



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** 

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



The screenshot shows the Databricks website header with navigation links: 'Why Databricks', 'Product', 'Solutions', 'Resources', 'About', 'Login', 'Contact Us', and a 'Try Databricks' button. The main content area features the heading 'Databricks + Informatica' and the text 'Learn why Informatica was chosen as the Databricks Data Integration Partner of the Year'. Below this is a 'Read announcement' button. To the right is a 'Winner' badge for the '2024 Databricks Partner Award' with the Informatica logo and 'Data Forward Award' text.

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

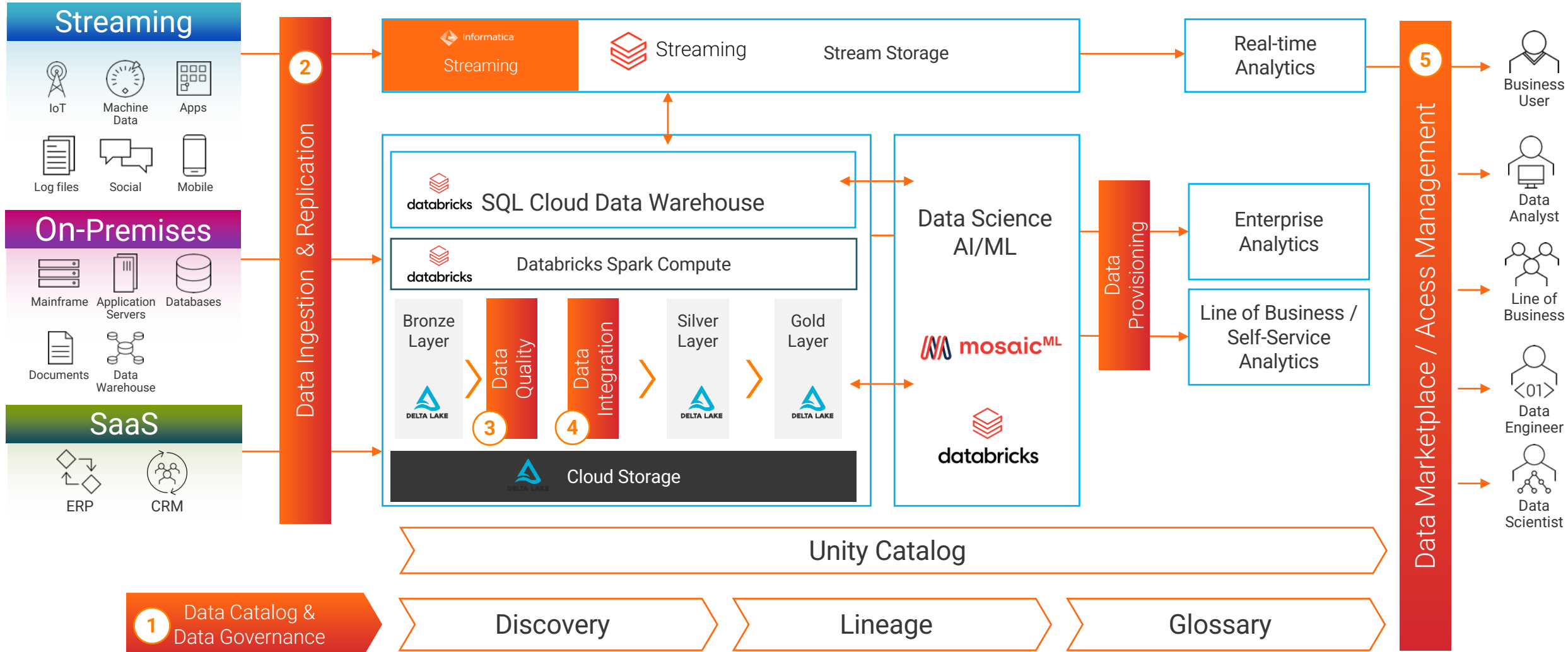
Profile & remediate data quality issues in Databricks

Ensure **trusted data** by **governing and securing** data across enterprise sources

Match, merge and consolidate to create a golden record (360° 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

The screenshot displays the Informatica Data Integration console interface. At the top, the task name is **ORCL_Ingestion_To_DBX**. Below it, there are four tabs: **1 Definition**, **2 Source**, **3 Target** (which is selected), and **4 Schedule and Runtime Options**.

The **Target** configuration section shows a dropdown menu for **Connection: *** set to **Demo-Databricks_mi (Databricks)**. Below this, a task summary card for **ORCL_Ingestion_To_DBX_4490** is shown, with a status of **Up and Running**. The summary includes a diagram of the data flow from the **Source** (Oracle Database In... AWS-Oracle-v1-RD...) to the **Target** (Databricks Delta Demo-Databricks...). A red box highlights the target component in the diagram.

The **Overview** section features a donut chart showing the task's progress: **83.33% Running : 5** and **16.67% Starting : 1**, with a total of **6** objects. Other details include **Runtime Environment: Azure_infanode**, **Task Name: ORCL_Ingestion_To_DBX**, **Load Type: Initial Load**, and **Task Type: Database Ingestion and Replication Task**.

The **Object Detail (6)** table provides a granular view of the data objects being processed:

Object	Target Object	Agent Name	Records Read	Records Written	Task Duration	Status	Log
AUSPOST_SRC.CUSTO...	AG_CUSTOMERS	master.azure.infa.world	0	0	00:00:45	Running	Select log
AUSPOST_SRC.IDM_S...	AG_IDM_SUBSTITUTIO...	master.azure.infa.world	103	103	00:01:09	Completed	Select log
AUSPOST_SRC.ORDERS	AG_ORDERS	master.azure.infa.world	1790	1790	00:01:07	Completed	Select log
AUSPOST_SRC.ORDER...	AG_ORDER_DETAILS	master.azure.infa.world	3241	3241	00:00:45	Running	Select log
AUSPOST_SRC.ORGAN...	AG_ORGANISATIONS	master.azure.infa.world	0	0	00:00:45	Running	Select log
INITIAL_LOAD_AUSPOST_CD	N/A	N/A	0	0	00:00:45	Running	Select log

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

New Asset

Select the type of asset you want. Some asset types include templates for common integration patterns.

- Tasks**
- Mappings**
- Mapplets**
- Taskflows**
- Components**

- Replicate a SQL Server table or view incrementally to a new target**
Extract data incrementally from a SQL Server database table or view based on its timestamp and load it to a dynamically created target.
- Replicate an Oracle table or view incrementally to a new target**
Extract data incrementally from an Oracle database table or view based on its timestamp and load it to a dynamically created target.
- Replicate an SAP table incrementally into a new target**
Extract data incrementally from an SAP table or view based on timestamp fields (for example, VBAK.ERDAT and VBAK.ERZET) and load it to a dynamically created target.

