Informatica

09 Sep, 2025

Maximizing Synergies between Informatica and Databricks

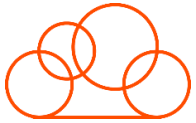
- Spoorthi Vaidya, Senior Consultant, IPS
- Sotha Ith, Principal Solutions Architect, IPS

Where data & AI come to LIFE

Housekeeping Tips

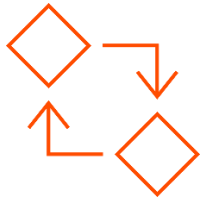
- Today's Webinar is scheduled for 1 hour
- The session will include a webcast and then your questions will be answered live at the end of the presentation
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to "All Panelists" via the Q&A option and we will respond at the end of the presentation
- The webinar is being recorded and will be available on our Success Portal - where you can download the slide deck for the presentation. The link to the recording will be emailed as well.
- Please take time to complete the post-webinar survey and provide your feedback and suggestions for upcoming topics.

Feature Rich Success Portal



Bootstrap trial and POC Customers

Enriched Customer Onboarding experience

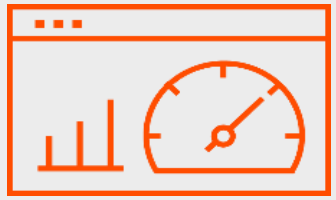


Product Learning Paths and Weekly Expert Sessions

Informatica Concierge

Tailored training and content recommendations

More Information



Success Portal

https://success.informatica.com

Communities & Support

https://network.informatica.com

Documentation

https://docs.informatica.com

University

https://www.informatica.com/in/services-and-training/informatica-university.html

Safe Harbor

The information being provided today is for informational purposes only. The development, release, and timing of any Informatica product or functionality described today remain at the sole discretion of Informatica and should not be relied upon in making a purchasing decision.

Statements made today are based on currently available information, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products or functionality in the future.

Agenda

1. Partnership
2. Capabilities Overview
3. Best Practices & Considerations
4. Case Studies
5. Q&A

Informatica + Databricks

Partnership


Informatica
databricks

Partnership Overview

Growth

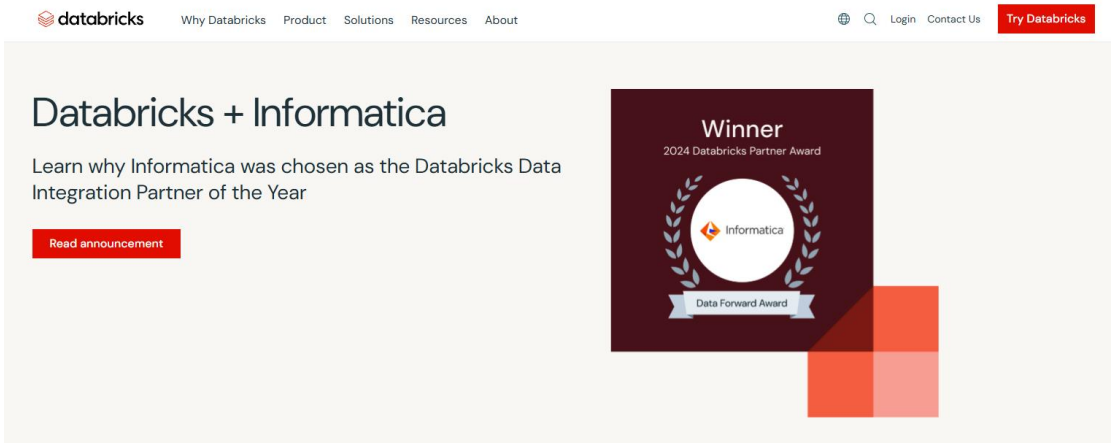
Informatica was named top 3 fastest growing services on Databricks over 170% growth

Partnership Alignment

Technology Partner Select

End to End Cloud Data Management

Full support across Informatica Data Management for Cloud (IDMC) for Databricks Unity Catalog



Informatica

Informatica + Databricks

Overview

End-to-end Enterprise Data Management

Extending Databricks to the enterprise data estate

Discover & understand enterprise data assets needed to bring to Databricks

Move & optimize more data from disparate enterprise sources to Databricks faster

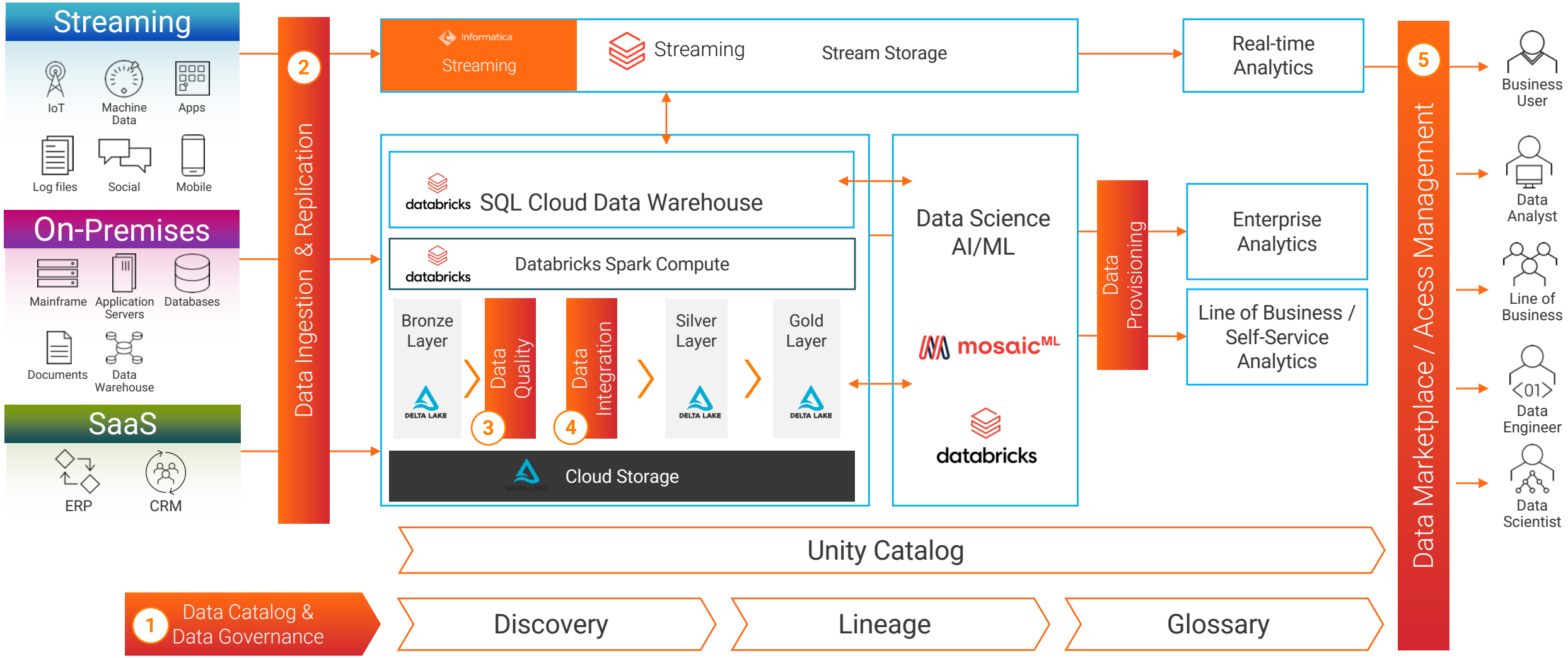
Profile & remediate data quality issues in Databricks

Ensure trusted data by governing and securing data across enterprise sources

Match, merge and consolidate to create a golden record (360º view) of multi-domain data to provide to Databricks

Share enterprise data assets securely to ensure compliance for internal data mesh users

