26th – 28th April 2022

Informatica Cloud Data Integration Bootcamp

Global Technical Alliances (GTA) and PTS Team

Informatica

Agenda – Day2

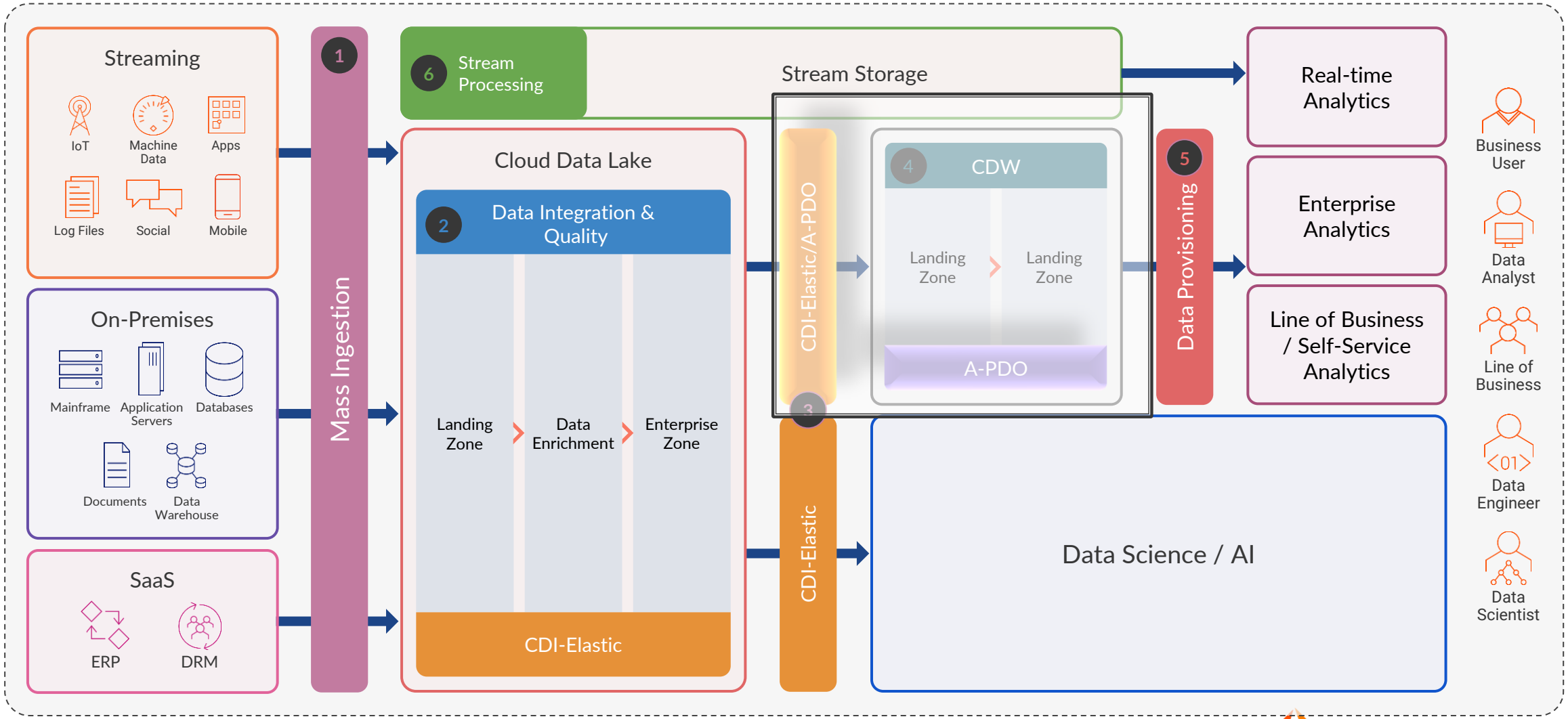
9:30 AM–11:30 AM BST | 10:30 AM CEST – 12:30 PM CEST | 2:00 PM–4:00 PM IST

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

Informatica

Advanced Pushdown Optimization(APDO)

Informatica Data Warehouse and Datalake Architecture



Informatica

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

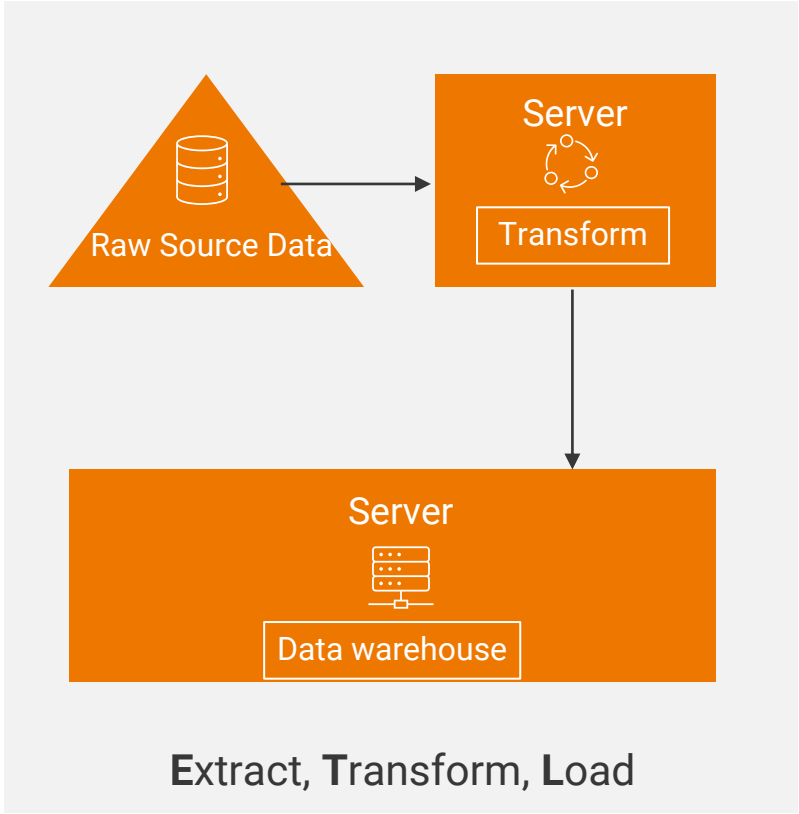
Research Circle Survey on data management strategy from Gartner -

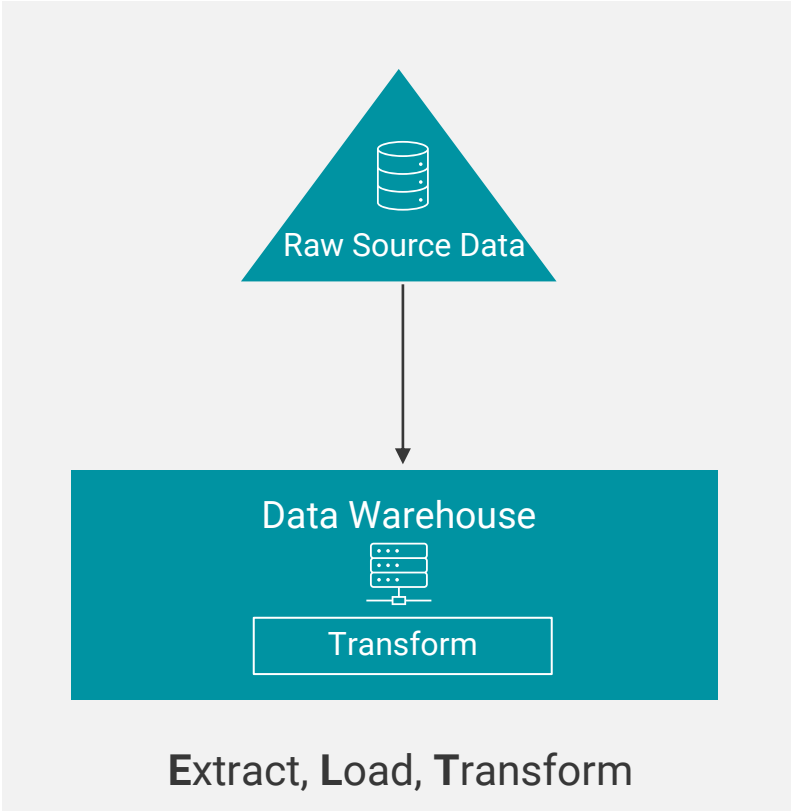
•By 2022, public cloud services will be essential for 90% of data and analytics innovation.
•By 2023, cloud-based AI will increase 5x from 2019 making AI one of the top workload categories in the cloud.
•Out of 200 survey respondents, more than one-third (38%) are already using cloud data warehouses (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.

Informatica

ETL vs ELT

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





Informatica

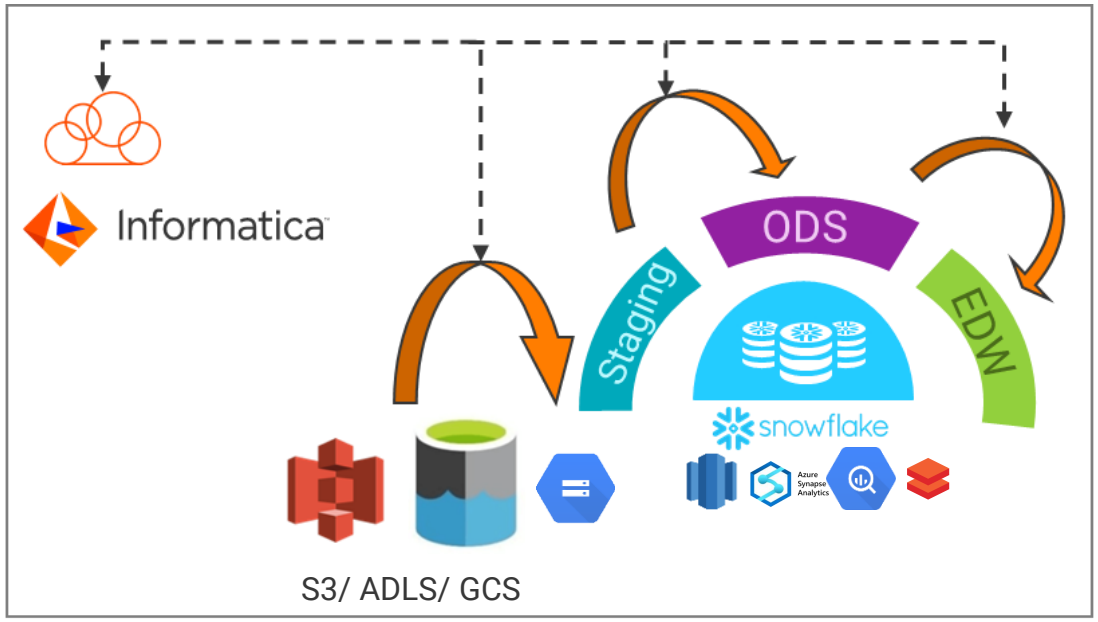
Advanced Pushdown Optimization(APDO)

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

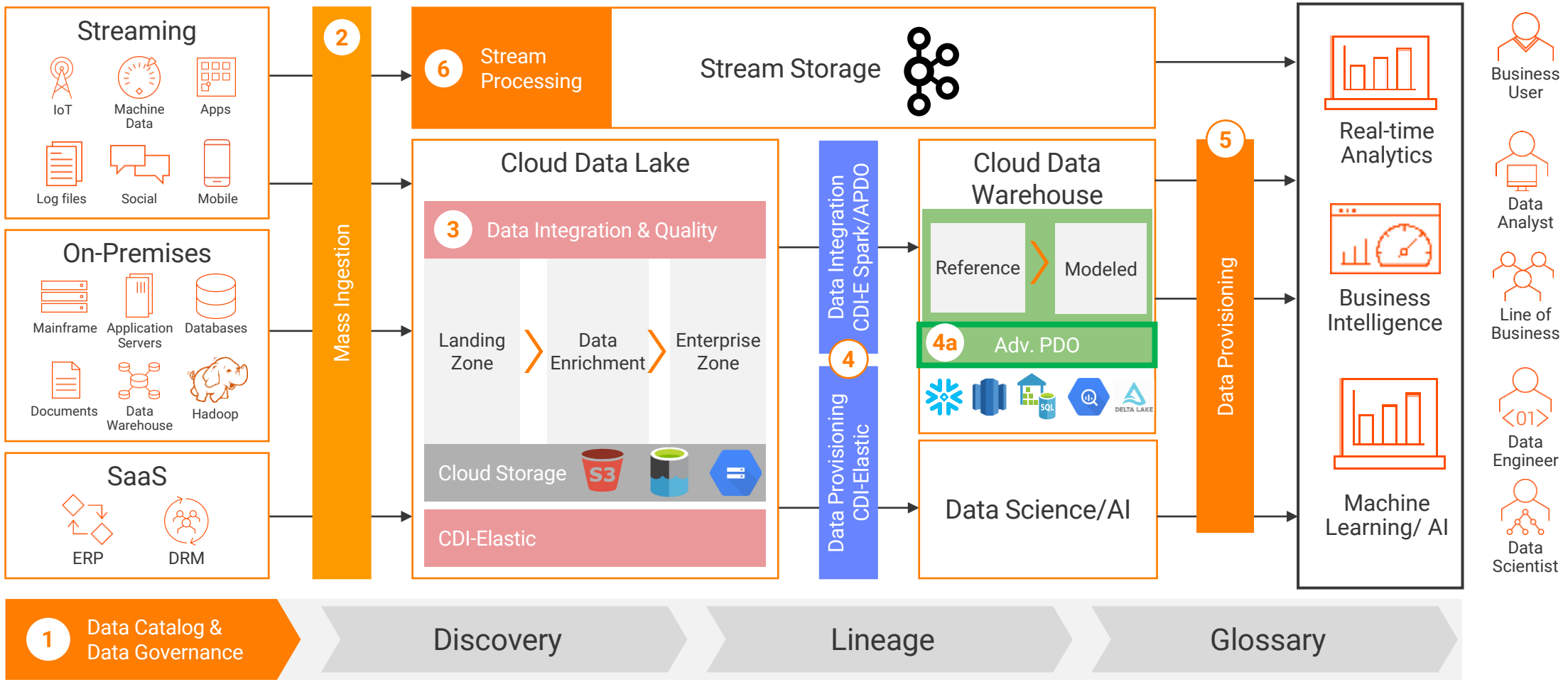
- Data pipeline logic gets translated into Cloud ecosystem based native SQL (SQL Based PDO) or native ecosystem API/ commands (Ecosystem PDO) based on the Data integration pattern
- Support for Full, Source, Partial PDO
- Broadest array of connectors and support for all major ecosystems (CDL/CDWs)
- 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

aws Microsoft Azure Google Cloud snowflake databricks Informatica

Informatica Reference Architecture for Data Management



Informatica

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.

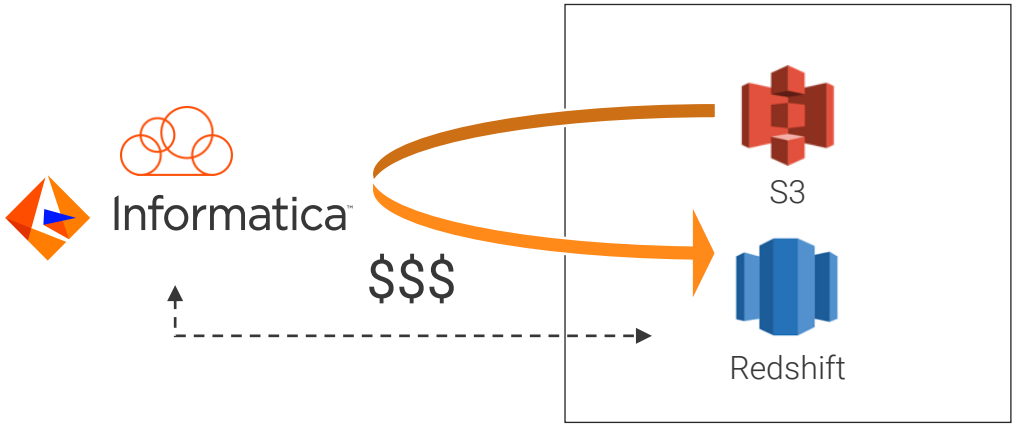
Use Case1

Ecosystem Pushdown

Use Case 1: Ecosystem Pushdown

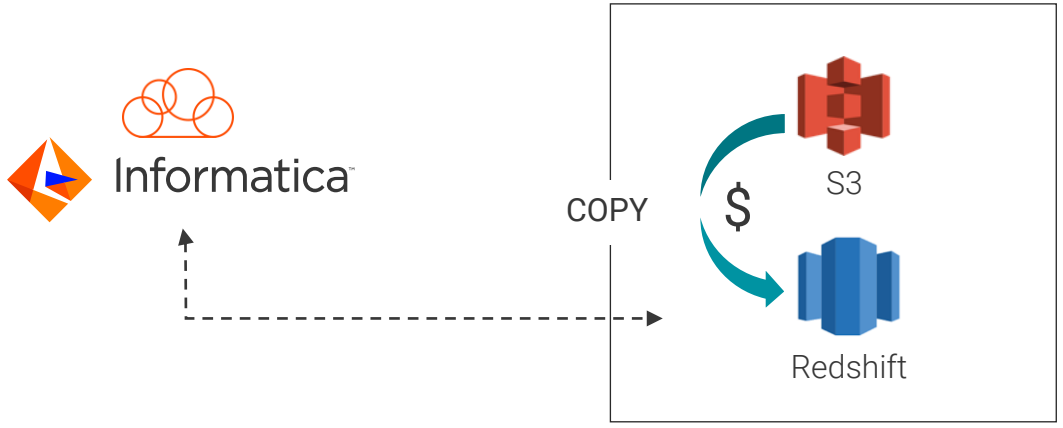
Ecosystem Pushdown transfers data from cloud data lake to data warehouse using native ecosystem commands

