Informatica

23 Sep, 2025

Automate CDGC Custom Data Lineage Creation Process

- Giri Chintarlapalli, Senior Principal Consultant, IPS
- Rajesh Chillakuru, Principal Consultant, IPS

Where data & AI come to LIFE

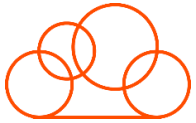
© Informatica. Proprietary and Confidential.

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.

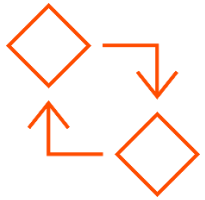
© Informatica. Proprietary and Confidential.

Feature Rich Success Portal



Bootstrap trial and POC Customers

Enriched Customer Onboarding experience



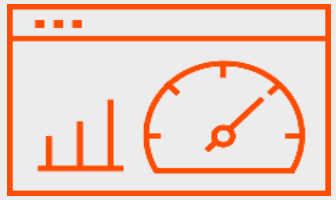
Product Learning Paths and Weekly Expert Sessions

Informatica Concierge

Tailored training and content recommendations

© Informatica. Proprietary and Confidential.

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

© Informatica. Proprietary and Confidential.

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.

© Informatica. Proprietary and Confidential.

Informatica

Introduction: API Lineage Automation Approaches

Where data & AI come to LIFE

© Informatica. Proprietary and Confidential.

Agenda

Introduction: Needs and Use Cases for API Lineage

Automating API Lineage in CDGC

Benefits and Future Directions

Informatica

© Informatica. Proprietary and Confidential.

Why API Lineage Matters

Data Traceability: Track the flow of data through APIs for transparency

Accountability: Identify origin and transformations of data for audit purposes

Impact Analysis: Understand downstream effects of API changes

Compliance: Ensure data handling meets regulatory requirements

Error Diagnosis: Quickly isolate and resolve issues across API chains

Security: Monitor data access and detect unauthorized modifications

Data Quality: Maintain integrity by tracking data transformations

Change Management: Manage API versions and dependencies effectively

Operational Efficiency: Provide visibility to optimize API ecosystem

Risk Management: Identify and mitigate data flow risks

© Informatica. Proprietary and Confidential.

Key Use Cases for API Lineage

- Tracking how data moves through multiple APIs and services.
- Identifying upstream and downstream data dependencies.
- Supporting migration and modernization projects.
- Facilitating data privacy and access controls.
- Improving collaboration between development and governance teams.

© Informatica. Proprietary and Confidential.

Challenges Without API Lineage Automation

- Manual documentation is error-prone and quickly outdated.
- Limited visibility into real-time data flows.
- Difficulty in identifying the root of data quality issues.
- Increased compliance and audit risks.
- Slower response to regulatory changes.

