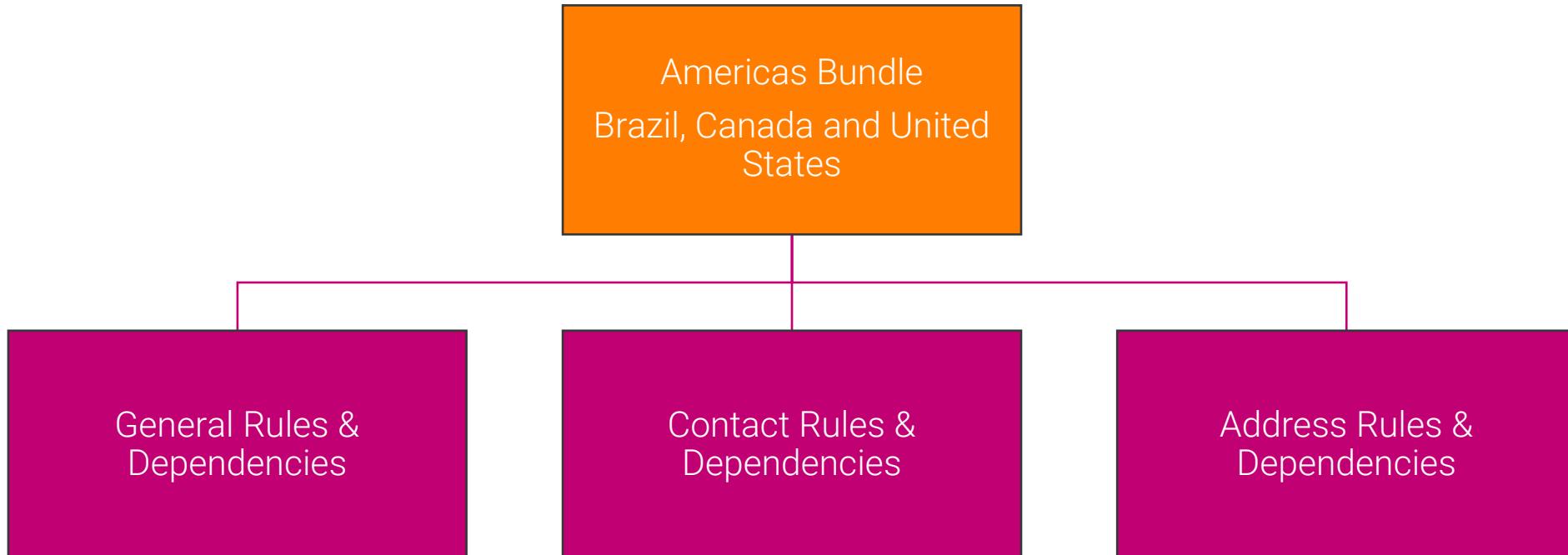
# Cloud Data Quality - Core Bundle

The Core accelerator includes rules that perform the following data quality processes:



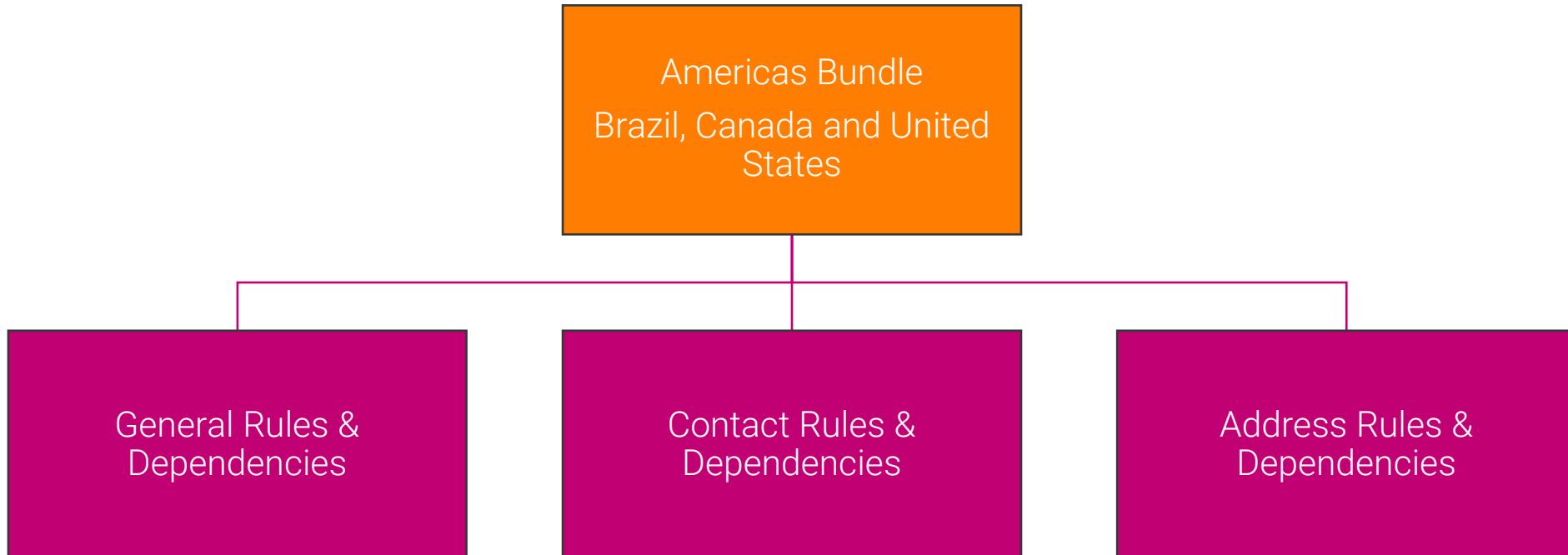
# Cloud Data Quality – Americas Bundle

The Americas accelerator includes rules that perform the following data quality processes:



# Cloud Data Quality – United Kingdom

The UK bundle accelerator includes rules that perform the following data quality processes:



# Cloud Data Quality Bundles

## Where to find them?

Administration->Add-On Bundles->Available Bundles  
Search for “CORE” – Core Bundle  
Search for “CRISIS” - Crisis Response Bundle

The screenshot shows the Informatica Administrator interface. The left sidebar includes links for Organization, Licenses, SAML Setup, Metering, Settings, Users, User Groups, User Roles, Runtime Environments, Connections, Add-On Connectors, Schedules, Add-On Bundles (which is the active tab), Swagger Files, and Logs. The main content area is titled "Add-On Bundles". It has three tabs: "Installed Bundles", "Copied Bundles", and "Available Bundles" (which is highlighted). Below the tabs, it says "Available Bundles (29)". A list of bundles is shown with columns for Name and Description. Some entries include "Informatica Cloud Connectivity as...", "Data Migration Quick Start", "Application Integration Quick Start", "Data Quality Quick Start", "SAP BAPI Quick Start", "Salesforce to Siebel Account Sync", "Siebel to Salesforce Activity Sync", "Siebel to Salesforce Asset Sync", "Siebel to Salesforce Quote Sync", and "Salesforce to Siebel Activity Sync". A callout box points to the "Available Bundles" tab and the "Data Quality Core Bundle" details page. The details page shows the following information:

Bundle Details	
Name:	Data Quality Core Bundle
Description:	The Data Quality Core Bundle is a set of pre-configured data quality assets that you can use to parse, standardize, validate and deduplicate data. These Data Quality Assets cover general, address, contact, corporate and product data. You can copy the content of the bundle to a project or folder and use as-is or create copies to adjust the logic to your business requirement. This bundle is updated from time to time and the modifications will not affect the changes that you make into the copied assets.
Publisher:	Informatica
Version:	1.0
Bundle Type:	Public
Allow:	Copy
Published On:	Feb 3, 2021, 4:16:01 PM

**Bundle Content (205)**

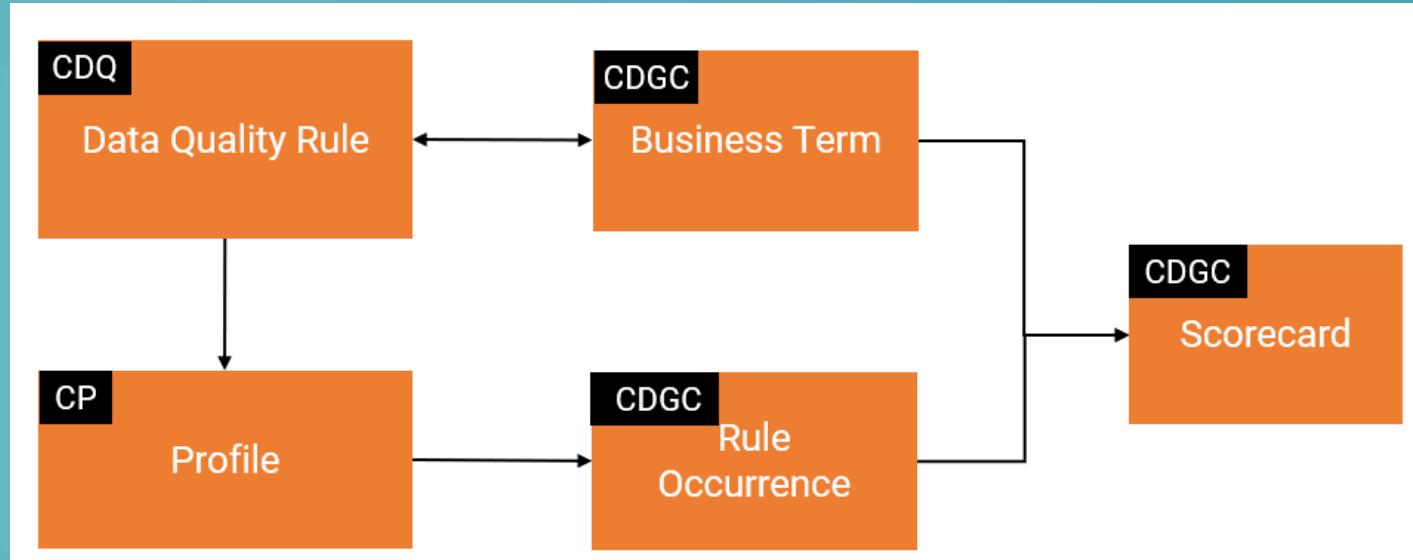
Name	Description	Type
dq_av_availabilityscores_info	This reference table has 'AddressDoctor' status values that indicates the likelihood of delivery to an address on a 0 through 5 scale.	Dictionary
dq_av_match_code_descriptions_info	This reference table has 'Match Code' output values and identifies the process modes that return the values.	Dictionary
Address_Validator_Discrete	DQ Asset(VERIFIER) published from PC to ICS: Address_Validator_Discrete	Verifier