Without Pushdown



Loading data from Data lake to Data warehouse using Informatica engine

With Pushdown



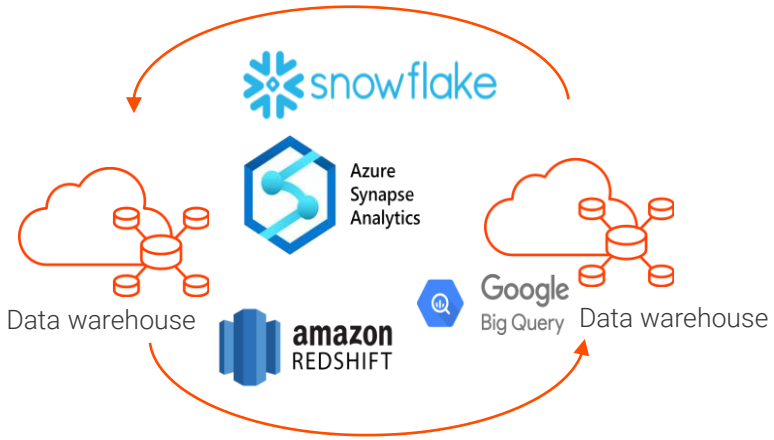
Loading data from Data lake to Data warehouse using AWS commands

Informatica

Use Case2

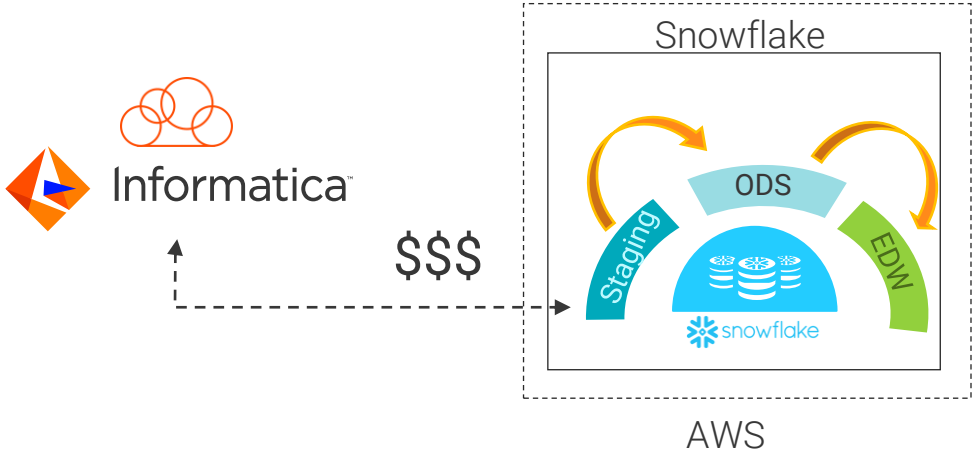
Datawarehouse Pushdown

Use Case 2: Data warehouse Pushdown



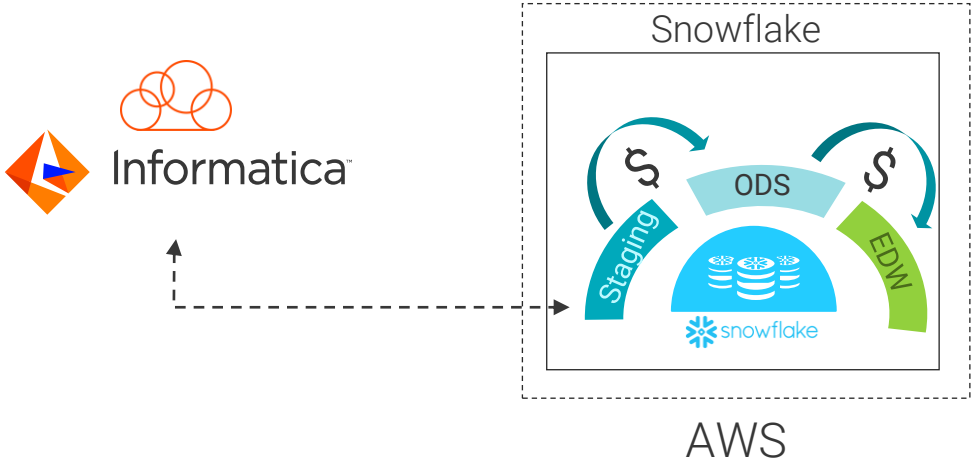
Data warehouse pushdown uses SQL queries to move data from staging area to ODS and ODS to EDW within a data warehouse

Without Pushdown



Loading data from staging to ODS in Snowflake using Informatica engine

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

Ecosystem agnostic

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

Extensive connector support

Support all major cloud ecosystems and connectivity

Faster than traditional ETL

Up to 20x faster than traditional ETL

Easy switching between runtime options

Enable pushdown at runtime with a click

No code experience

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

Informatica

DEMO

Informatica

Informatica

Cloud Data Quality

Informatica Intelligent Cloud Services

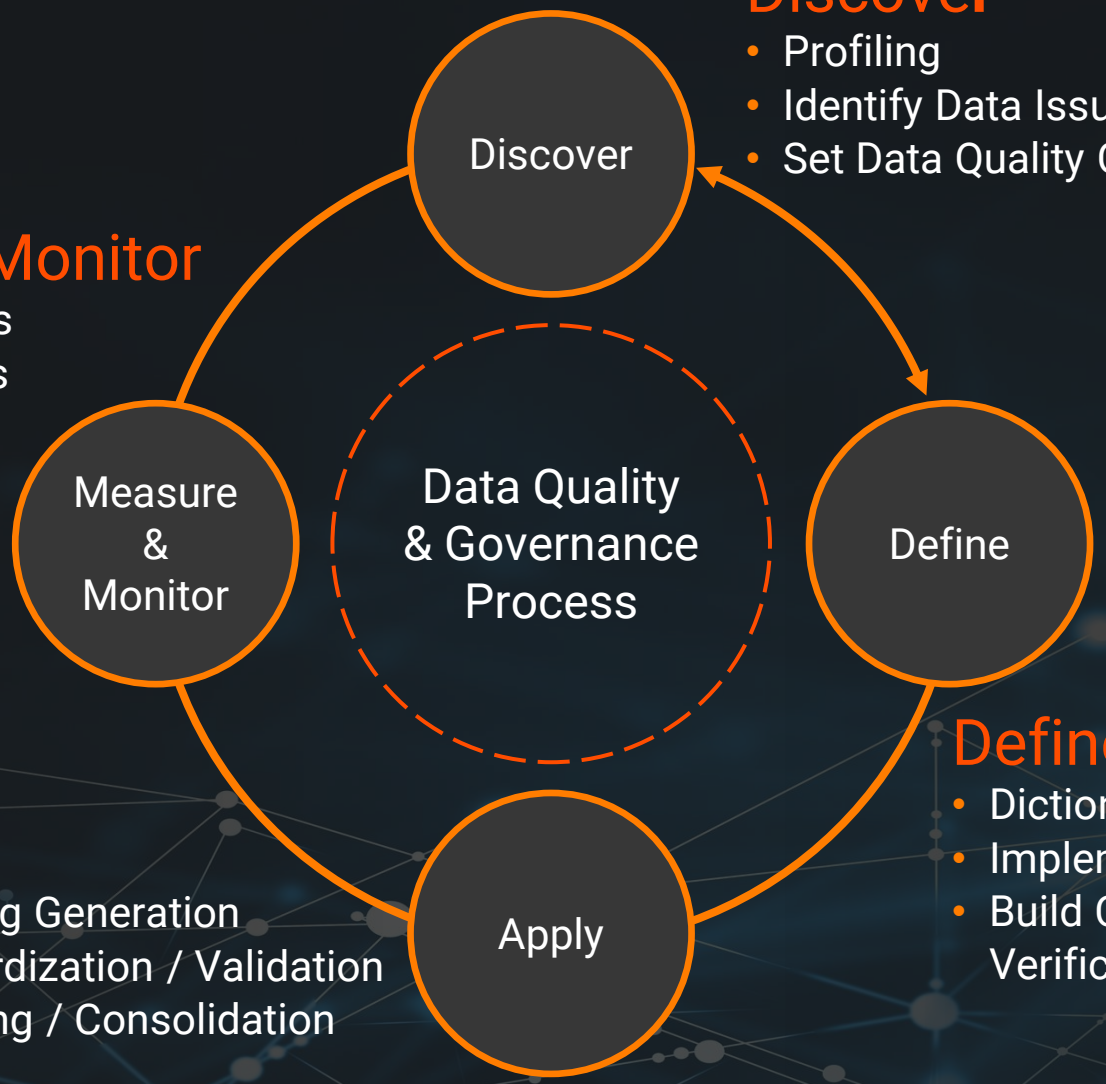
Enabling **Any User** with Self-Service Data Quality

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



Discover

- Profiling
- Identify Data Issues
- Set Data Quality Goals

Measure & Monitor

- Review Progress
- Threshold Alerts
- Scorecards

Define

- Dictionaries
- Implement Rules
- Build Cleanse, Parse, Verification, etc. processes

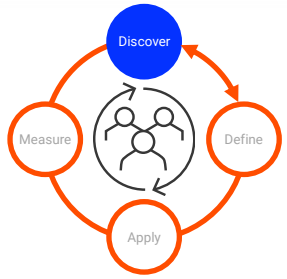
Apply

- Mapping Generation
- Standardization / Validation
- Matching / Consolidation

Informatica

Discover

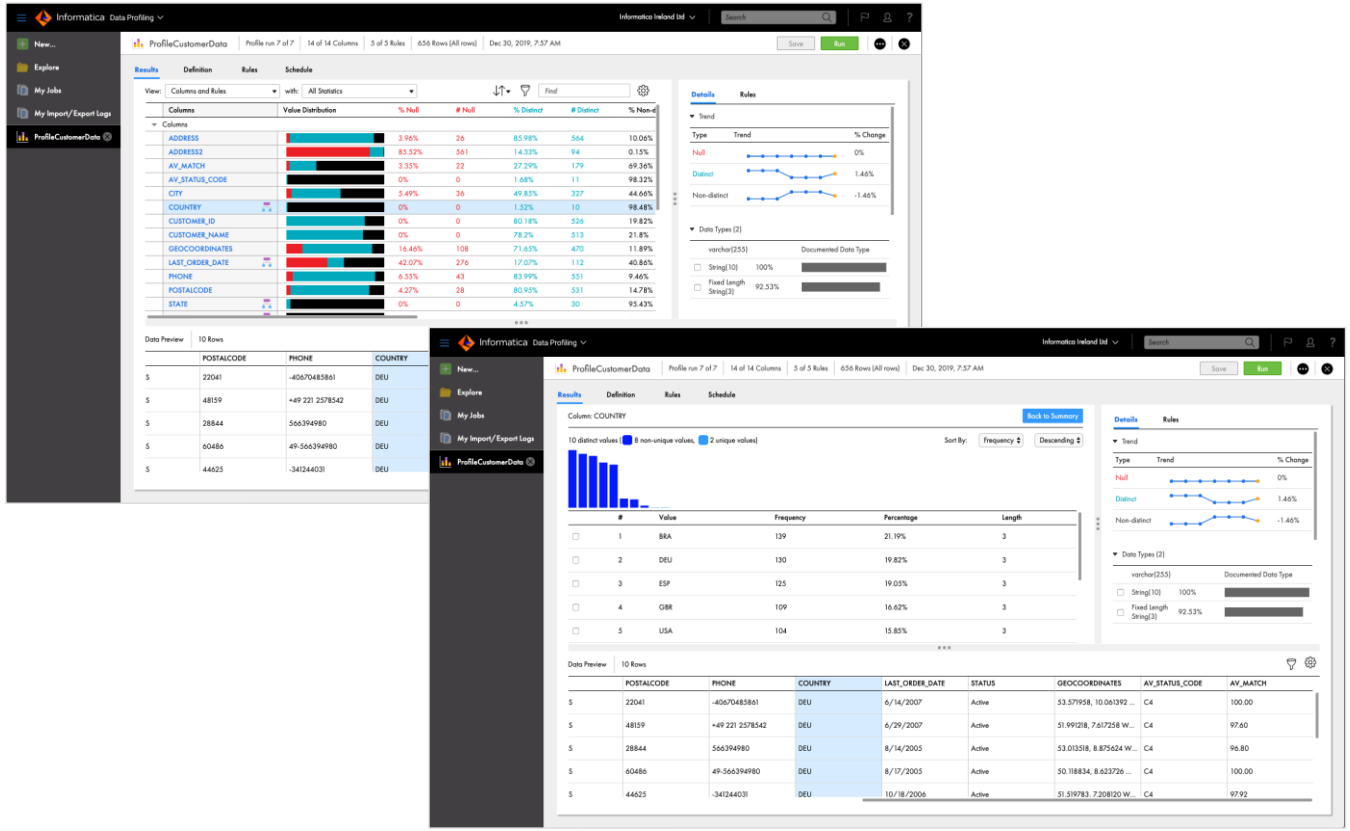
Enterprise Discovery and Profiling



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

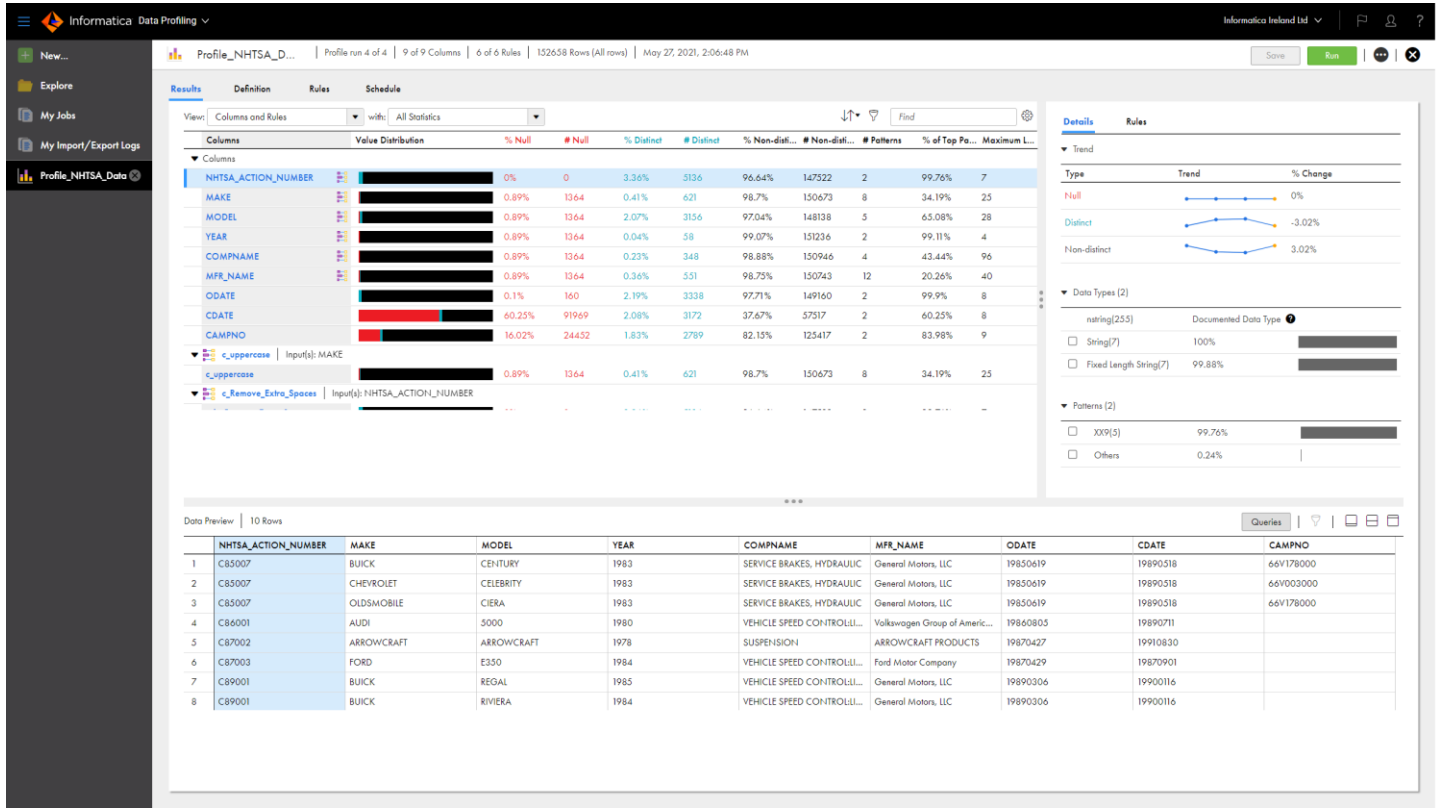


Informatica

Cloud Data Profiling

Get to Know Your Data

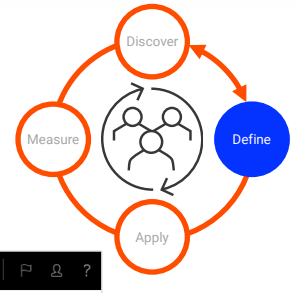
- Identify potential data issues
- Track aggregated changes on data over time
  - Provide statistics on data
  - Identify uniqueness or repeating values (value frequencies)
  - Identify patterns and formats
- Can profile Data Quality Assets
  - What-if analyses to determine rule fit