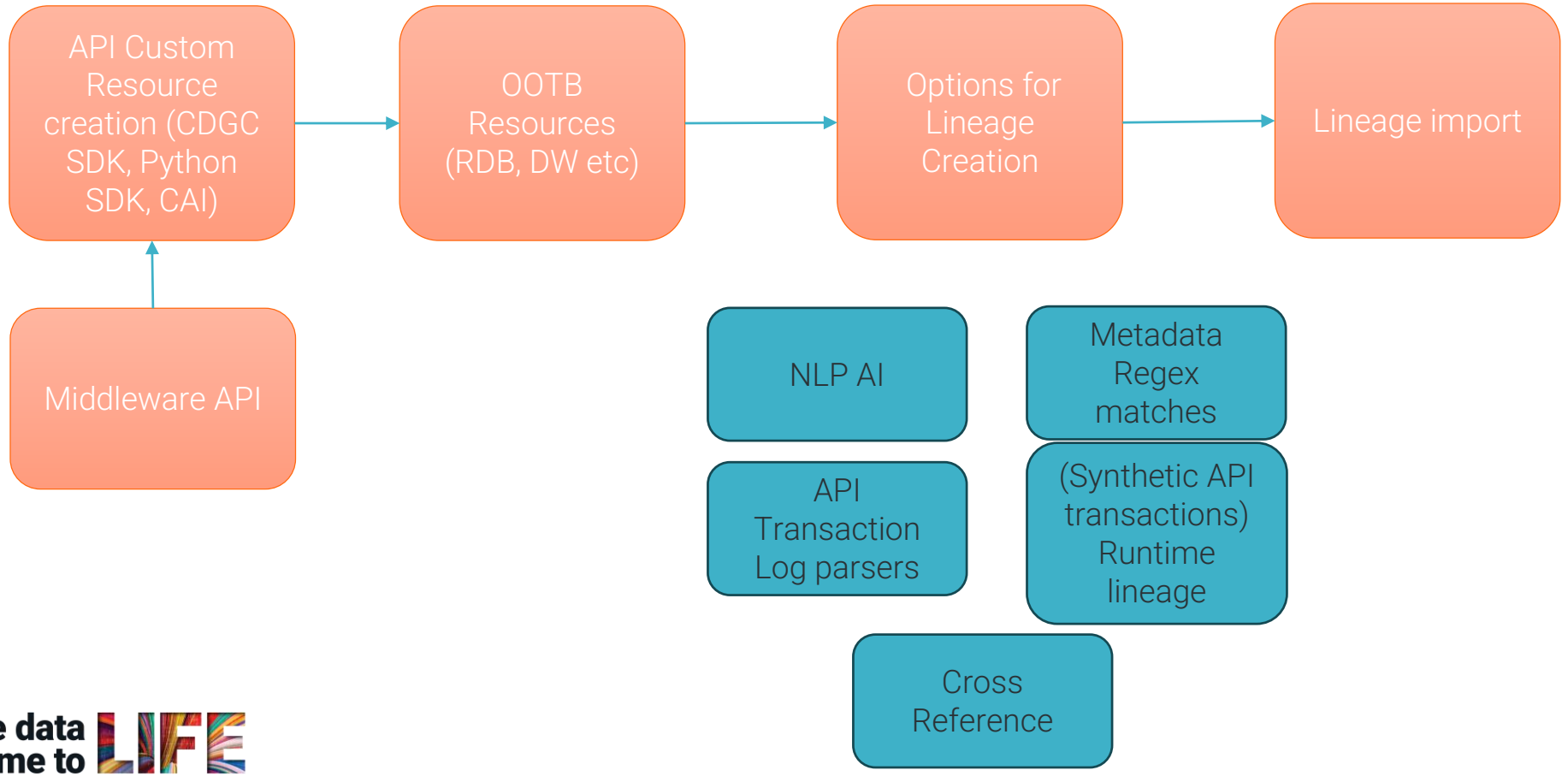
© Informatica. Proprietary and Confidential.

Informatica

Automating API Lineage in CDGC

Where data & AI come to LIFE

© Informatica. Proprietary and Confidential.