# DEMO

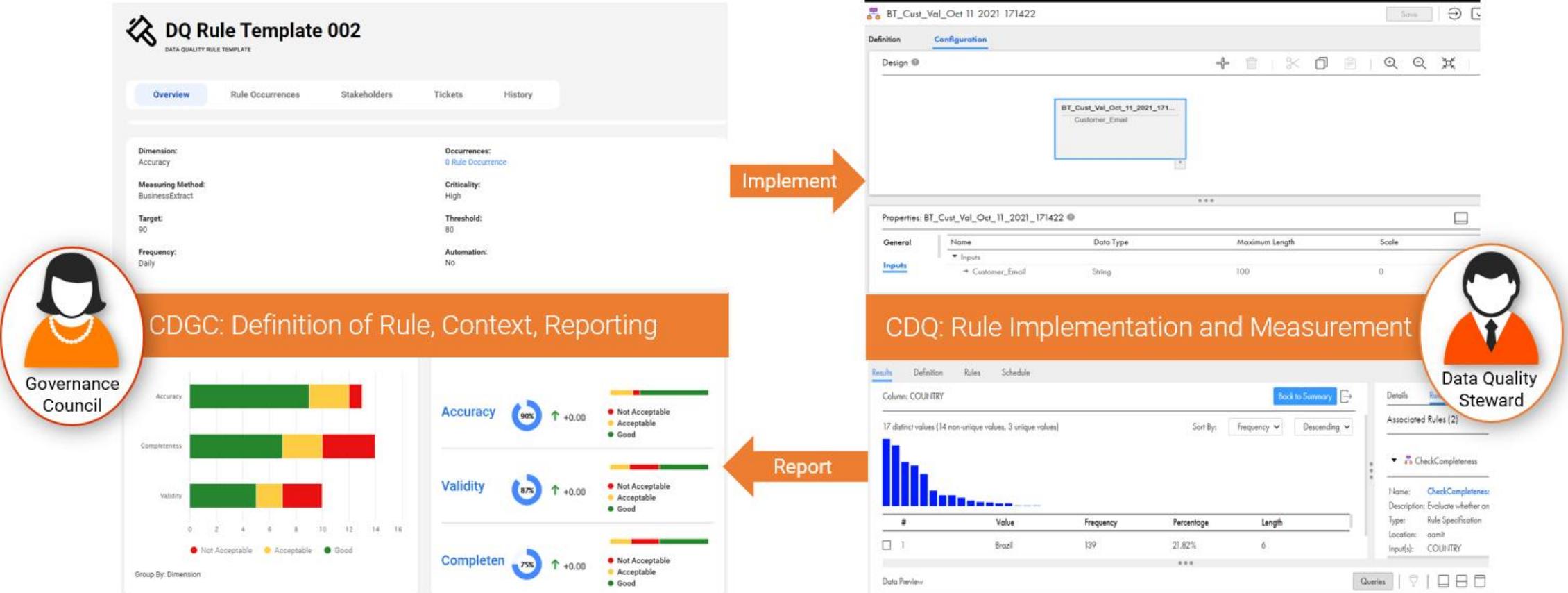


# Cloud DGC – DQ Integration

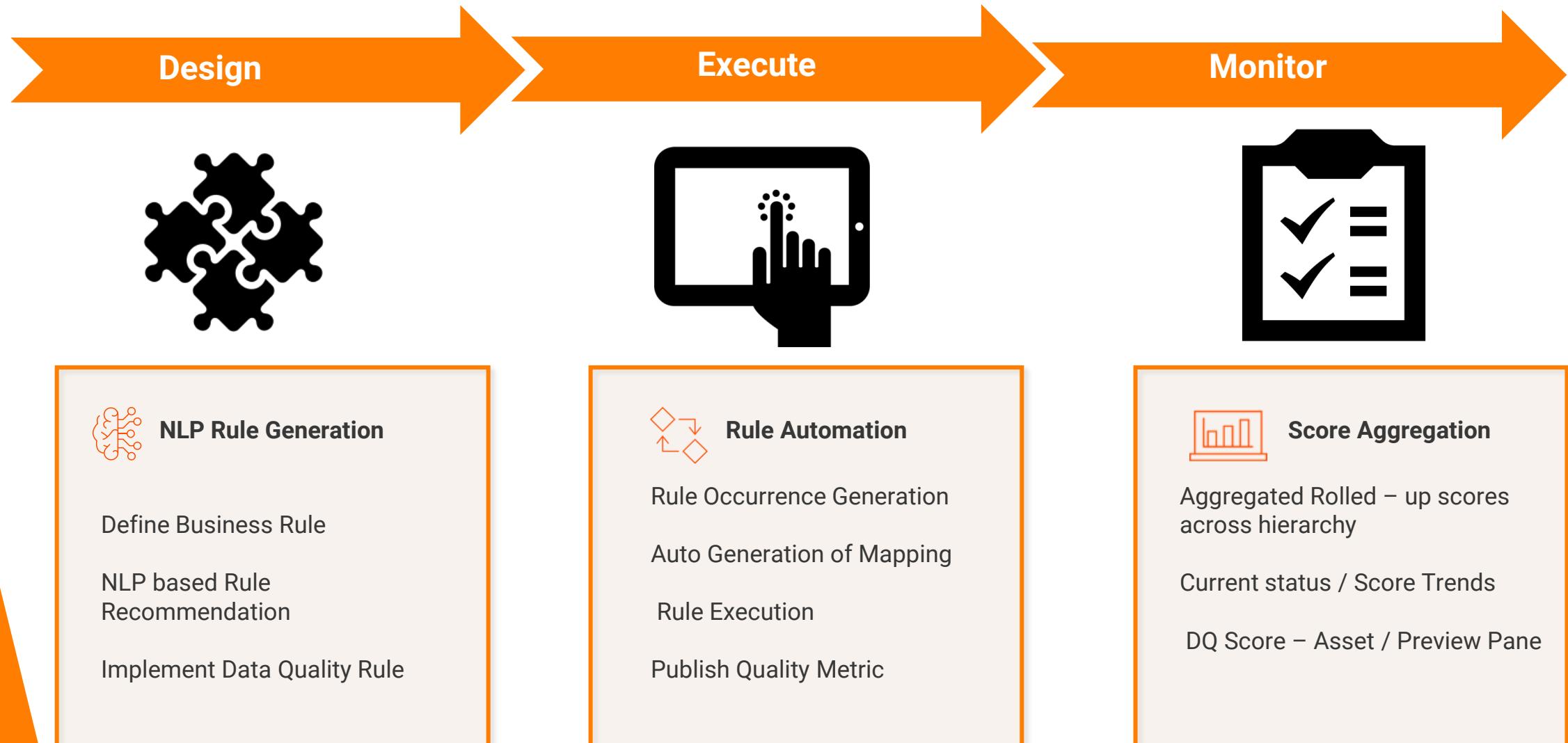
- Data Quality Rules Built using CDQ
- Glossary/ Business Term from CDGC are linked to Data Quality Rules (through Rule Templates)
- DQ Rules are run as part of Profile executions which carry out the statistical counting in order to generate the metrics
- Scorecards and Dashboards based on Cloud Profile Metrics and CDGC Rule Occurrences with Scorecards and Dashboards visualised in CDGC



# Intelligent and Contextual Data Quality



# Highlight at a Glance



# New Data Quality Rule Template

Secondary Glossary:

Name

Type

No data to display

Technical Rule Reference:<sup>\*</sup>

Attach a rule from the connected data quality application.

## Select Technical Rule Reference

Create a Rule     Pick an existing rule

Describe the rule [?](#)    Click here to use the description from the rule template.

First Name cannot be null

[View Recommendations](#)

### CLAIRE Recommendations

Select to view details and create a new rule:

IF First Name != 'NULL' THEN TRUE

CLAIRE will generate and attach the following rule

Rule Name: BT\_Firs\_Acc\_Oct 22 2021 1829

Description: This is a CLAIRE generated rule  
"First Name cannot be null"

Input: First\_Name

IF First Name = 'NULL' THEN TRUE

Existing rule

Create a Rule     Pick an existing rule

Explore | All Assets

All Assets (44)

Name	Type	Updated On	Location	Description	Tags
BT_ACCO_Acc_Oct 07 2021 135551	Rule	Oct 7, 2021, 1:56 PM	Default	x is not null	
BT_ACCO_Acc_Oct 07 2021 142447	Rule	Oct 7, 2021, 2:24 PM	Default	x is not null	
BT_ACCO_Acc_Oct 08 2021 153752	Rule	Oct 8, 2021, 3:37 PM	Default	id not null	
BT_ACCO_Acc_Oct 08 2021 154421	Rule	Oct 8, 2021, 3:44 PM	Default	x is not null	
BT_ACCO_Acc_Oct 08 2021 160809	Rule	Oct 8, 2021, 4:08 PM	Default	x is not null	
BT_ACCO_Acc_Oct 08 2021 170108	Rule	Oct 8, 2021, 5:01 PM	Default	X is not null	

1 - 25 of 44

< 1 of 2 >

Items Per Page: 25

[OK](#) [Cancel](#)

Rule  
Recommendation



## Check address

DATA QUALITY RULE TEMPLATE

**Overview** Rule Occurrences Stakeholders Tickets History

**Primary Glossary:** Address

**Secondary Glossary:**

Name	Type
No data to display	

**Dimension:** Completeness **Occurrences:** 1 Rule Occurrence **Criticality:** High **Measuring Method:** BusinessExtract **Target:**

**Properties**

## Auto-ADDRESS\_LINE1-Completeness-Check address

DATA QUALITY RULE OCCURRENCE

**Overview** Score Stakeholders Tickets History

**Score Trend**

Score Status: Good  
Volume: 24  
Exceptions: 0  
[View Trend Details](#)

**Dimension:** Completeness **Frequency:** Adhoc **Criticality:** High **Measuring Method:** BusinessExtract **Target:**

**Rule Occurrence**

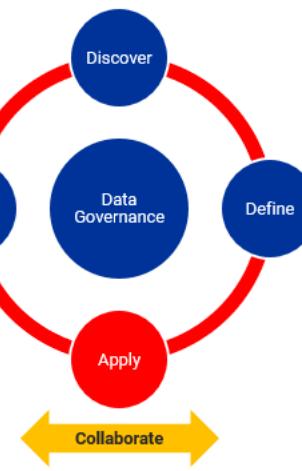
**Check address**  
DATA QUALITY RULE TEMPLATE