Informatica

Define

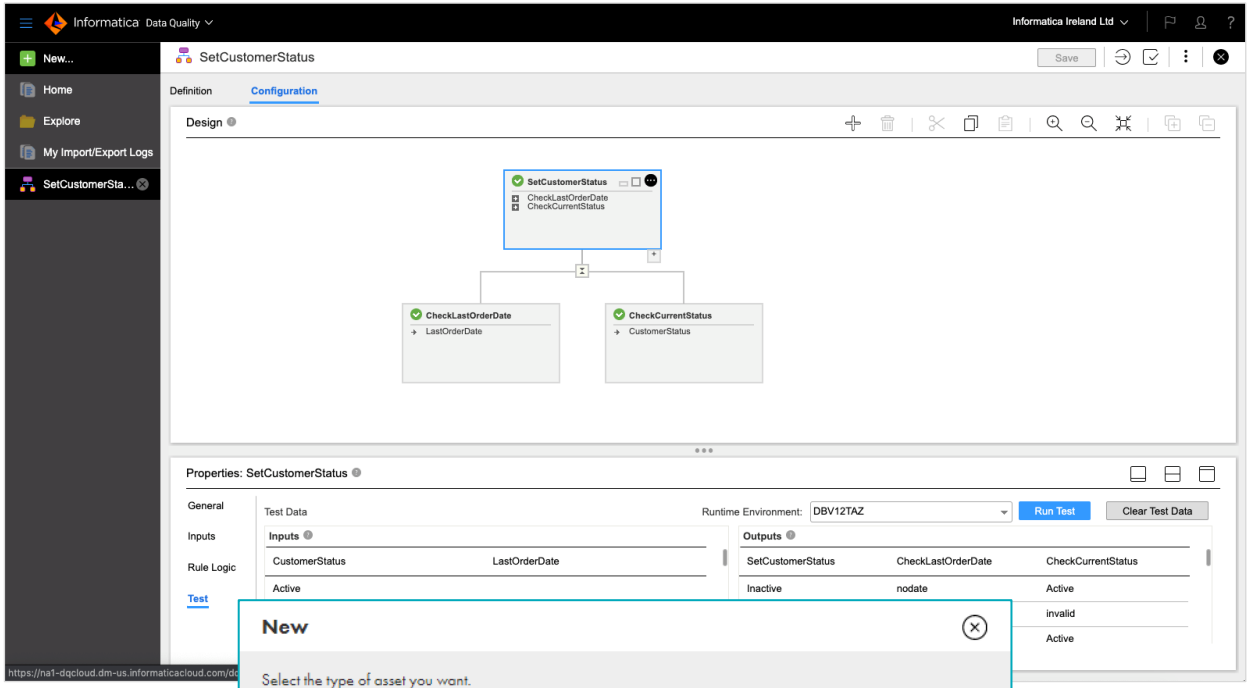
Business Rule Definition

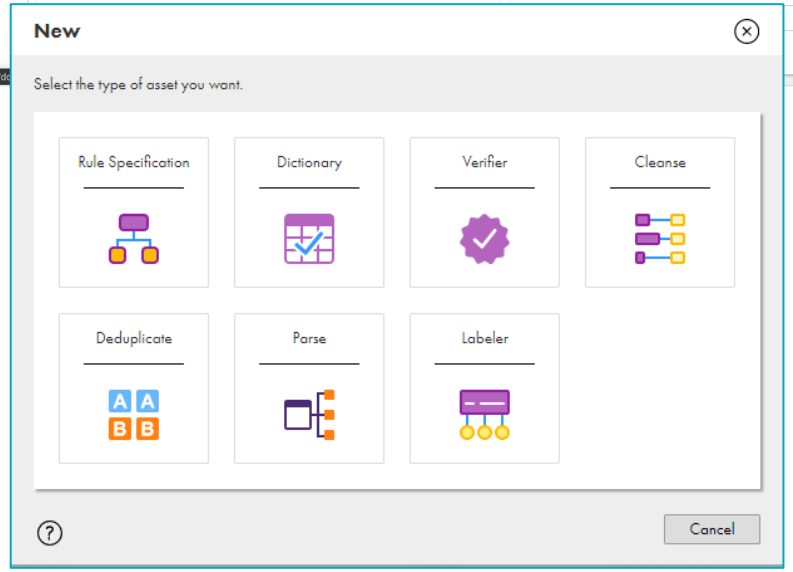


Empower the business to lead data quality initiatives

Reduce project cycles

Enable IT to focus on strategic projects

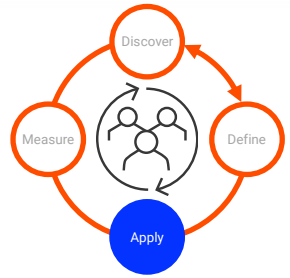




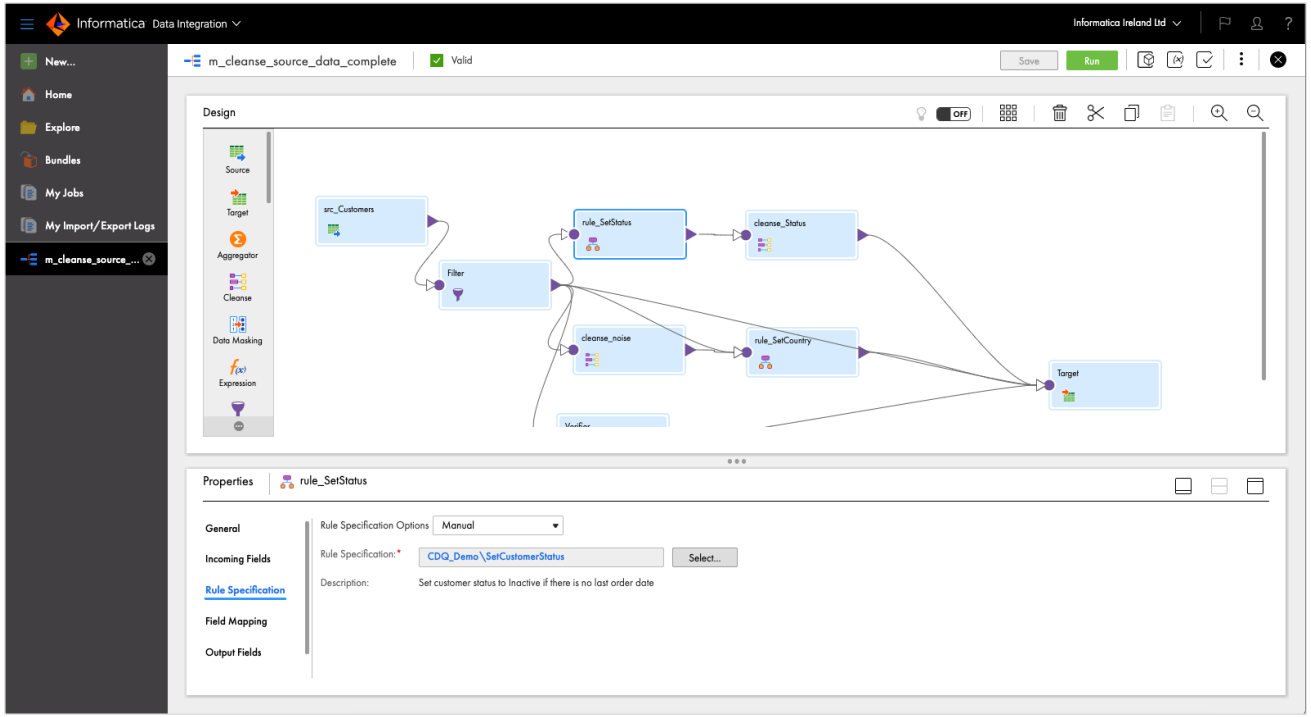
Informatica

Apply

Centralized Re-usable Rules



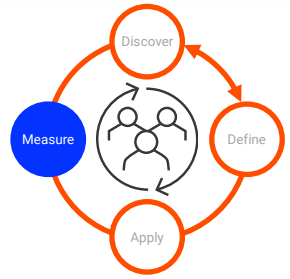
- Consistently apply data quality rules across the enterprise in support of data governance
- Reduce cost through re-use of centrally managed data quality rules
- Streamline the resolution of data quality issues



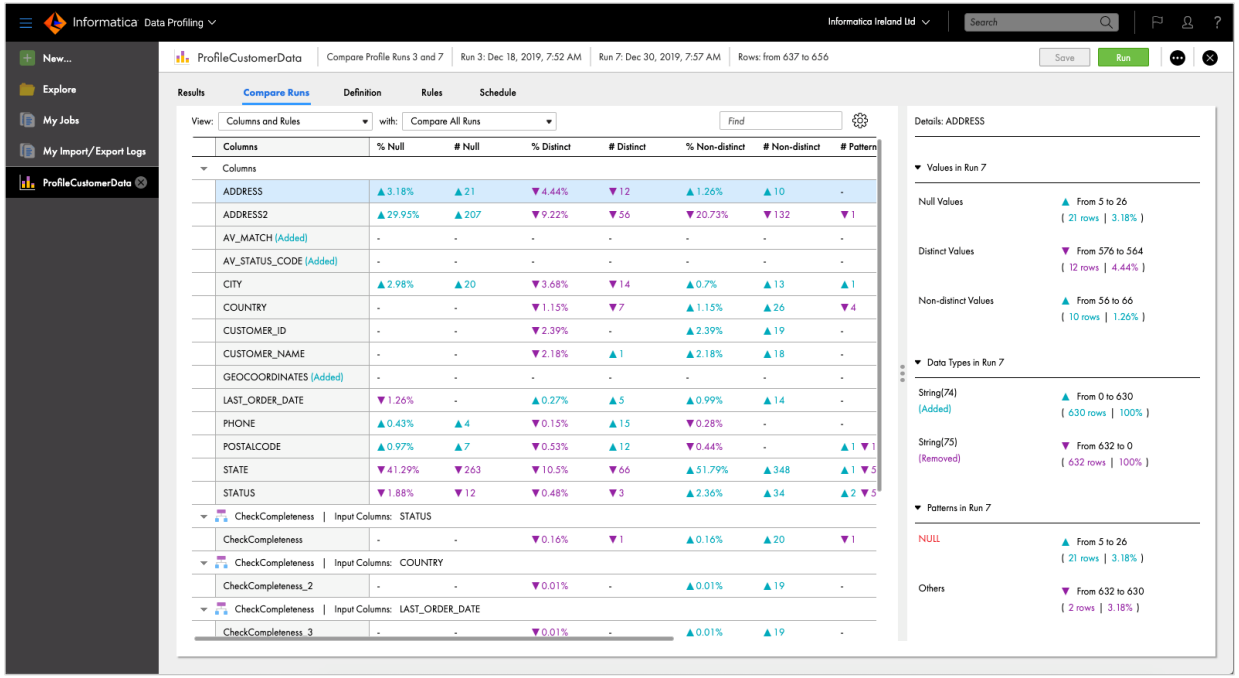
Informatica

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



Informatica

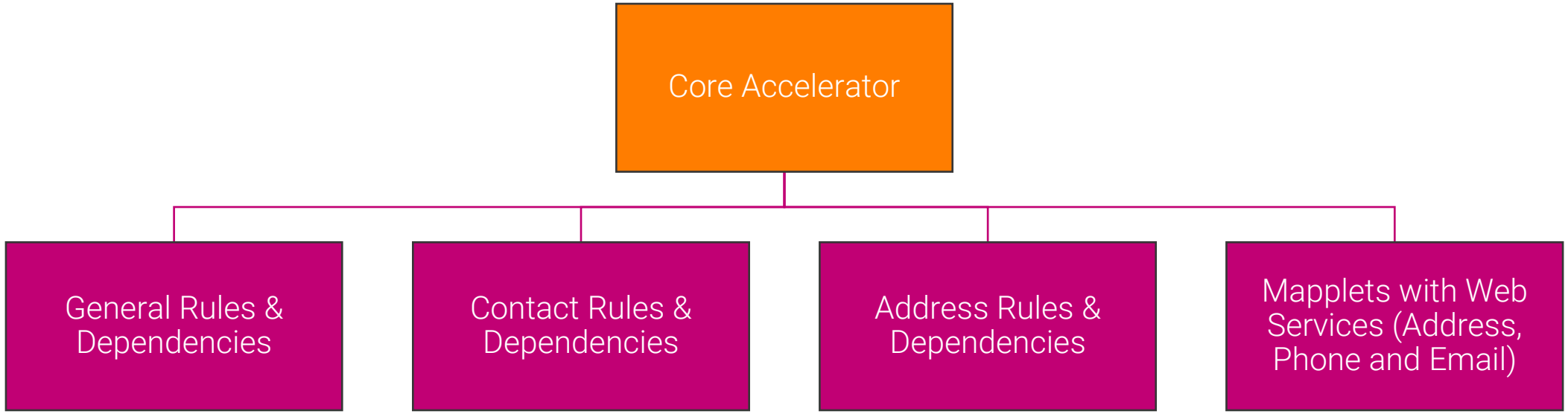
Informatica

Data Quality Bundles

Pre-built Data Quality content available to accelerate adoption

Cloud Data Quality - Core Bundle

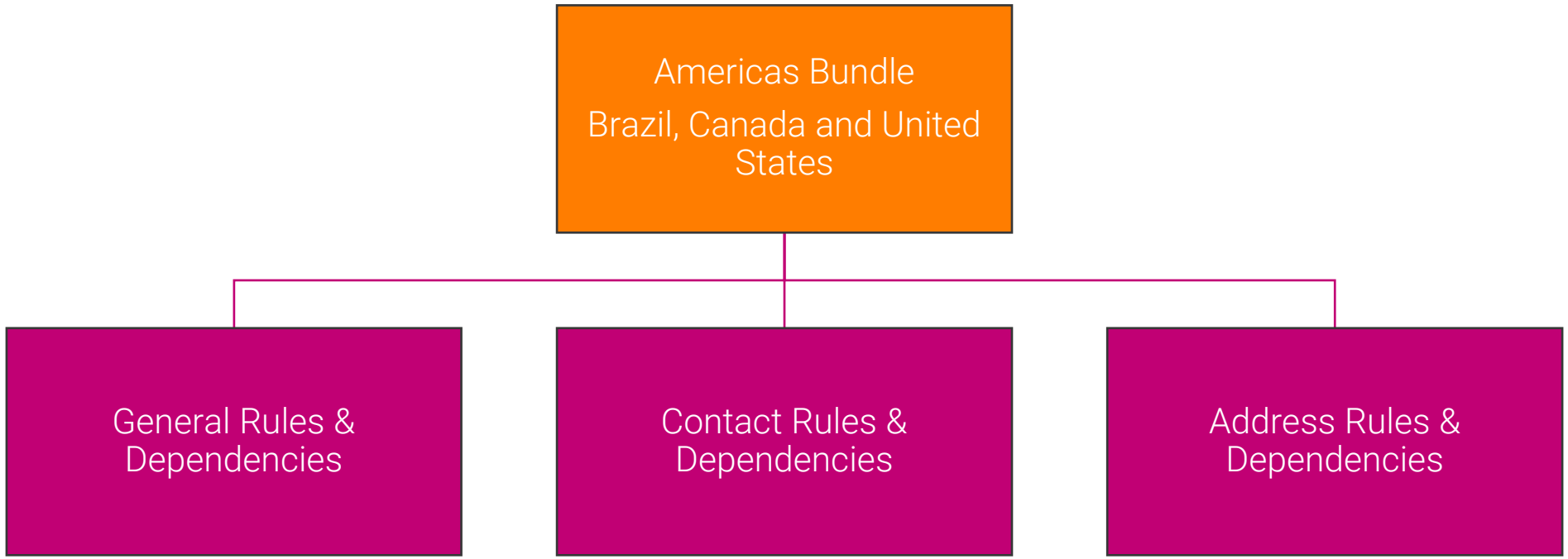
The Core accelerator includes rules that perform the following data quality processes:



Informatica

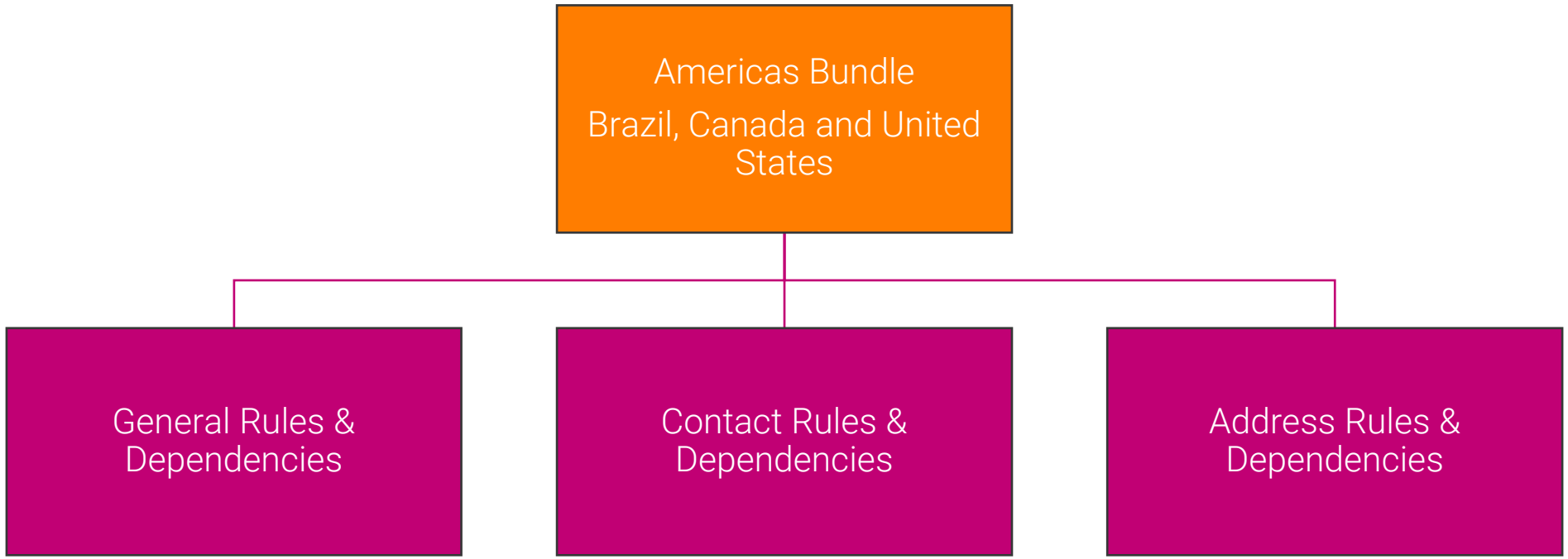
Cloud Data Quality – Americas Bundle

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



Cloud Data Quality – United Kingdom

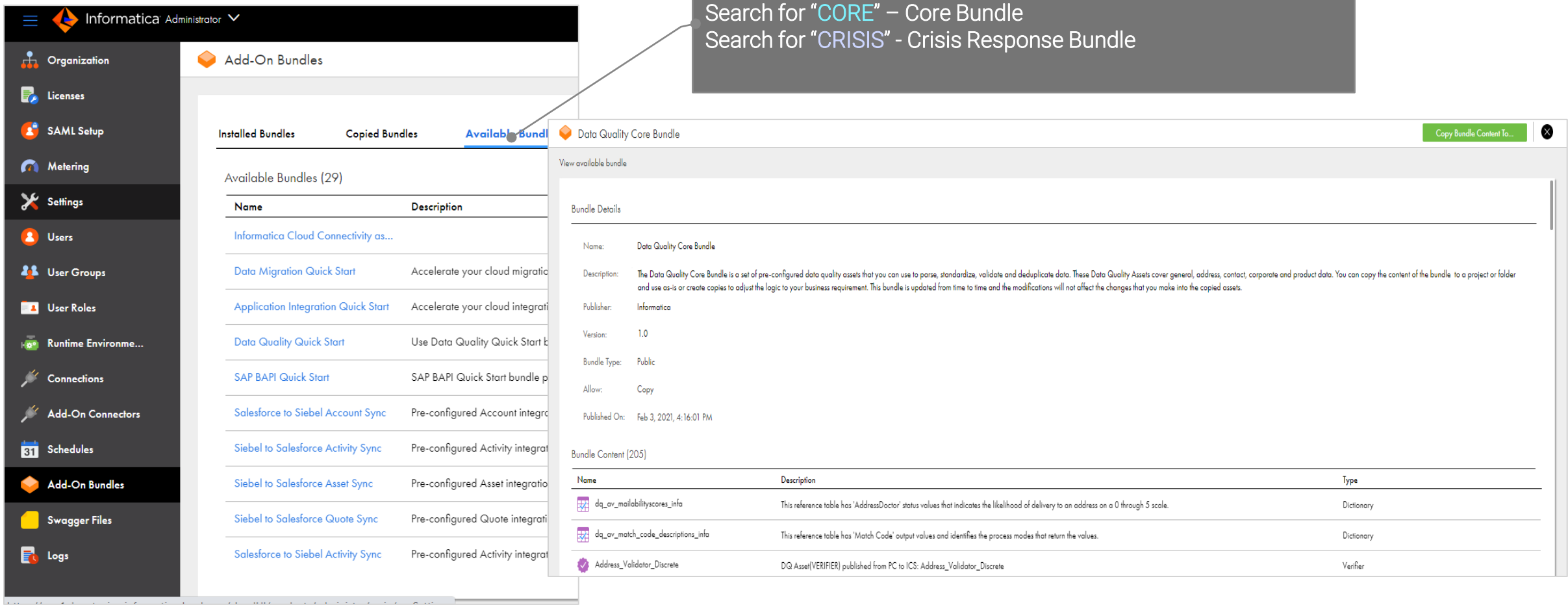
The UK bundle accelerator includes rules that perform the following data quality processes:



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



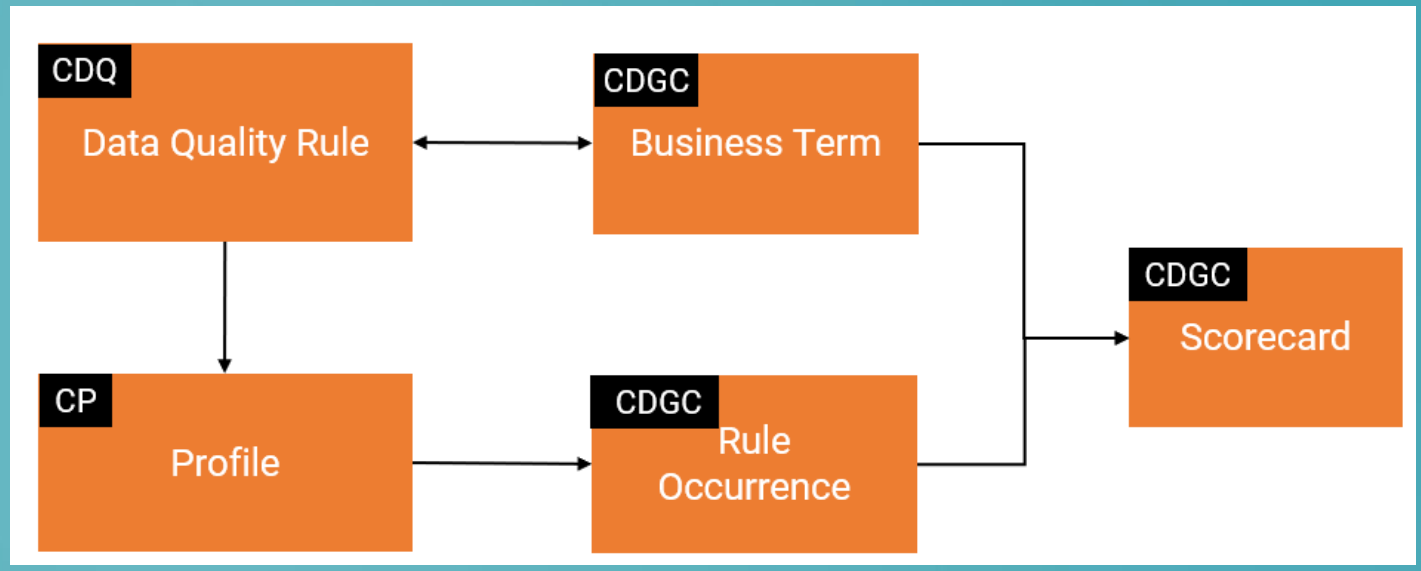
Informatica

DEMO

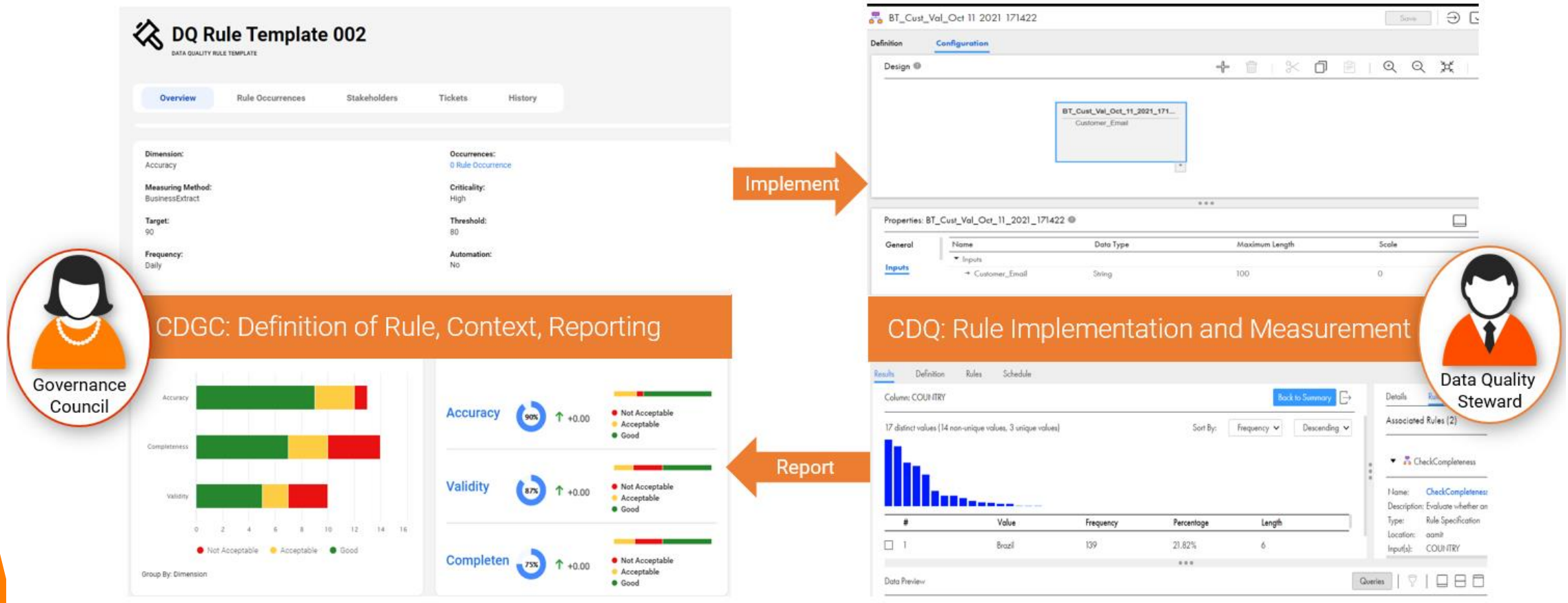
Informatica

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



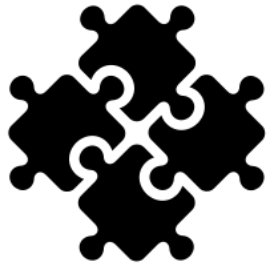
Intelligent and Contextual Data Quality

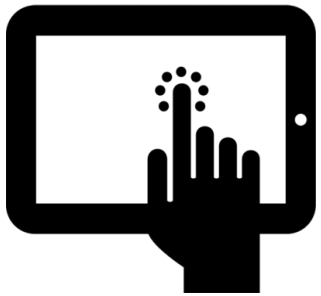


Informatica

Highlight at a Glance

Design → **Execute** → **Monitor**





NLP Rule Generation

Define Business Rule

NLP based Rule Recommendation

Implement Data Quality Rule

Rule Automation

Rule Occurrence Generation

Auto Generation of Mapping

Rule Execution

Publish Quality Metric

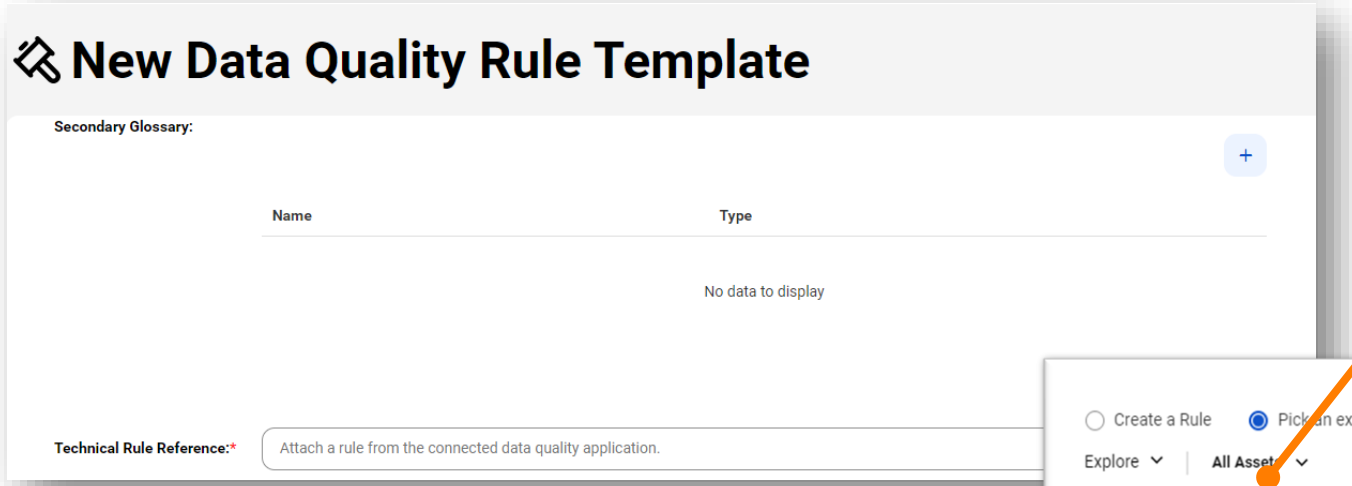
Score Aggregation

Aggregated Rolled – up scores across hierarchy

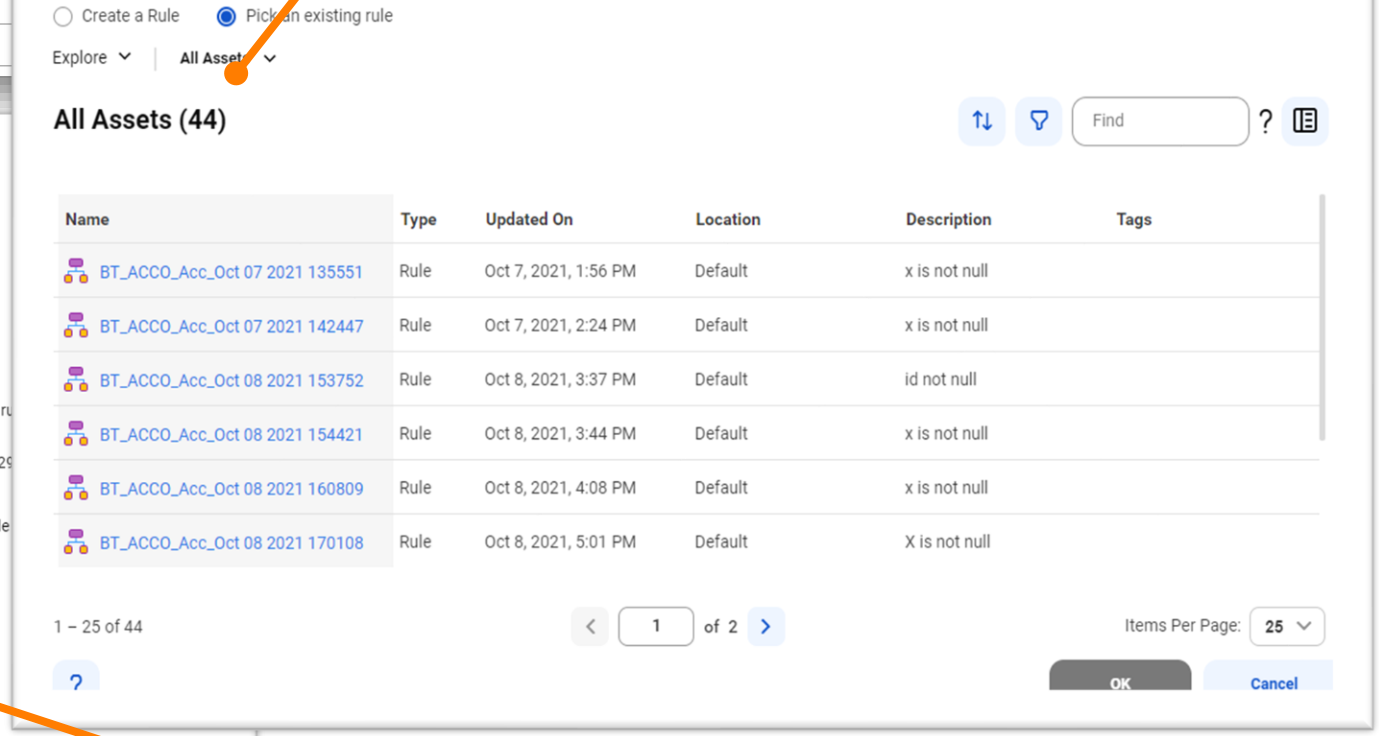
Current status / Score Trends

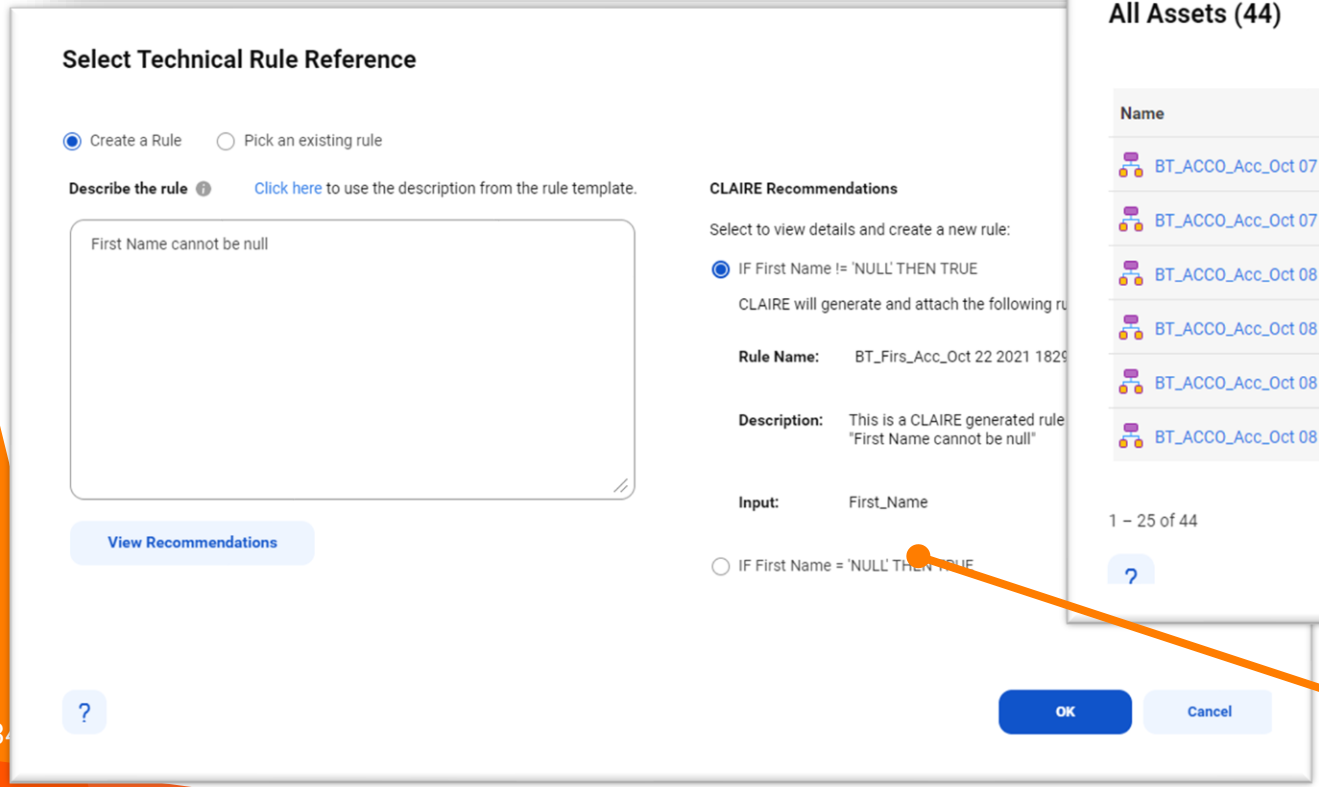
DQ Score – Asset / Preview Pane

Informatica



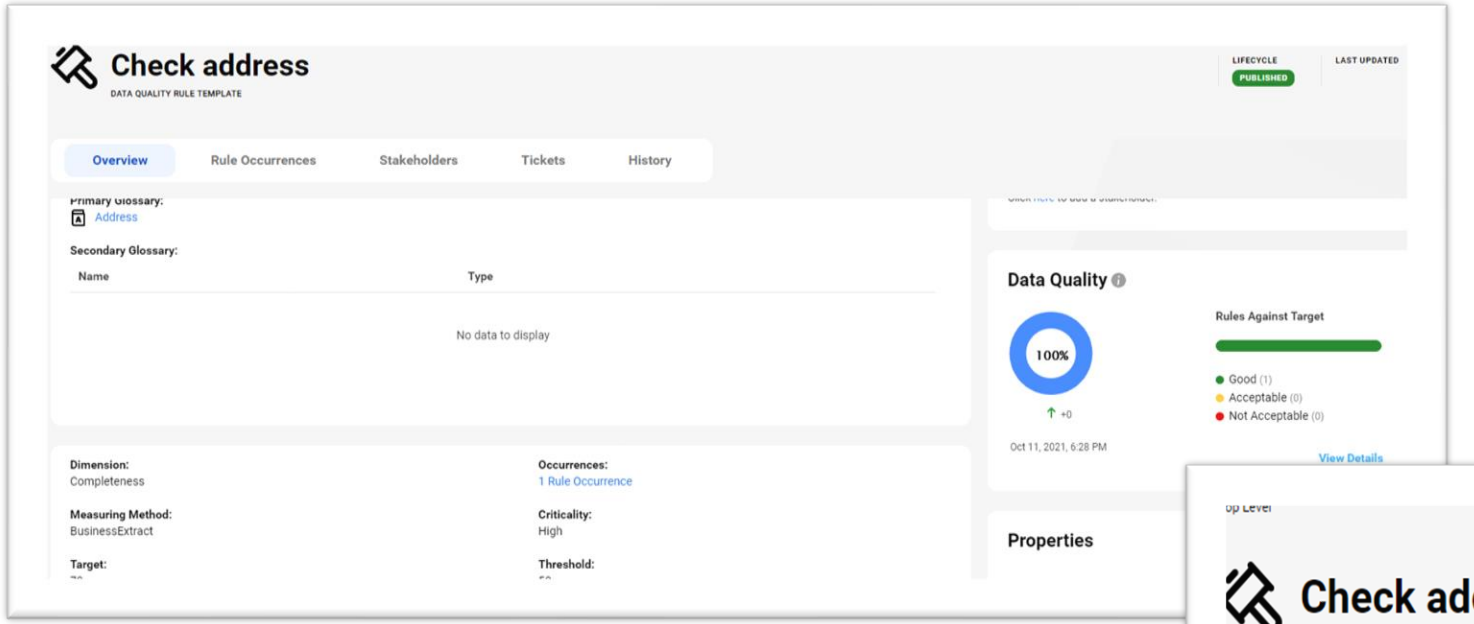
Existing rule

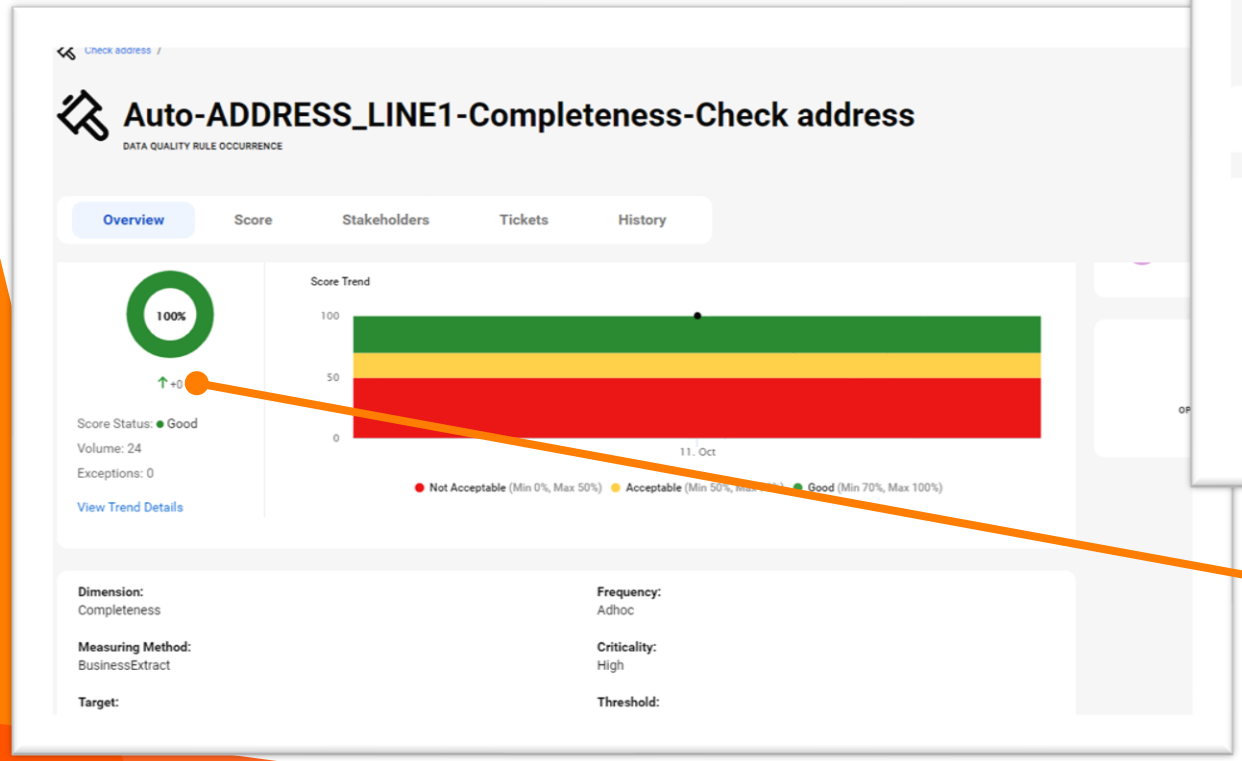


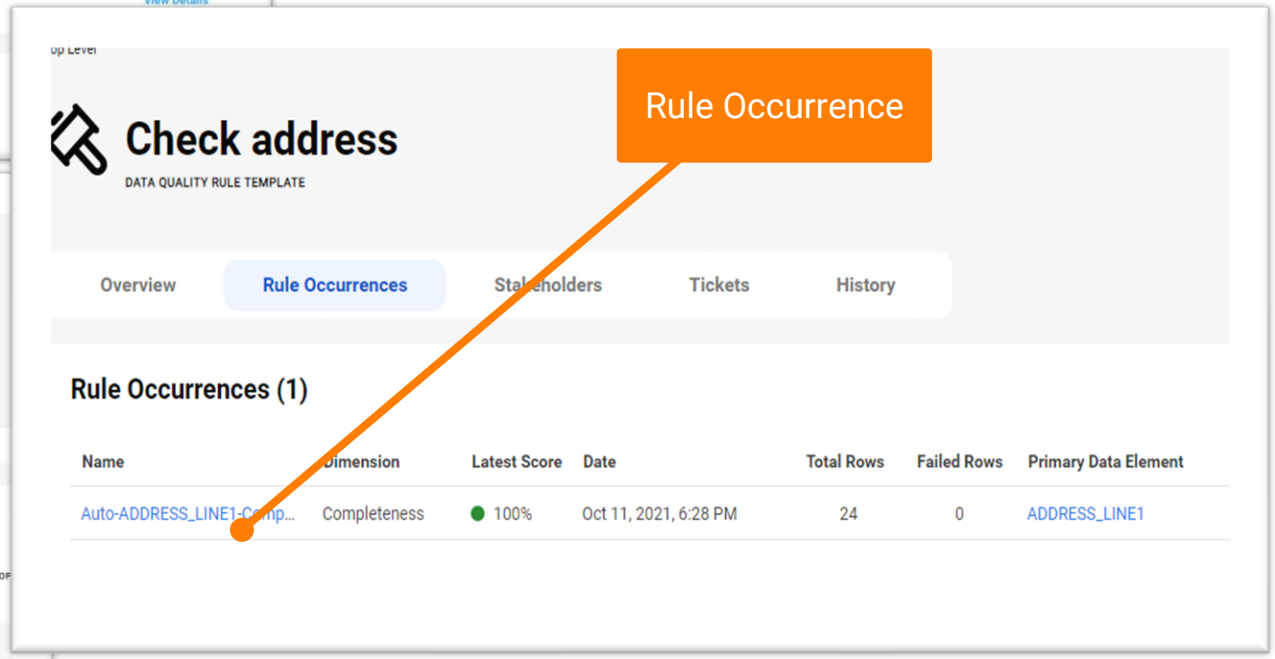


Rule Recommendation

Informatica



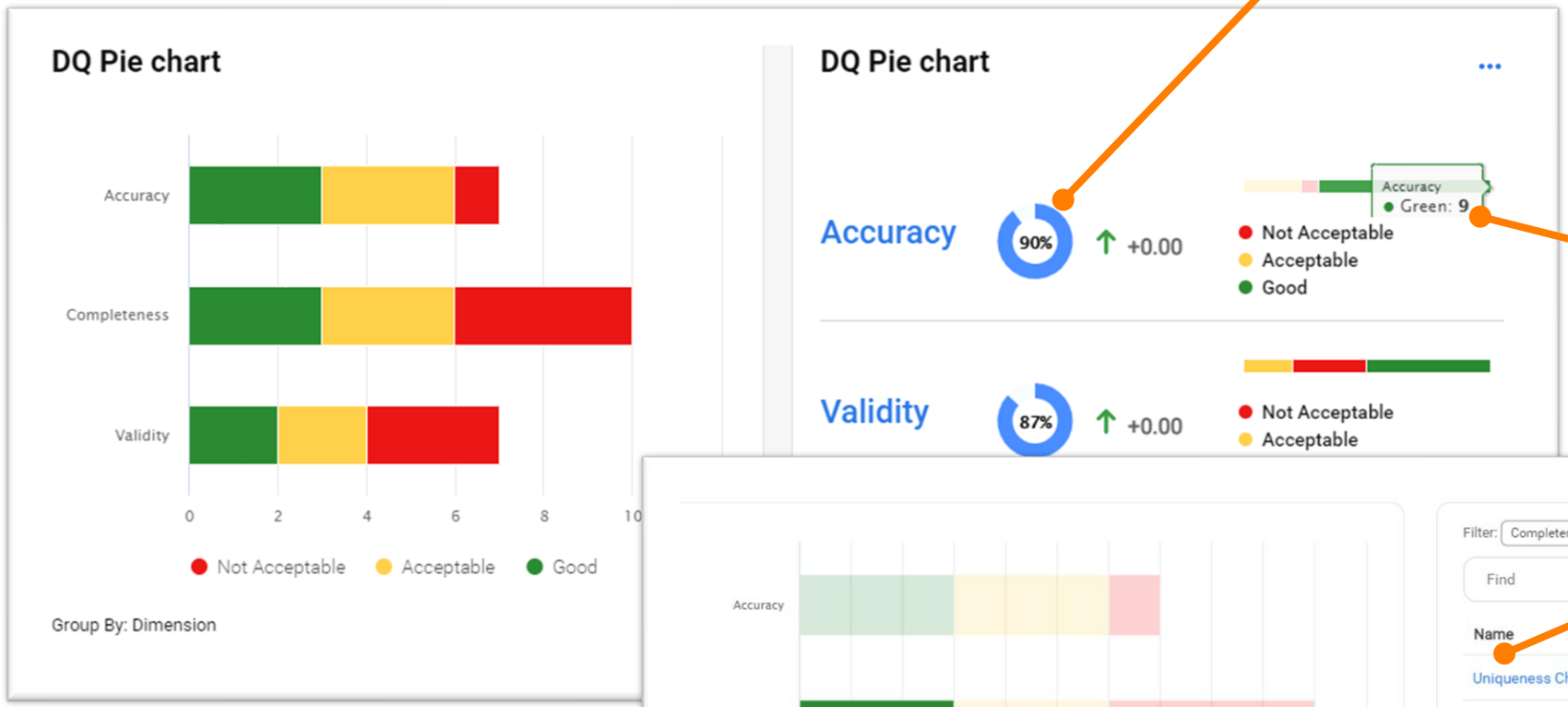




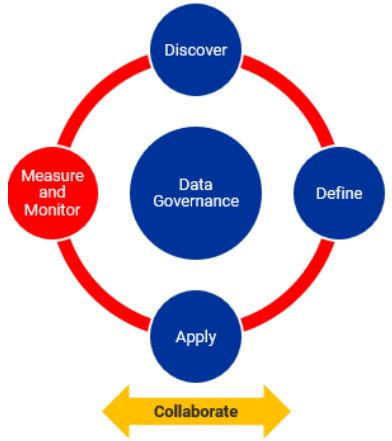
Rule Occurrence

Score Trends

Informatica



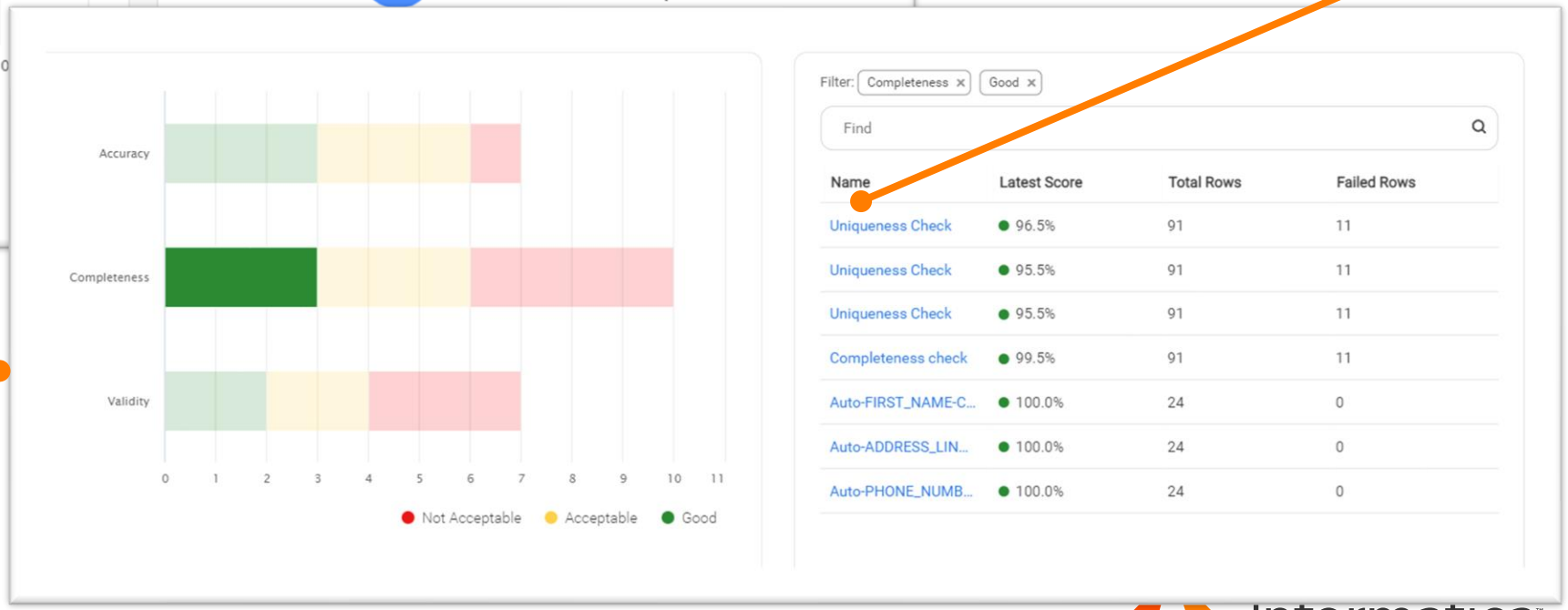
Scores and Trends



Counts

Rule Occurrences

Drill Down



Informatica

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

HANA

SAP

AWS Glue

S3

amazon REDSHIFT

aws

Azure Data Catalog