Informatica Empowers the Lakehouse Architecture

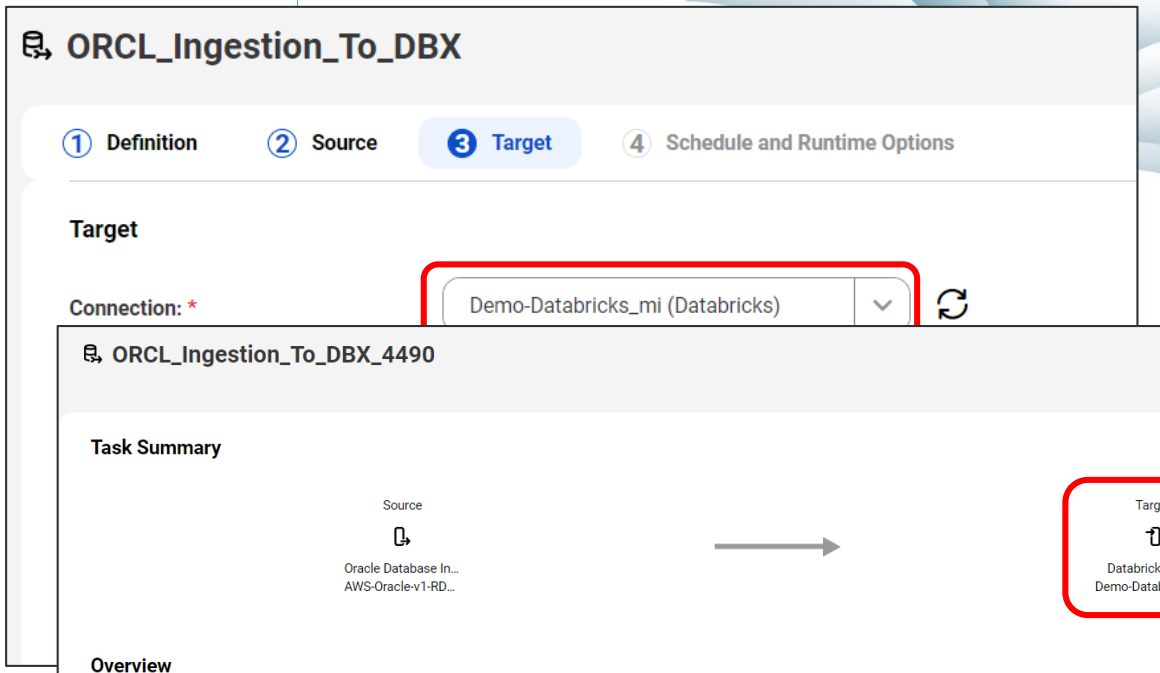


Data Ingestion & Replication

Template Driven Experience for Standardization & Simplification

Apps, Streaming, File & DBs from on-premises to cloud solutions

Real-Time Monitoring & Alerting

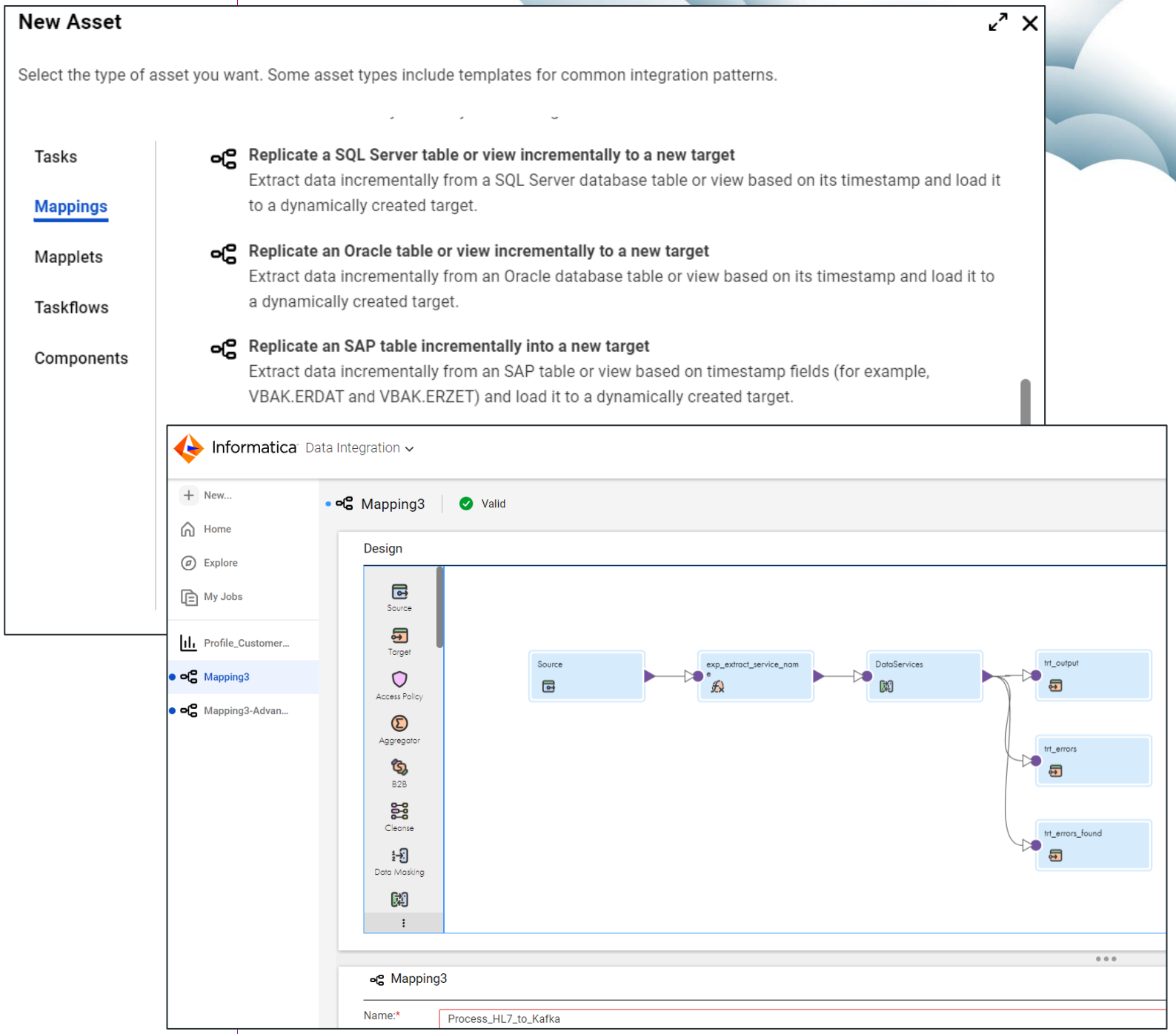


Data Integration

Templates to create design assets from scratch for standardization

Pre-defined templates for popular patterns to enable faster onboarding

300+ purpose-built cloud-native connectors

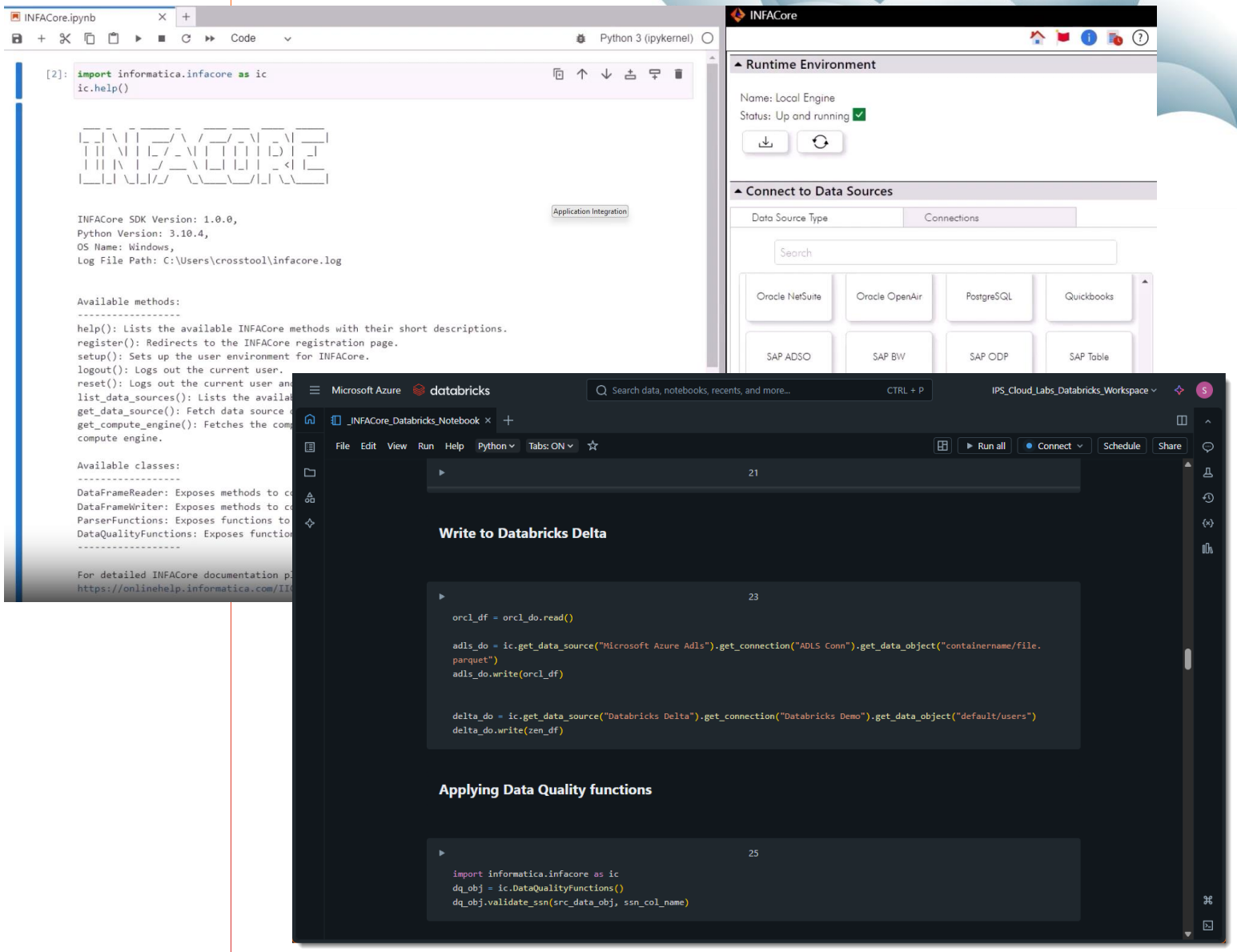


INFACore

Low-Code Intelligent Data Management

AI-Powered Automation with CLAIRE Engine

Seamless Integration via SDKs and Extensions



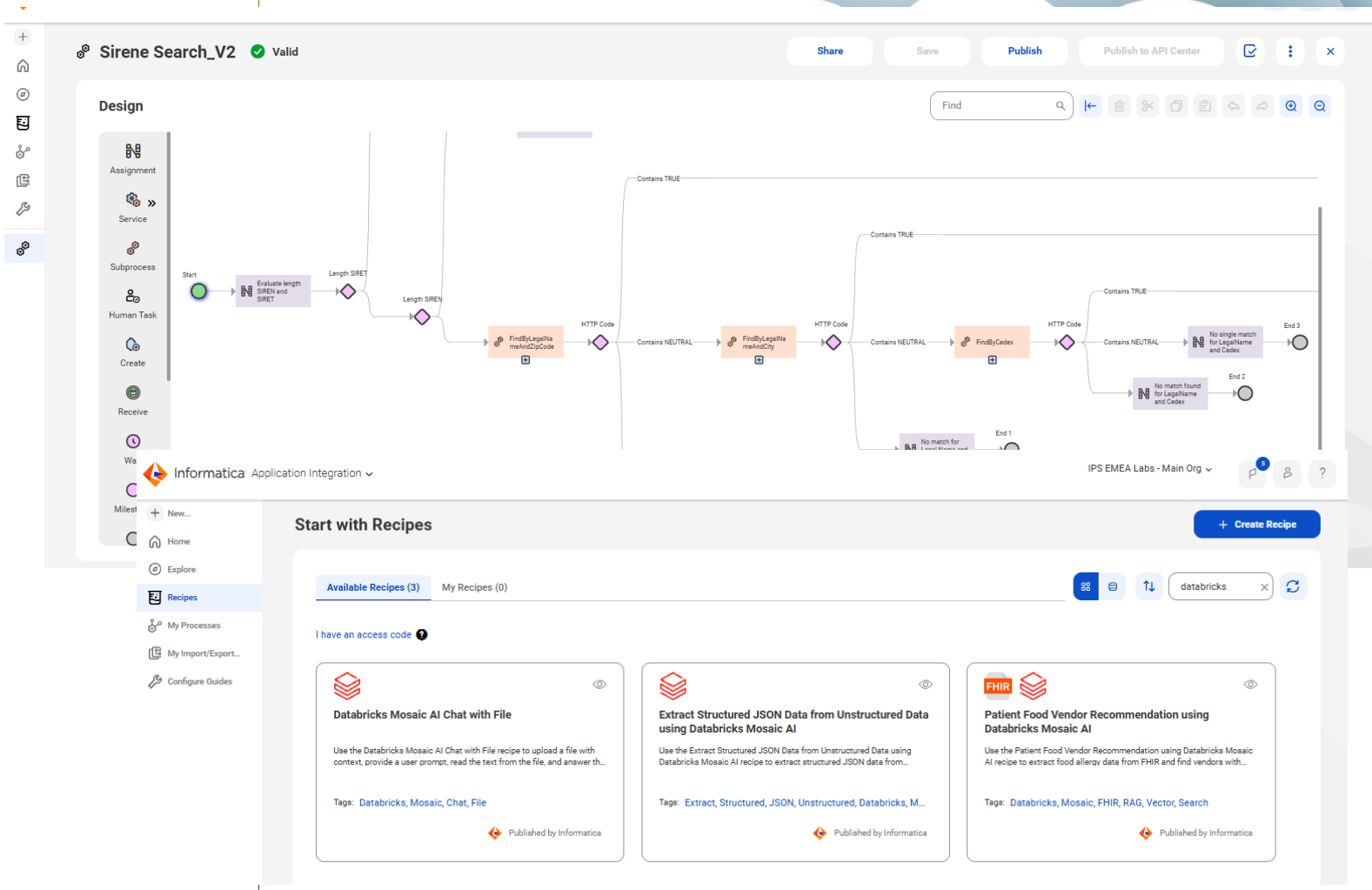
Informatica

Application Integration

Enables automation of complex business processes and user workflows

