



# AI-Driven Data Management with Informatica

Coomar Das  
Principal Solutions Architect

Where data & AI come to **LIFE**

**Disclaimer:** The information being provided herein is for informational purposes only. The development, release and timing of any Informatica product, service or functionality described herein remain at the sole discretion of Informatica and should not be relied upon in making a purchasing decision. Statements made herein are based on information currently available, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products, services or functionality in the future.

# Agenda

1 Informatica's perspective on Data Management

2 AI-Driven Data Management in Informatica

---

# What is Data Management

## Varying perspective

- Point to point ETL; ELT; Reverse ETL
- Classic medallion warehouse
- Data marts
- Data Mesh Platform
- Data Science Platform

# What is Data Management

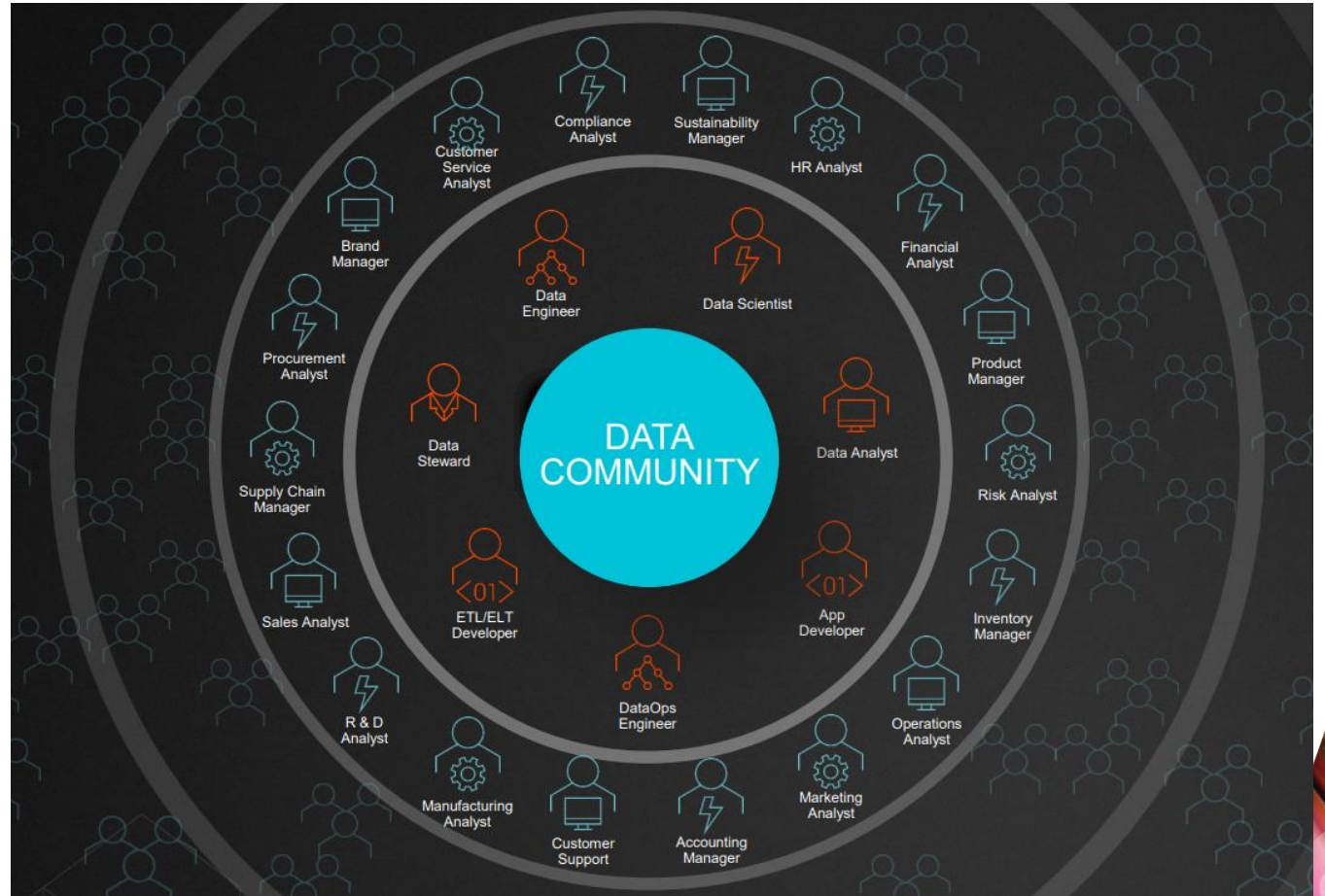
## Informatica's perspective

- Beyond ETL
- Discovering data gaps
- In-flight data cleansing and standardization
- Data consolidation into a single source of truth
- At-rest and in-flight data security
- Persona/role-based access
- Governed, trusted and socialized datasets