SQL

ADLS

Microsoft Azure

Google Data Catalog

BigQuery

Cloud Storage

Google Cloud

Informatica

BROADEST, MOST COMPLETE METADATA CONNECTIVITY

File Formats

CSV | Delimited | XML | JSON | Avro | Parquet | MS Excel | Adobe PDF | Flat File | MS PowerPoint | MS Word

Informatica

PowerCenter | DQ
MDM | DEI | DIH |
BG | ILM | IICS |
| Axon Data Governance |

BI & Analytics

Tableau | IBM Cognos |
QlikView | MSFT Power BI
| SAP BusinessObjects |
MicroStrategy | OBIEE

Big Data

HIVE (Cloudera, Hortonworks, MapR, IBM BigInsights, EMR, HDI)
HDFS | MapRFS |
Cloudera Navigator | Atlas |
Cassandra | MongoDB | Kafka

Databases

Oracle | DB2 | DB2 for z/OS
SQL Server | Sybase | Teradata
Netezza | JDBC | SQL Scripts |
SAP HANA

Cloud Platforms

AWS S3 | AWS Redshift | AWS Glue | Azure SQL DB | Azure SQL DW | Azure ADLS | Azure Blob | ADLS Gen 2 | Google BigQuery | Google Cloud Storage | Snowflake

Applications

SAP ECC | SAP S4/HANA
| Salesforce | Workday

Other

Erwin Models | SAP Power Designer
Custom Scanner Framework

EDC Advanced Scanners

Code and Scripting

- Oracle
- SQL Server
- Netezza
- DB2
- Teradata
- Sybase ASE

ETL Tools

- IBM Datastage
- Microsoft SSIS
- Oracle Data Integrator
- Talend DI

Mainframes

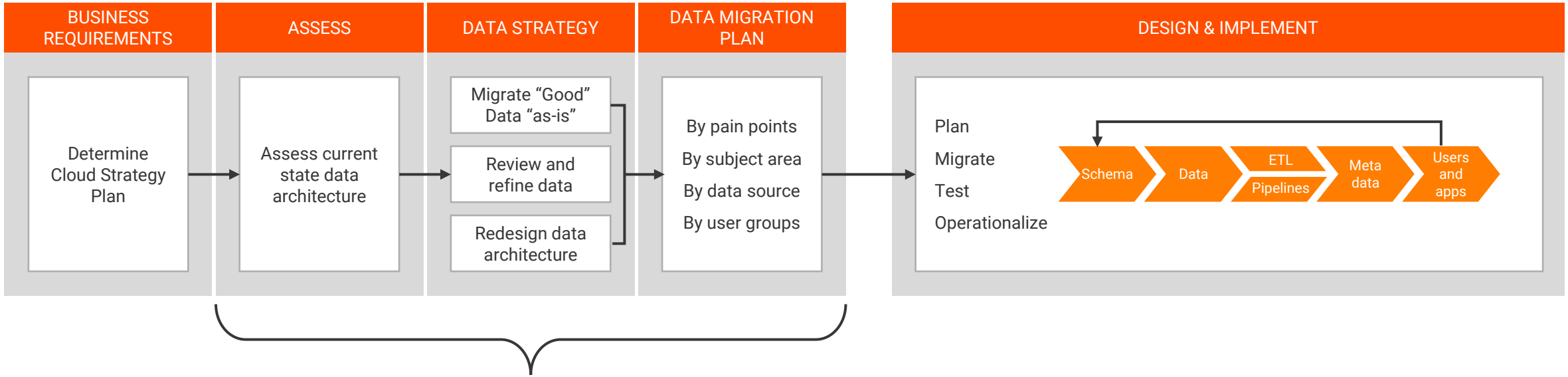
- COBOL
- JCL

Statistical/BI Tools and Complex Systems

- SAS
- Microsoft SSAS
- Microsoft SSRS
- SAP BW
- SAP BW4HANA

How Can a Catalog Help with Cloud **Data** Migration?