Easy integration to REST, SOAP APIs and others

Recipes available as pre-configured sets of assets

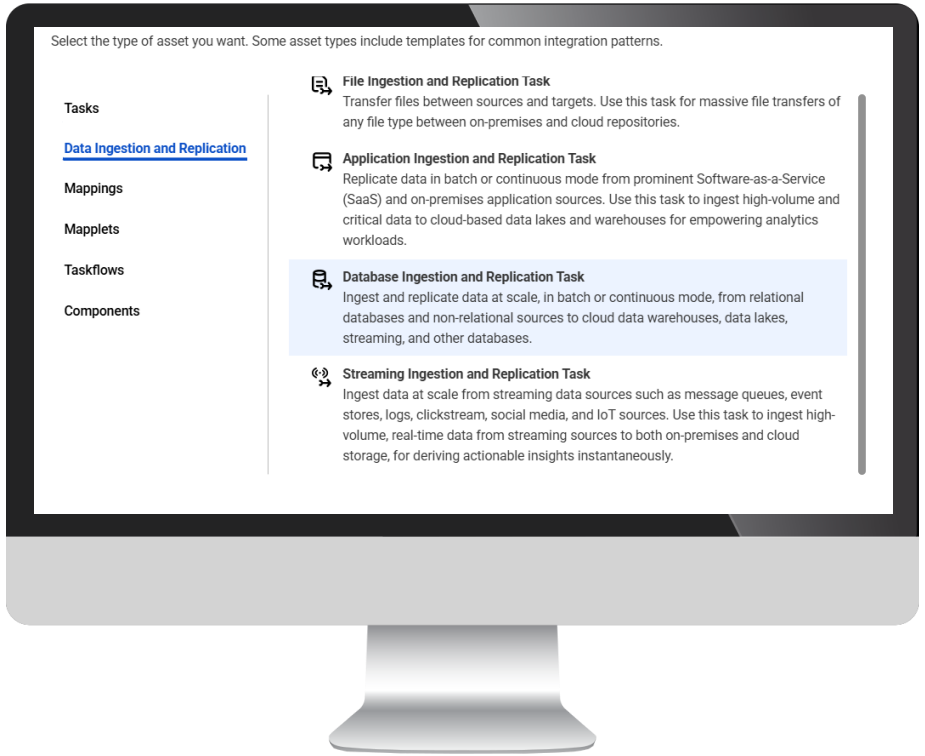


Informatica

Informatica

DEMO

Database Ingestion and replication - Databricks as Destination



Where data & AI come to LIFE™

Informatica

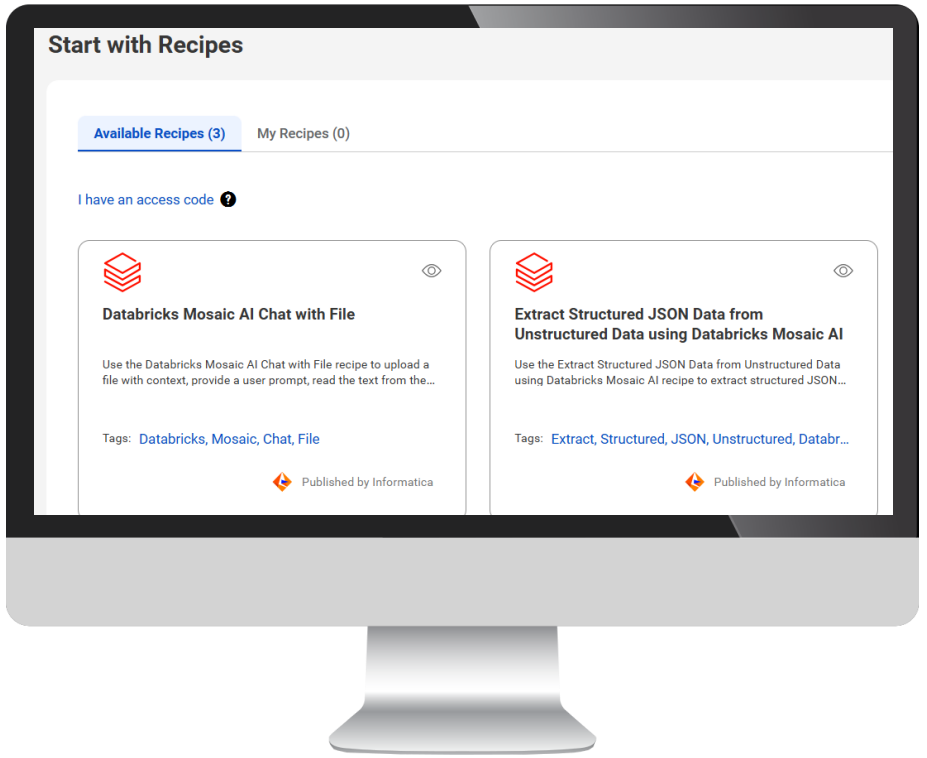
DEMO

Databricks Integration with INFACore

Where data & AI come to LIFE

DEMO

Databricks Mosaic AI Recipes – Cloud Application Integration

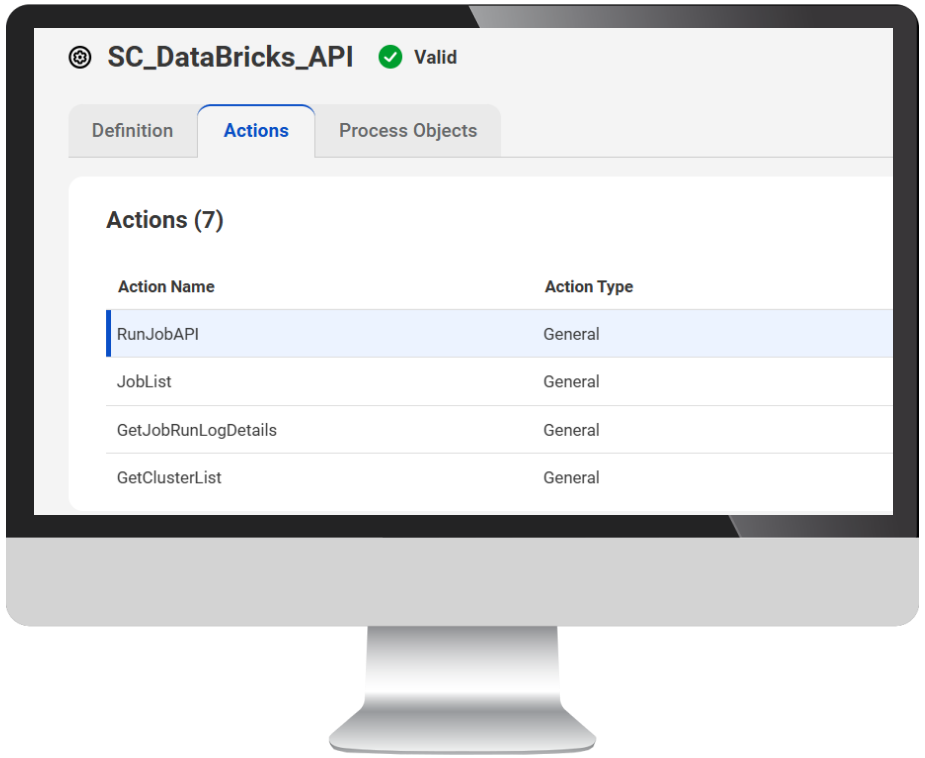


Where data & AI come to LIFE

Informatica

DEMO

Connect to Databricks Rest API using CAI Service Connector



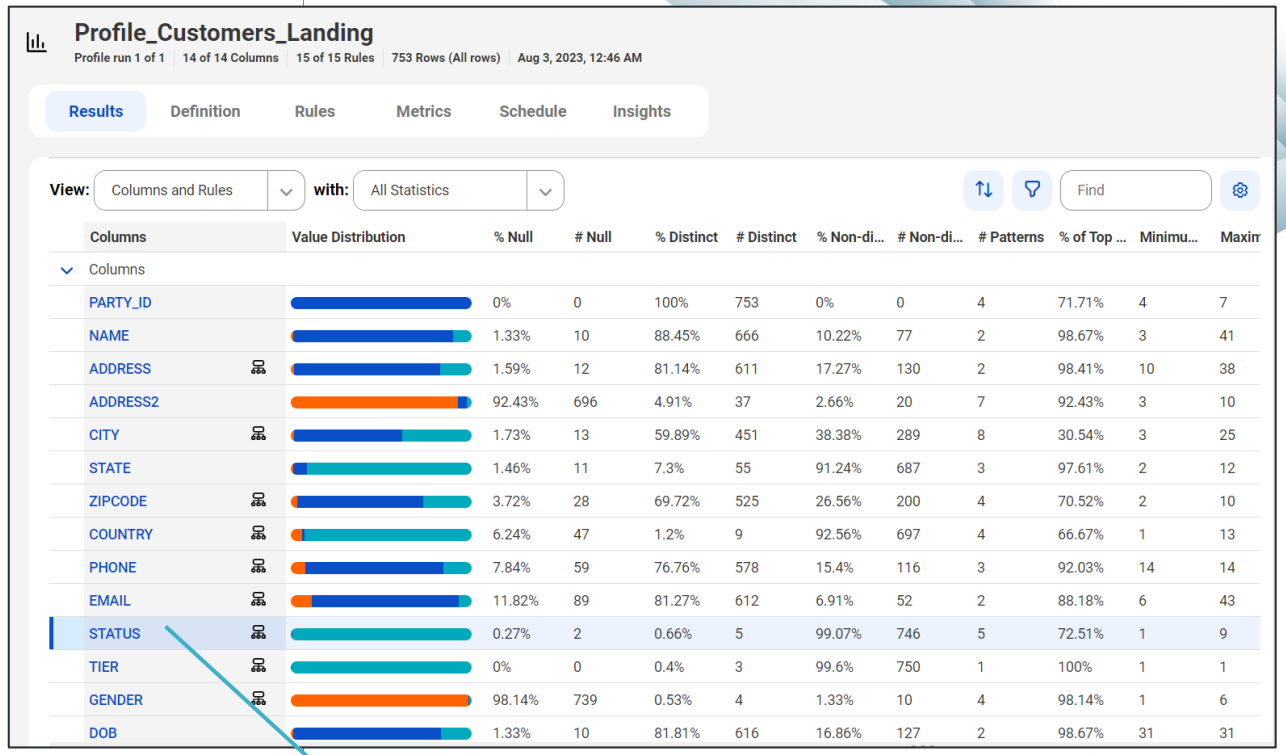
Where data & AI come to LIFE

Data Profiling

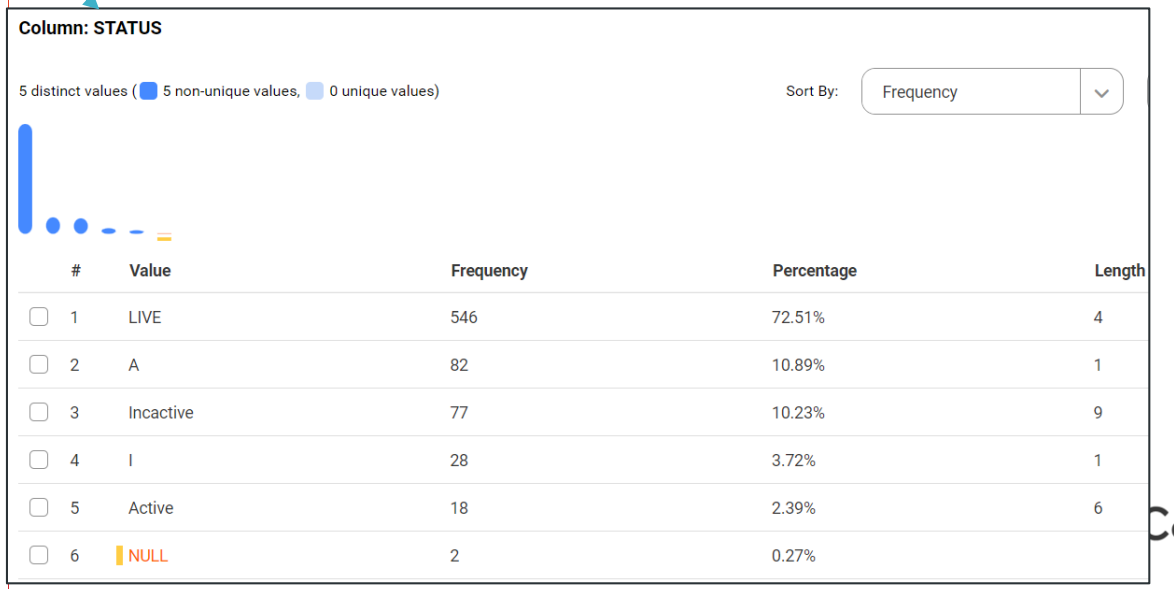
Profile data with zero coding and using OOTB templates

Analyze data frequencies, anomalies, etc…
Drill down to see details

Identify potential data issues