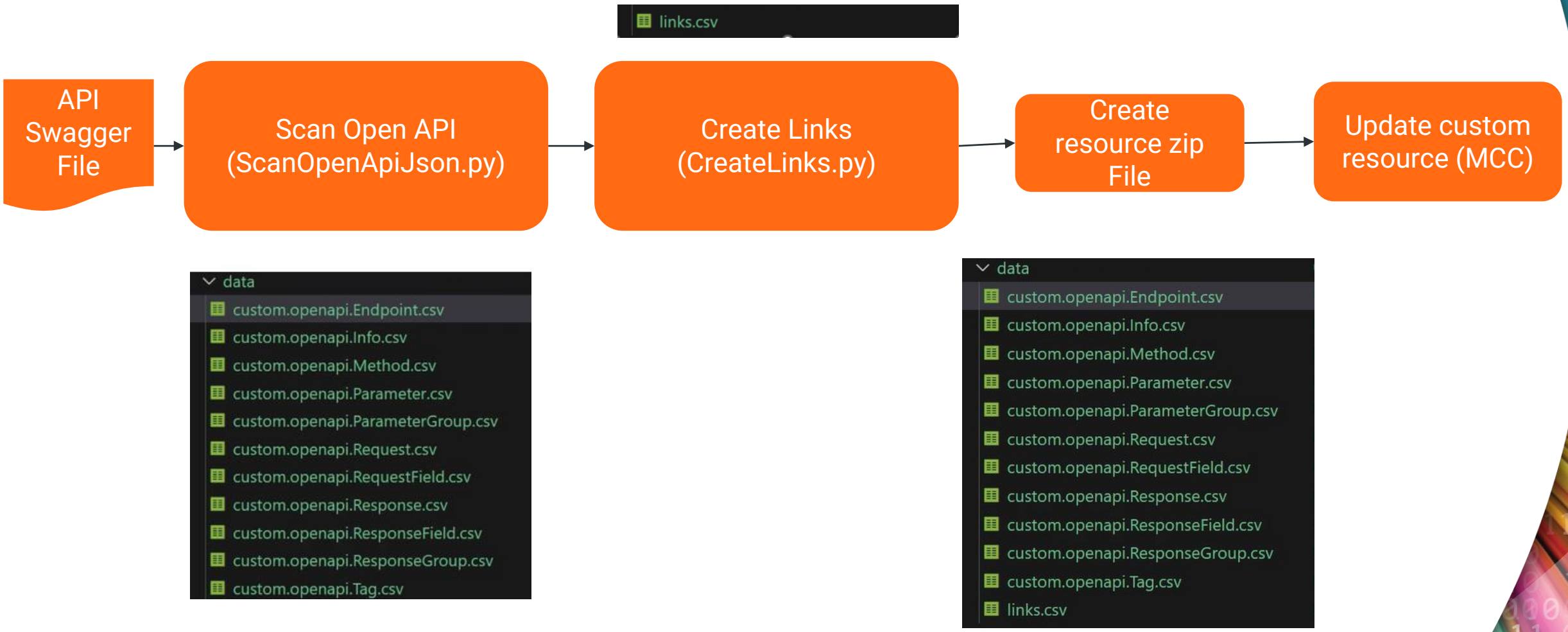

API Custom Resource creation (CDGC SDK, Python SDK, CAI)
OOTB Resources (RDB, DW etc)
Options for Lineage Creation
Lineage import
Middleware API
NLP AI
Metadata Regex matches
API Transaction Log parsers
(Synthetic API transactions) Runtime lineage
Cross Reference

API Lineage Integration – Process flow

Regex-Based Data Lineage Mapping	metadata-defined regular expression patterns to match and map Swagger API fields with database fields
AI NLP Tools	Create base lineage by mapping Swagger file attributes to database resources
Runtime Lineage Transaction Logs	Capture unique transaction IDs, payloads, and responses during API execution for precise data flow tracing
Synthetic payload generation	Create synthetic but realistic API payloads covering all fields defined in the Swagger API
Manual Cross-Reference Validation	Manual review and maintain a curated mapping between API fields and database columns to ensure accuracy