- Typical High-level Cloud Data Migration Process



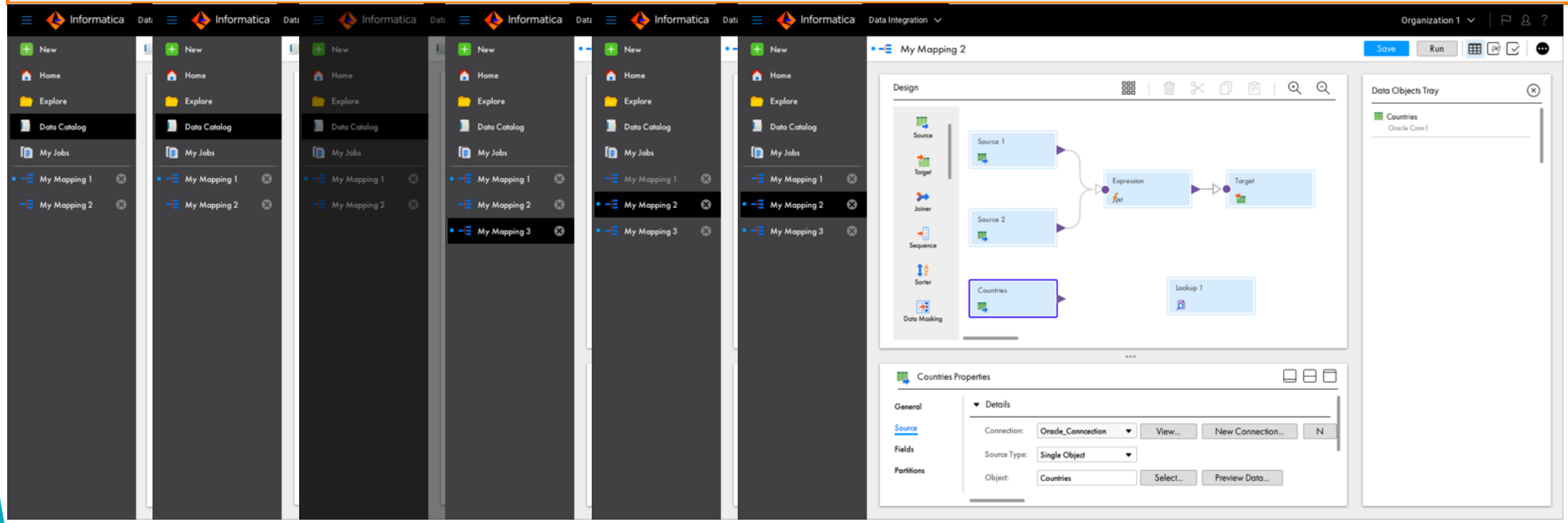
A Data Catalog can help assess part of the cloud data migration process to better understand the nature of the data

Informatica

Enterprise Data Catalog Integration

ML Driven Data Discovery and Integration

Claire



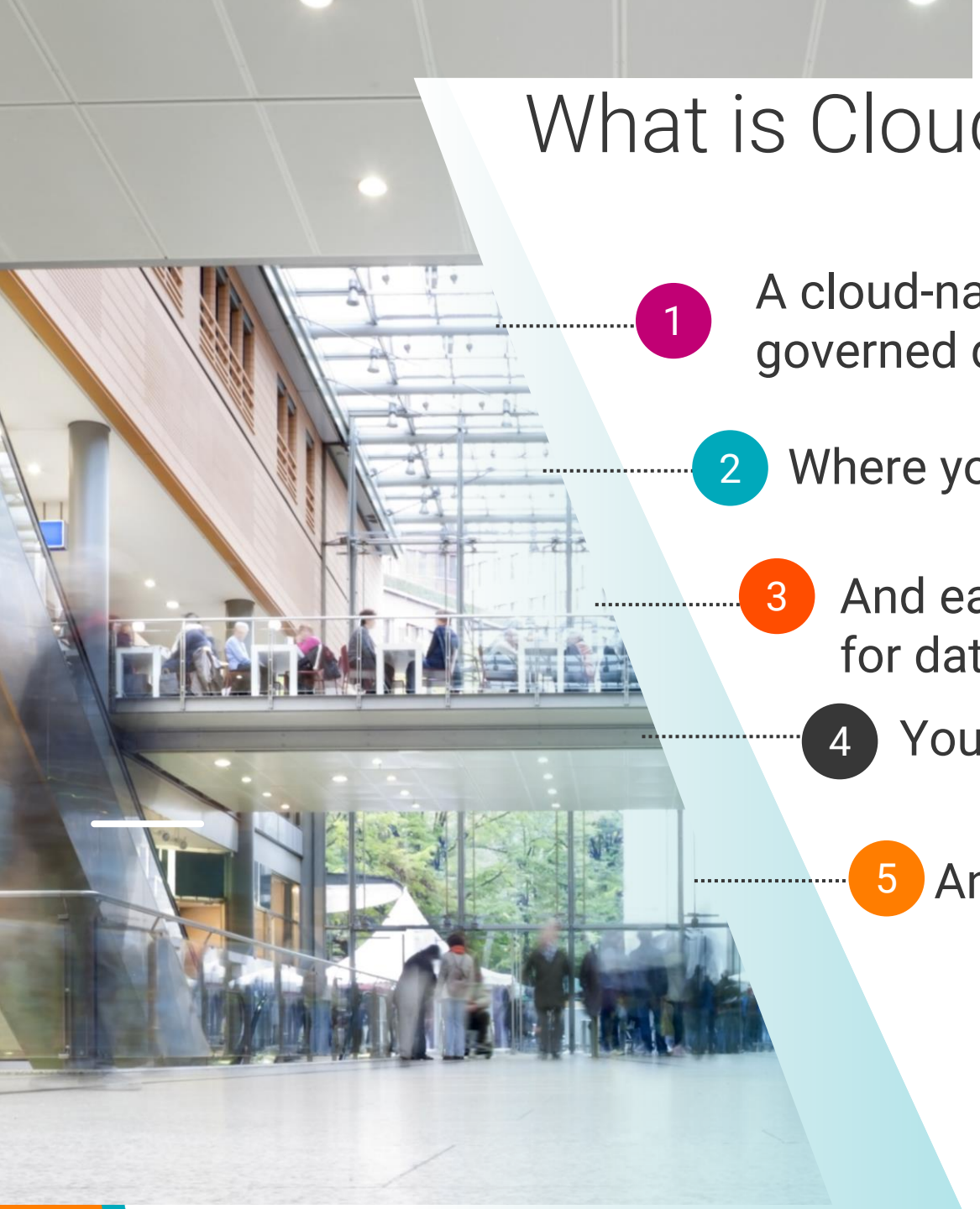
Informatica

DEMO

Informatica

Informatica

Cloud Data Marketplace



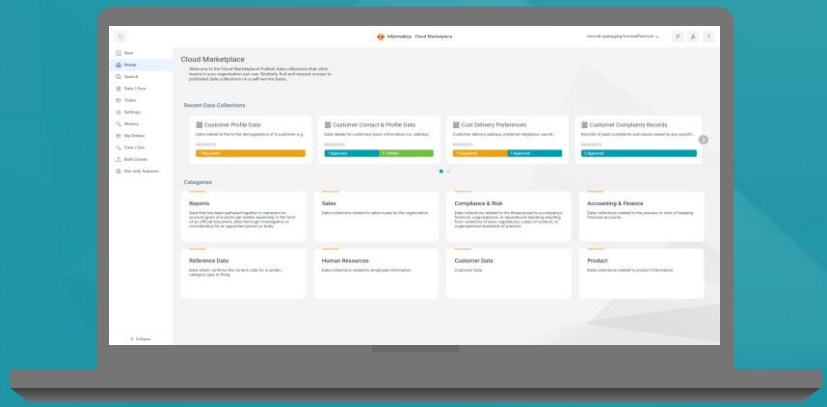
What is Cloud Data Marketplace?

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

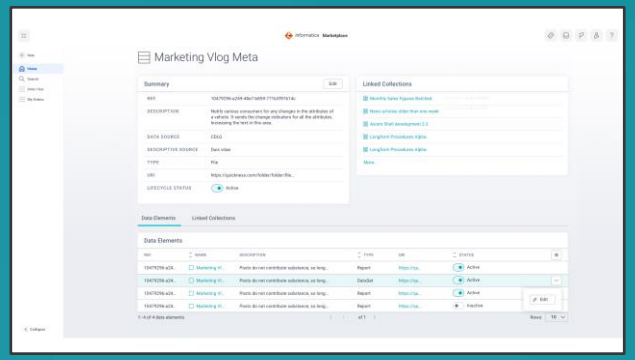
Informatica

Cloud Data Marketplace

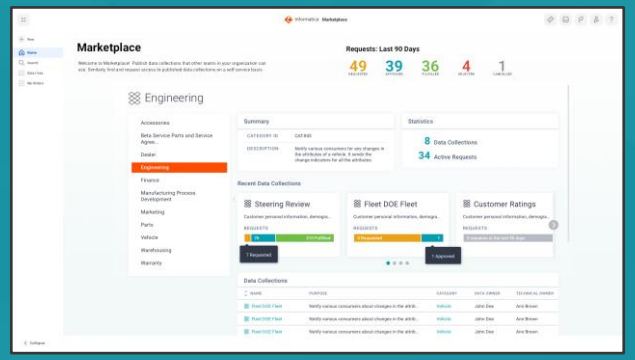
A storefront, with order management, and governed delivery capabilities



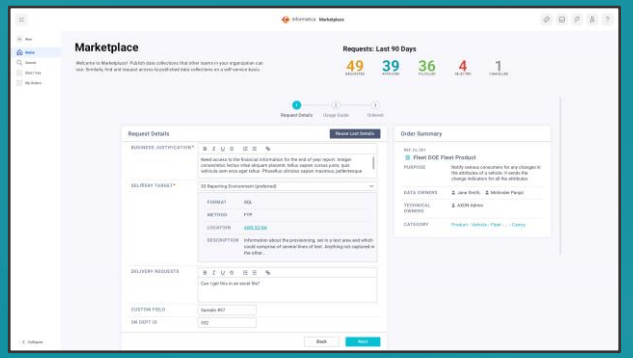
Package Data Assets



Shop & Checkout



Fulfill & Track



Informatica

DEMO

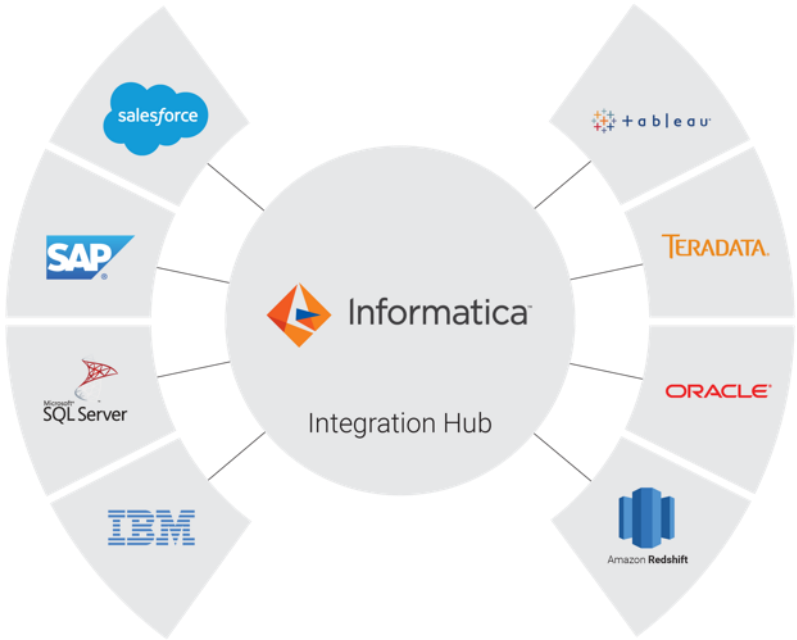
Informatica

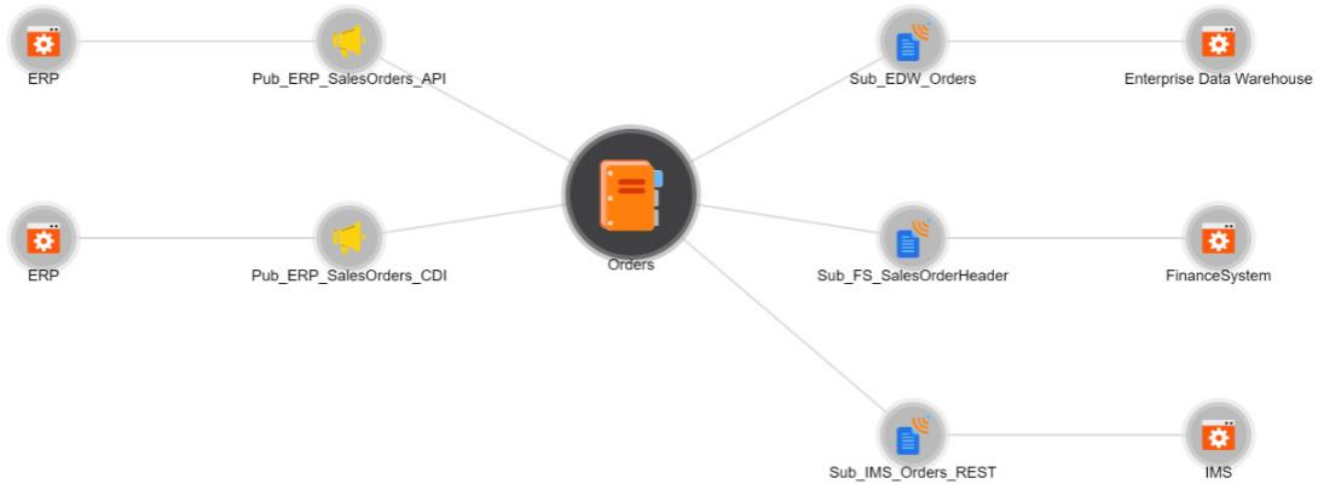
Informatica

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.