Execute What-If scenarios



Columns	Value Distribution	% Null	# Null	% Distinct	# Distinct	% Non-di...	# Non-di...	# Patterns	% of Top ...	Minimu...	Maxim
Columns											
PARTY_ID		0%	0	100%	753	0%	0	4	71.71%	4	7
NAME		1.33%	10	88.45%	666	10.22%	77	2	98.67%	3	41
ADDRESS		1.59%	12	81.14%	611	17.27%	130	2	98.41%	10	38
ADDRESS2		92.43%	696	4.91%	37	2.66%	20	7	92.43%	3	10
CITY		1.73%	13	59.89%	451	38.38%	289	8	30.54%	3	25
STATE		1.46%	11	7.3%	55	91.24%	687	3	97.61%	2	12
ZIPCODE		3.72%	28	69.72%	525	26.56%	200	4	70.52%	2	10
COUNTRY		6.24%	47	1.2%	9	92.56%	697	4	66.67%	1	13
PHONE		7.84%	59	76.76%	578	15.4%	116	3	92.03%	14	14
EMAIL		11.82%	89	81.27%	612	6.91%	52	2	88.18%	6	43
STATUS		0.27%	2	0.66%	5	99.07%	746	5	72.51%	1	9
TIER		0%	0	0.4%	3	99.6%	750	1	100%	1	1
GENDER		98.14%	739	0.53%	4	1.33%	10	4	98.14%	1	6
DOB		1.33%	10	81.81%	616	16.86%	127	2	98.67%	31	31