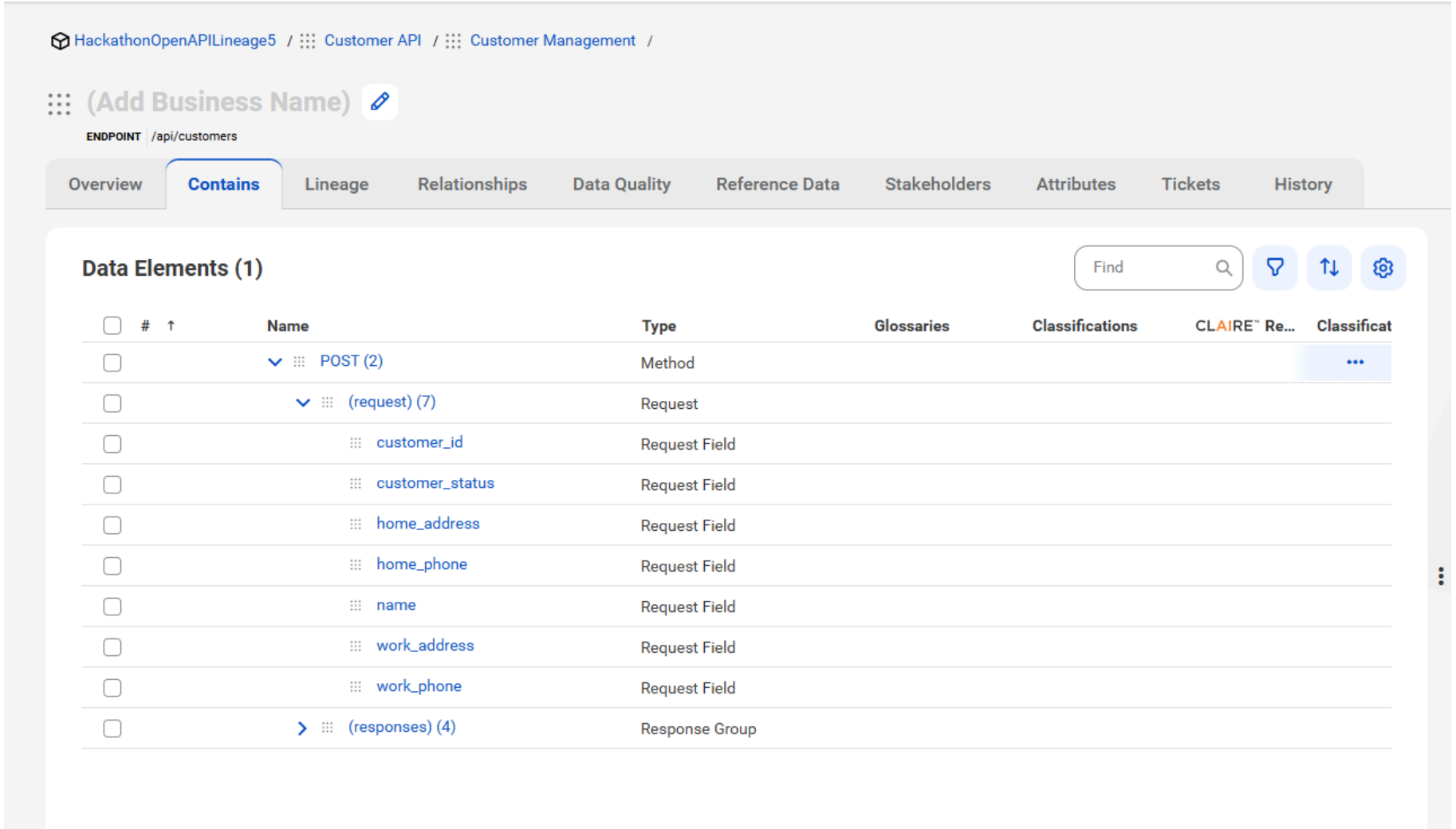
© Informatica. Proprietary and Confidential.