Informatica Data Integration

New... | Mapping3 | Valid

Home | Explore | My Jobs | Profile_Customer... | Mapping3 | Mapping3-Advan...

Design

Source | Target | Access Policy | Aggregator | B2B | Cleanse | Data Masking

Source → exp_extract_service_nam → DataServices → ttt_output, ttt_errors, ttt_errors_found

Mapping3

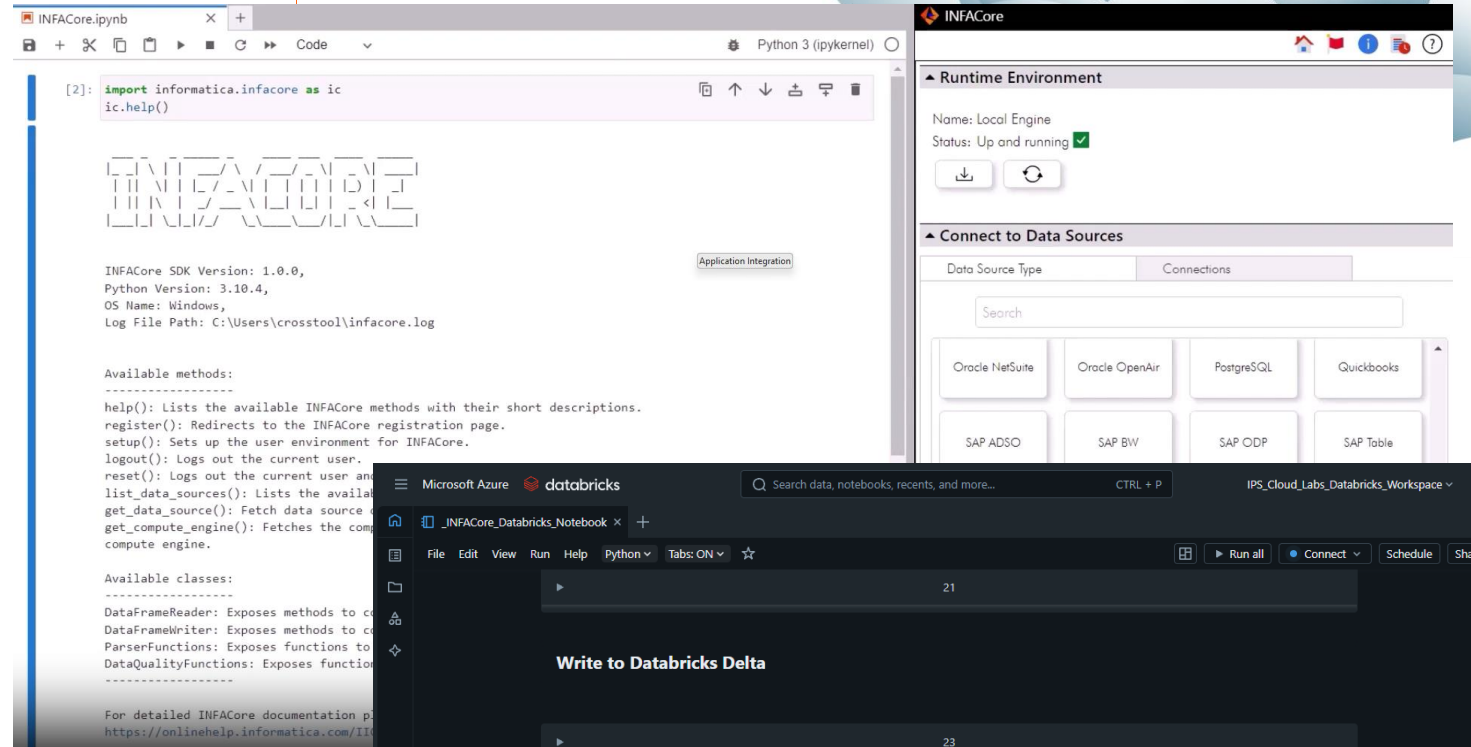
Name* | Process_HL7_to_Kafka

INFACore

Low-Code Intelligent Data Management

AI-Powered Automation with CLAIRE Engine

Seamless Integration via SDKs and Extensions



```
INFACore.ipynb Python 3 (ipykernel) Code
```

```
[2]: import informatica.infacore as ic
ic.help()
```

INFACore SDK Version: 1.0.0,
Python Version: 3.10.4,
OS Name: Windows,
Log File Path: C:\Users\crosstool\infacore.log

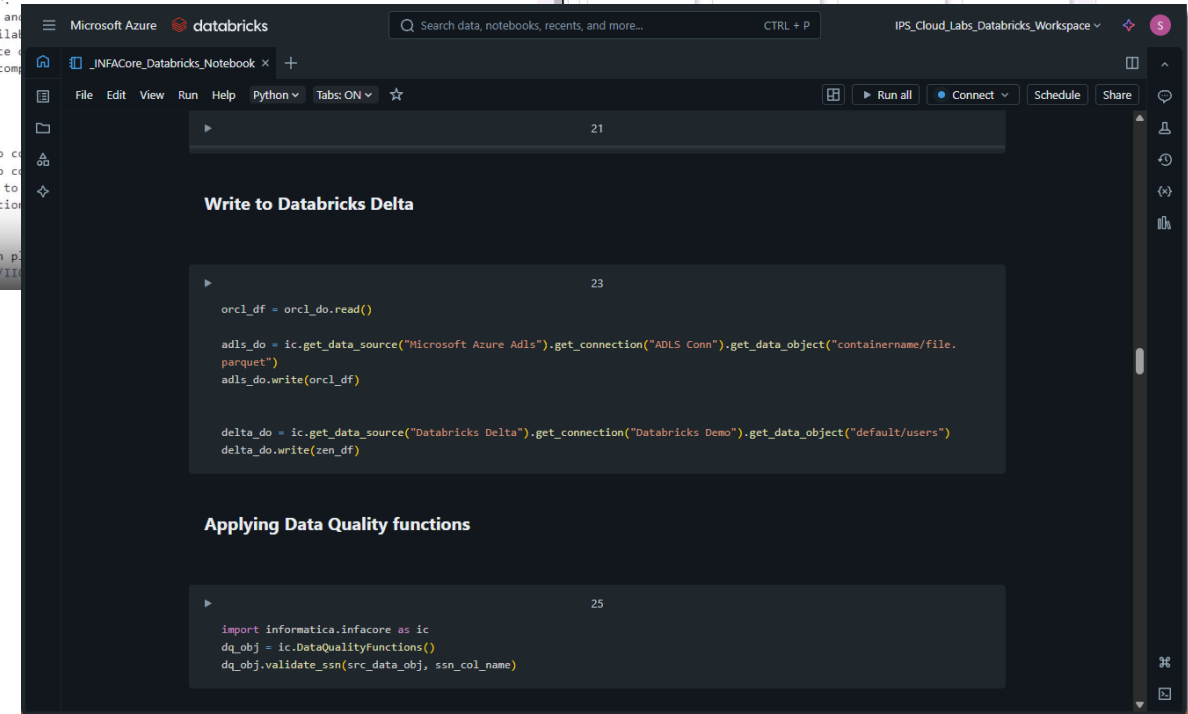
Available methods:

```
help(): Lists the available INFACore methods with their short descriptions.
register(): Redirects to the INFACore registration page.
setup(): Sets up the user environment for INFACore.
logout(): Logs out the current user.
reset(): Logs out the current user and sets up the user environment.
list_data_sources(): Lists the available data sources.
get_data_source(): Fetches data source details.
get_compute_engine(): Fetches the compute engine details.
```