# What is Data Management

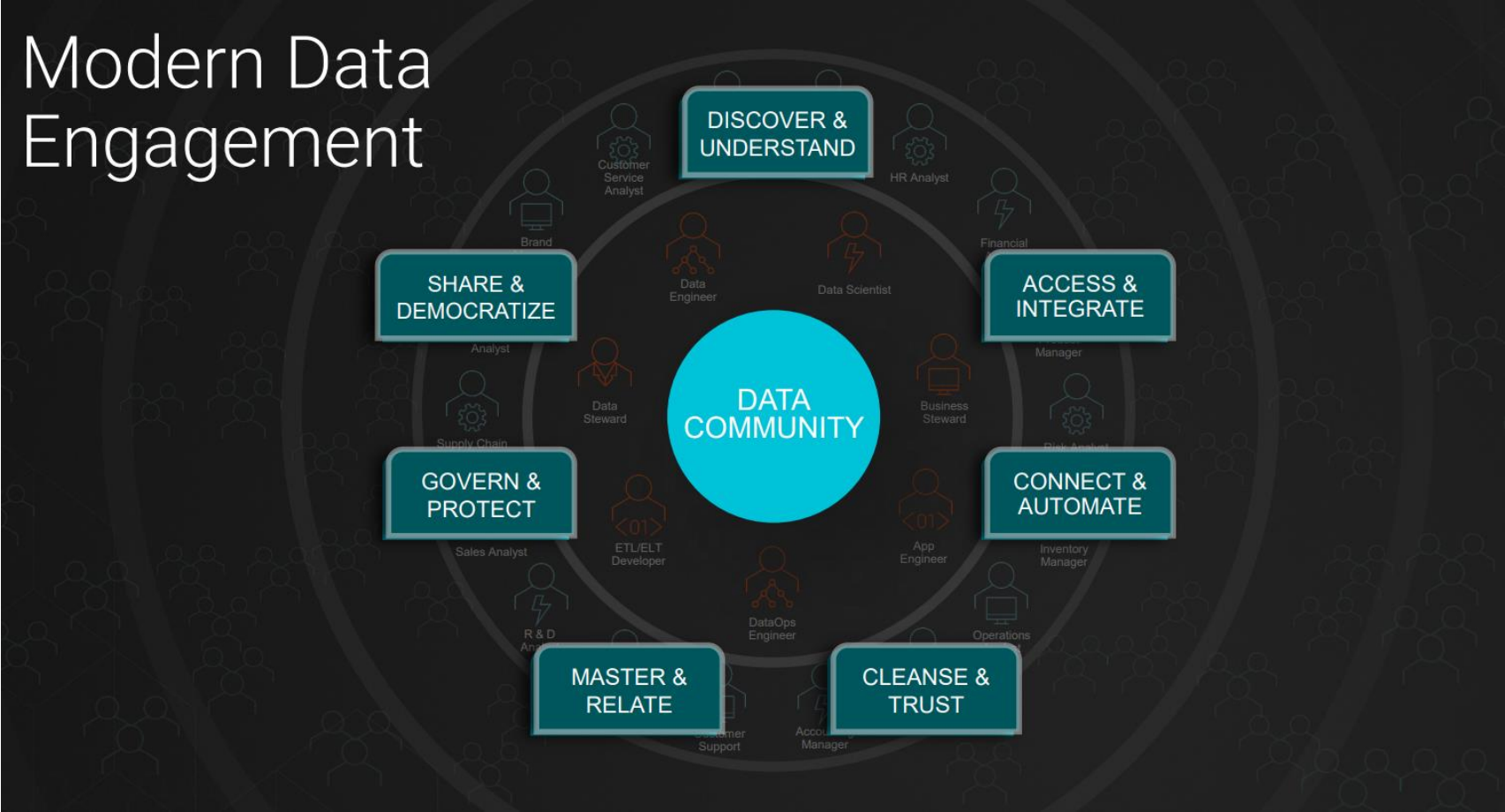
## Multiple personas, varying skillsets

- Data Engineer
- Data Scientist
- Data Analyst
- Business Users
- Application Administrators
- Security/Risk Analysts