CDGC – API Custom resources



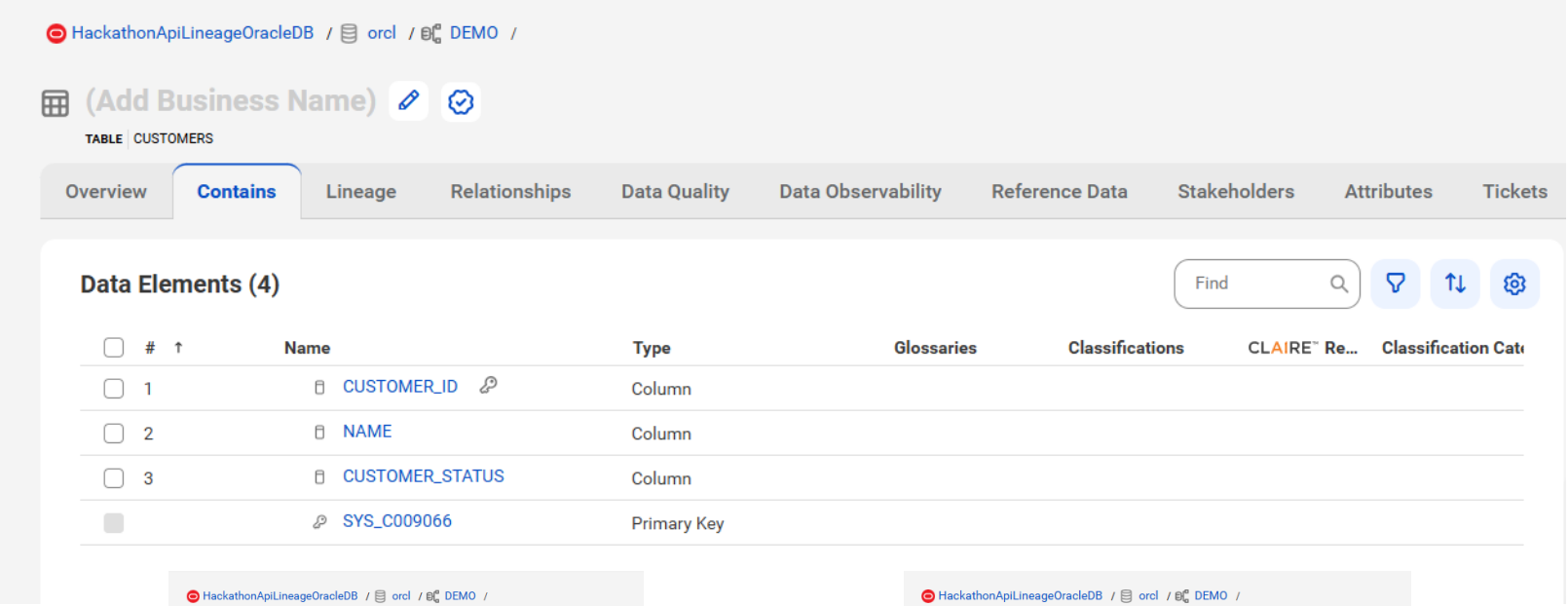
© Informatica. Proprietary and Confidential.

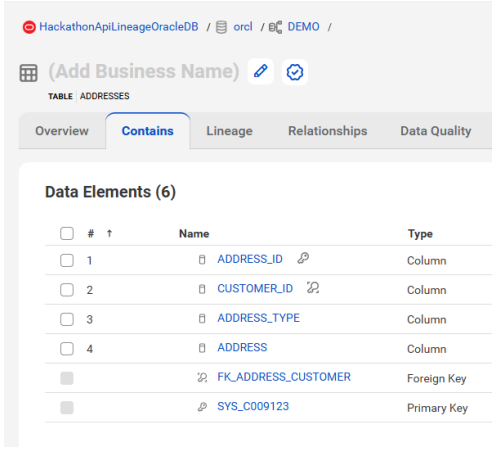
CDGC – API Custom resources (Continued)

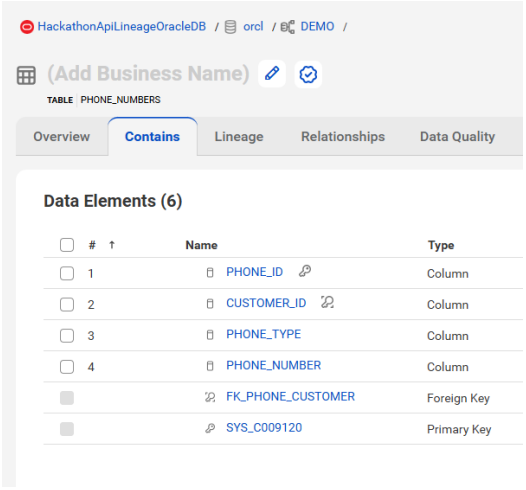


© Informatica. Proprietary and Confidential.