Available classes:

```
DataFrameReader: Exposes methods to read data from various sources.
DataFrameWriter: Exposes methods to write data to various destinations.
ParserFunctions: Exposes functions to parse and transform data.
DataQualityFunctions: Exposes functions to perform data quality checks.
```

For detailed INFACore documentation please visit: <https://onlinehelp.informatica.com/II>



```
Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P IPS_Cloud_Labs_Databricks_Workspace
```

```
File Edit View Run Help Python Tabs: ON ☆ Run all Connect Schedule Share
```

```
21
```

```
Write to Databricks Delta
```

```
23
```

```
orc1_df = orc1_do.read()
adls_do = ic.get_data_source("Microsoft Azure Adls").get_connection("ADLS Conn").get_data_object("containername/file.parquet")
adls_do.write(orc1_df)

delta_do = ic.get_data_source("Databricks Delta").get_connection("Databricks Demo").get_data_object("default/users")
delta_do.write(zen_df)
```

```
Applying Data Quality functions
```

```
25
```

```
import informatica.infacore as ic
dq_obj = ic.DataQualityFunctions()
dq_obj.validate_ssn(src_data_obj, ssn_col_name)
```

Application Integration

Enables automation of complex business processes and user workflows

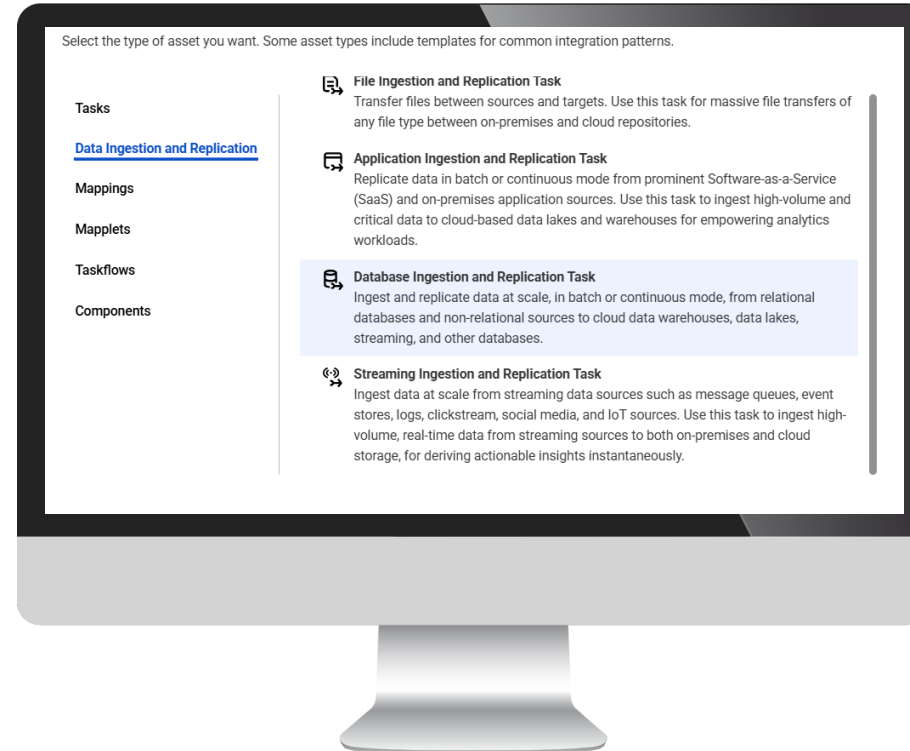
Easy integration to REST, SOAP APIs and others

Recipes available as pre-configured sets of assets

The screenshot displays the Informatica Integration Cloud interface. The top section shows a workflow design for 'Sirene Search_V2', which is marked as 'Valid'. The workflow starts with a 'Start' node, followed by an 'Evaluate length SIREN and SIRET' task, then a 'Length SIRET' decision diamond. The flow continues through 'FindByLegalNameAndCity' and 'HTTP Code' tasks, leading to another 'Length SIRET' decision diamond. From there, it branches into three paths based on 'Contains TRUE', 'Contains NEUTRAL', and 'No match found' conditions, each leading to different 'FindBy' tasks and 'HTTP Code' tasks, eventually ending at 'End 1', 'End 2', and 'End 3' nodes respectively. The bottom section, titled 'Start with Recipes', shows a list of available recipes under the 'databricks' filter. The recipes include: 'Databricks Mosaic AI Chat with File', 'Extract Structured JSON Data from Unstructured Data using Databricks Mosaic AI', and 'Patient Food Vendor Recommendation using Databricks Mosaic AI'. Each recipe card provides a brief description and tags.