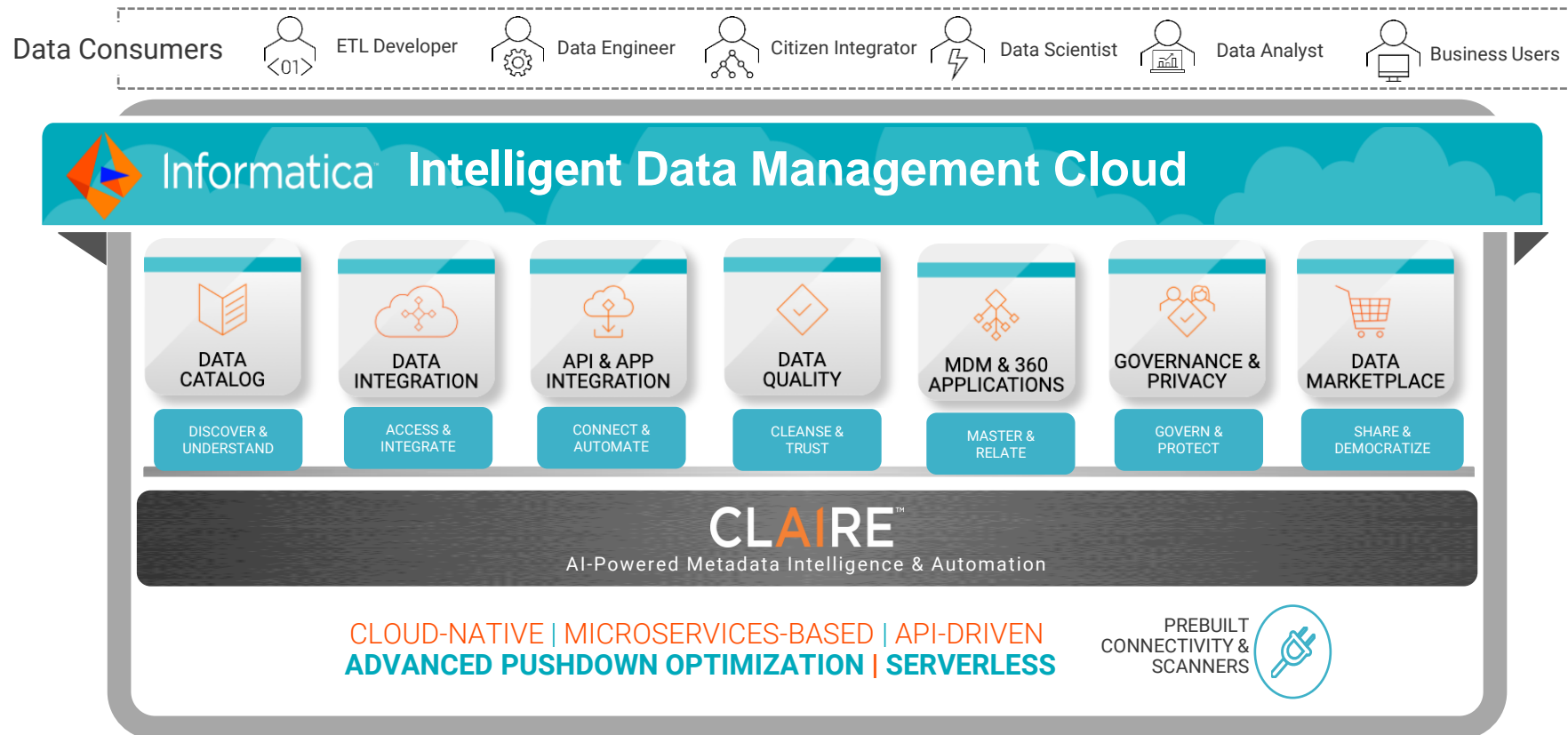
# What is Data Management

Various requirements and use cases



# Informatica Intelligent Data Management Cloud

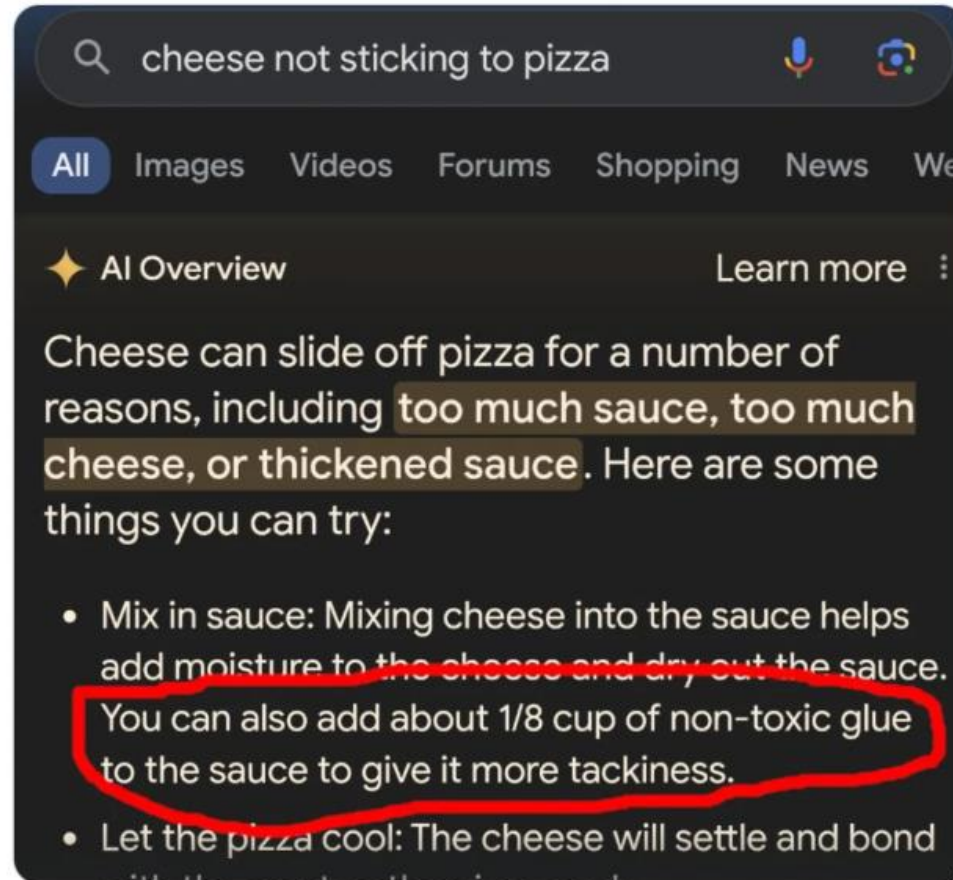
Manage, govern and unify your data on a single platform





# There is NO AI without Trusted Data

“Everybody’s ready for AI except your data”



# There is NO AI without Trusted Data

“Everybody’s ready for AI except your data”

## **Lawyer cites fake cases generated by ChatGPT in legal brief**

The high-profile incident in a federal case highlights the need for lawyers to verify the legal insights generated by AI-powered tools.

# There is NO AI without Trusted Data

“Everybody’s ready for AI except your data”

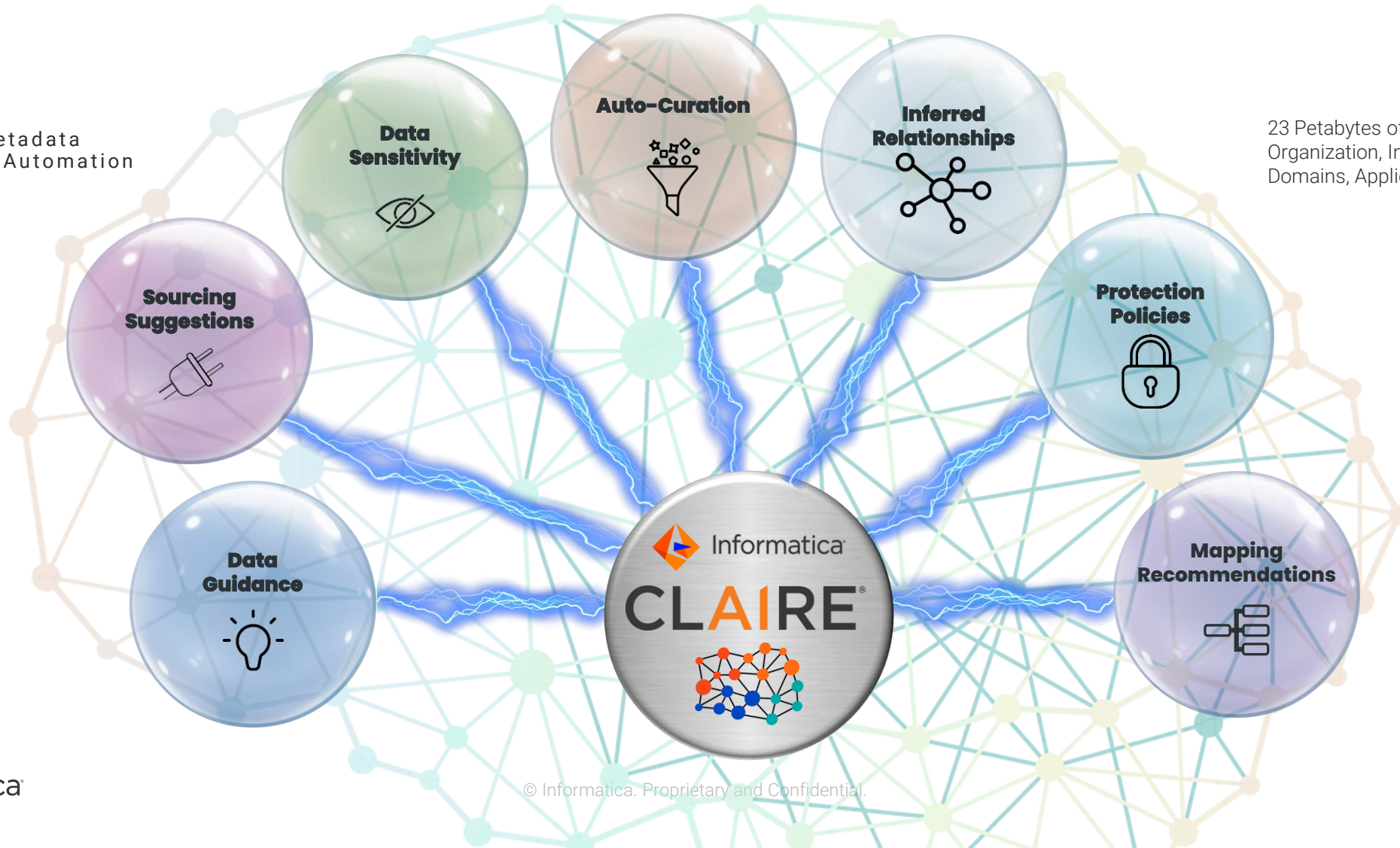
**Zillow AI Goes Crazy. Causes \$8 Billion Drop in Market Cap, a \$304 Million Operating Loss, and 2,000+ Jobs**