CDGC – Database resources

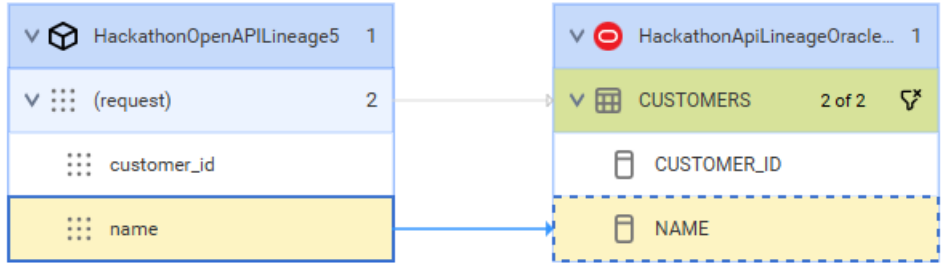


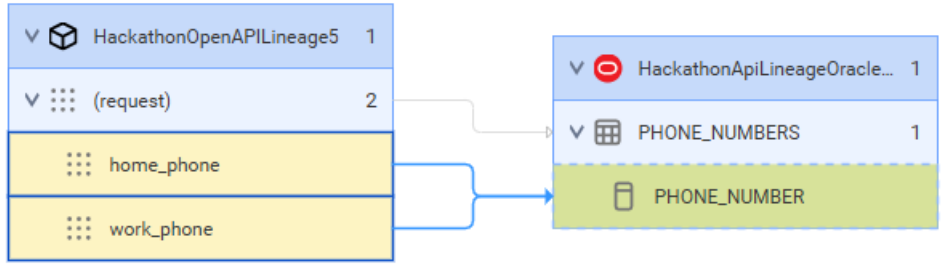




© Informatica. Proprietary and Confidential.

API to Database lineage





© Informatica. Proprietary and Confidential.

Regex-Based Data Lineage Mapping

Extract API resource metadata (fields/attributes) from CDGC

Extract DB resources metadata (columns) from CDGC

Process API fields through regex matcher

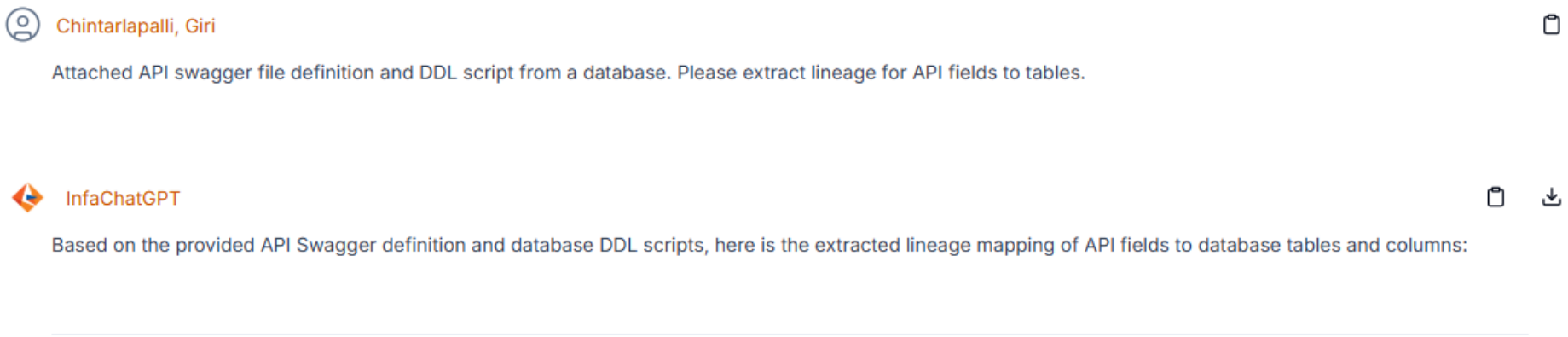
Process DB resources metadata through regex matcher

Create lineage links file

© Informatica. Proprietary and Confidential.