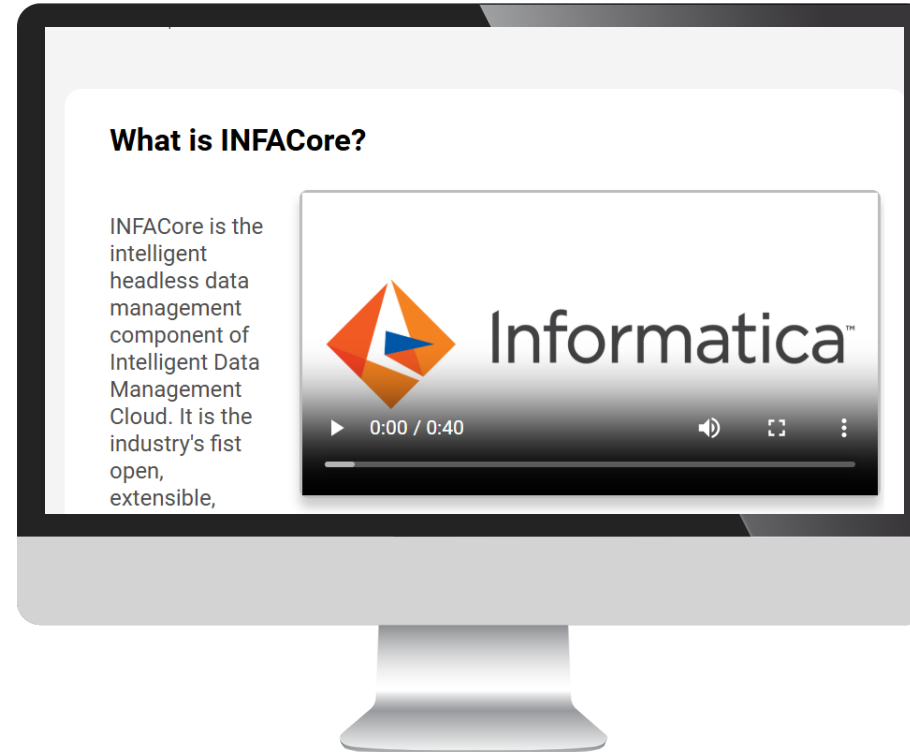
DEMO

Database Ingestion and replication - Databricks as Destination



DEMO

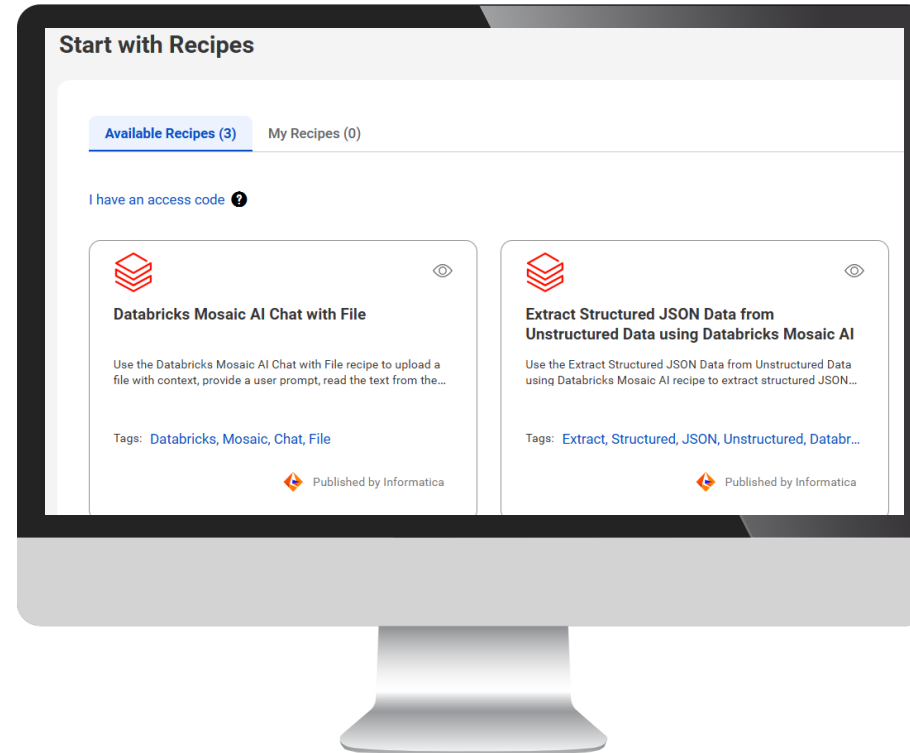
Databricks Integration with INFACore



Where data & AI come to 

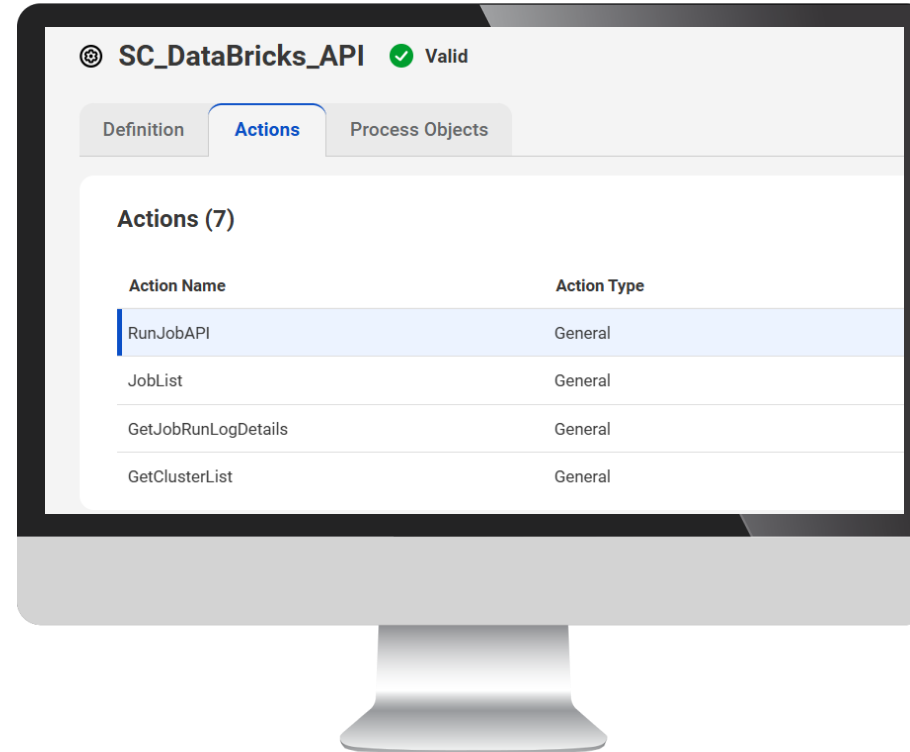
DEMO

Databricks Mosaic AI Recipes – Cloud Application Integration



DEMO

Connect to Databricks Rest API using CAI
Service Connector



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

Profile_Customers_Landing
Profile run 1 of 1 | 14 of 14 Columns | 15 of 15 Rules | 753 Rows (All rows) | Aug 3, 2023, 12:46 AM

Results | Definition | Rules | Metrics | Schedule | Insights

View: Columns and Rules | with: All Statistics

Columns	Value Distribution	% Null	# Null	% Distinct	# Distinct	% Non-di...	# Non-di...	# Patterns	% of Top ...	Minimu...	Maxim...
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

Column: STATUS

5 distinct values (5 non-unique values, 0 unique values)

Sort By: Frequency

#	Value	Frequency	Percentage	Length
<input type="checkbox"/> 1	LIVE	546	72.51%	4
<input type="checkbox"/> 2	A	82	10.89%	1
<input type="checkbox"/> 3	Inactive	77	10.23%	9
<input type="checkbox"/> 4	I	28	3.72%	1
<input type="checkbox"/> 5	Active	18	2.39%	6
<input type="checkbox"/> 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