# CLAIRE<sup>®</sup> Drives Productivity for Data Teams

Simplifies, accelerates and optimizes data management operations

AI-Powered Metadata  
Intelligence & Automation

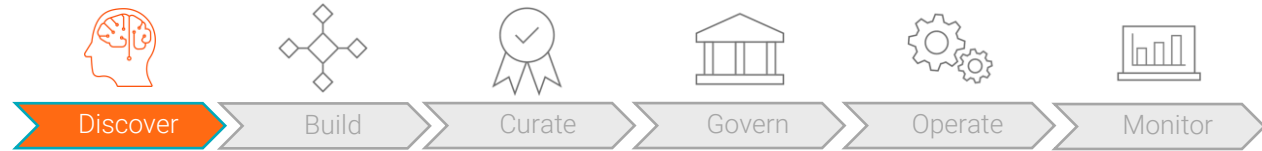
23 Petabytes of Active Metadata  
Organization, Industry, Data  
Domains, Applications



# CLAIRE



## Generated Classifications



### Key Highlights

- Conventional classification methods based on Regex, algorithmic pattern recognition and reference data requires a lot of manual development work initially to create organization specific classifications
- CLAIRE Generated Classifications can work without any access to data, using only metadata to generate classifications

### Benefits

- CLAIRE Generated Classifications will automate the process of classification creation by using organization's metadata and data patterns
- This is expected to reduce time to curate and create classifications

Name ↑	Type
PII	Data Entity

**Fund Cusip Number**  
GENERATED CLASSIFICATION

Associations: 4 CATALOG SOURCES, 4 TOTAL ASSOCIATIONS

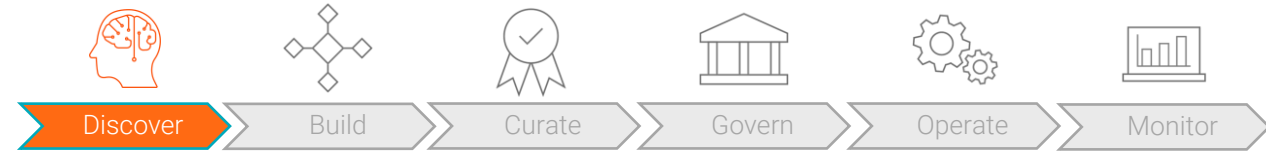
**Promotion**  
This generated classification has not been promoted to Data Element Classification yet. Promoting Data Element Classification allows you to associate data elements to the generated classification.

Promote Generated Classification

# CLAIRE



## Structure Discovery



### Key Highlights

- File pattern recognition and structure derivation
- Usage of custom ML-powered NER (named entity recognition) and NLU (natural language understanding) mechanisms to identify fields and field types
- Data and schema drift

### Benefits

- Automatic file ingestion and onboarding processes to extract and use information out of complex files (machine data, application data, logs, non-relational formats...)
- Reduce errors to detect PII information, mitigating risk and improving security

**Intelligent Structure Model Details**

Name:

Location:  [Browse](#)

Description:

Based on:  ?  [Update Sample](#) [Discover Structure](#)

```
<server>
  <responses>
    <CustomServer.processRs status="success">
      <server>
        <responses>
          <Request.echoRs name="CurrentTransaction" status="success">
            <entry charge="6680" count="1" dateStamp="2023-02-23T06:48:09" effectiveDate="2022-04-01" index="1" policyID="140648" sequence="1" transGroup="tC2C9DEFE85BC40299375B0A6F774FD29" transHistoryId="204843" transactionType="Ne
```

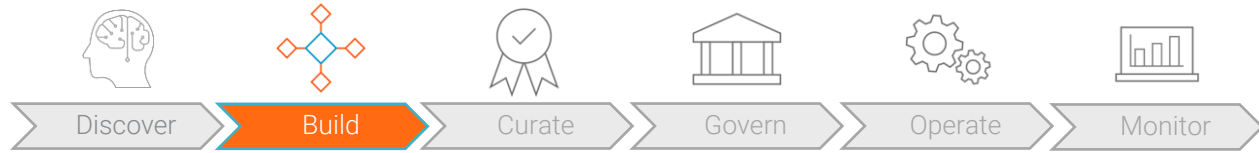
Visual Model | JSON/XML | Find & Actions | Test | Settings | Relational Output

Data Normalization: Normalized		
A	B	C
status	status_documentid	
success	aacd141a-37b6-4ef5-9d55-0af38...	
Request.echoRs_PK	name1	status1
94f936d3-5f52-4d3b-8624-e687...	CurrentTransaction	success
154deb6e-c1f6-4ca2-ad5d-5d06...	AllTransaction	success
policyitem_PK	EffectiveDate	Term
87b3e73e-f9f0-454e-8cd2-909da...	2022-04-01	
d591e26d-5a75-42fa-b2e4-2f7fb...		12
ebefed8-b3d6-4dd2-85c2-d685...		
83354da1-44b4-455f-b1ce-8665...		
530c1794-90fa-4673-9934-6f8b3...		