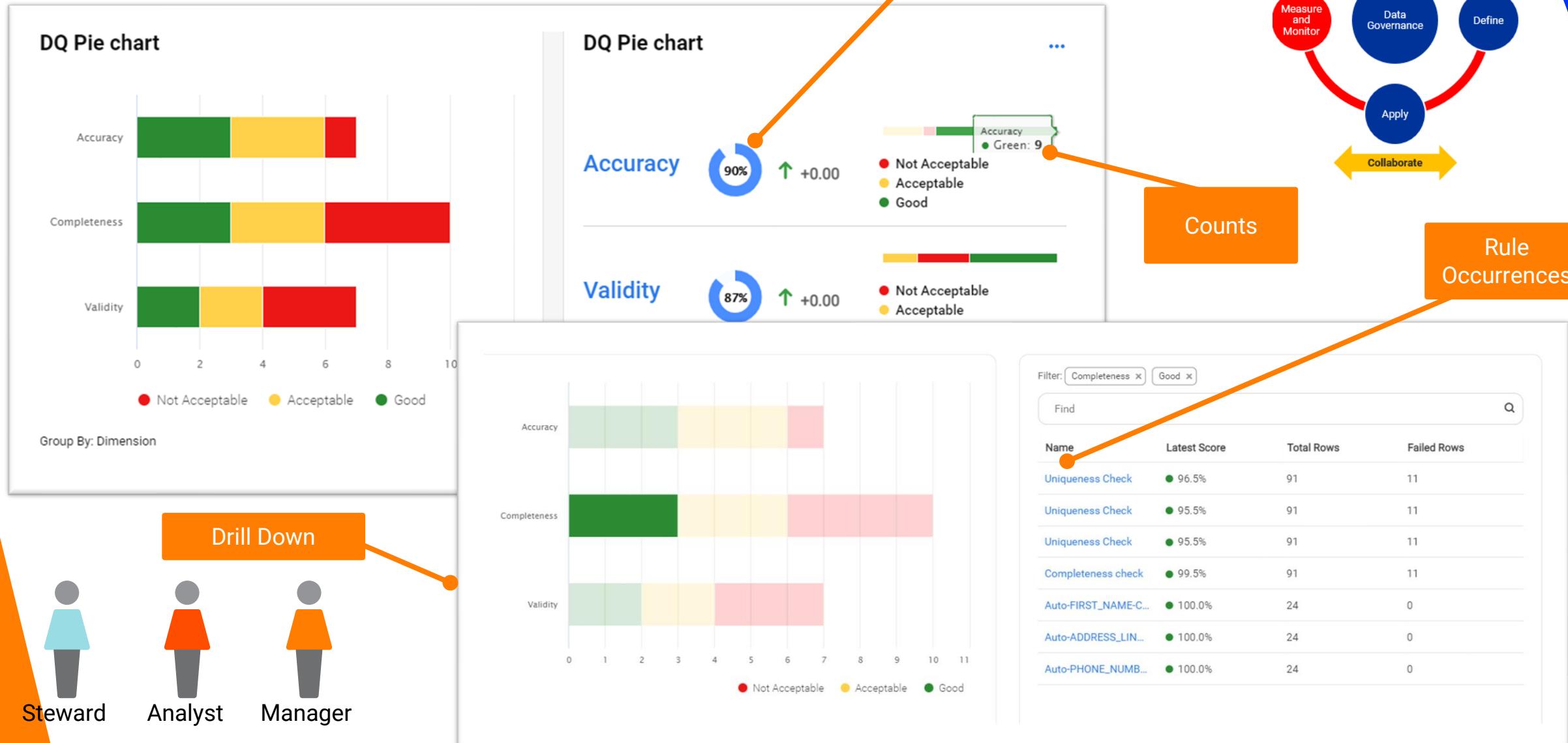
**Overview** Rule Occurrences Stakeholders Tickets History

**Rule Occurrences (1)**

Name	Dimension	Latest Score	Date	Total Rows	Failed Rows	Primary Data Element
Auto-ADDRESS_LINE1-Completeness-Check address	Completeness	100%	Oct 11, 2021, 6:28 PM	24	0	ADDRESS_LINE1

Score Trends







# Enterprise Data Catalog Integration

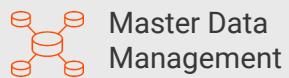
# The Catalog of Catalogs



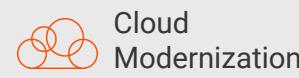
Self-Service  
Analytics



Data  
Governance



Master Data  
Management



Cloud  
Modernization



Metadata  
Intelligence



Data  
Integration



Data  
Quality

Open APIs, Full Access

- Discovery
- Profiling
- Lineage
- Impact Analysis

- Semantic Search
- Domain Discovery
- Similarity Clustering
- Business Term Association



Enterprise Data Catalog  
Metadata System of Record  
for The Enterprise

- Relationships
- PK-FK Discovery
- Business Context
- Custom Annotations

- Reviews/Ratings
- Questions/Answers
- Data Certifications
- Change Notifications



Knowledge Graph + AI/ML

**CLAIRE™**



Breadth of Active Metadata



On-prem  
Databases



File  
Systems



BI  
Tools



On-prem/  
SaaS Apps



ETL



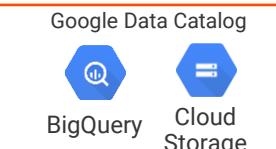
SAP



aws



Microsoft Azure

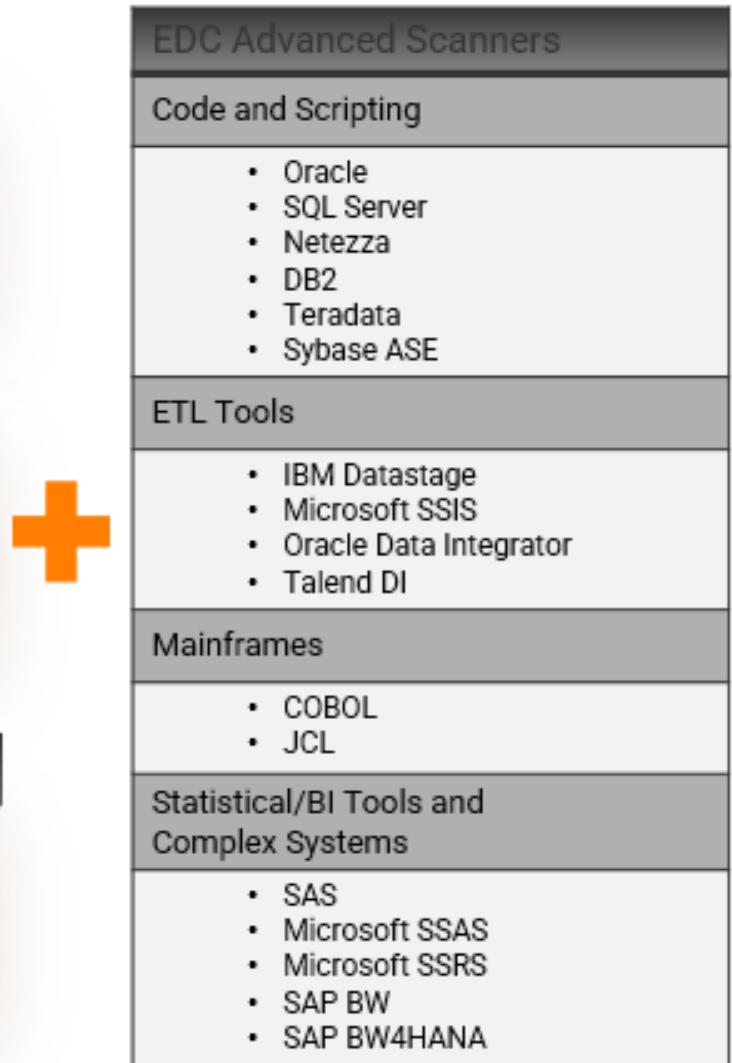
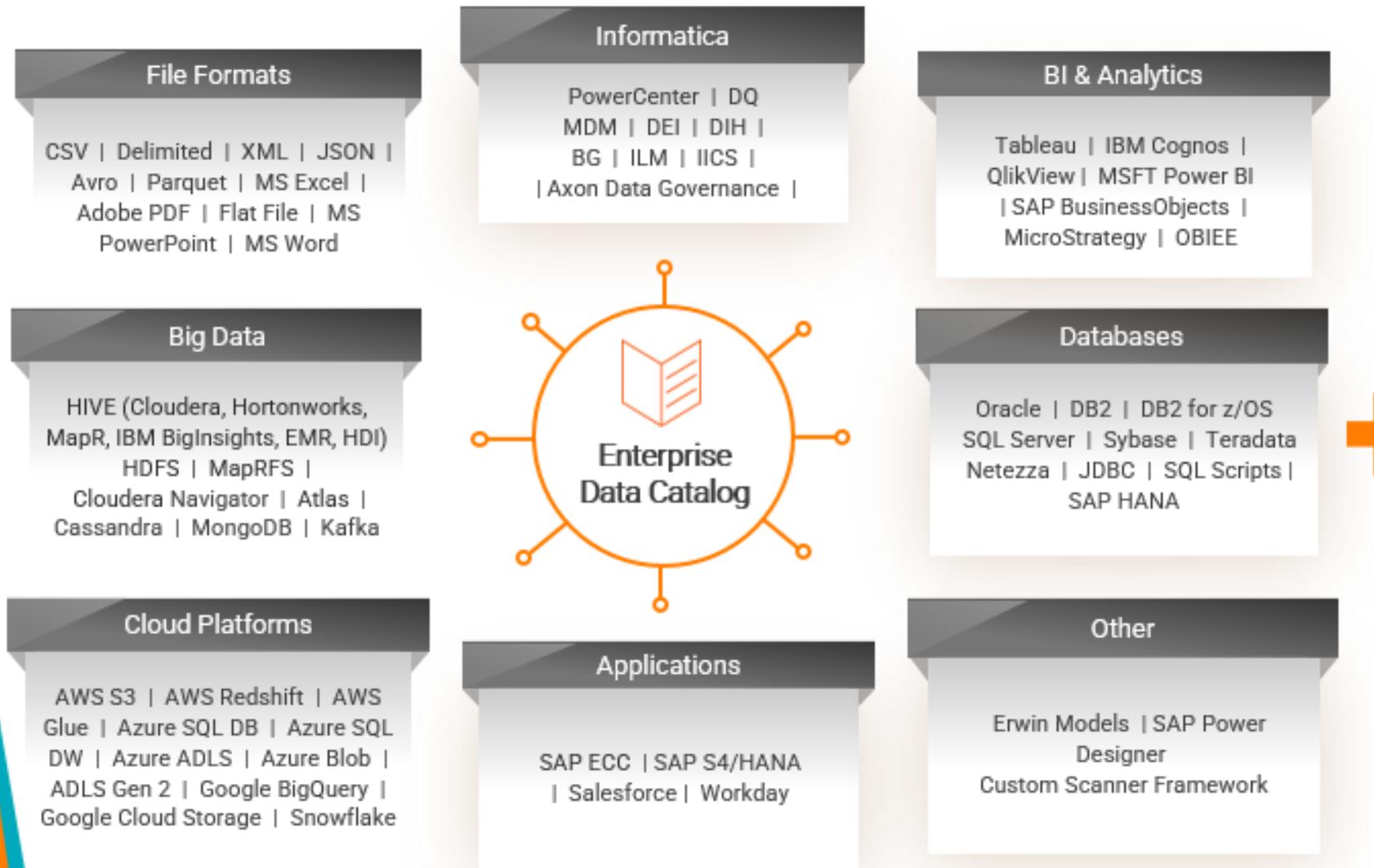