#	Step	Options
1	Replace Values	Gender_Dictionary

Step Properties: Replace Values

Mode: Replace Input Values with Dictionary Values

Dictionary: Gender_Dictionary

Valid Column: Column 1

Casing: Case Insensitive

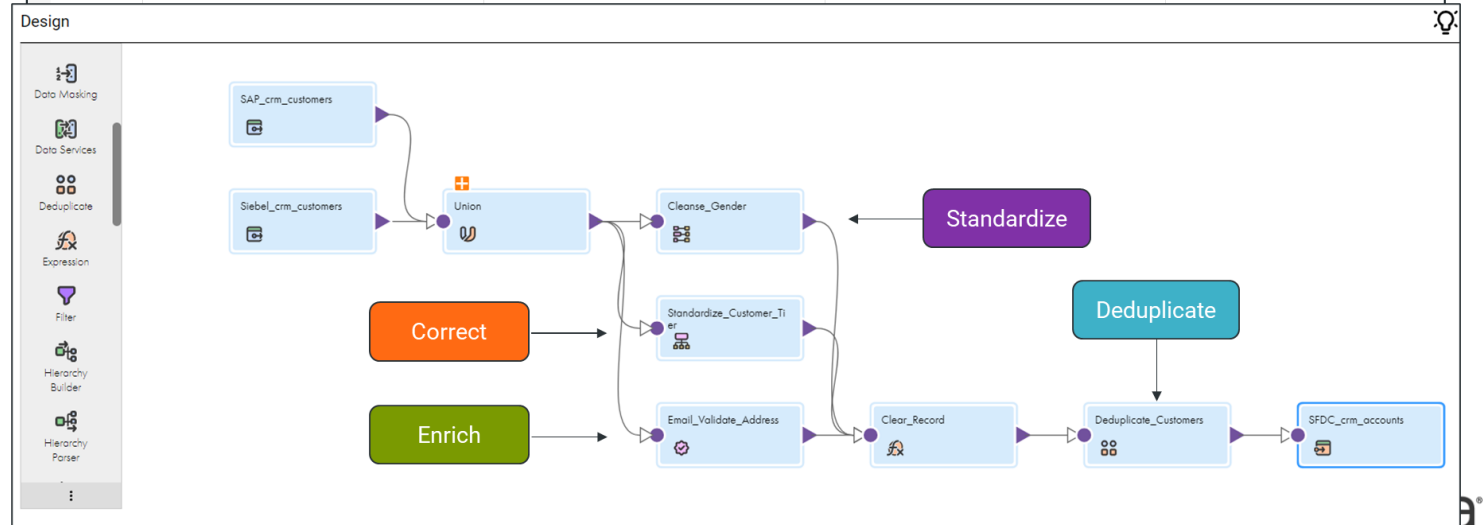
Delimiter: Space

Scope: Anywhere

Dictionary Preview: Gender_Dictionary

Column 1	Column 2	Column 3	Column 4
M	Male	Men	Men

	Column 1	Column 2	Column 3	Column 4
1	M	Male	Man	Men
2	F	Female	Woman	Women
3	U	Unknown	Unk	N/A
4	F	Femall		

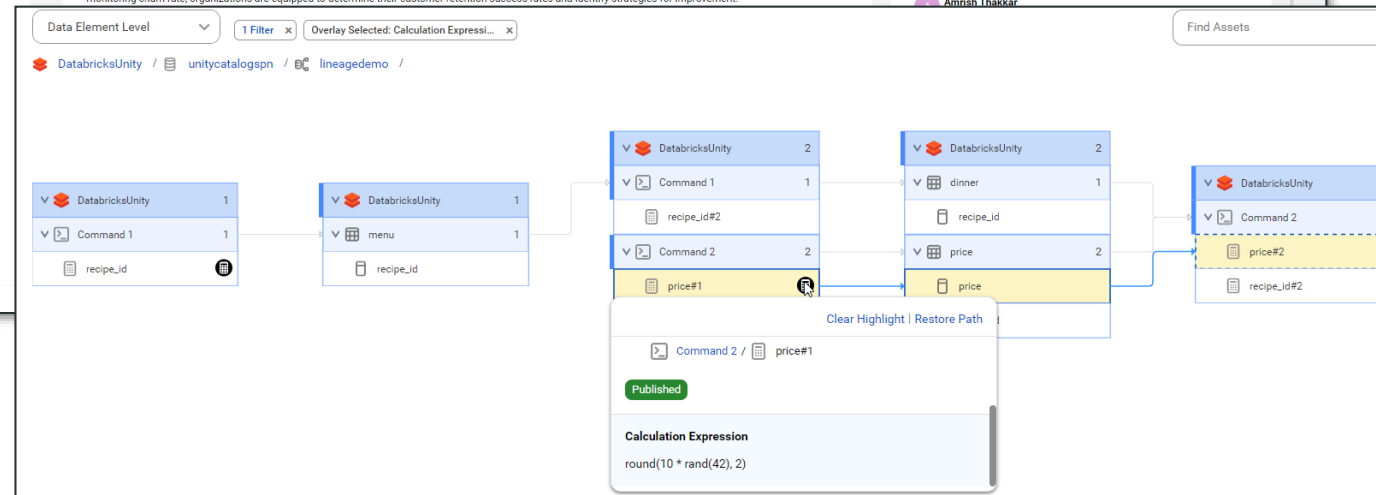
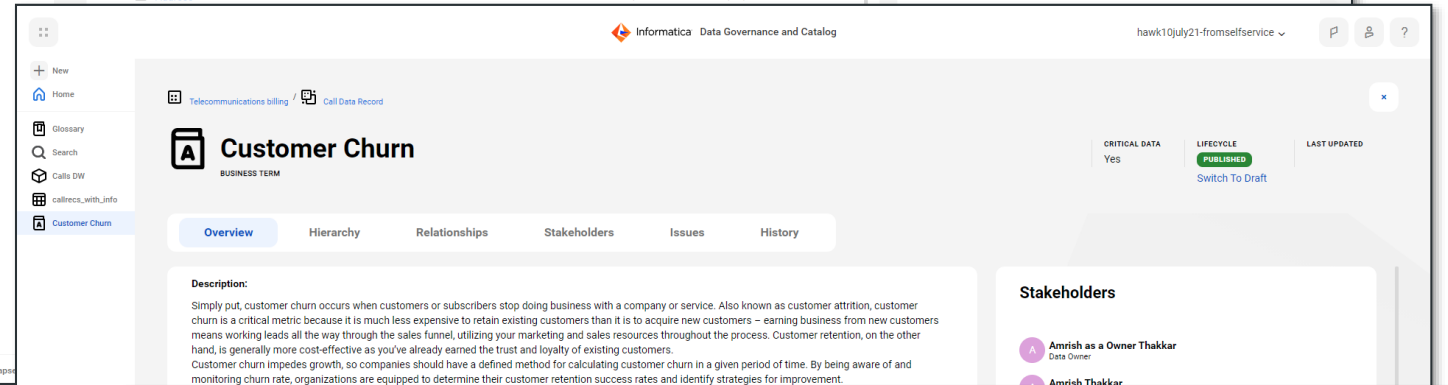
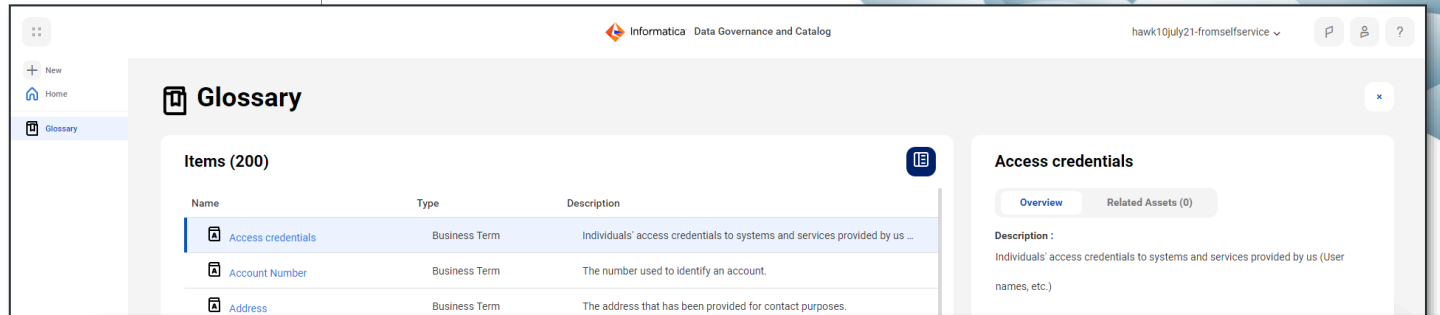