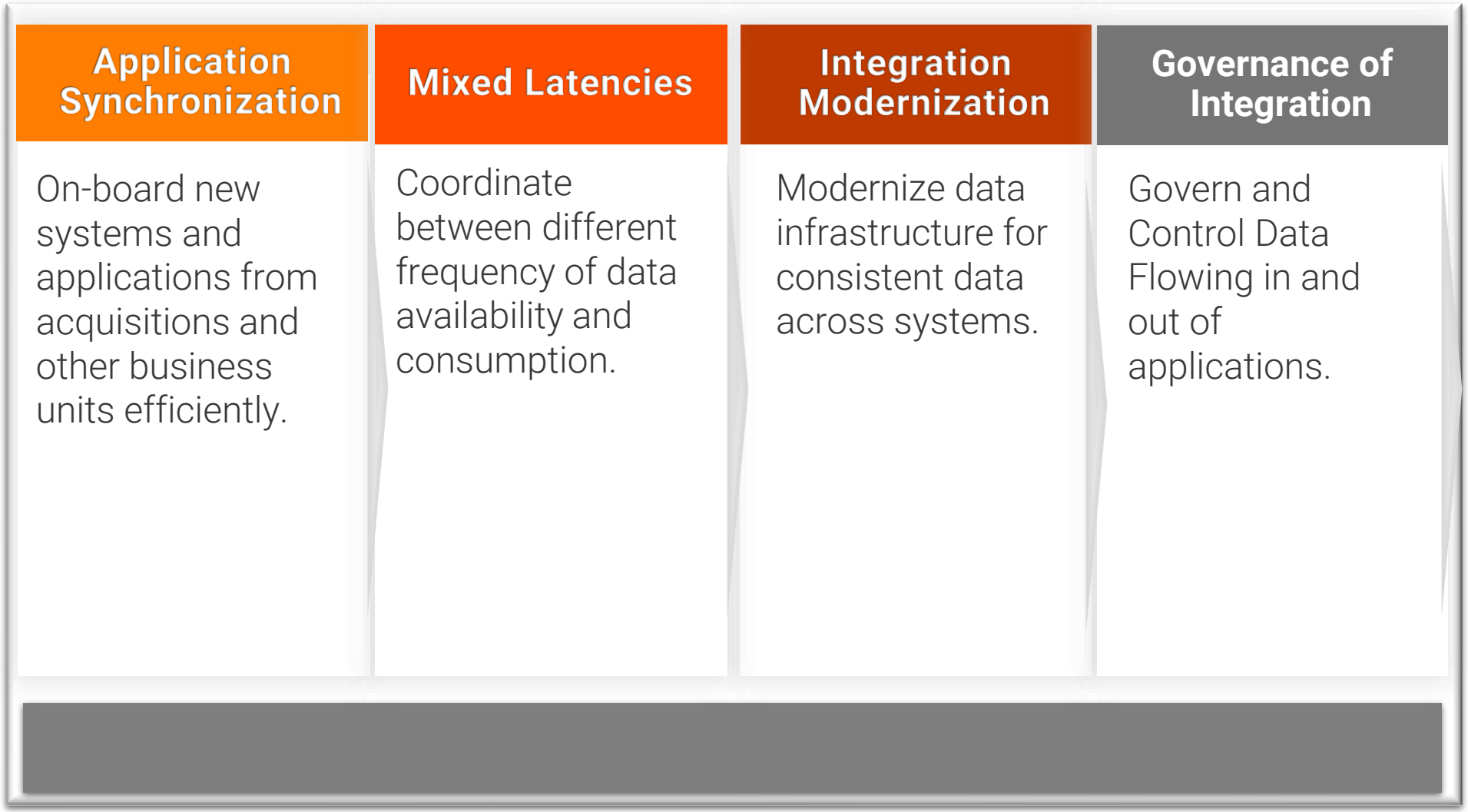




Informatica

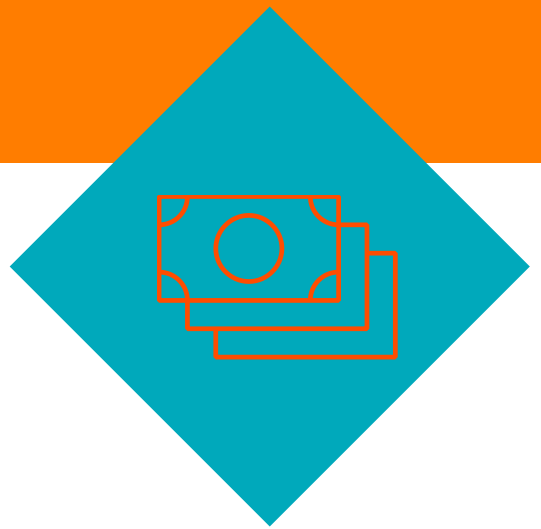
Use Cases

Use Cases Driving Integration Hub Initiatives



Informatica

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

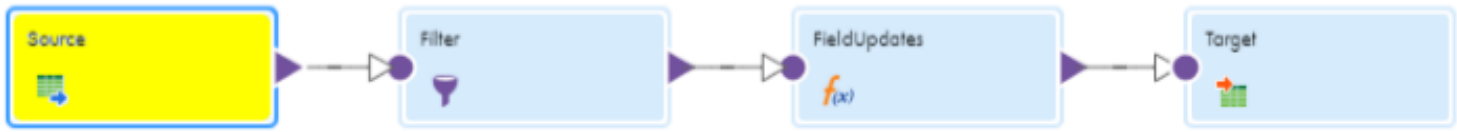
Downtime getting costlier by the minute

Network bandwidth is limited

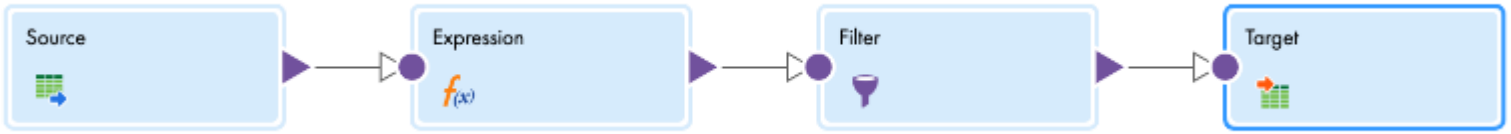
Example 1: Cost per integration – how it's going up

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

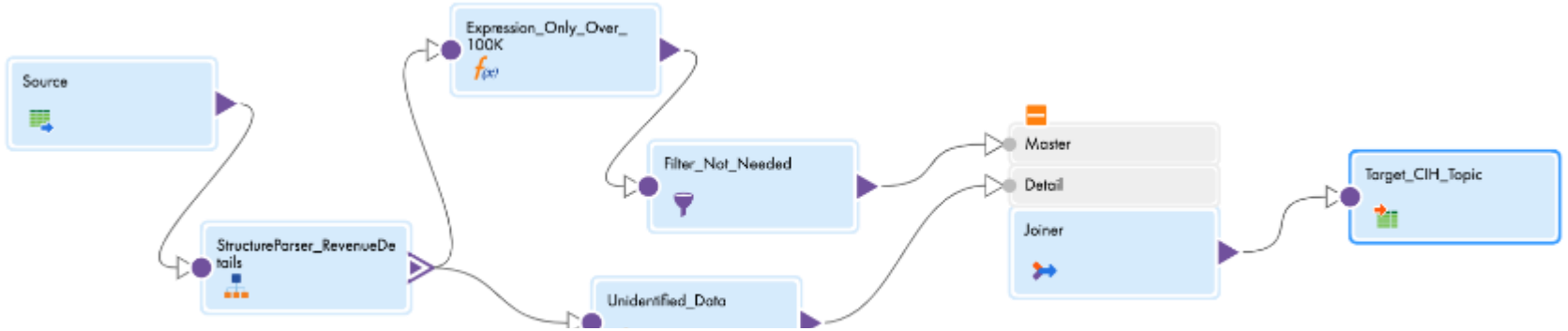
Dataflow Itr-1



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

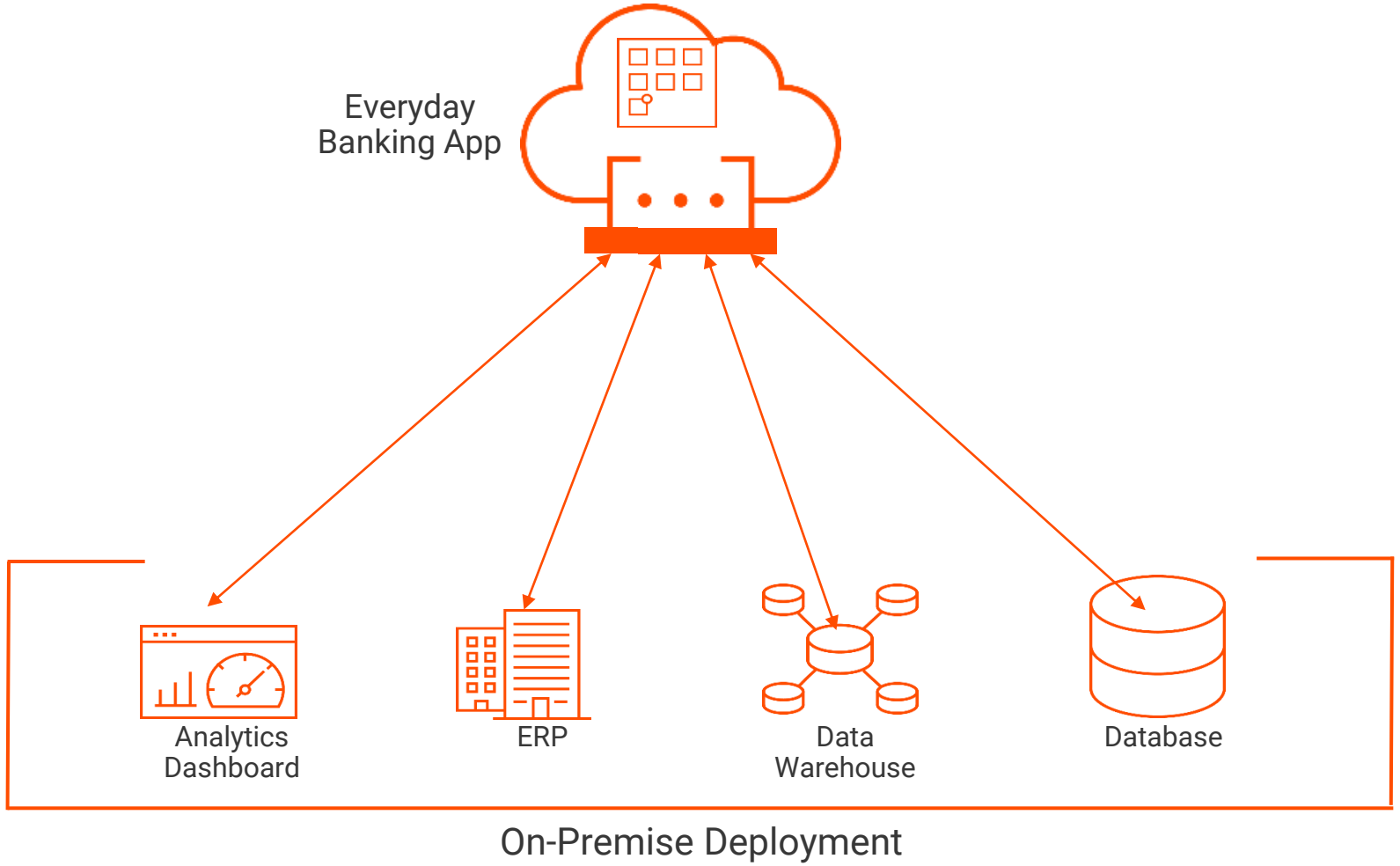
- Simple data flow mappings turn into a "hairball" as the integration tasks gets repetitive and complex

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

Informatica

Example 2: Cost is going up due to integration

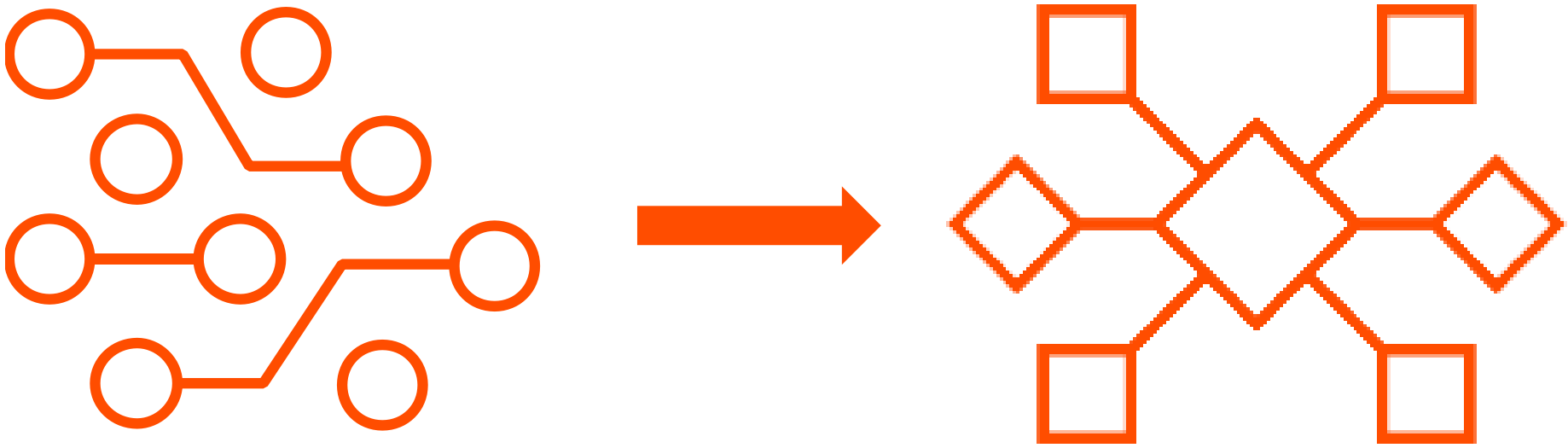
You pay API charges and transfer fees every time you fetch data from the cloud



Informatica

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

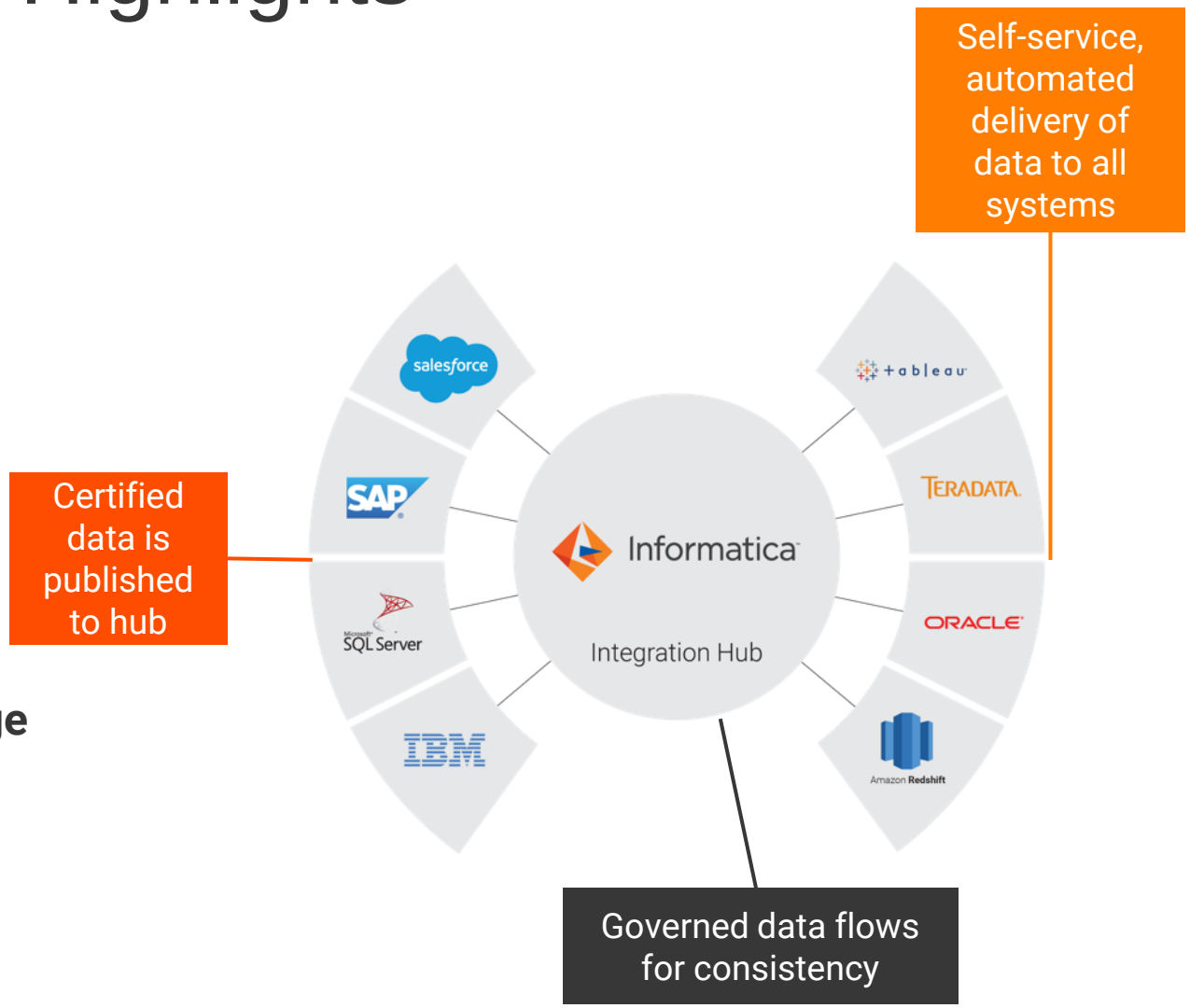
Hub n Spoke Integration

10 sources and 10 targets mean 20 connections

Informatica

Integration Hub – Features Highlights

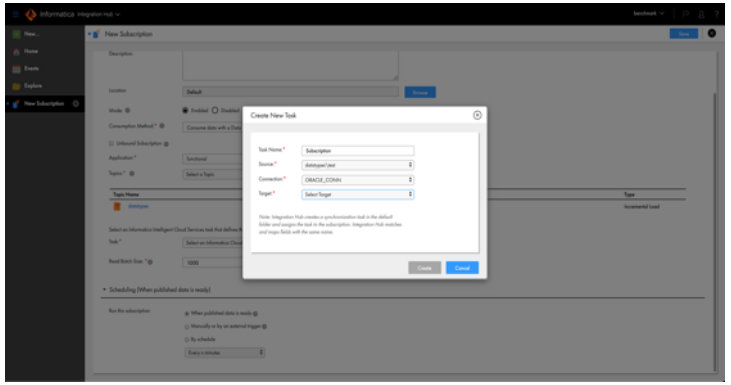
- **Decouple** source from target
- **Standardize** data access with canonical Topics
- **Hosted** or **Private** Publication Repository
- **REST APIs**
- Internal **scheduler** for Pubs/Subs
- View end to end data flow with integration **lineage**
- Track and **monitor** events with granular search



Informatica

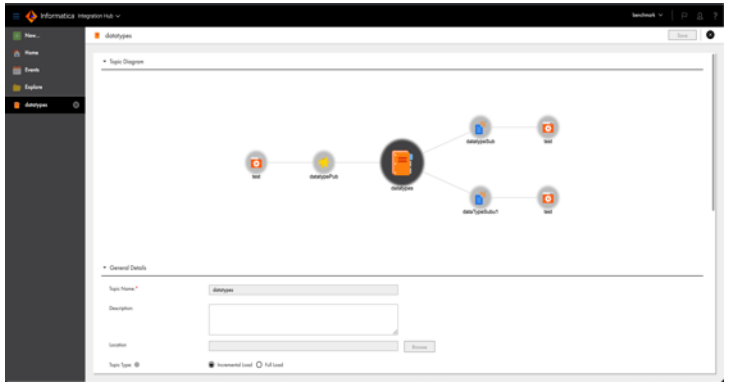
Cloud Integration Hub tightly coupled with CDI as part of IICS

Self Service Integration and Management



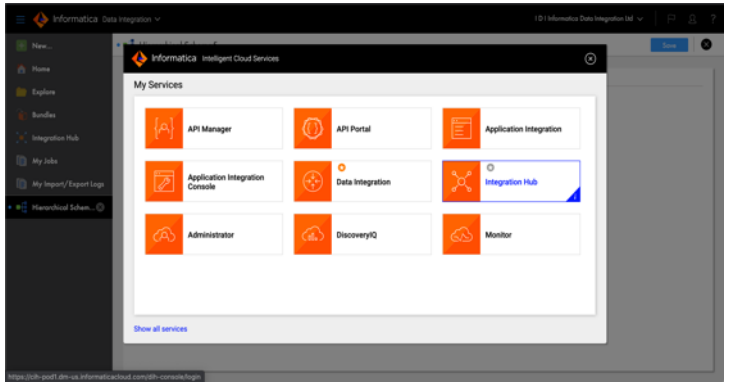
- Simple interface to manage Topics, Applications, Publications and subscriptions
- Pre-configured Accelerators for commonly used cloud application
- Associate CDI tasks with CIH assets (pub/sub – Mappings/DSS)