Google Cloud



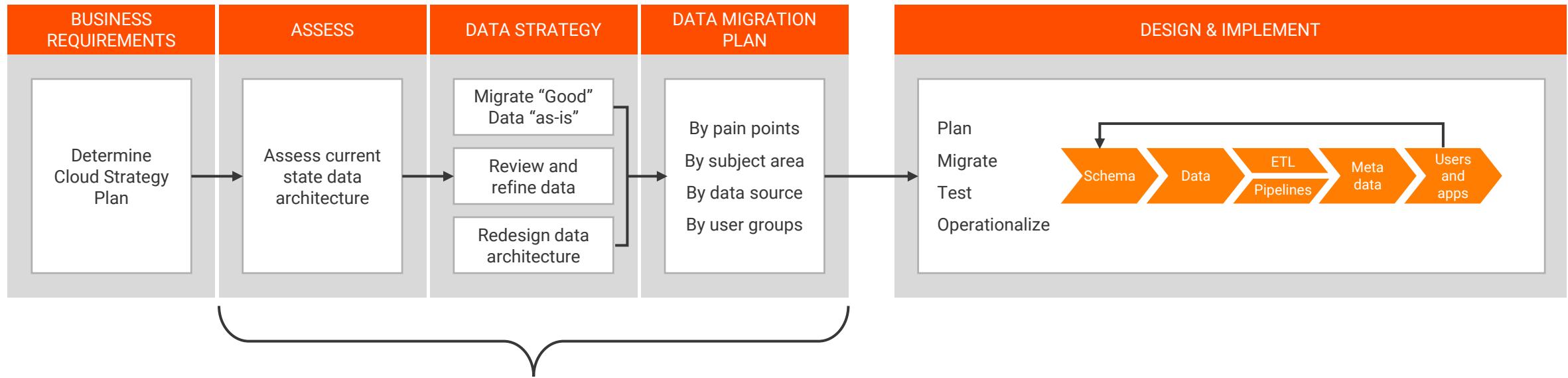
Informatica

# BROADEST, MOST COMPLETE METADATA CONNECTIVITY



# How Can a Catalog Help with Cloud Data Migration?

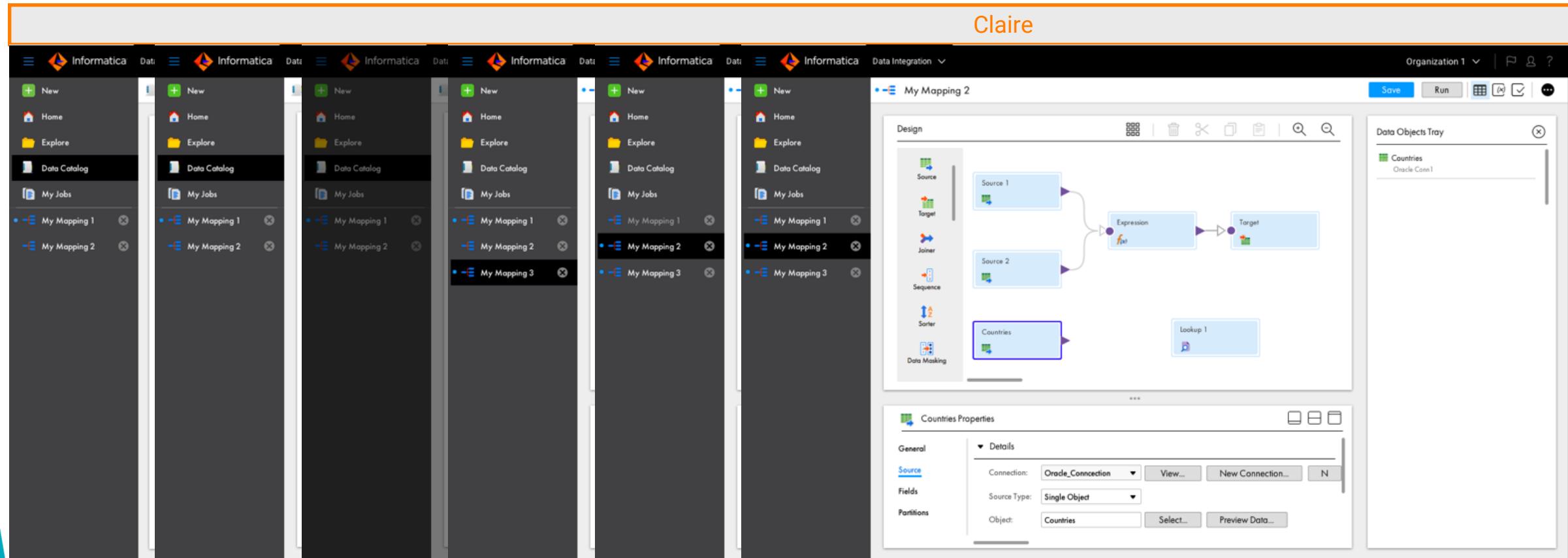
- Typical High-level Cloud Data Migration Process



A **Data Catalog** can help assess part of the cloud data migration process to better understand the nature of the data

# Enterprise Data Catalog Integration

ML Driven Data Discovery and Integration

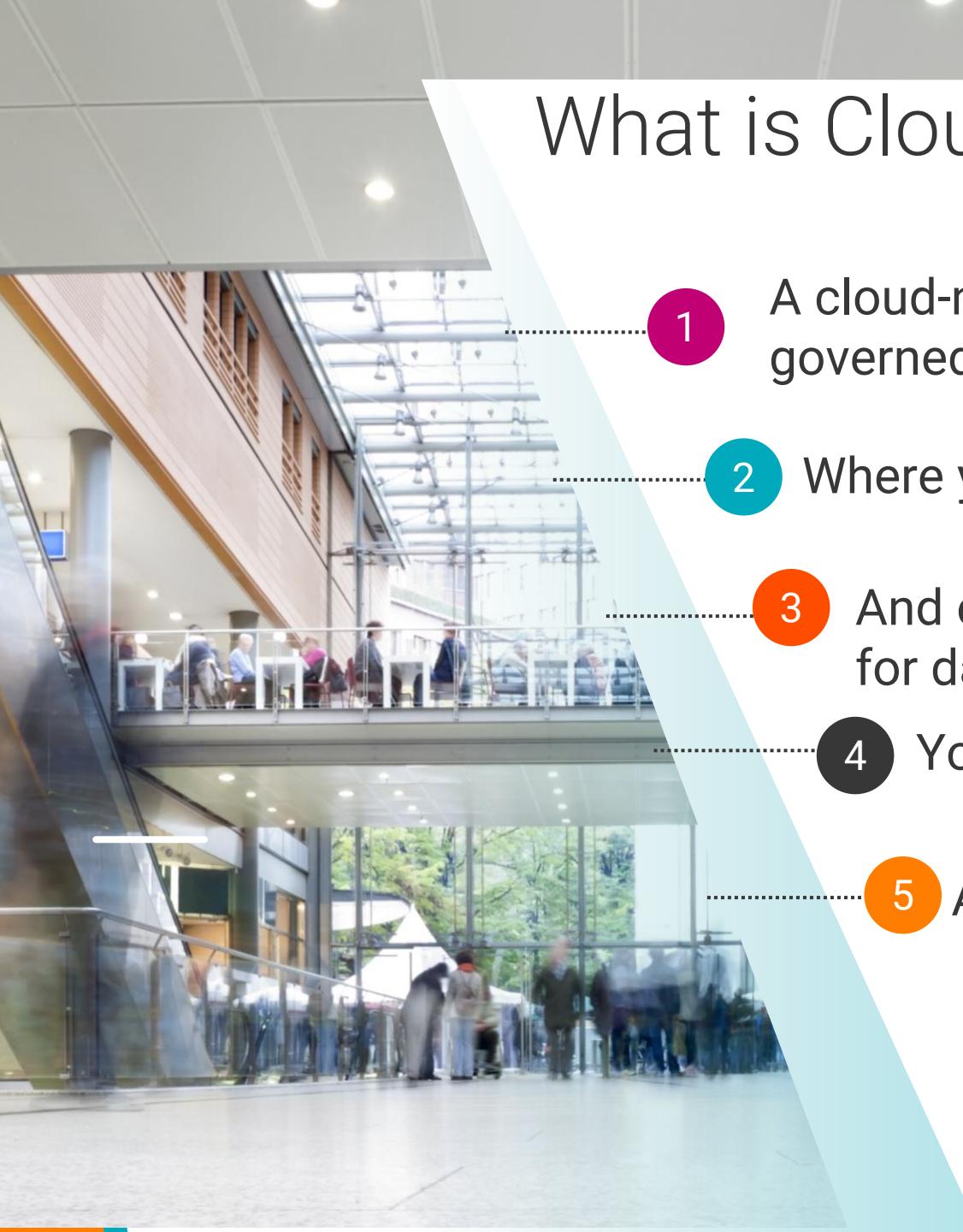


# DEMO





# Cloud Data Marketplace

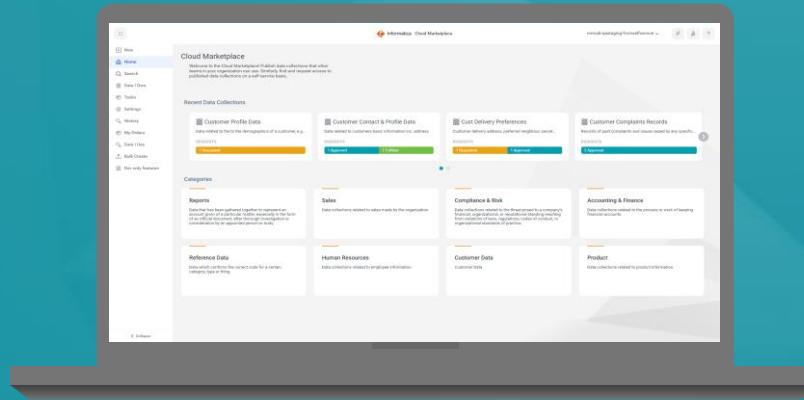


# What is Cloud Data Marketplace?

- 1 A cloud-native storefront, with order management, and governed delivery capabilities
- 2 Where you can package data assets from any data source
- 3 And easily ensure your data consumers can shop for data and AI models
- 4 Your teams can quickly fulfill their requests
- 5 And track requests, data usage and more

# Cloud Data Marketplace

A storefront, with order management, and governed delivery capabilities



Package Data Assets

A screenshot of the "Marketing Vlog Meta" data asset details page. The top navigation bar shows the asset name and a "Marketplace" tab. The main content area is divided into sections: "Summary" (with fields like ID, DESCRIPTION, DATA SOURCE, and LIFECYCLE STATUS), "Data Elements" (listing various fields like "Marketing ID", "Video URL", and "Thumbnail URL"), and "Linked Collections" (listing "Marketing Vlog Meta" and "Marketing ID"). A sidebar on the right shows "Recent Data Collections" and "Data Assets".