Data Governance and Catalog

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



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

The screenshot displays the Informatica Data Marketplace interface. At the top, a progress bar shows three steps: 1. Find Data Assets (checked), 2. Prepare and Create (active), and 3. Summary. Below this, the 'Prepare Data Assets' section includes a filter for status (Ready: 1, Error: 0) and a table with columns for Name, Description, Status, and Data Source. A table entry shows 'Sales Order Details' with a description 'This is a sales order details data set', status 'Enabled', and data source 'Snowflake'. Below the table, a 'Data Marketplace' dashboard shows a welcome message and four data collection cards: 'Expense Data', 'Sensitive Employee Info', '2022 Reports', and 'EJ Collection'. Each card displays request and consumer counts. The bottom section is a 'Checkout' page with a 'Usage Guide' containing four sections: 'GDPR & PRIVACY CODE', 'GDPR & PRIVACY CODE > PRIVACY COMPLIANCE', 'EMPLOYEE BEHAVIORAL CHARTER', and another 'GDPR & PRIVACY CODE' section. Each section has a text block and an 'I accept' checkbox. To the right is an 'ORDER SUMMARY' section with fields for REF. DC 001, Fleet DOE Fleet Product, PURPOSE, DATA OWNERS (Jane Smith, Mohinder Panjal), TECHNICAL OWNERS (AXON Admin), CATEGORY (Product > Vehicle > Fleet > ... > Camry), DELIVERY, DELIVERY REQUESTS (Can I get this in an excel file?), and BUSINESS JUSTIFICATION (Need access to the financial information for the end of year report).

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

cdam_FIRSTNAME
DATA ELEMENT CLASSIFICATION

Classification Rule
NAME LIKE '%FIRSTNAME%'

Associations
6 CATALOG SOURCES | 230 TOTAL ASSOCIATIONS

CLAIRE Data Classification

CDAM Access Control Policy

Rule Names	Conditions	Transformations
Ofuscate_PersonName	Trigger this rule if: → User Group is any of <input type="text" value="Data_Consumers"/>	Assign all of the following transformations to data classes: → Replace <input type="text" value="T cdam_BUSINESS_ID"/> with regex → Replace <input type="text" value="T cdam_FIRSTNAME"/> with consistent regex → Replace <input type="text" value="T cdam_LAST_NAME"/> with *****

User with Privileges

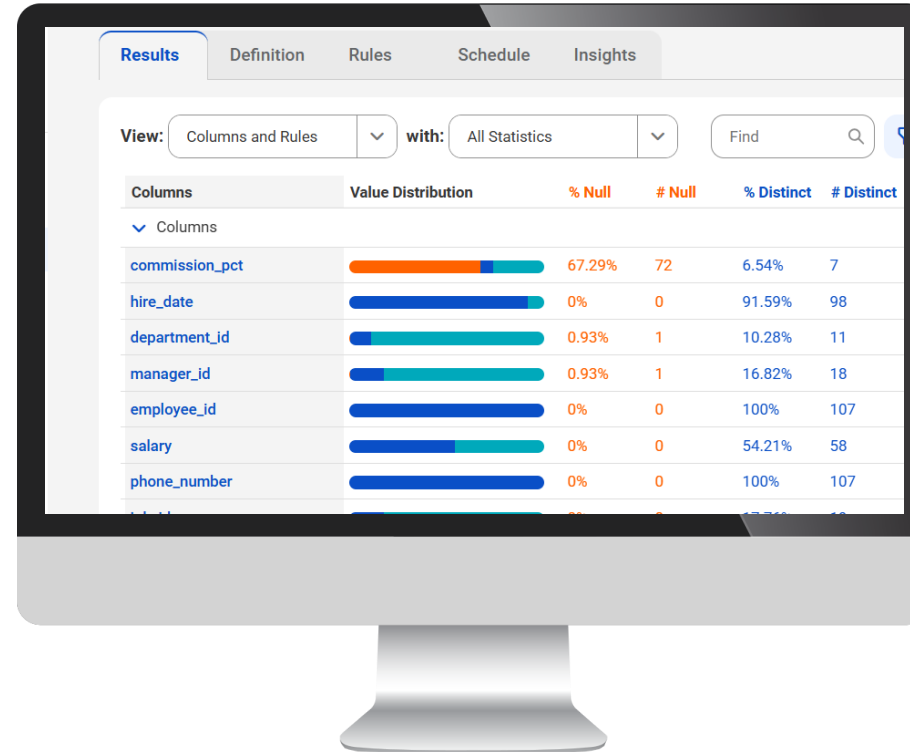
ABC FIRST_NAME	ABC SECOND_NAME	ABC EMPLOYEE_NUMBER	ABC COUNTRY	ABC JOB_TITLE
Salvador	Meyer	84047	Ireland	Legal Administrator
Sara	Townsend	55686	France	Network Administrator
Alexis	Deleon	55801	Ireland	Compliance Officer
Gabrielle	Cuevas	13962	Germany	Media Buyer
Brecken	Berg	95088	Germany	Electrical Engineer
Emmalyn	Benjamin	88647	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	Marketing Research Analyst
Osman	Camacho	72220	United Kingdom	Compliance Officer