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

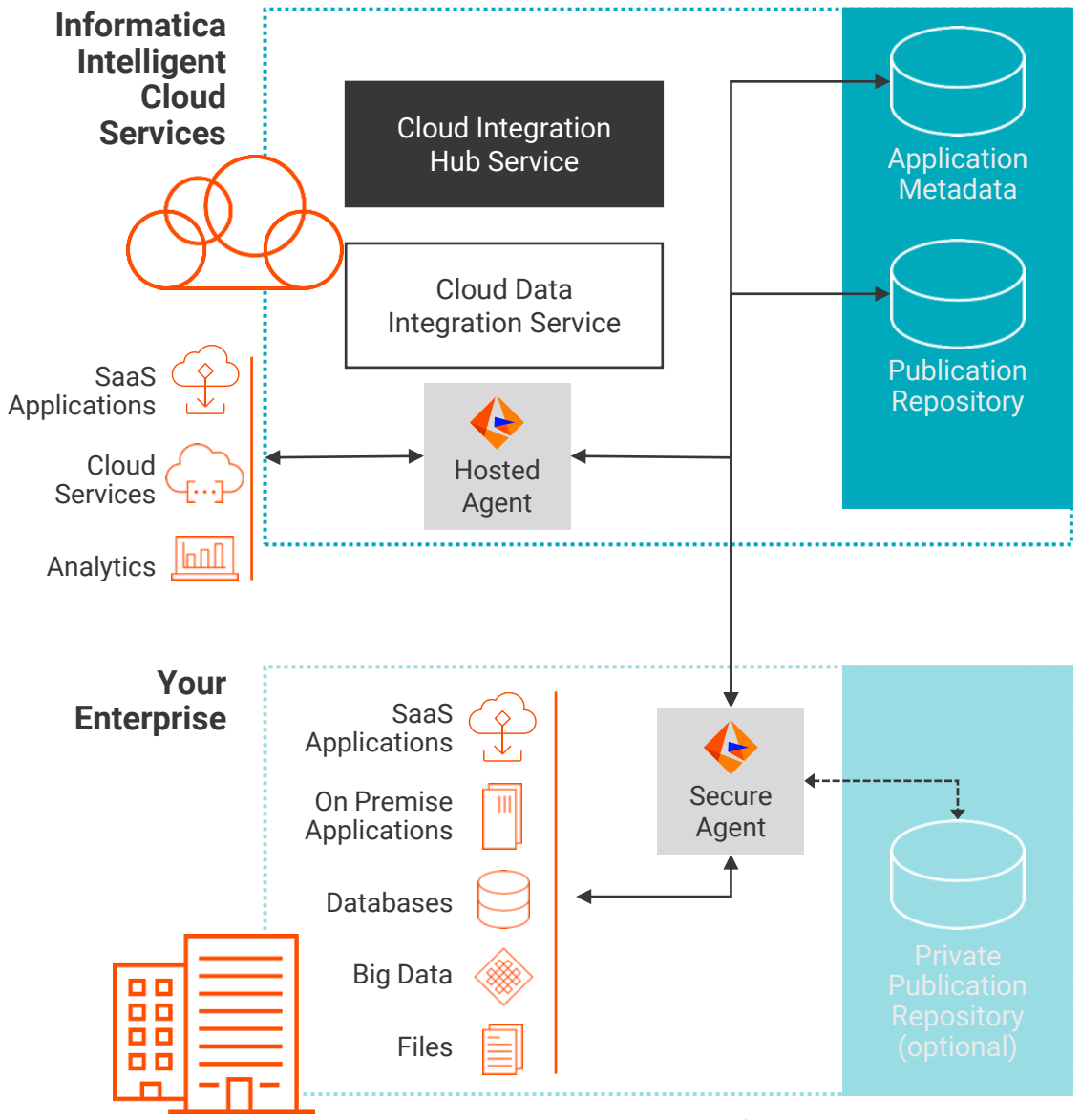


- Utilize Informatica Cloud mappings and tasks for publish/subscribe workflows
- Auto create simple DI tasks from CIH interface
- Use Pub/Sub APIs with CAI as well as use CAI as a scheduler

Informatica

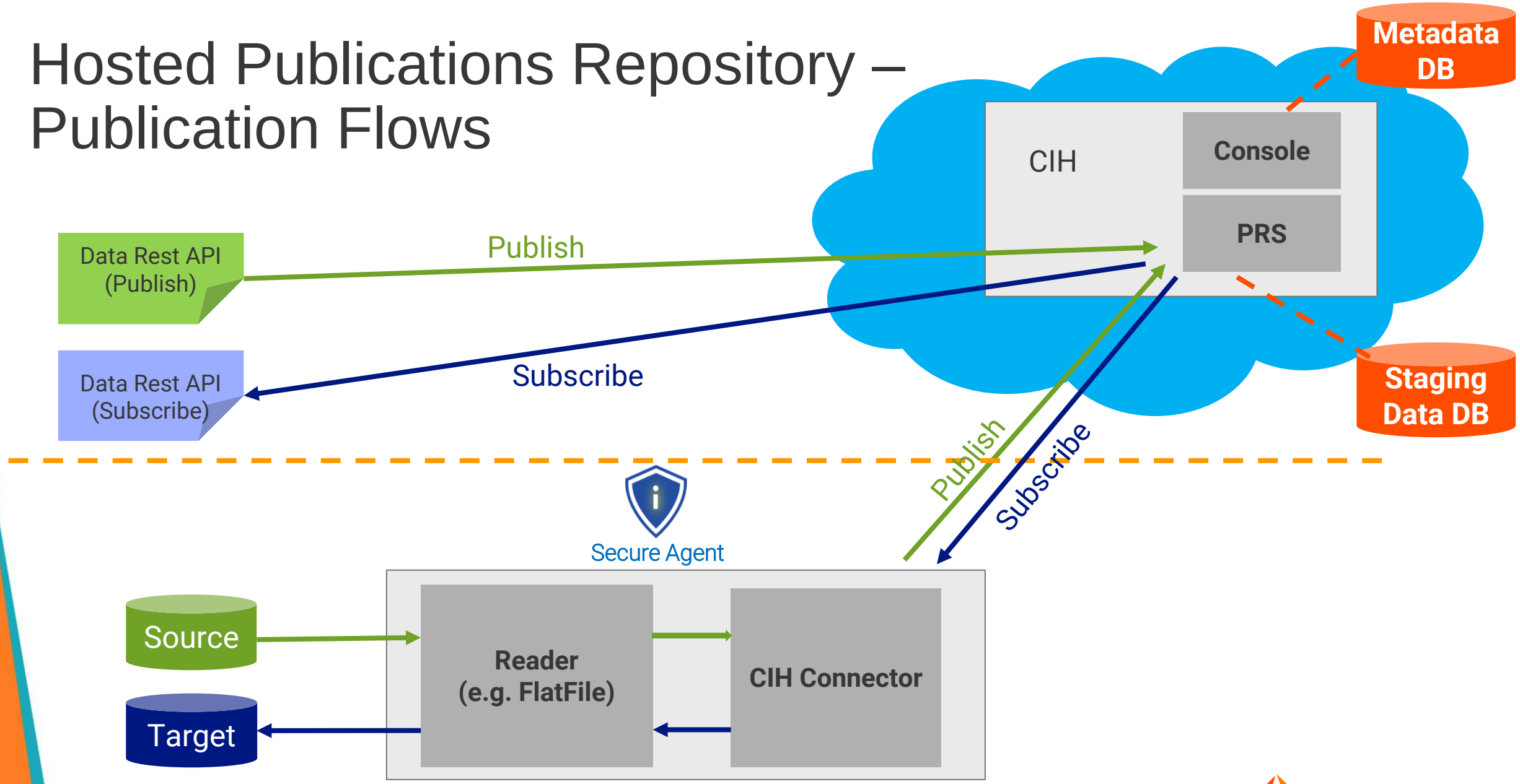
Cloud Integration Hub Architecture

- 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



Informatica

Hosted Publications Repository – Publication Flows

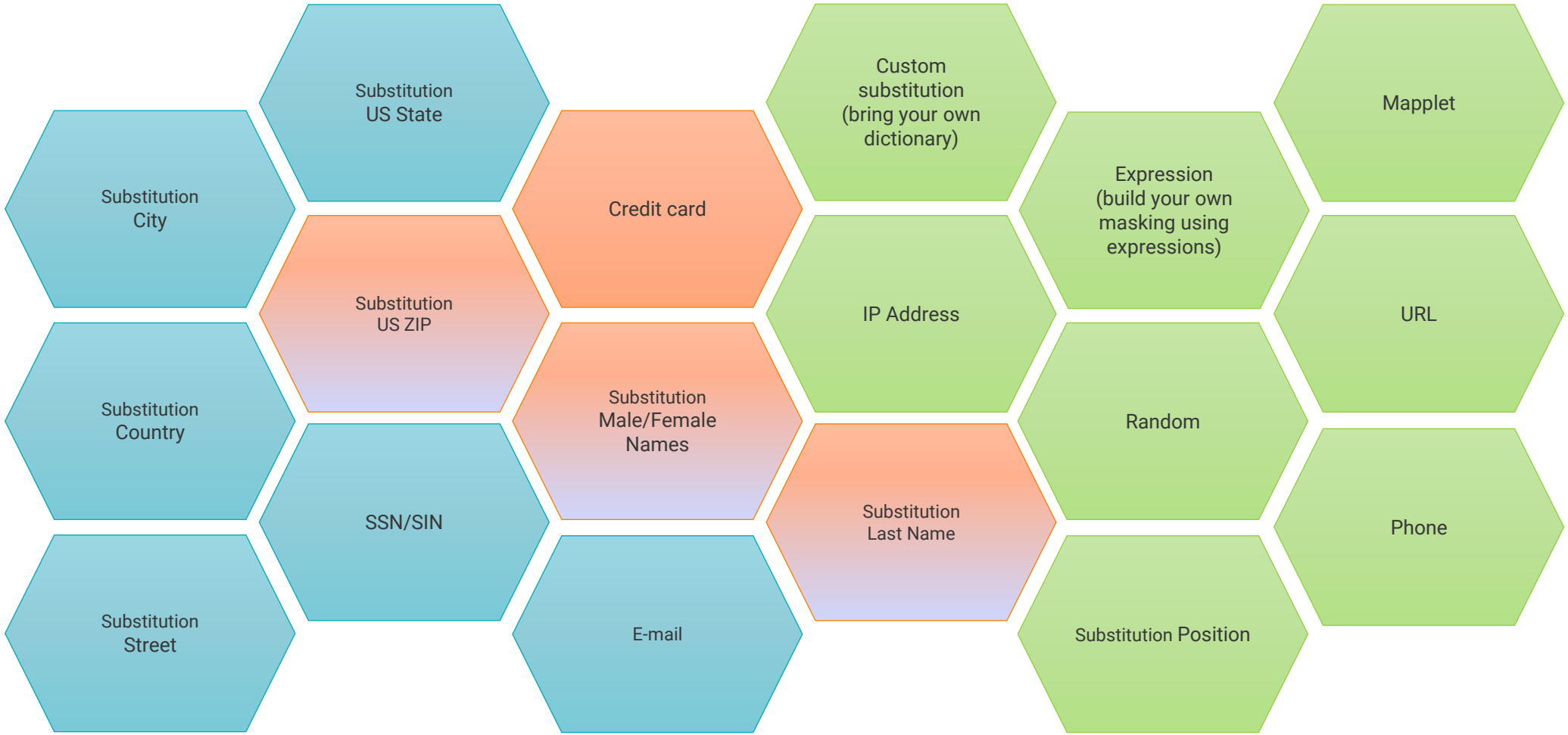


Informatica

Informatica

Cloud
Test Data Management

Out-of-the-Box Masking Techniques



Regulations: PII, PHI, PCI

Informatica

Consistent Masking

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

Repeatable: same masked result for the same input

Original	Masked
John Adams	**Tim Jones**
Nicholas Cage	Jack Nicholson
John Adams	**Tim Jones**

Repeatable ON

Original	Masked
John Adams	**Tim Jones**
Nicholas Cage	Jack Nicholson
John Adams	**Fred Mercury**

Repeatable OFF

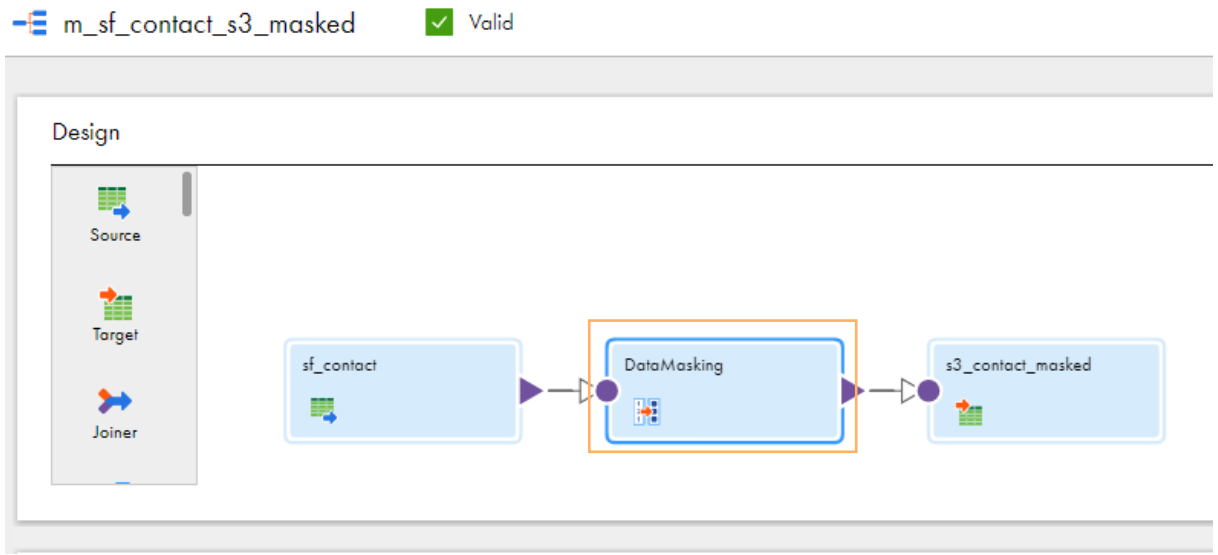
Seed value: point for generating masked values. If the same seed value and same dictionary (for substitutions) is used, same result will be obtained

Informatica

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

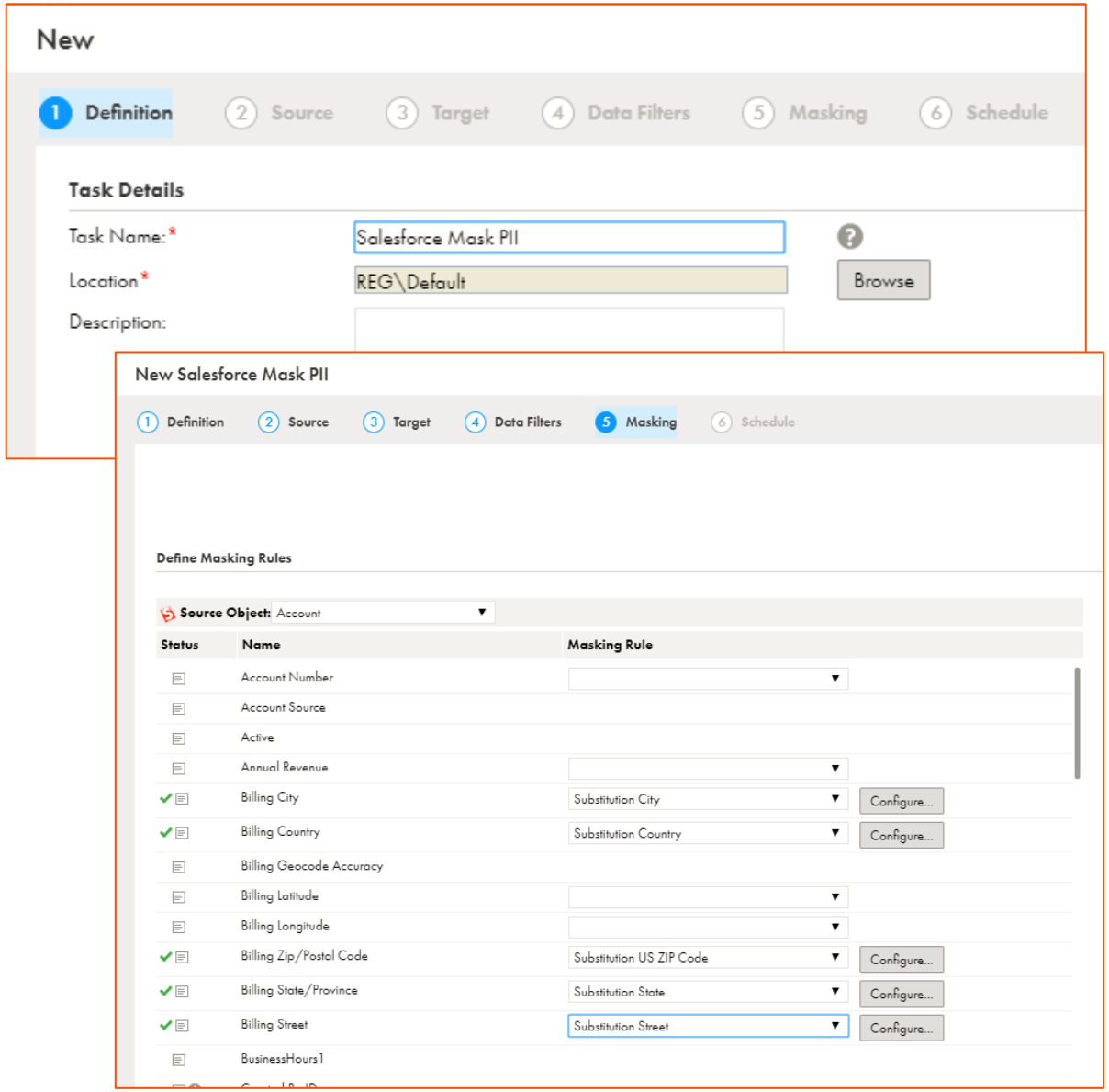


Informatica

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



Question?

Thank You

Global Technical Alliances (GTA) and PTS Team

Informatica®