Informatica

# Artificial Intelligence for the Data-Driven Enterprise

How the machine learning-based innovations in CLAIRE are driving new advances in data management

Where data & AI come to LIFE

# Contents

## The Importance of AI

Artificial intelligence (AI) and machine learning (ML) are powering the digital transformations happening in every industry worldwide. As a strategy to transform their businesses, generative AI is top of mind for boardroom executives. It has become pervasive in enhancing our daily lives, from the movies we watch to the cars we drive. AI and ML have played a critical role in helping to discover new therapies in life sciences, reducing fraud and risk in the financial services sector and delivering genuinely personalized customer experiences.

For business leaders, AI/ML may seem a bit like magic — while its potential impact is clear, they may not quite understand it or how best to wield these powerful innovations. AI/ML is the underpinning technology for many new business solutions — next-best actions, customer satisfaction tracking, efficient operations, innovative products and more.

And now that generative AI is poised to turbocharge these initiatives, it's important to note that it is fueled by large language models (LLMs), making it very data hungry. LLMs need vast amounts of domain-specific data for training to provide the required accuracy. This data must accurately reflect the current state of the business. AI trained with bad or limited data will have a terrible impact on business initiatives, to the point where it has a reverse effect on the desired outcome.

For effective AI, in which the right features are used and trained, we need to tap into a wide variety of holistic, high-quality, governed data from inside and outside the organization. This data must be combined to build and train AI models and fuel LLMs through **data management**.

But it's more than a question of dealing with the scale and complexity

**"AI is a valuable resource that can dramatically improve both productivity and the value companies obtain from their data."[1]**

— MIT Sloan Management Review, How AI Is Improving Data Management

[1] **MIT Sloan Management Review, "How AI Is Improving Data Management"**

— it is also about trust. Is the data used to train the model coming from the right systems? Have we removed personally identifiable information (PII) and followed all regulations? Are we transparent, and can we prove the lineage of the data the model uses? Can we document and be ready to show regulators or investigators that the data has no bias? All of this requires adequate control and a basis for data management. Without a solid data management foundation, AI is incomprehensible and unreliable — in other words, without data management, AI won't be ready even when everybody else is.

## AI Needs Data

The success of AI is dependent on the effectiveness of the AI models designed by data scientists and engineers to train and scale it, which is dependent on the availability of trusted and timely data.

Why do data scientists and engineers building AI/ML models need holistic, high-quality, governed data? Take, for example, an LLM tasked with anticipating a consumer's behavior. A valuable feature for such a model could be consumer location as indicated by the postal ZIP code. But what if the ZIP code data is missing, incomplete or inaccurate? The model's behavior will be adversely affected during training and deployment, which could lead to incorrect predictions and reduce the value of the entire effort.

In addition, an accurate, complete and verified ZIP code could also help to predict an individual's market segmentation, income class, age, life expectancy and more — even more reason to get it right. We should all expect "explainable AI" to become a regulated mandate, not just an option. AI-powered applications and insights cannot be deployed into production without metadata-driven lineage and traceability.

For AI to be successful, it needs access to an intelligent data management cloud. This data management cloud should be capable of quickly identifying all the necessary features for the model, automatically adjusting the data to meet the requirements of the AI model (such as feature scaling and standardization), removing duplicate data and providing reliable master data about customers, patients, partners and products.

It should also offer a complete **data lineage** within the model and its operations. The effectiveness of the models designed by data scientists to train and scale the AI is critical to its success, and the availability of trusted and timely data is crucial to the success of those models.

## Data Needs an AI Copilot

AI/ML also plays a critical role in scaling data management practices. Due to the massive volumes of data needed for digital transformation and AI initiatives, organizations must discover and catalog their most relevant data and metadata to certify the relevance, value and security — and to ensure transparency. They must cleanse and master this data. And they must effectively govern and protect this data. If data is not managed effectively — and to scale — AI/ML models will suffer the same fate as every traditional data warehousing initiative over the past 30 years: when you use poor-quality data, you will deliver untrustworthy insights.

According to recent research, 328.77 million terabytes of data are created daily, and it is estimated that 90% of the world's data was generated in the last two years alone. This figure has increased by an estimated 60x from just two zettabytes in 2010. The 120 zettabytes generated in 2023 are expected to increase by over 150% in 2025, hitting 181 zettabytes.[2] All this data needs to be processed and made usable, trustworthy and democratized for use across the organization while adhering to governance policies.

Adding to all this is the requirement to move quickly and respond to business strategy and process changes. The effort involved in preparing the data for digital transformation initiatives has increased in complexity with the amount of data growth. A recent report by DICE predicted that data engineering would be one of the fastest-growing jobs in technology, forecasting a 42% year-over-year growth in open positions.[3] But more than hiring alone is needed to manage the increase in data volume.

### Don't Take a Linear Approach to an Exponential Challenge

We cannot solve these challenges by throwing more engineers and developers at the problem — this issue cannot be solved at a linear, human scale. Traditional approaches are riddled with inefficiencies. Projects are implemented in silos with little end-to-end metadata visibility and limited automation. There is no automated learning, the processing is expensive, and governance and privacy steps are repeated on an endless loop. So how can organizations move at the speed of business, enable self-service, better serve their customers, increase operational efficiency and rapidly innovate?