User without Privileges

ABC FIRST_NAME	ABC SECOND_NAME	ABC EMPLOYEE_NUMBER	ABC COUNTRY	ABC JOB_TITLE
Jxkbladhinfv	*****	936042276038	Germany	Court Reporter
Xwlqscrfcspj	*****	444607627894	United Kingdom	Marketing Research Analyst
Pkvvxbrtcain	*****	5484484031640	Germany	Project Manager
Msduoqicgext	*****	6872587489830	United States	Copywriter
Xqmzuogzjemn	*****	2201799735622	Germany	Cloud Solutions Architect
Lmonphcbeepzc	*****	7088887775853	Germany	Quality Control Analyst
Zrvxqmyvaadga	*****	3309725977403	Germany	Marketing Research Analyst
Ypmxmpukmbad	*****	332215463410	France	Web Developer
Bpwjiqcyvi	*****	569455264405	France	Cybersecurity Analyst
Bvexvmtobow	*****	0122095421122	United Kingdom	Software Developer
Nuzzrodlezafj	*****	3854298161402	Germany	Media Buyer
Xpiczswgxrnrl	*****	6649559732424	United Kingdom	Civil Engineer

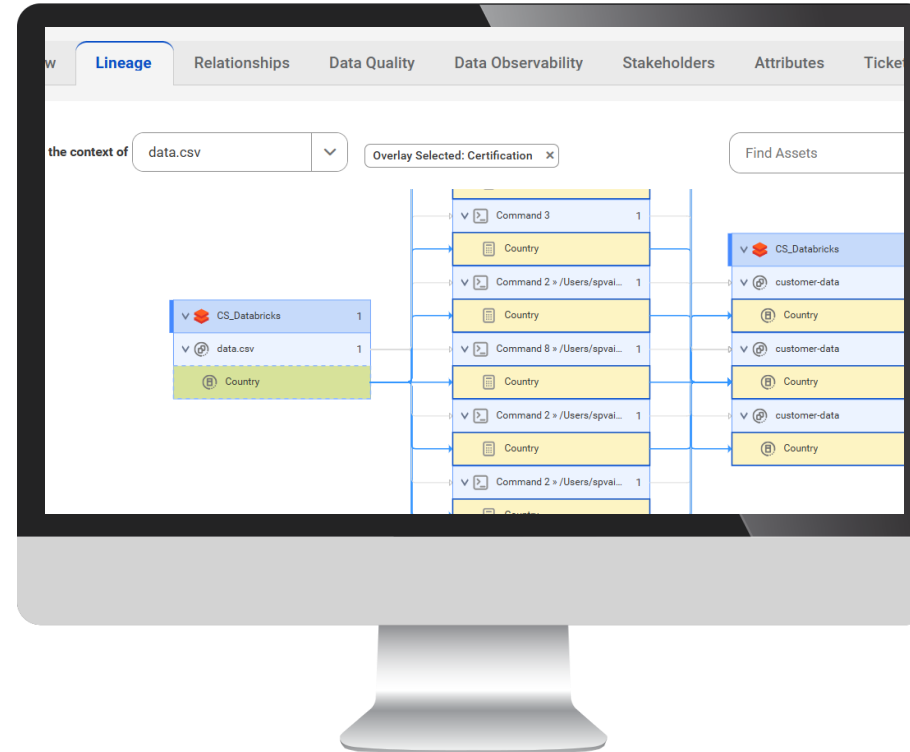
DEMO

Data Profiling on Databricks Connection Object



DEMO

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



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

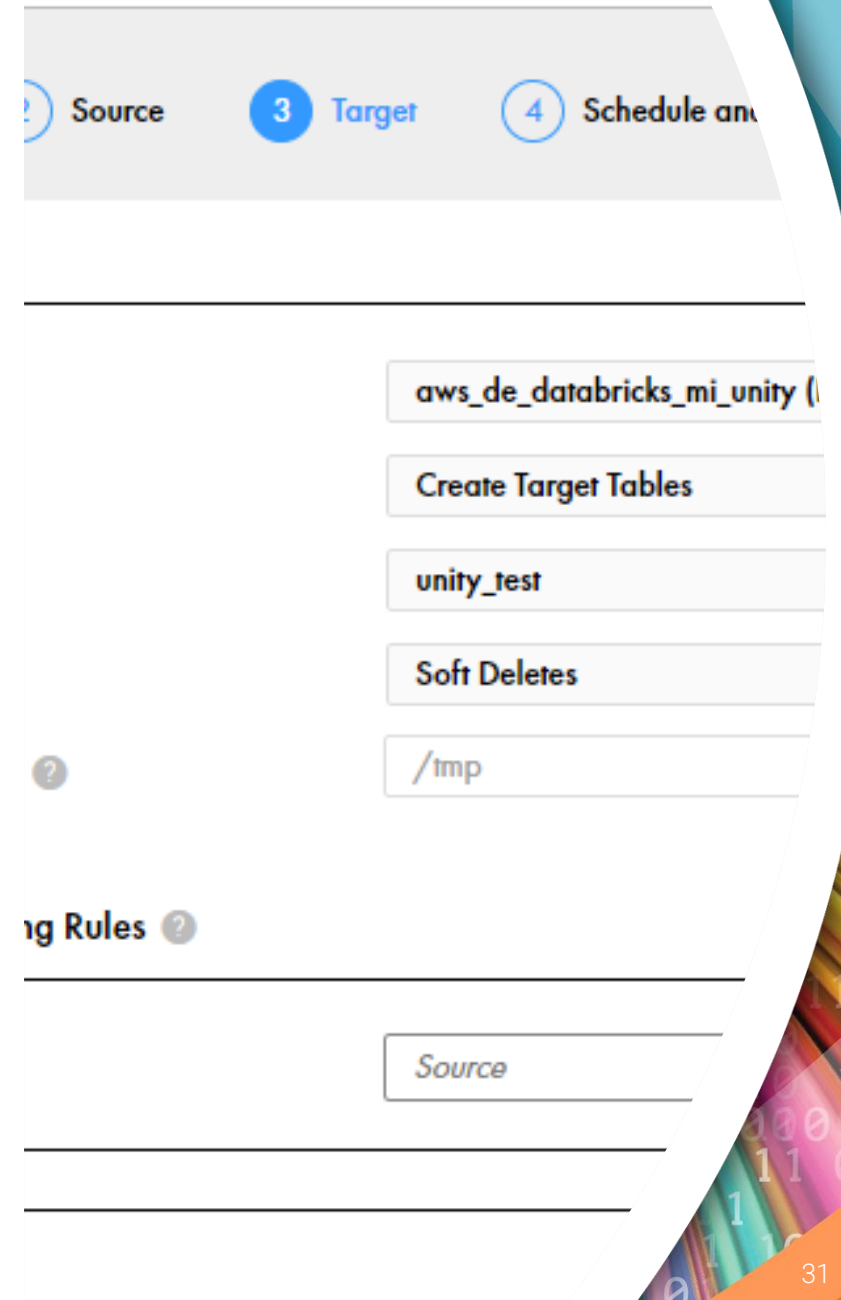
Supported Function Category	Examples (not exhaustive)
Aggregate	AVG(), SUM(), COUNT() and specialized functions like APPROX_COUNT_DISTINCT() and REGR_SLOPE()
AI	AI_ANALYZE_SENTIMENT(), AI_TRANSLATE(), and AI_SUMMARIZE()
Cast	Conversion-related operations such as TO_CHAR(), TO_NUMBER(), and TRY_TO_NUMBER()
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