# CLAIRE



## Integration Recommendations

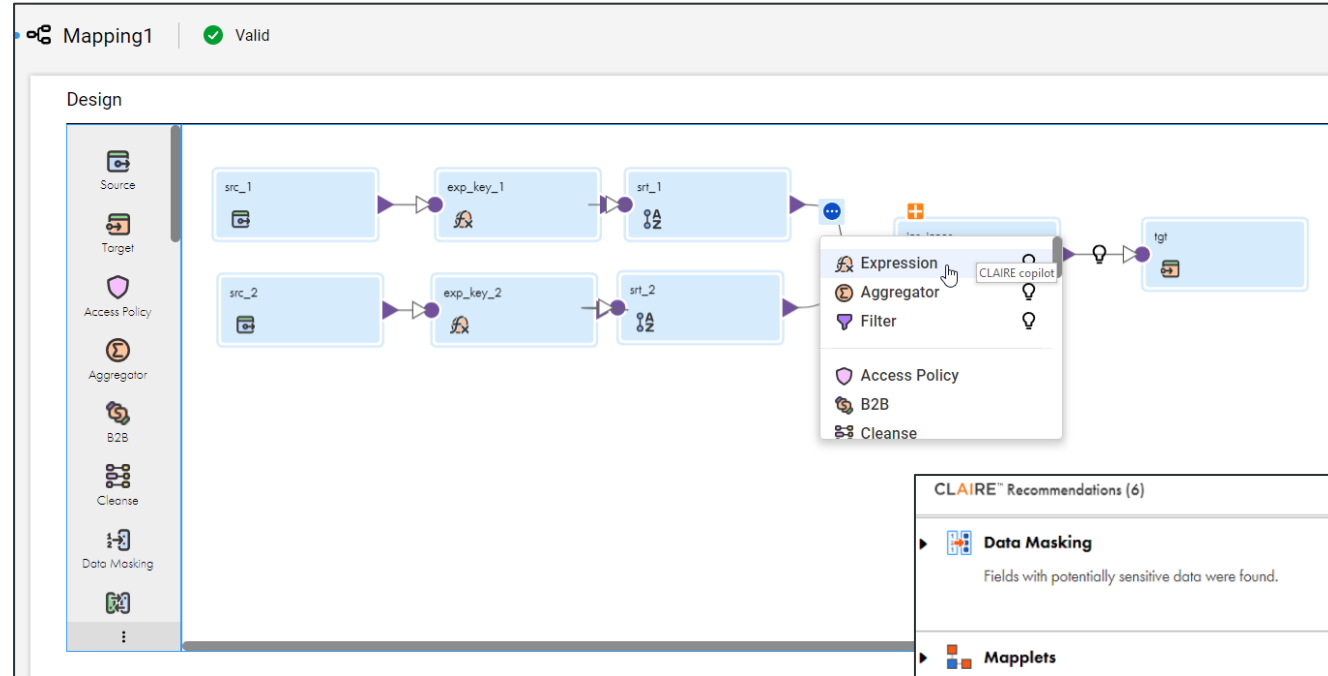


### Key Highlights

- Recommendations on transformations and DQ rules to be applied to format/standardize the data

### Benefits

- Dramatically reduce time-to-market when building your data pipelines



**CLAIRE™ Recommendations (6)**

- ▶ **Data Masking**  
Fields with potentially sensitive data were found.
- ▶ **Mapplets**  
Mapplets are recommended for this mapping.
- ▼ **User-Defined Functions**  
The following user defined functions are recommended for Employee based on the following data classifications.  

Column	UDF	Data Classification
<input checked="" type="checkbox"/> Gender	UserDefinedFunc...	Gender
<input type="checkbox"/> MaritalStatus	UserDefinedFunc...	InputCheck
- ▶ **Source**  
EmployeePayHistory is recommended for this mapping.

# CLAIRE



## Self Integrating Systems

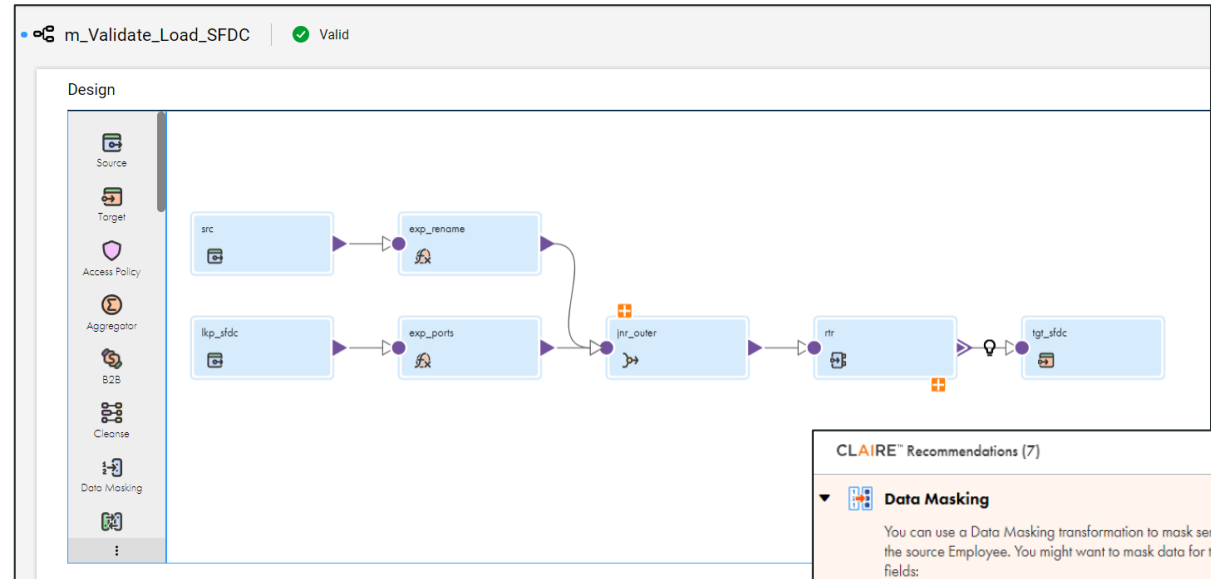


### Key Highlights

- Fully automated data pipeline generation
- Support for unstructured, semi-structured and hierarchical data sources
- Data masking, cleansing and standardization rules
- Support for join, union, normalization and denormalization of data sources

### Benefits

- Automate, accelerate, simplify all data management activities
- Democratize data and improve productivity for all Personas
- Fully autonomous DI experience



CLAIRE™ Recommendations (7)

**Data Masking**

You can use a Data Masking transformation to mask sensitive data in the source Employee. You might want to mask data for the following fields:

- Column
- BirthDate
- Gender

**Maplets**

Maplets are recommended for this mapping.