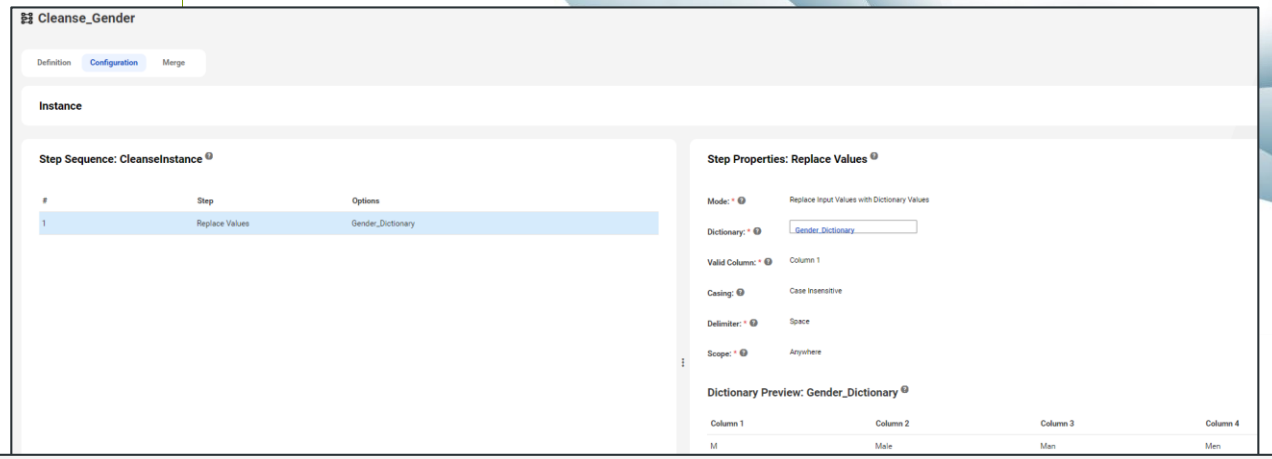
#	Value	Frequency	Percentage	Length
1	LIVE	546	72.51%	4
2	A	82	10.89%	1
3	Incactive	77	10.23%	9
4	I	28	3.72%	1
5	Active	18	2.39%	6
6	NULL	2	0.27%	

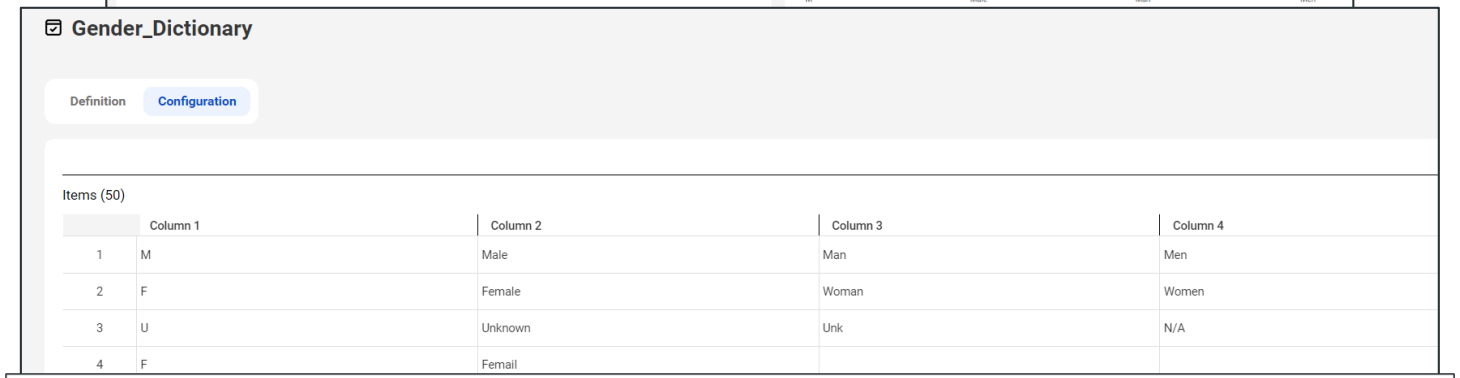
Data Quality

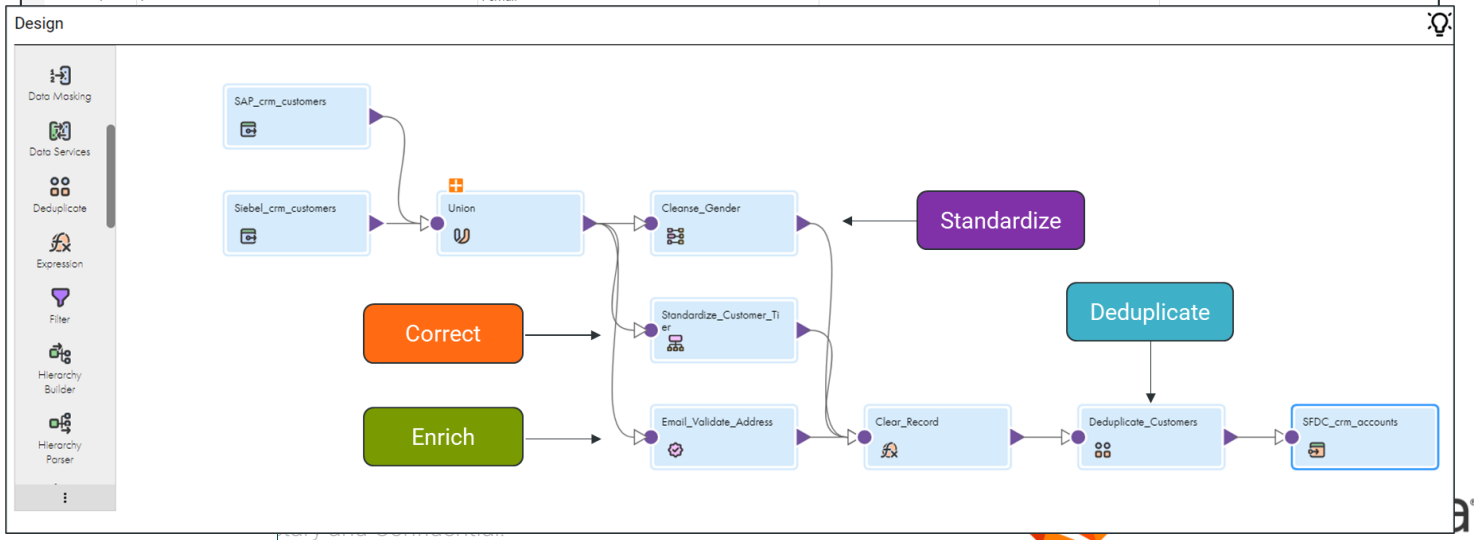
Build Reusable Assets for Standardization & Cleansing (applicable to DI processes)

Leverage Reference Tables to Refine & Detect Data Outliers

Unit Test to Ensure Accuracy & Expected Outcomes





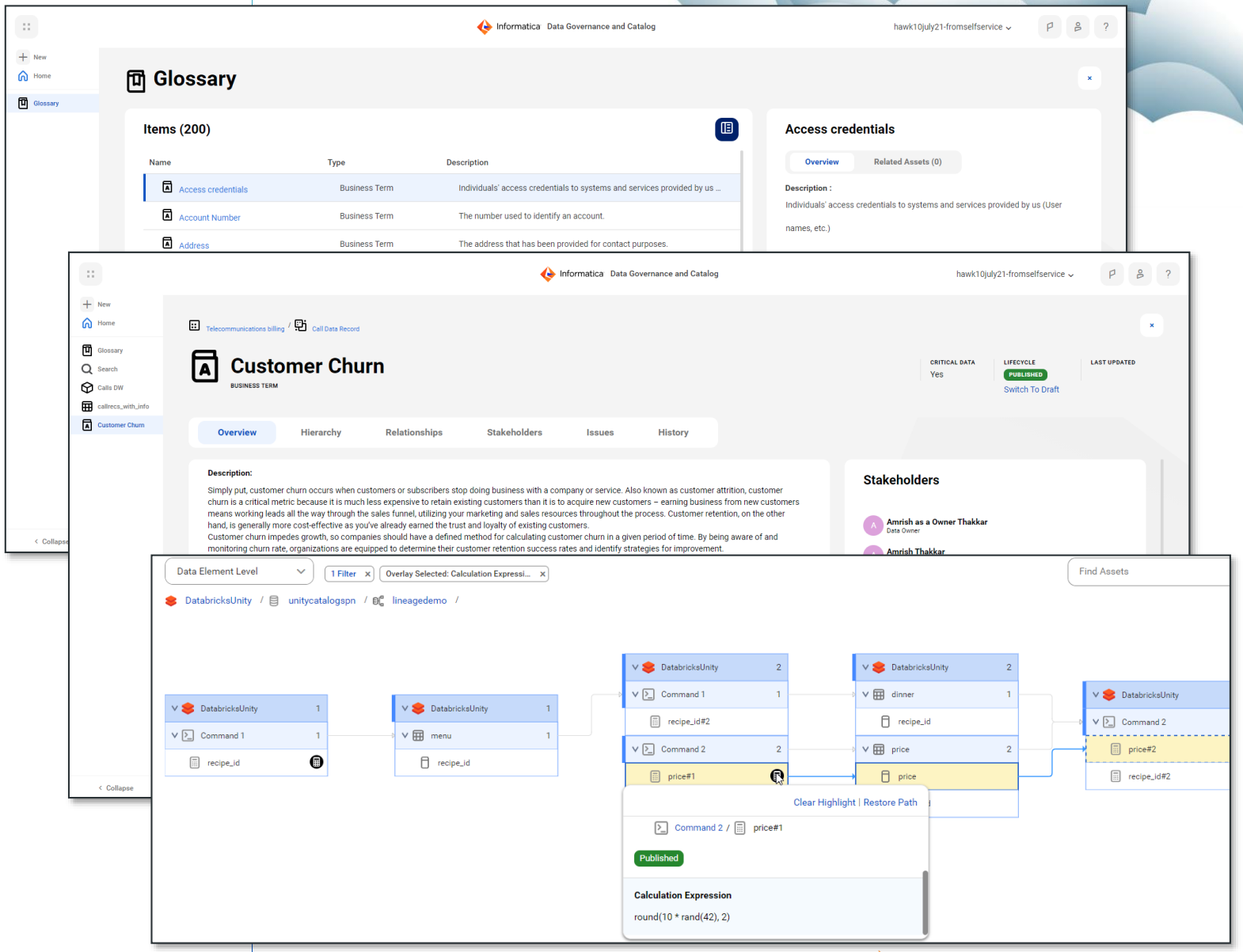


Data Governance and Catalog

Access business and technical assets in a single UI featuring universal search, relationships, meta-model and lineage