Shop & Checkout

A screenshot of the "Marketplace" storefront for the "Engineering" department. The top navigation bar shows the department name and a "Marketplace" tab. The main content area includes a summary section with "Requests: Last 90 Days" (49 Requests, 39 Active, 36 Inactive, 4 Pending, 1 Late), a "Recent Data Collections" grid (Steering Review, Fleet DOE Fleet, Customer Ratings), and a "Data Collections" table. A sidebar on the right shows "Recent Data Collections" and "Data Assets".

Fulfill & Track

A screenshot of the "Marketplace" storefront for a "Fleet DOE Fleet Product". The top navigation bar shows the product name and a "Marketplace" tab. The main content area includes a summary section with "Requests: Last 90 Days" (49 Requests, 39 Active, 36 Inactive, 4 Pending, 1 Late), a "Request Details" form (Business Justification: "Need access to the finance or information for the end of year review - major contractual factor, factor rate inquiry, planned fuel cost, current parts, gas consumption, etc.", Delivery Task: "Get Fleet Management info"), and an "Order Summary" table. A sidebar on the right shows "Recent Data Collections" and "Data Assets".



Informatica®

# DEMO

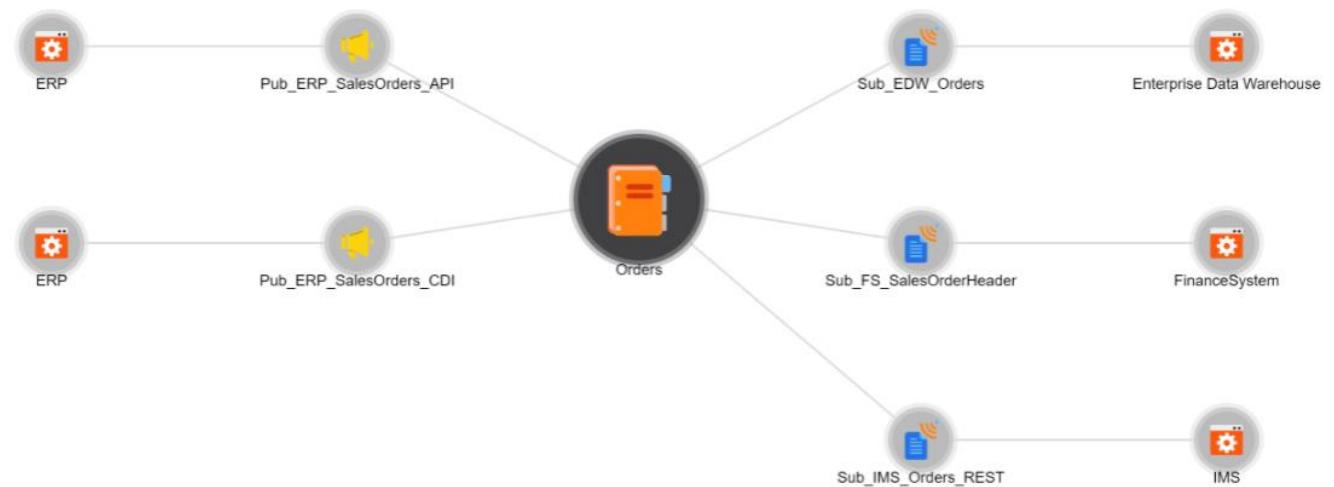
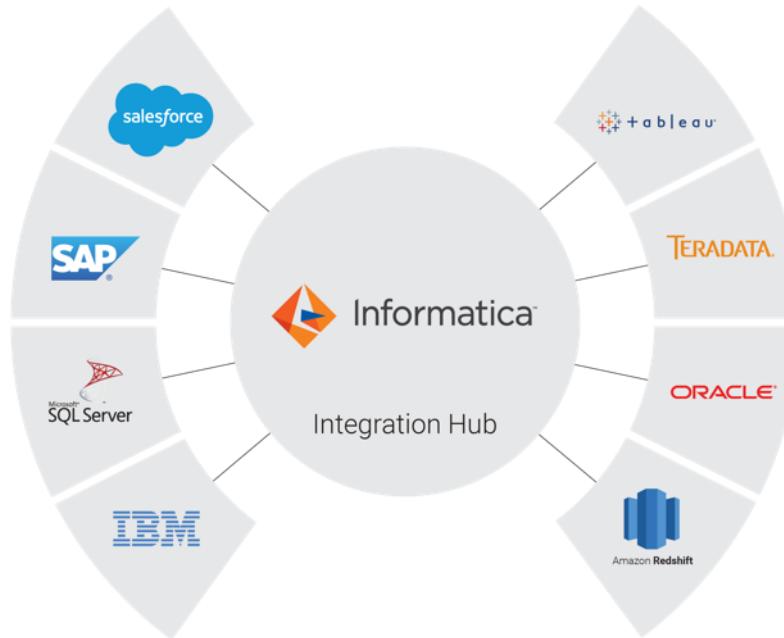




# Cloud Integration Hub (CIH)

# What is Cloud Integration Hub (CIH)

The Cloud Integration Hub is a great solution to cater data distribution scenarios with a store and forward mechanism.



# Use Cases

# Use Cases Driving Integration Hub Initiatives

Application Synchronization	Mixed Latencies	Integration Modernization	Governance of Integration
On-board new systems and applications from acquisitions and other business units efficiently.	Coordinate between different frequency of data availability and consumption.	Modernize data infrastructure for consistent data across systems.	Govern and Control Data Flowing in and out of applications.

# Challenges that companies are facing



## Cost per integration going up

Business users do not have access to right data

Not able to bring systems on-board fast enough

Data stored far from consuming applications



## Data loss causes business to slow or stop

Too many developers hand coding integrations

Draining resources on point to point integration tasks

Data transfer fees and API call charges going up



## Sub-optimal integration design delaying time to market

Data loss/delay occurs as tightly coupled applications fail and stop communicating

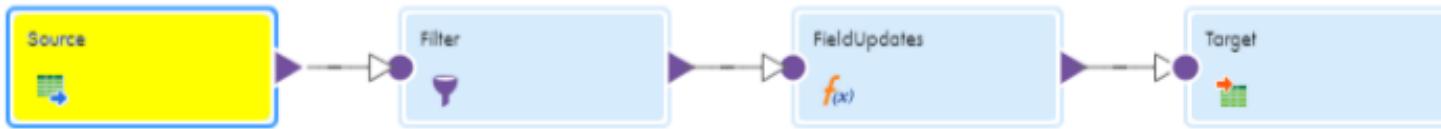
Downtime getting costlier by the minute

Network bandwidth is limited

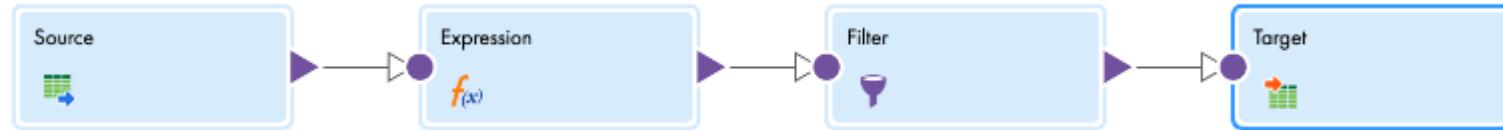
# Example 1: Cost per integration – how it's going up

- Simple data flow mappings turn into a “hairball” as the integration tasks gets repetitive and complex

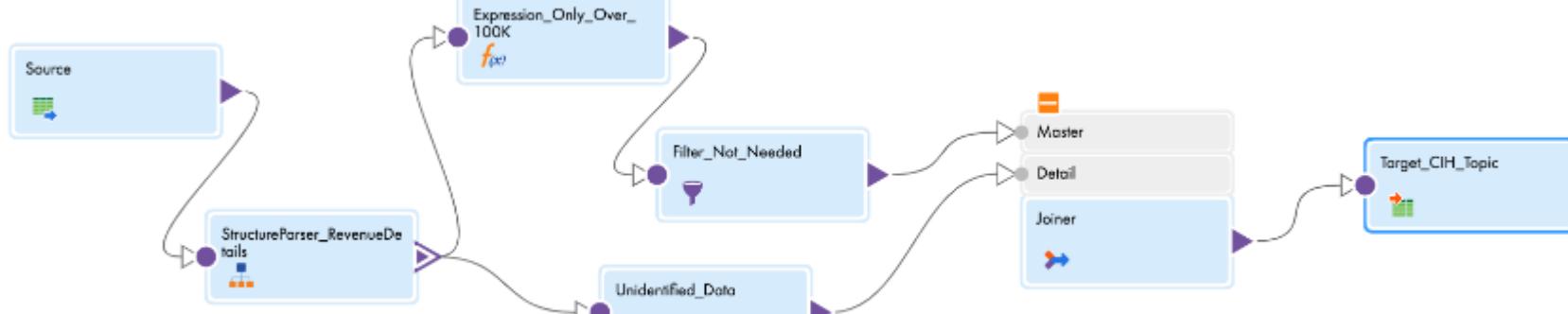
Dataflow Itr-1



Dataflow Itr-2



Dataflow Itr-3



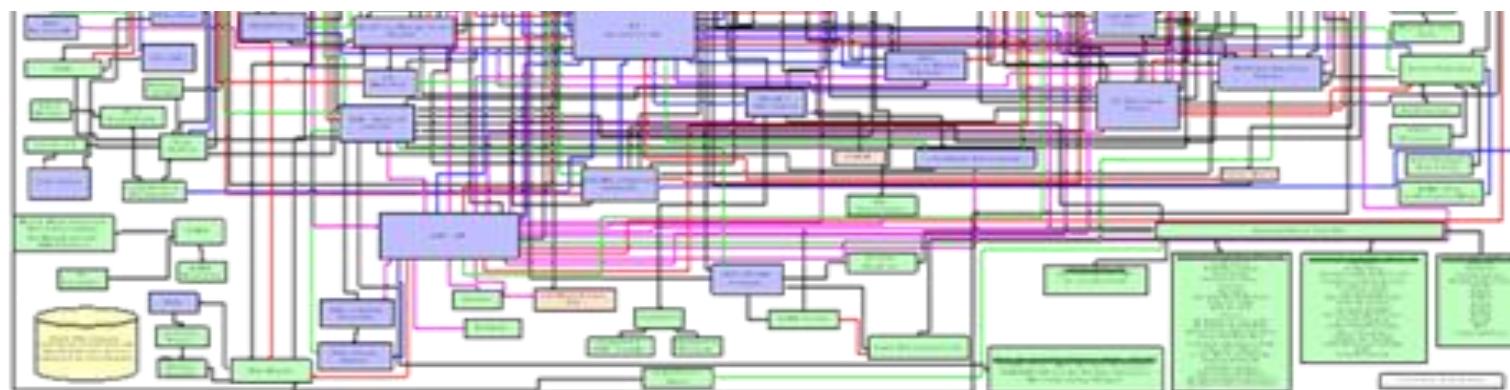
Point to point data integration architecture increases the number of connections and the cost