Azure Open AI LLM – Inferring Lineage

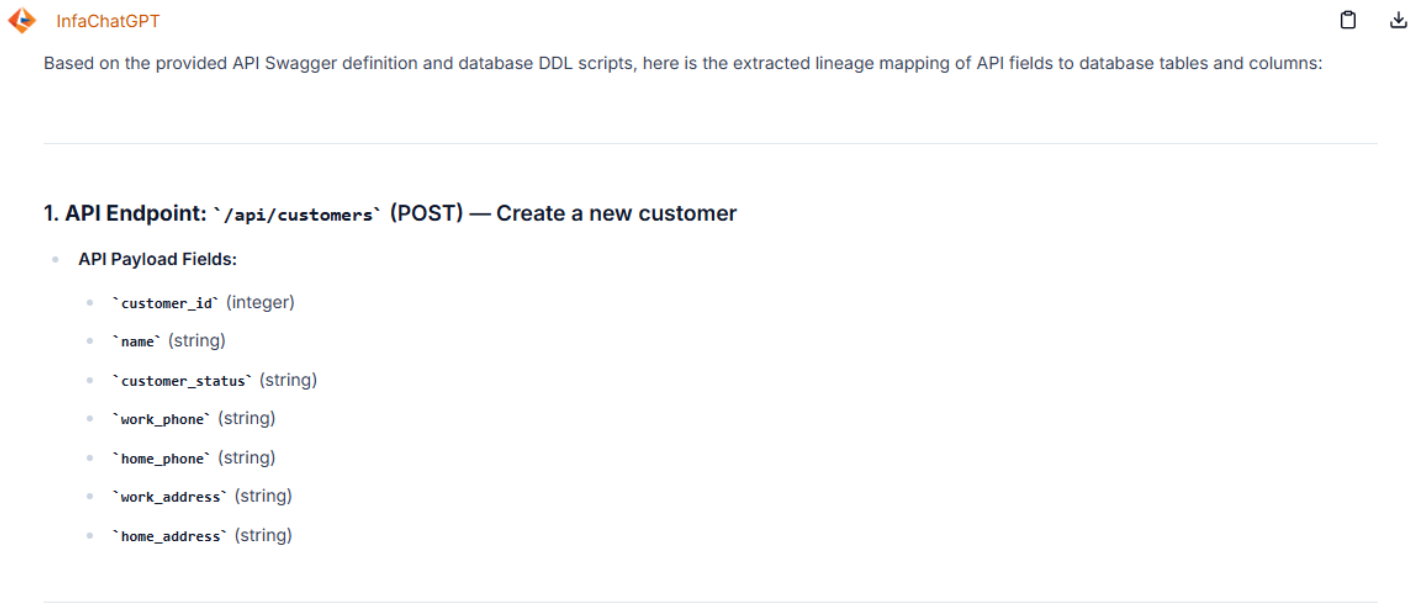
- Input:
  - Swagger file
  - Database DDL
- Output – Startup lineage
- Sample Input prompt



© Informatica. Proprietary and Confidential.

Azure Open AI Extracting Lineage - Continued

- LLM Output



© Informatica. Proprietary and Confidential.

Azure Open AI Extracting Lineage - Continued

- LLM Output

2. Database Tables and Columns:

Table: CUSTOMERS

Column	Mapped API Field(s)	Notes
CUSTOMER_ID	`customer_id`	Primary key for customer
NAME	`name`	Customer name
CUSTOMER_STATUS	`customer_status`	Status of customer

Table: PHONE_NUMBERS

Column	Mapped API Field(s)	Notes
CUSTOMER_ID	`customer_id`	Foreign key to CUSTOMERS
PHONE_TYPE	derived from API key	Set to "work" for `work_phone`, "home" for `home_phone`
PHONE_NUMBER	`work_phone`, `home_phone`	Phone numbers for work and home phones

Table: ADDRESSES

Column	Mapped API Field(s)	Notes
CUSTOMER_ID	`customer_id`	Foreign key to CUSTOMERS
ADDRESS_TYPE	derived from API key	Set to "work" for `work_address`, "home" for `home_address`
ADDRESS	`work_address`, `home_address`	Customer's work and home addresses

© Informatica. Proprietary and Confidential.