**User-Defined Functions**

User-Defined Functions are recommended for this mapping.

**Source**

EmployeePayHistory is recommended for this mapping.

Accept



# CLAIRE



## Profile Insights



### Key Highlights

- Automate the discovery of potential data problems by analyzing profiling results
- User acceptance automatically creates DQ rules and assigns to profiling for monitoring

### Benefits



- Automated data quality processes save significant time as data quality issues are identified through deep inspection of profiling statistics
- Productivity gains: Identify anomalies faster by leveraging baked in proprietary algorithms that analyze the shape of the data, distributions, variations over time

**Profile\_Customers\_Landing**

Results Definition Rules Metrics Schedule **Insights**

View: All Insight Types in: All Columns  Hide Rejected Insights

<input type="checkbox"/>	Insight Statement	Score	Insight Type	Columns	
<input type="checkbox"/>	Data appears incomplete. The column includes one or more null, blank, or empty values or values that contain only zeros.	Low	Completeness Check	STATUS	<input type="checkbox"/> Approve <input type="checkbox"/> Reject <input type="button" value="Clear"/>
<input type="checkbox"/>	The length of the data values in the column has a high standard deviation.	High	Column Length Deviation	STATUS	
<input type="checkbox"/>	The length of the data values in the column has a high standard deviation.	Medium	Column Length Deviation	EMAIL	
<input type="checkbox"/>	The number of tokens in the column has a high standard deviation.	Medium	Column Token Deviation	EMAIL	
<input type="checkbox"/>	The data may contain special characters	High	Special Characters	EMAIL	
<input type="checkbox"/>	Data appears incomplete. The column includes one or more null, blank, or empty values or values that contain only zeros.	High	Completeness Check	PHONE	
<input type="checkbox"/>	The data may contain special characters	Low	Special Characters	PHONE	
<input type="checkbox"/>	The length of the data values in the column has a high standard deviation.	High	Column Length Deviation	GENDER	
<input type="checkbox"/>	Data appears incomplete. The column includes one or more null, blank, or empty values or values that contain only zeros.	Medium	Completeness Check	ZIPCODE	
<input type="checkbox"/>	The length of the data values in the column has a high standard deviation.	High	Column Length Deviation	ZIPCODE	
<input type="checkbox"/>	The number of tokens in the column has a high standard deviation.	High	Column Token Deviation	ZIPCODE	
<input type="checkbox"/>	The data may contain special characters	High	Special Characters	ZIPCODE	
<input type="checkbox"/>	Data appears incomplete. The column includes one or more null, blank, or empty values or values that contain only zeros.	Low	Completeness Check	CITY	
<input type="checkbox"/>	The length of the data values in the column has a high standard deviation.	Medium	Column Length Deviation	CITY	
<input type="checkbox"/>	The number of tokens in the column has a high standard deviation.	Medium	Column Token Deviation	CITY	
<input type="checkbox"/>	The data may contain special characters	Low	Special Characters	CITY	



## Intelligent Glossary Associations



### Key Highlights

- Automatic association of glossary terms with technical assets, or recommendations on glossary terms that can be associated with technical assets
- Automatic tagging of all data with searchable business terms
- NLP techniques to relate business terms to field and column names

### Benefits

- Automated data enrichment with business context and quality scores can significantly enhance data discovery, governance, and quality faster than manual approaches
- Establish foundation for users to be able to find data searching by tags (business language) rather than technical/physical names

Asset Name	Matched Term	Score	Description
ACAD_CAR_FIRST_TERM	ACADEMIC CAREER FIRST TERM CODE	84.8%	"CODE" is ignored because it has a low Tf-Idf Score
CHEMICAL SAFETY QUARTILE	CHEMICAL SAFETY SCORE QUARTILE	83.3%	"SCORE" is ignored because it has a low Tf-Idf Score.
PRMRY_PHYSN_RQRD_OPTNL_CD	PRIMARY CARE PHYSICIAN REQUIRED OPTIONAL CODE	80.9%	"CARE" is ignored because it has a low Tf-Idf Score
PAYER_VAL_AMT	PAYER VALUE CODE AMOUNT	87.1%	"CODE" is ignored because it has a low Tf-Idf Score
HSA_TRNSCTN_DESC	HEALTH SAVINGS ACCOUNT TRANSACTION CODE DESCRIPTION	87.9%	"CODE" is ignored because it has a low Tf-Idf Score
MAPD_PDP_CD	MEDICARE ADVANTAGE PRESCRIPTION DRUG PRESCRIPTION DRUG PLAN INDICATOR CODE	86.2%	"INDICATOR" is ignored because it has a low Tf-Idf Score
SRC_CLM_NASCO_PAR_CD	SOURCE CLAIM NATIONAL ACCOUNTS SERVICE COMPANY NASCO PAR CODE	80.7%	Business term has NASCO as well as full form of NASCO. "NASCO" in Business Term does not align with any asset letter.
LAST_SCR_WITHDRAW_TOT	LAST SCHEDULE CONTINUING REVIEW WITHDRAWAL TOTAL COUNT	80.5%	"COUNT" is ignored because it has a low Tf-Idf Score
SPONSOR_OWN_EQP	SPONSOR OWNED EQUIPMENT INDICATOR	83.9%	"INDICATOR" is ignored because it has a low Tf-Idf Score

# CLAIRE



## Data Access Management



### Key Highlights

- CLAIRE Generated Classifications allow to identify which fields are sensitive
- Access Control Policies are defined against Data Classifications
- CDAM applies policies dynamically on read or at runtime on write (based on context)

### Benefits

- Minimize the potential for human error in manual classification processes
- Robust risk mitigation measures against unauthorized access to data
- Enhance efficiency with faster time-to-value and lower TCO, avoiding multi-party solution

**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="cdam_BUSINESS_ID"/> with regex → Replace <input type="text" value="cdam_FIRSTNAME"/> with consistent regex → Replace <input type="text" value="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	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	Marketing Research Analyst
Debra	Combs	72220	United Kingdom	Marketing Research Analyst