# Example 1: Cost per integration – how it's going up

- Simple data flow mappings turn into a “hairball” as the integration tasks gets repetitive and complex

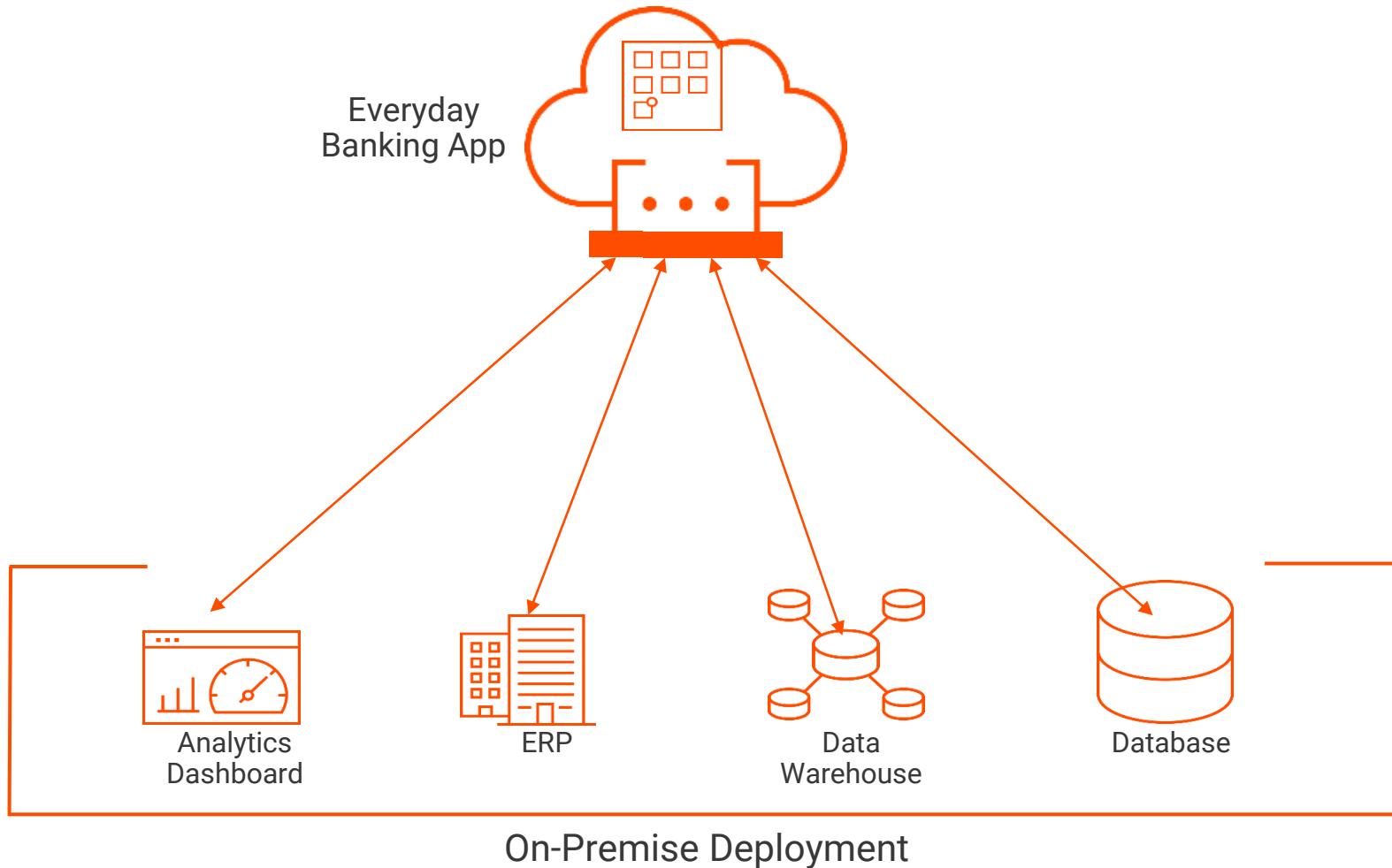


Point to point data integration architecture increases the number of connections and the cost



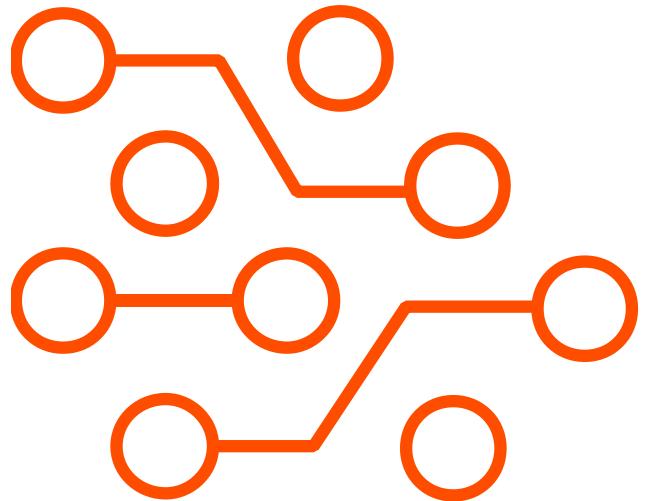
# Example 2: Cost is going up due to integration

You pay API charges and transfer fees every time you fetch data from the cloud



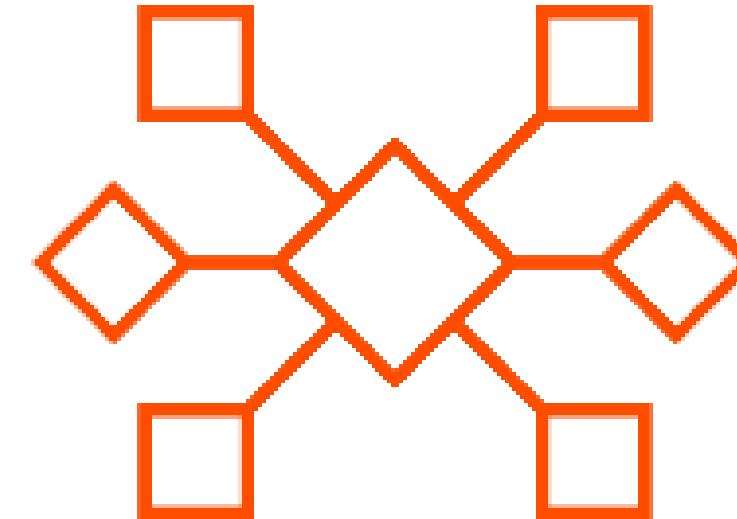
# A Hub and a Spoke integration pattern is what's needed