Azure Open AI Extracting Lineage - Continued

- LLM Output

3. Field-Level Lineage Summary:

API Field	Target Table	Target Column	Transformation / Mapping
`customer_id`	CUSTOMERS	CUSTOMER_ID	Direct mapping, primary key
`name`	CUSTOMERS	NAME	Direct mapping
`customer_status`	CUSTOMERS	CUSTOMER_STATUS	Direct mapping
`work_phone`	PHONE_NUMBERS	PHONE_NUMBER	Mapped to phone_number where phone_type = 'work'
`home_phone`	PHONE_NUMBERS	PHONE_NUMBER	Mapped to phone_number where phone_type = 'home'
`work_address`	ADDRESSES	ADDRESS	Mapped to address where address_type = 'work'
`home_address`	ADDRESSES	ADDRESS	Mapped to address where address_type = 'home'

4. Additional Notes:

- The middleware normalizes the payload by splitting denormalized fields such as `work_phone` and `home_phone` into multiple rows in the `phone_numbers` table, setting `phone_type` appropriately.
- Similarly, `work_address` and `home_address` are split into the `addresses` table with corresponding `address_type`.
- Foreign key constraints ensure data integrity linking `phone_numbers` and `addresses` to `customers` via `customer_id`.

This lineage mapping connects each API field to the corresponding physical table columns, showing how a single API payload results in inserts across three normalized tables in the database.

Hackathon_Lineage_swagger_v20250909.docx

1

tables_definition.docx

1

© Informatica. Proprietary and Confidential.

Azure Open AI Extracting Lineage – with API

- LLM Payload


```
{
  "messages": [
    {
      "role": "system",
      "content": "You are an assistant that extracts API to database field lineage from given documents."
    },
    {
      "role": "user",
      "content": "### Swagger API Definition ###\n{\n\"openapi\": \"3.0.0\",\n\"info\": {\n\"title\": \"Customer API\",\n\"version\": \"1.0.0\",\n\"description\": \"API for managing customers with create, nc
    },
    {
      "role": "user",
      "content": "Based on the above, please extract a summary of the data lineage between API fields and database tables/columns."
    }
  ],
  "max_tokens": 500,
  "temperature": 0
}

```


© Informatica. Proprietary and Confidential.

API Transaction log-based approach

- Retrieve Payload from API Transaction Log
- Parse and Normalize Payload
- Identify Database Tables and Columns
- Compare Payload Fields with Database Schema
- Generate lineage report

© Informatica. Proprietary and Confidential.

Synthetic Payload approach

- Generate Synthetic Payloads
- Invoke API to Insert Data
- Transactions are logged
- Query and inspect Database
- Extract lineage

© Informatica. Proprietary and Confidential.

Manual cross validation

Manual review and maintenance of mappings between API fields and database columns

Corrects **false positives** generated by automated lineage extraction methods

Leverages **domain expertise** to ensure data mapping accuracy and consistency

Acts as a **trusted source** for validating and refining data lineage

Complements heuristic or algorithmic approaches for **higher confidence**

Essential for **resolving ambiguities** and complex data relationships

© Informatica. Proprietary and Confidential.

Informatica

Benefits and Value of Automated Lineage

Where data & AI come to LIFE

© Informatica. Proprietary and Confidential.

Key Benefits of Automated API Lineage

- Reduces manual effort and operational overhead.
- Improves accuracy and completeness of lineage data.
- Accelerates compliance and audit processes.
- Enhances trust in data assets and APIs.
- Drives faster decision-making with real-time insights.

© Informatica. Proprietary and Confidential.

Summary: The Value of Automated API Lineage

Essential for modern data governance and transparency.

Addresses critical compliance and operational needs.

Enables organizations to scale data operations efficiently.

Paves the way for future innovation in data management.

© Informatica. Proprietary and Confidential.

Thank You!

Questions?

Where data & AI come to LIFE

© Informatica. Proprietary and Confidential.