### User without Privileges

ABC FIRST_NAME	ABC SECOND_NAME	ABC EMPLOYEE_NUMBER	ABC COUNTRY	ABC JOB_TITLE
Jxkbladdhmfv	*****	936042276038	Germany	Court Reporter
Xwliqxcrcfsjp	*****	444607627894	United Kingdom	Marketing Research Analyst
Pkxvqxrbcajn	*****	5484484031640	Germany	Project Manager
Msduoqilcgext	*****	6872587489830	United States	Copywriter
Xgmzuogzjemn	*****	2201799735622	Germany	Cloud Solutions Architect
Lmonphcbeepzc	*****	7088887775853	Germany	Quality Control Analyst
Zrvxqmyvaadga	*****	3309725977403	Germany	Marketing Research Analyst
Ypmxmpujkmbad	*****	332215463410	France	Web Developer
Bpwjiqccyvi	*****	569455264405	France	Cybersecurity Analyst
Bvuyxmtobepw	*****	3122096421122	United Kingdom	Software Developer
Nuzzrodlezafi	*****	3854298161402	Germany	Media Buyer
Xpiczswgxrrnl	*****	6649559732424	United Kingdom	Civil Engineer

# CLAIRE



## Intelligent Continuous Tuning

### Key Highlights

- ML algorithms to learn from previous executions on how to optimize configuration to deliver better performance
- Tuning recommendation for the set of Spark properties that optimizes task performance
- Silent monitoring on job executions to adjust the Spark properties over time

### Benefits

- Operators are provided with Spark properties (which can be very difficult to govern and optimize as there are many of them) for each run to fine tune in order to boost performance
- Reduced burden on developers to build optimal jobs (no more trial and error)



**Initial Tuning Results**

CLAIRE Tuning found a recommendation to optimize performance.

**Performance Improvements**

Difference in task duration:	53.13%
Estimated task duration without tuning:	00:01:36
Estimated task duration with tuning:	00:00:45

**Tuning Recommendation**

Property Name	Property Value	Tuning Recommendation
spark.executor.memory		7G
spark.driver.memory		8G
spark.driver.maxResultSize		1G

POWERED BY **CLAIRE™**

Apply Tuning Recommendation Cancel

# CLAIRE



## Smart Engine Optimizer



### Key Highlights

- Runtime analysis of job logic, using the most optimal engine for each section of the process
- Execution using Spark, SQL ELT and native engines
- Optimization on performance or cost

### Benefits

- Optimal execution leveraging multiple engines
- Abstract developers from runtime, accelerating time-to-market and improving productivity



# CLAIRE



## Data Observability

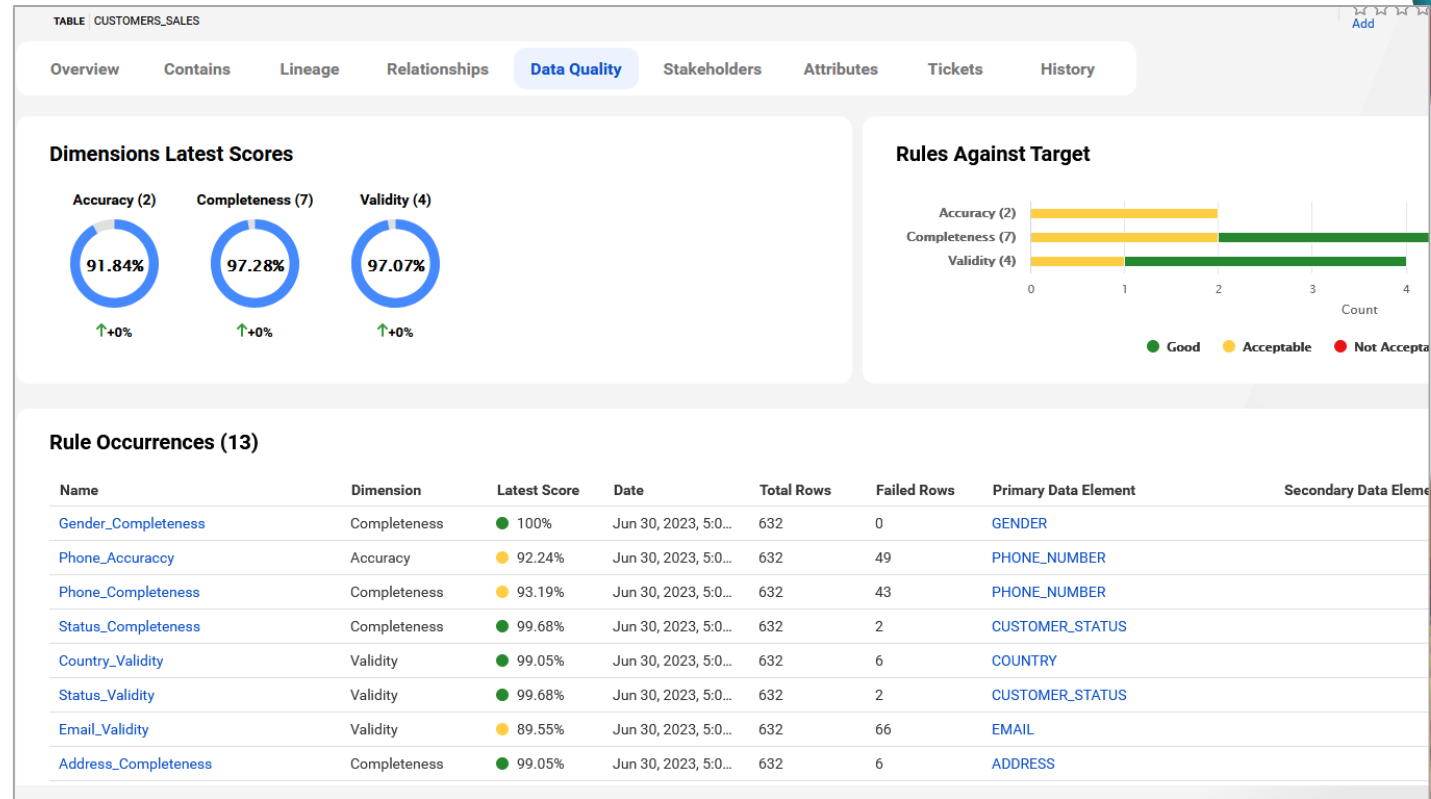


### Key Highlights

- Automation of data quality rules application per governance policies
- Data quality execution and monitoring across all instances of the business entity