Govern AI Models, exposing what is available for reuse and highlighting the use of types of data and related policies

Drilldown from Business to Technical Lineage and overlay with key metrics



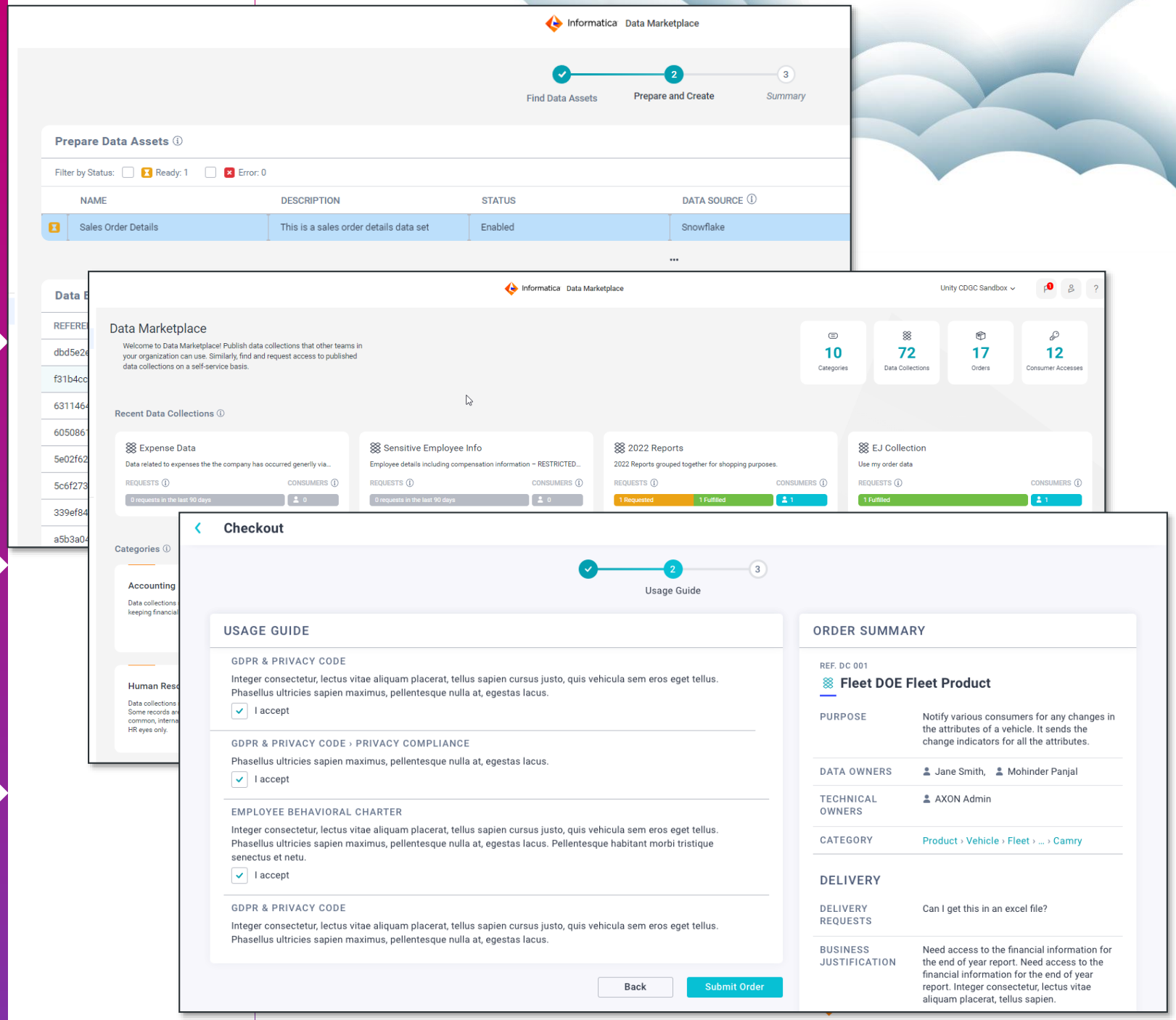
Informatica

Data Marketplace

Data Producers populate marketplace with governed datasets

Data Consumers shop for data and checkout if data fits their needs

Data Owners deliver and auto-provision data after approval of request

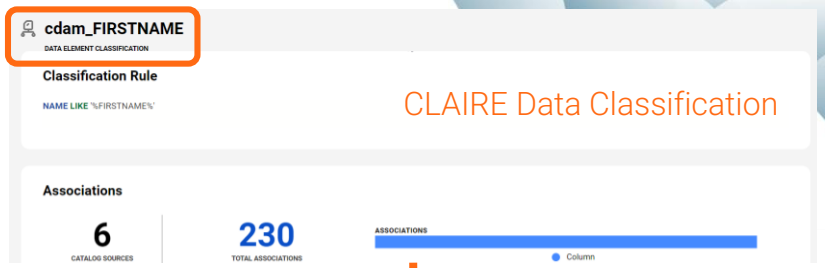


Data Access Management

Centrally define data access control policies with a no-code approach

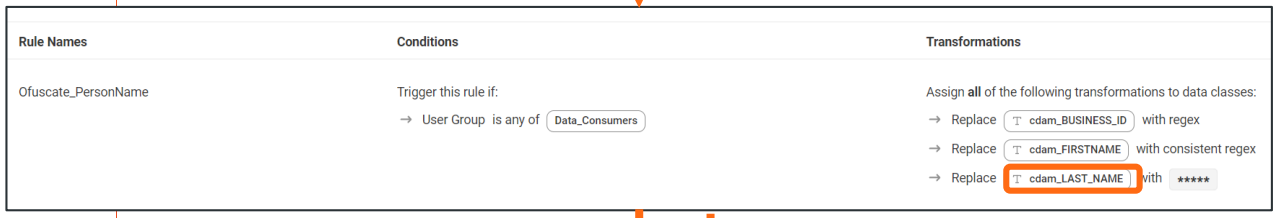
Several de-identification techniques such as substitution, tokenization, generalization to mask sensitive data

Offers granular control over data access



CLAIRE Data Classification

CDAM Access Control Policy



User with Privileges

FIRST_NAME	SECOND_NAME	EMPLOYEE_NUMBER	COUNTRY	JOB_TITLE
Salvador	Meyer	84047	Ireland	Legal Administrator
Sara	Townsend	55686	France	Network Administrator
Alexis	Deleon	55801	Ireland	Compliance Officer
Gabrielle	Cuevas	13962	Germany	Media Buyer
Brecken	Berg	95088	Germany	Electrical Engineer
Emmalyn	Benjamin	88642	United Kingdom	IT Project Manager
Kyro	Mayo	91949	Germany	Media Buyer
Aarya	Booth	76846	United Kingdom	Mechanical Engineer
Chaim	Ford	61428	Ireland	Compliance Officer
Alexandra	Sloan	49395	Ireland	Mob...
Ocean	Camacho	73339	United Kingdom	Con...

User without Privileges

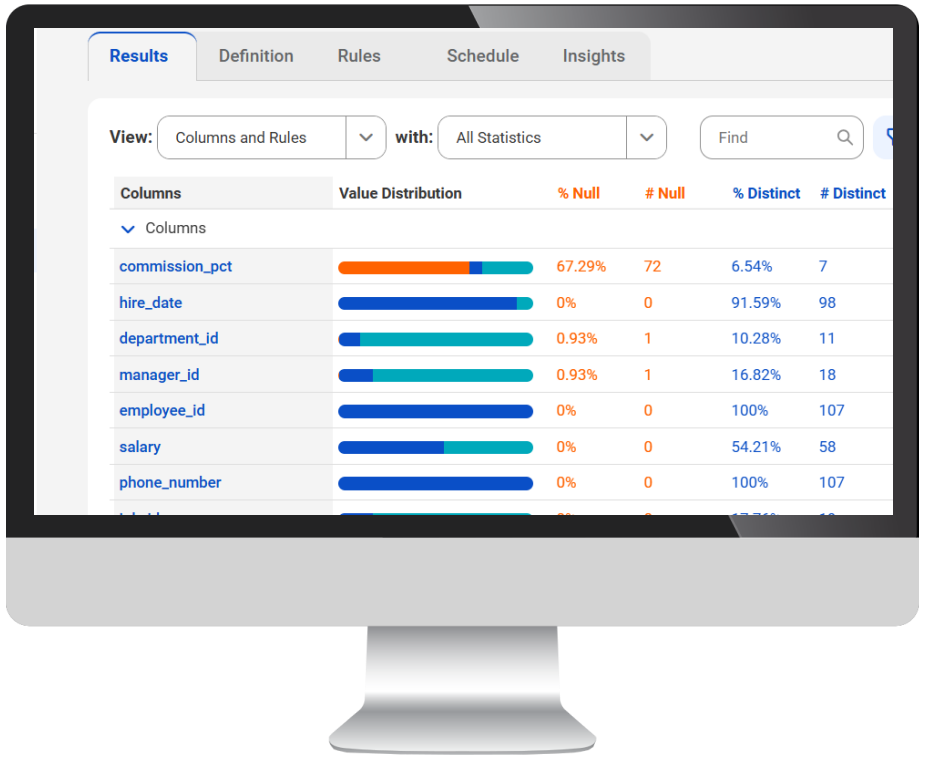
FIRST_NAME	SECOND_NAME	EMPLOYEE_NUMBER	COUNTRY	JOB_TITLE
Jxkbladdhinfv	******	936042276038	Germany	Court Reporter
Xwlqsxcrfcsjp	******	444607627894	United Kingdom	Marketing Research Analys
Pkvxqxbrtcajn	******	5484484031640	Germany	Project Manager
Msduoqilcgext	******	6872587489830	United States	Copywriter
Xqmzuogzjemn	******	2201799735622	Germany	Cloud Solutions Architect
Lmonphcbeepzc	******	7088887775853	Germany	Quality Control Analyst
Zrvxqmyvaadga	******	3309725977403	Germany	Marketing Research Analys
Ypmxmpujkmbad	******	332215463410	France	Web Developer
Bpwjiqccyvi	******	569455264405	France	Cybersecurity Analyst
Bpvzxpmtqbogw	******	2133995421122	United Kingdom	Software Developer
Nuzzrodlezafl	******	3854298161402	Germany	Media Buyer
Xpiczswgxrnnl	******	6649559732424	United Kingdom	Civil Engineer