- It improves internal efficiency by reducing the number of connections



Point to Point Integration

10 sources and 10 targets mean 100 connections

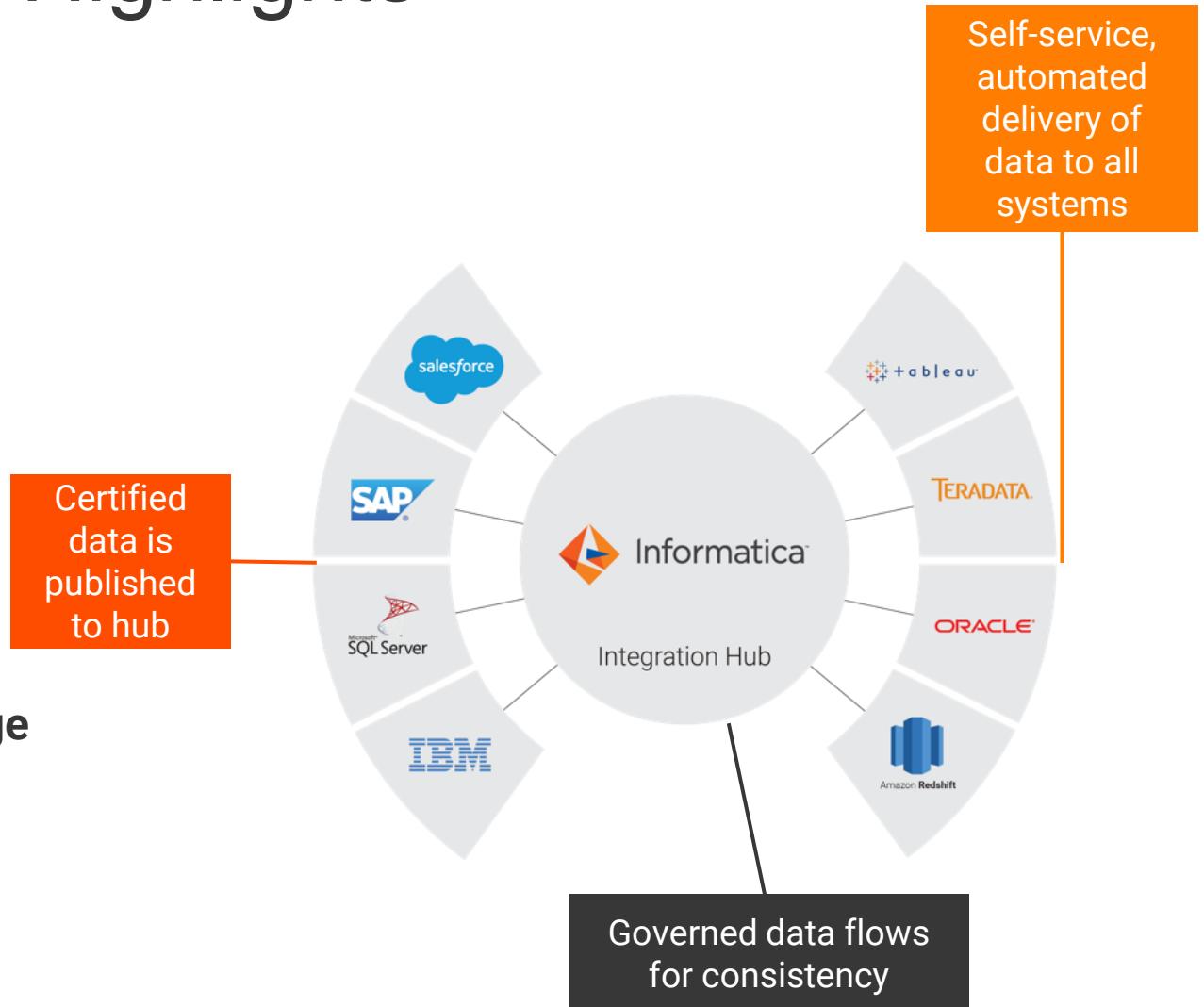


Hub n Spoke Integration

10 sources and 10 targets mean 20 connections

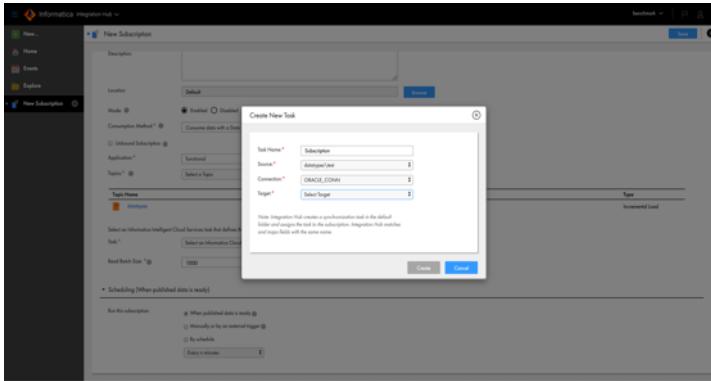
# Integration Hub – Features Highlights

- **Decouple** source from target
- **Standardize** data access with canonical Topics
- **Hosted or Private** Publication Repository
- **REST APIs**
- Internal **scheduler** for Pubs/Subs
- View end to end data flow with integration **lineage**
- Track and **monitor** events with granular search



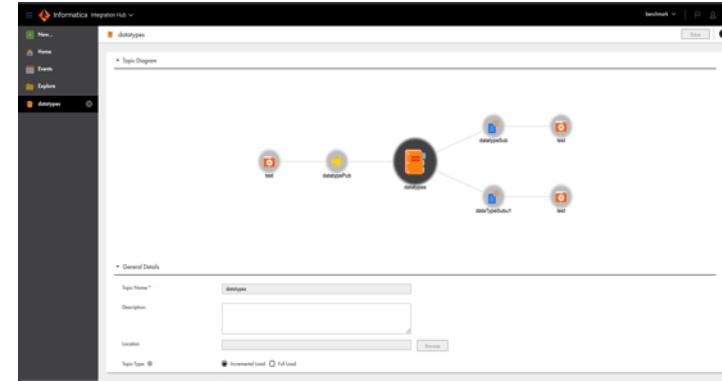
# Cloud Integration Hub tightly coupled with CDI as part of IICS

## Self Service Integration and Management



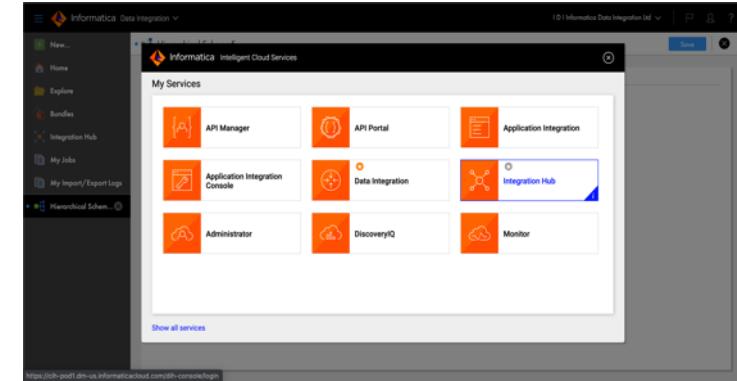
- Simple interface to manage Topics, Applications, Publications and subscriptions
- Pre-configured Accelerators for commonly used cloud application
- Associate CDI tasks with CIH assets (pub/sub – Mappings/DSS)

## Visibility across workflows



- Manage your integration workflows
- Monitor data flows and see data lineage
- Track events and trace problems when they occur and prevent data loss
- Custom proactive alerts for 24/7 control
- Reprocess failed events and for full resilience of data flows

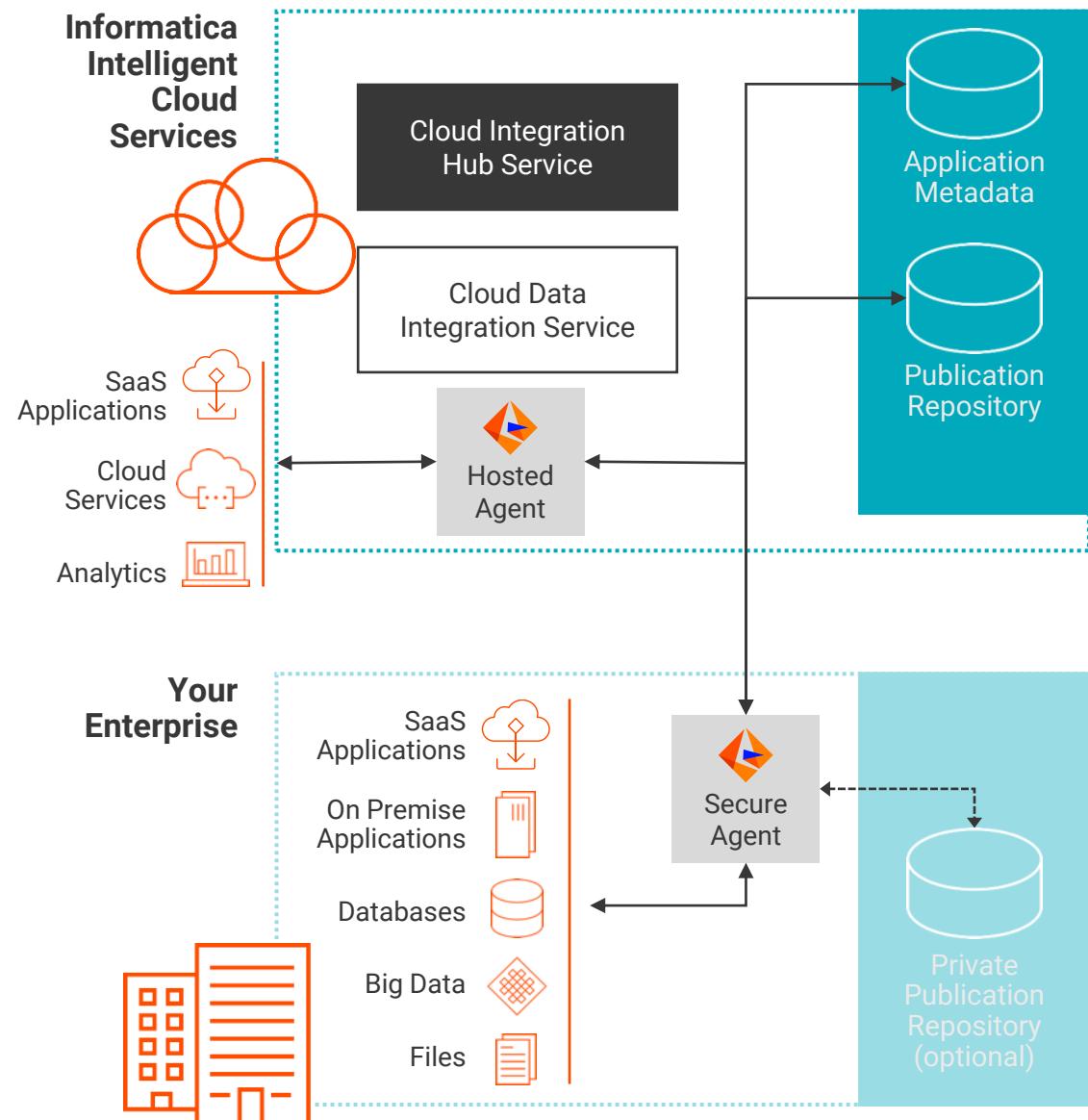
## Integrated with iPaaS Tasks



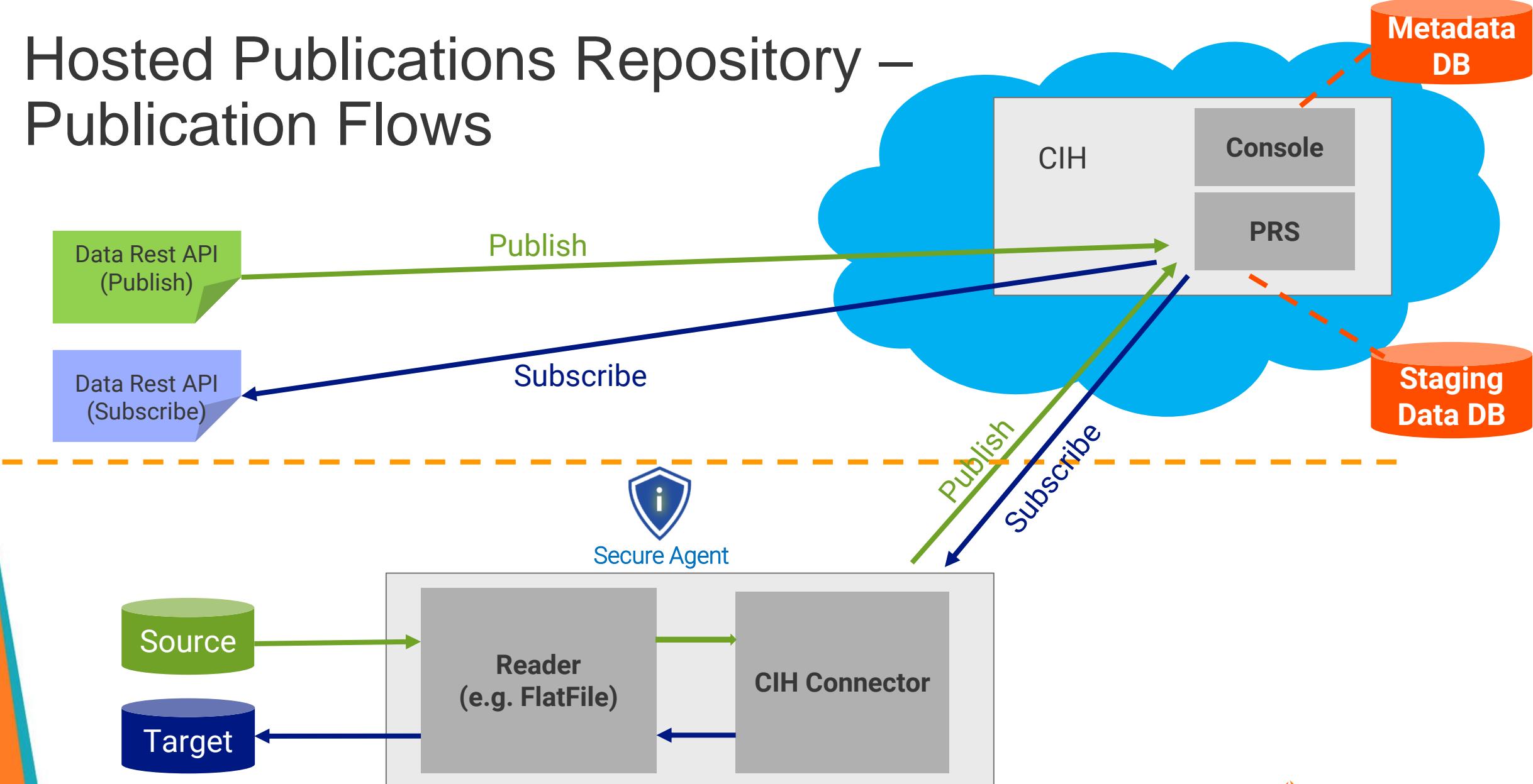
- Utilize Informatica Cloud mappings and tasks for publish/subscribe workflows
- Auto create simple DI tasks from CIH interface
- Use Pub/Sub APIs with CAI as well as use CAI as a scheduler