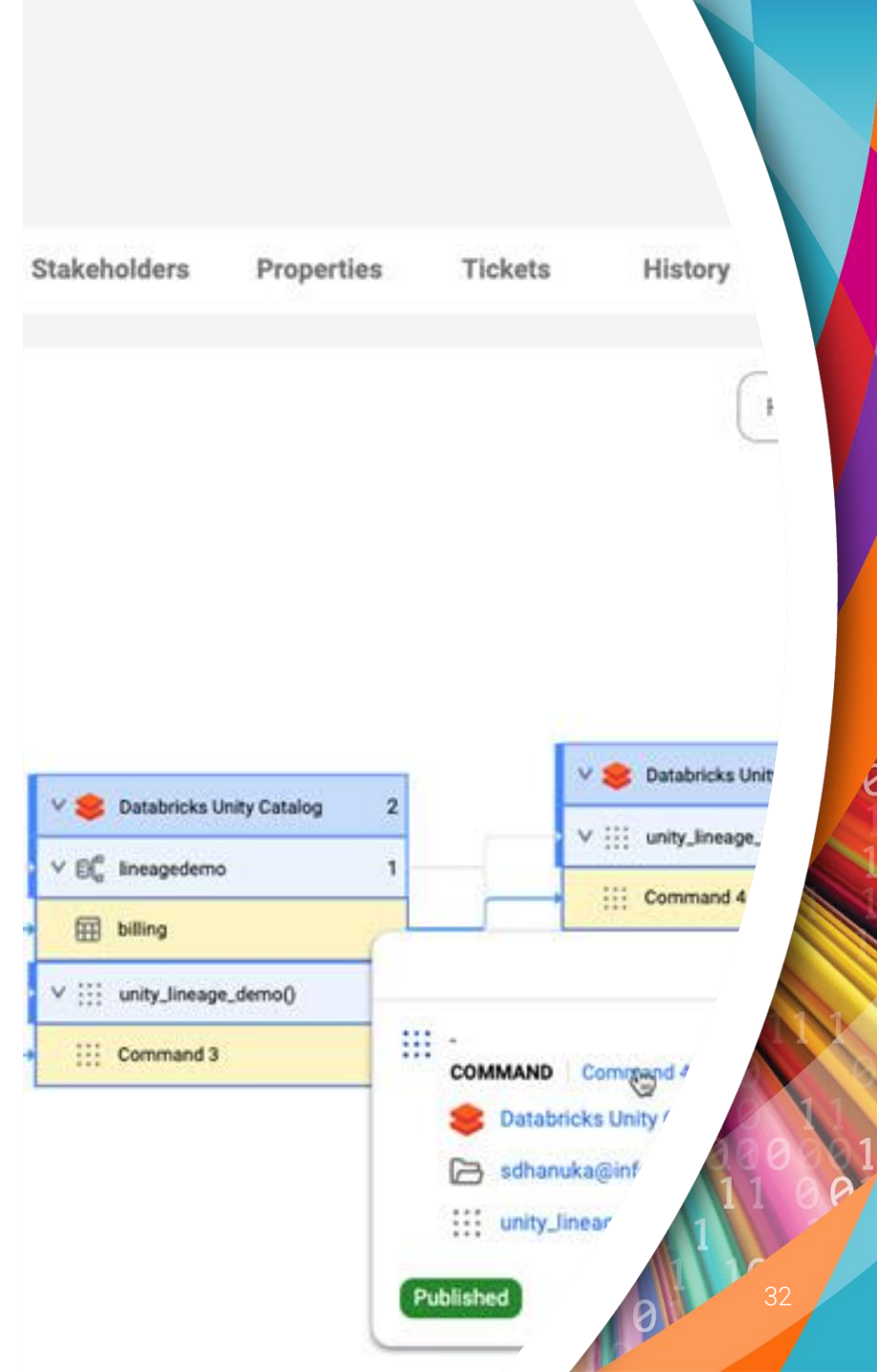
- You can only access Databricks tables created on top of S3 or ADLS Gen2
- Data ingestion jobs wait for 10 minutes for DBX cluster to be available. After that time, connection will time out and job will fail
 - This 10 minutes timeout value can be customized
- By default, Database Ingestion and Replication 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

Top Level

Read Access for HR Analysts.

DATA ACCESS CONTROL

Overview

Rules

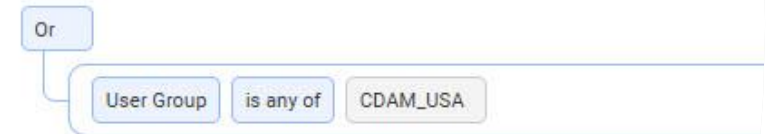
Relationships

Stakeholders

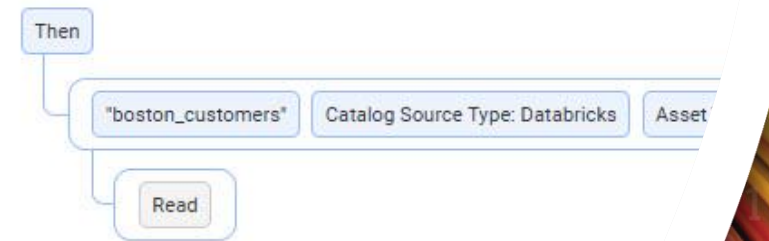
Rules (1)

Read Access for HR Analysts

Conditions



Access Controls



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