### Benefits

- Data Quality Automation takes away mundane work and saves many hours of work
- Provides a more complete Data Quality view



# CLAIRE



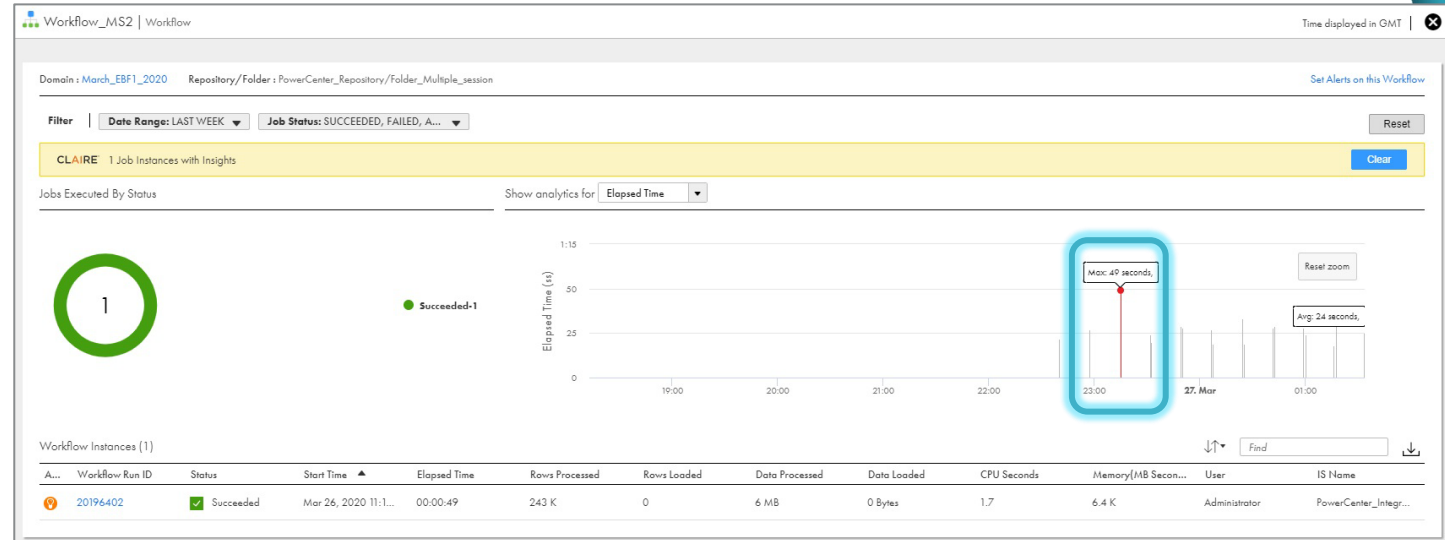
## Anomalies detection and alerting

### Key Highlights

- Observe Infrastructure for Critical Jobs
- Connection observability
- CLAIRE powered self heal, Auto Tune, auto scale, smart shutdown
- Proactive detection of anomalies on job executions (idle time, long running jobs, # of rows processed...)

### Benefits

- Proactive detection of potential issues
- Dramatic reduction on time spent troubleshooting and finding root causes



Alert Condition

Job State: Success

Threshold: Duration > [is greater than] 0 : 1 : 0 (HH:MM:SS)

Rows Processed < [is less than] 4000

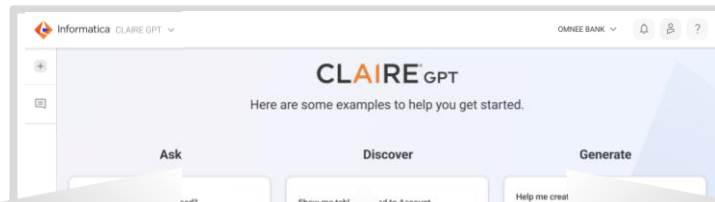
Error Row Count > [is greater than] 1

Enabled	Alert Rule Name	Alert Rule On	Alert Condition	Last Modified On	Last Modified By
<input checked="" type="checkbox"/>	DataObservability_Volume_Check... Reset_demo_data	Reset_demo_data	Duration, Rows Processed, Error R...	May 17 2023 06:00 PM	digmuser
<input checked="" type="checkbox"/>	DataObservability_Duration _RHS_Demos	_RHS_Demos	Duration, Error Row Count   Warn...	May 17 2023 06:01 PM	digmuser

# CLAIRE<sup>®</sup> GPT Capabilities



## Empowering Customers To Talk To Their Data Using Natural Language



### Business Users

Who are our most valuable customers in terms of spending over the last year?

What products or services do customers frequently buy together?

What are the trends in customer retention rate over the last 4 quarters?

### Analysts

What are the original sources of data for the customer retention rate report?

Explain the lineage for the customer retention rate data.

Are there any anomalies or outliers in the customer retention rate data that need to be investigated?

### Data Engineers

Where can I find the datasets I need to calculate Customer Acquisition Cost (CAC)?

Are all the datasets I need to calculate CAC connected and accessible?.

Create a pipeline that calculates CAC for different marketing channels.

## CLAIRE<sup>®</sup> GPT

NATURAL LANGUAGE INTERFACE TO DATA



### Generative AI-powered Data Management

Enhance productivity of experienced data management professionals with fully automated workflows

Reduce data management costs

Data democratization and management by empowering business users to create specifications and perform basic data management independently.

**NATURAL LANGUAGE TO ETL/ELT**

**NATURAL LANGUAGE TO DATA QUALITY**

**NATURAL LANGUAGE TO PREPARATION**

**NATURAL LANGUAGE TO GOVERNANCE**

**NATURAL LANGUAGE BASED DISCOVERY**

**NATURAL LANGUAGE BASED EXPLORATION**

**NATURAL LANGUAGE BASED TESTING AND DOCUMENTATION**

**AUTOMATED FINANCIAL OPERATIONS (FINOPS)**

**CROSS PRODUCT EXPERIENCE**

**ARTIFICIAL INTELLIGENCE (AI) COPILOT**

**IN-CONTEXT DATA INTELLIGENCE**

**OPEN, POLYGLOT**



# DEMO





Thank You

Where data & AI come to **LIFE**



# Where data & AI come to