# Cloud Integration Hub Architecture

- Hosted on IICS
- Support full Cloud and Hybrid mode
- Hosted or Private Publication Repository
- Secure and Encrypted, SOC-2 Compliant
- Source and Target De-coupling
- Multi-Latency, Batch & APIs
- Connectivity to any service or application



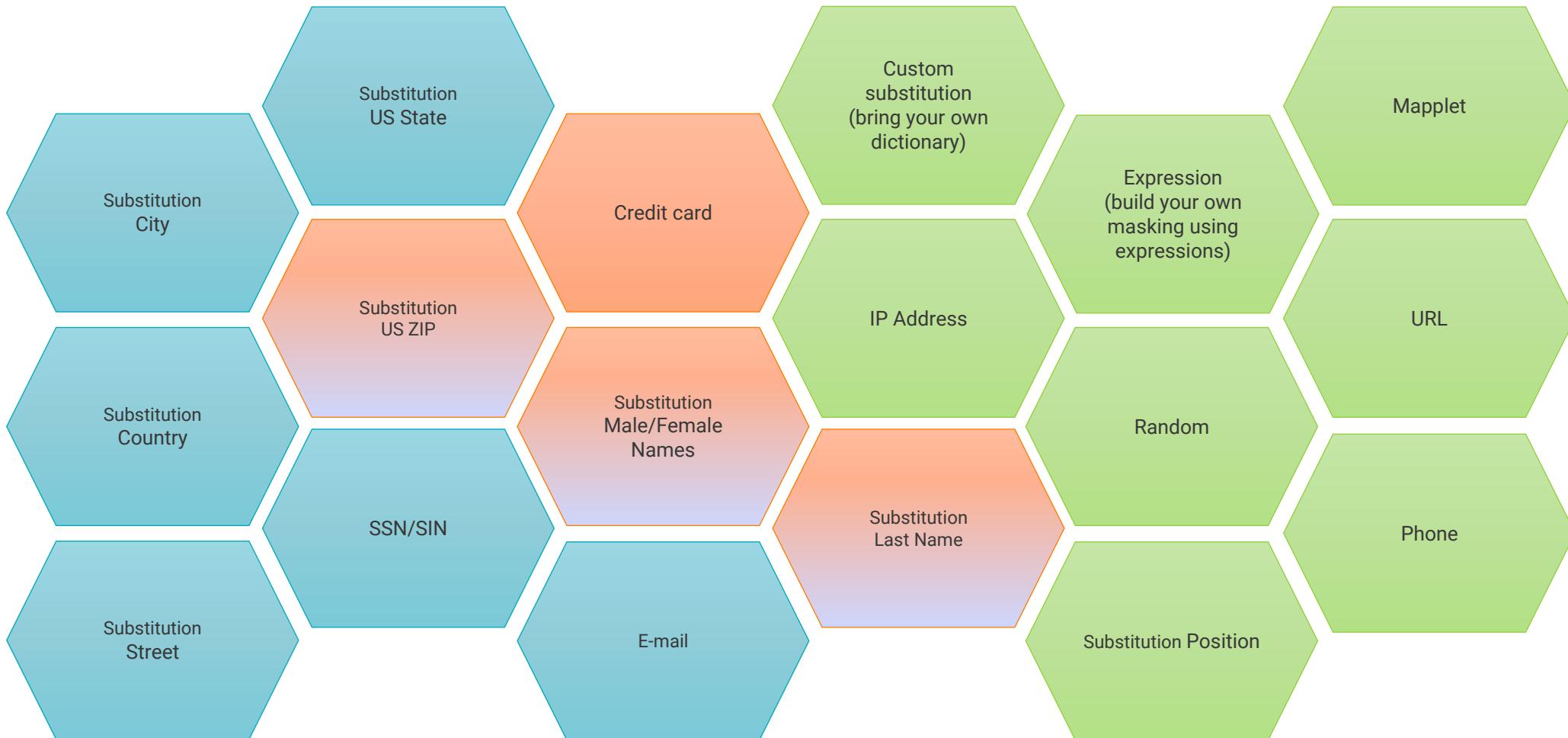
# Hosted Publications Repository – Publication Flows





# Cloud Test Data Management

# Out-of-the-Box Masking Techniques



Regulations: PII, PHI, PCI

# Consistent Masking

Cloud Data Masking can mask consistently across different sources and different runs

**Repeatable**: same masked result for the same input

Original
John Adams
Nicholas Cage
John Adams

Masked
Tim Jones
Jack Nicholson
Tim Jones

Repeatable ON

Original
John Adams
Nicholas Cage
John Adams

Masked
Tim Jones
Jack Nicholson
Fred Mercury

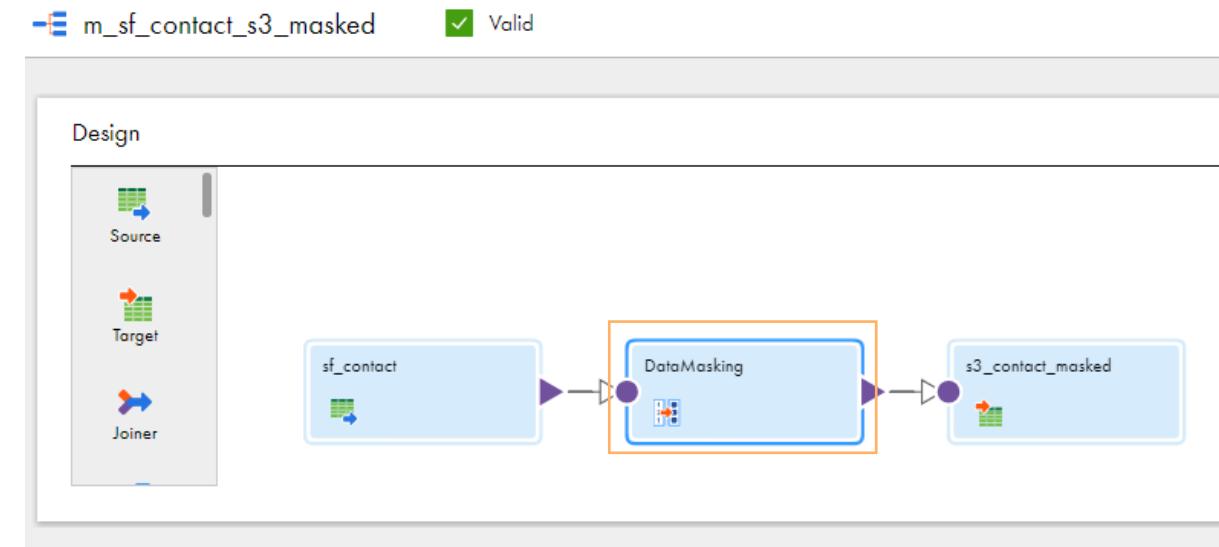
Repeatable OFF

**Seed value**: point for generating masked values. If the same seed value and same dictionary (for substitutions) is used, same result will be obtained

# Data Masking Transformation

Available on Cloud Mapping Designer

- Designed to mask sensitive data in data integration use cases
- **Supports all sources supported by Cloud Mapping Designer**
- Embeds several masking techniques: SSN, credit card, substitution, random and others
- Not designed for creation of new test environment systems



# Data Masking and Subset for salesforce.com

## Secure and Populate Sandbox Copies

- Masks existing sandboxes
  - Ensures data privacy
  - Out of the box PII, PHI, PCI data masking rules
- Create test data sets for dev sandboxes
  - Populate empty sandboxes with referentially intact data sets
- Rationalize existing SFDC investment
  - Saves cost of additional full sandbox copies

The screenshot shows the Informatica Data Masking interface. At the top, there is a navigation bar with tabs: 1 Definition, 2 Source, 3 Target, 4 Data Filters, 5 Masking, and 6 Schedule. The 'Definition' tab is selected. Below the tabs, there is a form for 'Task Details' with fields for 'Task Name' (set to 'Salesforce Mask PII'), 'Location' (set to 'REG\Default'), and a 'Description' field. A 'Browse' button is also present. A large orange box highlights the 'Definition' tab and the 'Task Details' section.

Below the main task details, there is a sub-section titled 'New Salesforce Mask PII' with its own set of tabs: 1 Definition, 2 Source, 3 Target, 4 Data Filters, 5 Masking, and 6 Schedule. The 'Masking' tab is selected here. This section is titled 'Define Masking Rules' and contains a table where masking rules are defined for various fields of the 'Account' object. The table has columns for 'Status', 'Name', and 'Masking Rule'. Fields listed include Account Number, Account Source, Active, Annual Revenue, Billing City, Billing Country, Billing Geocode Accuracy, Billing Latitude, Billing Longitude, Billing Zip/Postal Code, Billing State/Province, Billing Street, and BusinessHours1. Most fields have their masking rule set to 'Substitution [Field Name]' and a 'Configure...' button. The 'Billing City' and 'Billing Street' fields have their masking rule set to 'Substitution Street' and a 'Configure...' button.



---

# Question?

# Thank You

---

*Global Technical Alliances (GTA) and PTS Team*



Informatica®