Informatica

Informatica

DEMO

Data Profiling on Databricks Connection Object

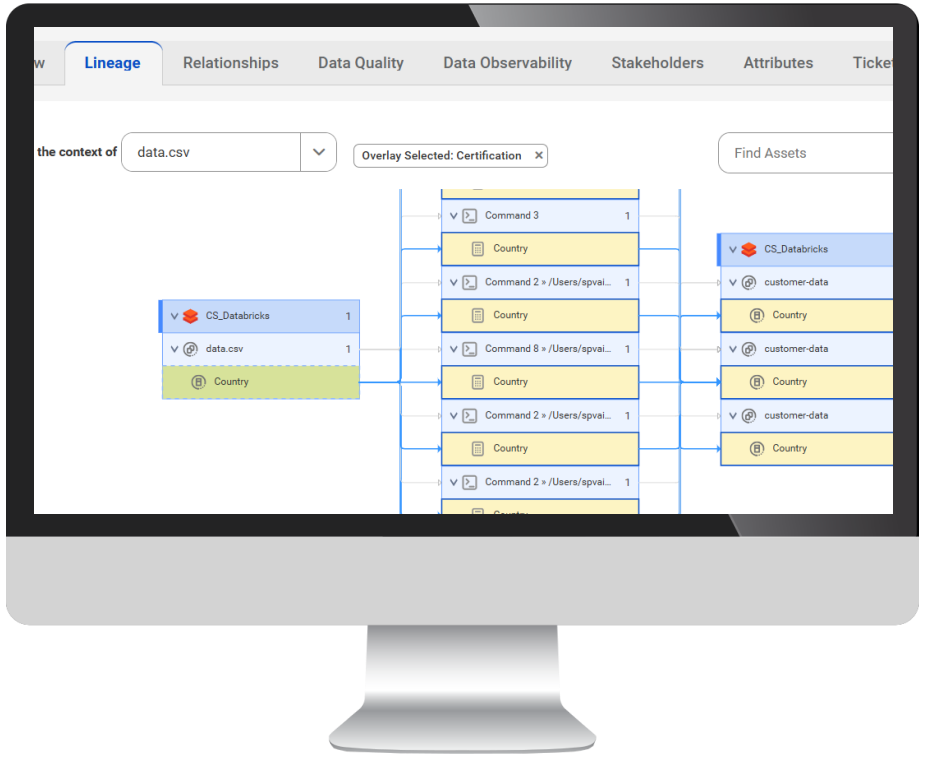


Where data & AI come to LIFE

Informatica

DEMO

Extract Metadata from Databricks Source System using Metadata Command Center & view the results in Data Catalog and Governance Service



Where data & AI come to LIFE

Informatica + Databricks

Best practices and Considerations

CDI : ETL versus ELT pattern comparison

Functionality / Feature	Informatica Secure Agent (ETL)	SQL ELT
Compute pattern / engine	Informatica Agent sitting on customer infrastructure (IaaS or onprem)	Databricks cluster via SQL WH endpoint (autogenerated commands and SQL queries)
Data Movement	Data needs to go from Databricks into Secure Agent to be transformed	No data leaves Databricks
Performance	Good	Optimal
Compatibility of functions and transformations	All functions and transformations are available	Not all functions/transformations can be pushed-down
Data Quality / Masking	Full availability of all DQ and masking transformations	DQ/masking transformations cannot be pushed-down to Databricks
Governance	Full lineage	Full lineage

CDI : Functions in Mappings in SQL ELT mode

Subtitle here

Supported Function Category	Examples (not exhaustive)
Aggregate	AVG(), SUM(), COUNT() and specialized functions like APPROX_COUNT_DISTINCT() and REGR_SLOPE()
AI	AI_ANALYZE_SENTIMENT(), AI_TRANSLATE(), and AI_SUMMARIZE()
Cast	Conversion-related operations such as TO_CHAR(), TO_NUMBER(), and TRY_TO_NUMBER()
Miscellaneous	Utility operations like NVL(), COALESCE(), IFNULL(), and GREATEST()
Window	ROW_NUMBER(), RANK(), DENSE_RANK(), and FIRST_VALUE()

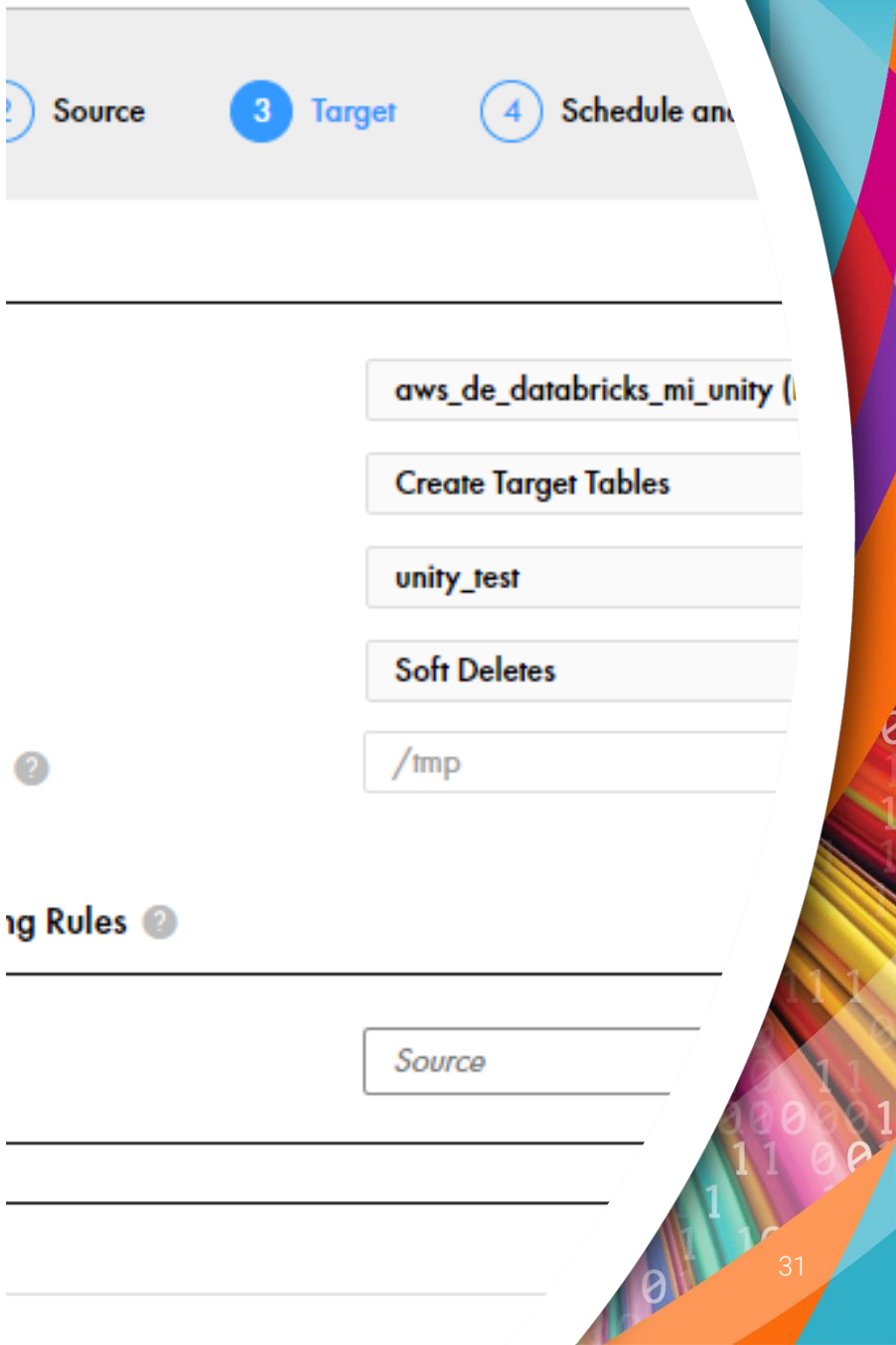
… and more

Please refer to Informatica documentation to get the full list.

CDI/CDIR : Considerations

Subtitle here

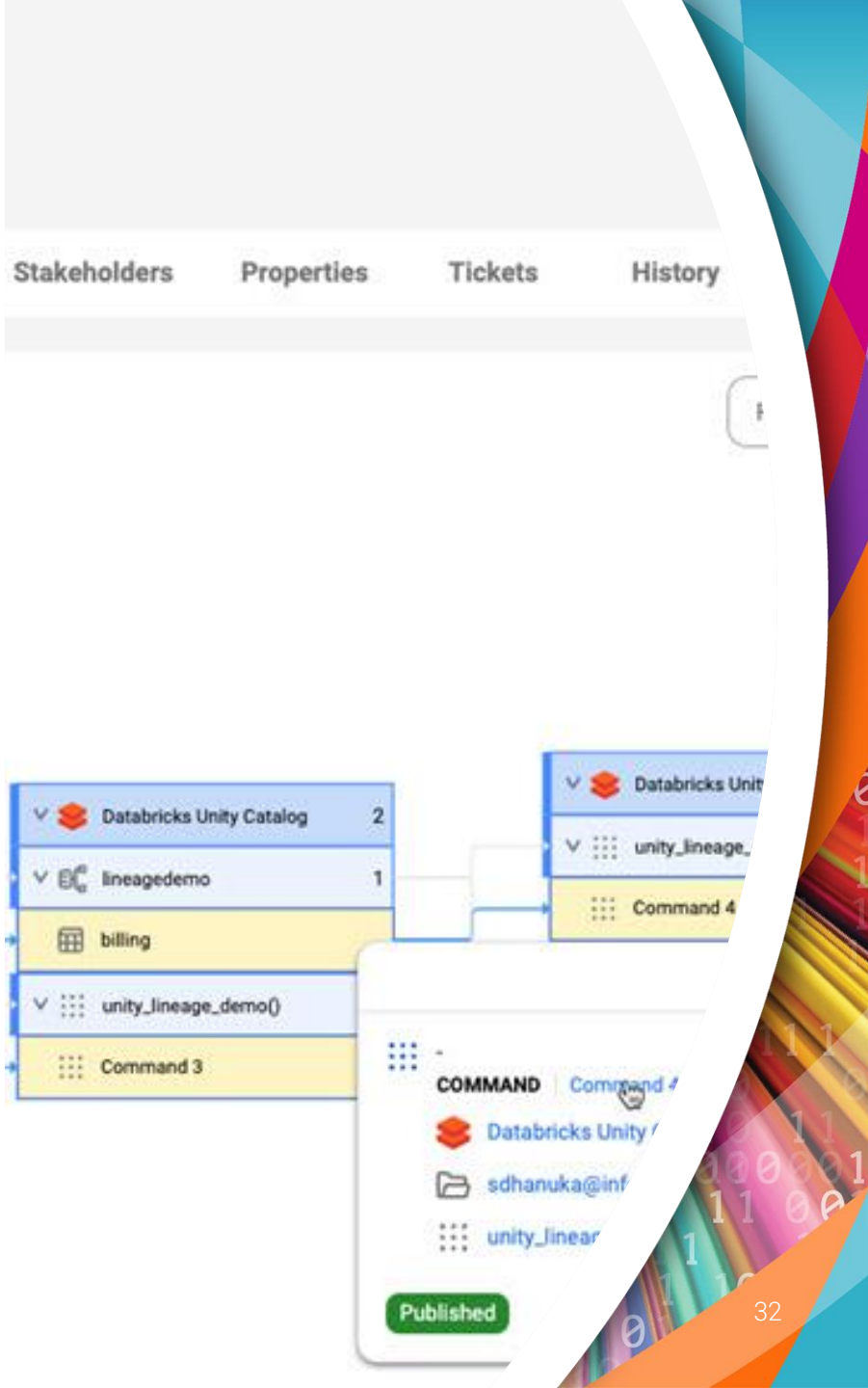
- You can only access Databricks tables created on top of S3 or ADLS Gen2
- Data ingestion jobs wait for 10 minutes for DBX cluster to be available. After that time, connection will time out and job will fail
  - This 10 minutes timeout value can be customized
- By default, Database Ingestion and Replication uses the Databricks COPY INTO feature to load data from the staging file to Databricks target tables
- Ingestion jobs that have Databricks targets can get schema information for generating the target tables from Databricks Unity Catalog



CDGC : Considerations

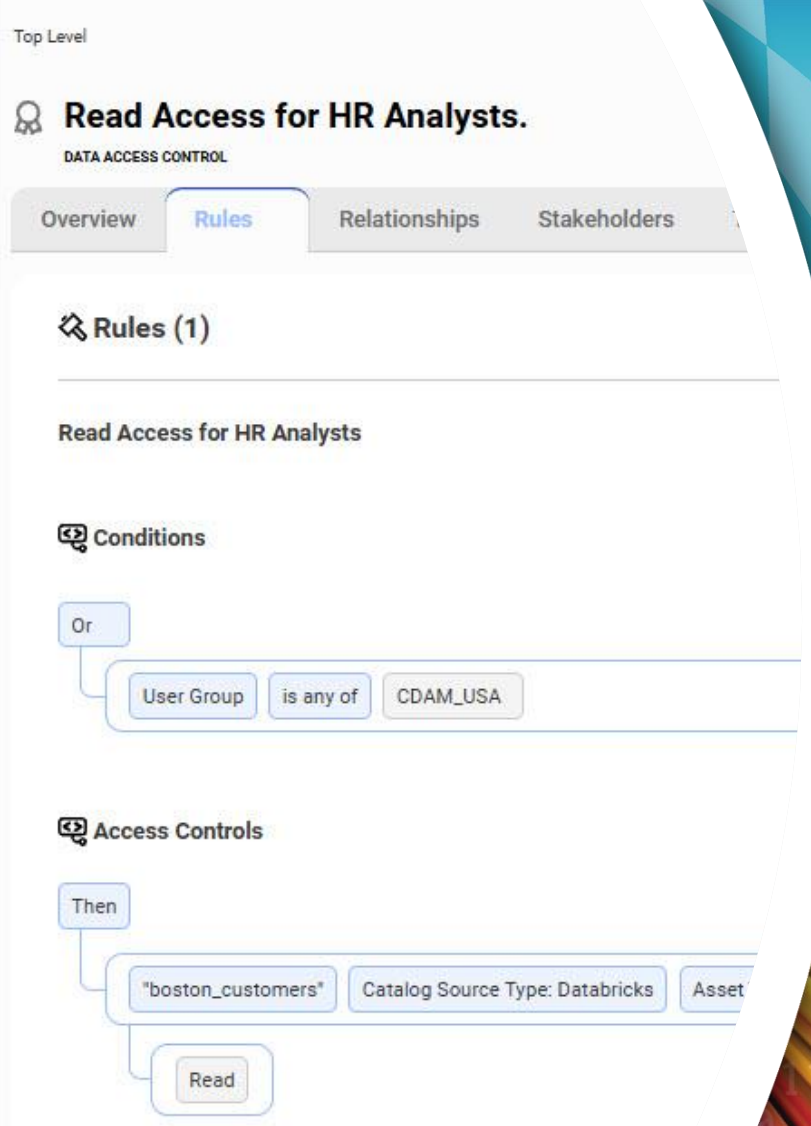
Subtitle here

- Unity Catalog requirement
  - Lineage extraction requires Unity Catalog to be enabled and run on the Premium tier.
  - Lineage is computed for tables registered in a Unity Catalog metastore, but it is not captured for data written directly to files in cloud storage, even if a table is defined at the storage location.
- Data Profiling filters
  - Profiling filters need to reference schema and table-level information accurately (e.g., Catalog.Schema_Name.TABLE_NAME). Errors can occur if an asset does not exist in the catalog or if filters are applied incorrectly



CDAM : Policy Push Down Considerations

- Ensure that the user who is configured in the catalog source connection, the one used to push the policies, has workspace admin privileges on the catalog source
- Use a personal access token to connect to your Databricks instance through Data Access Management
- Because views are read-only objects, a source system will ignore permissions other than read when a policy applies to a view
- If you grant write permission to a Databricks object, you are also implicitly granting delete permission
- Data Access Management grants usage permissions to Databricks catalogs and schemas
- CDAM continuously runs (every X minutes) synchronization jobs to keep CDAM policy permissions in sync with Databricks rules and permissions. Execution interval can be set per-org basis
- Whenever a new data access policy is created or there are changes on existing policies in CDAM, an event is triggered that uses Databricks Unity Catalog APIs to reflect changes



CDAM Policy Permission	Databricks Permissions
Read	select
Write	modify / delete (implicit)

Informatica + Databricks

Case studies

Success Stories with Informatica + Databricks

Point72 improved data discovery efficiency by 75% using Informatica's IDMC to scan data on AWS S3 and integrate it with Databricks, creating a centralized "Catalog of Catalogs" that enhanced productivity and trust in data insights

Takeda Pharma overcame scalability and cost issues with their on-premises Hadoop by migrating to a next-gen data platform using Informatica and Databricks. This enabled large-scale data ingestion and processing with Databricks' optimized Spark engine, supported MLOps collaboration, and simplified analytics via Delta Lake for tools like Tableau and Qlik. Informatica ensures continuous updates of Delta Lake from enterprise sources, facilitating efficient cloud data management and innovation

This **biotech** customer switched from a competitor to Informatica IDMC for better stability in handling massive datasets. Using Databricks to process high-volume data, they improved data quality, supported regulatory compliance, and enhanced clinical research accuracy, reducing inconsistencies and streamlining research to improve patient outcomes

Thank You

Spoorthi Vaidya

Senior Consultant
Services Experience (APJ)

Sotha Ith

Principal Solutions Architect
Services Experience (EMEA)

Where data & AI come to

™

Informatica