This is where AI for data management shines. AI can automate and simplify tasks related to data management — whether you need to discover, integrate, cleanse, govern or master data. Generative AI for data management can learn and take over mundane, repetitive tasks, freeing developers and users to work on high-value, innovative projects. AI improves data understanding and identifies data privacy and quality anomalies. AI is a perfect partner for developers, analysts, stewards and business users, increasing productivity by automating tasks and augmentation with recommendations and next-best actions.

[2] ExplodingTopics.com, "Amount of Data Created Daily."
[3] Dice.com, "The Dice Tech Job Report."

AI is most effective when you think about how it can help you accelerate end-to-end processes across your entire data environment. That's why we consider AI essential to data management and why Informatica® has focused our innovation investments so heavily on the CLAIRE® engine, our metadata-driven AI capability to advise and guide you. CLAIRE acts as a copilot by leveraging virtually all enterprise-unified metadata to automate and scale routine data management and stewardship tasks.

## Four Major Benefits of AI for Data Management

AI benefits data management teams in four significant ways: by improving data professionals' productivity; enhancing operations' efficiency; providing a more intelligently guided data experience and more profound understanding; and speeding up data governance processes. Below are a few examples to show what's possible today.

**Productivity:** A recommender system for data integration helps data engineers rapidly build mappings to extract, transform and deliver data. The recommender learns from existing mappings, understands the business content of databases and file systems, and suggests appropriate transformations for standardizing and cleansing data before delivering to target systems and data consumers.

**Efficiency:** In a typical enterprise, thousands of data integration processes run daily. Monitoring these processes is mainly passive, with administration tools just logging time taken and CPU and memory consumed. AI can learn from historical values of time-series data in log and monitoring files to predict issues that may occur if not handled beforehand.

**Data Democratization:** When a real-world entity (e.g., a patient record or sales order) is stored in a database or a set of files, its data gets shredded and distributed into multiple tables or files — optimizing it for storage and performance. AI can detect relationships among data and reconstitute the original entity quickly. Users don't have to remember or look up outdated primary-key/foreign-key relationship documentation and manually join the various datasets by hand. Furthermore, AI can identify similar datasets and make recommendations based on usage patterns, data quality and crowd-sourced collaboration.

**Data governance:** An ordinary but tedious step in data governance is to associate business terms with physical data elements to establish business context and relevance for data elements and make data understandable to users. In many cases, AI can automatically link business terms to physical data using natural language processing (NLP) techniques and business-type identification. This can dramatically reduce the drudgery of this error-prone task.

In this cloud era, it is essential to note that this approach also works for SaaS applications. Metadata can be gathered from SaaS applications such as Salesforce and Workday and added to the enterprise catalog.

### AI-Driven Data Management: An Example From Banking

To illustrate why AI needs data management and why data needs AI, let's consider this example from the financial services industry.

By applying AI to more and more data for advanced, predictive and real-time analytics, banks can:

- Offer more personalized services that increase customer retention
- Reduce fraudulent transactions at the point of sale
- Increase consumer investor results while reducing the cost of wealth advisors
- Reduce the cost of project-related regulatory compliance

From a data management perspective, AI can automatically discover and catalog virtually all relevant data such as ERP, CRM, cloud and web apps, machine and log files and third-party data. This gives data scientists a head start in accessing the data they need to run hundreds of experiments searching for patterns that reveal insights related to consumer behavior, fraudulent activity, investment opportunities matched with consumer propensity for risk and more.

Regarding data management, AI can automatically enrich a **360-degree view of customers** and persons of interest (POI) by discovering relationships between customer data and matching insights to specific people. This helps organizations better engage with their customers with more relevant offers and provide a seamless experience across various online, mobile or phone channels. A 360-degree view of POIs helps banks discover patterns and networks of fraudulent activity much faster, potentially saving millions of dollars.

AI can automate and guide data integration and data quality tasks to combine and cleanse data from hundreds of data sources, thereby increasing the predictive power of analytics models and algorithms. More and better data, combined with AI/ML and advanced analytics, has yielded significant results, such as improving next-best offers and identifying fraud.

AI also powers data governance that ensures policies are documented and enforced. This helps information security professionals comply with data privacy regulations such as the General Data Protection Regulation (GDPR), Sarbanes-Oxley Act (SOX), Basel II and Basel III and more.

## Informatica CLAIRE: The AI Copilot in the Intelligent Data Management Cloud

Informatica's approach to driving data management productivity with ML is:

1. The **Intelligent Data Management Cloud**™: We developed an all-in-one cloud-native data management platform to help streamline the entire data management process for maximum efficiency. It provides seamless connectivity, **metadata management** and operations management and facilitates the swift development and deployment of new data management projects. The platform offers robust and consistent capabilities that enable efficient data management across various sources, including on-premises, cloud, multi-cloud and multi-hybrid environments.

2. This platform is modular: Start with any single capability and grow at your own pace:

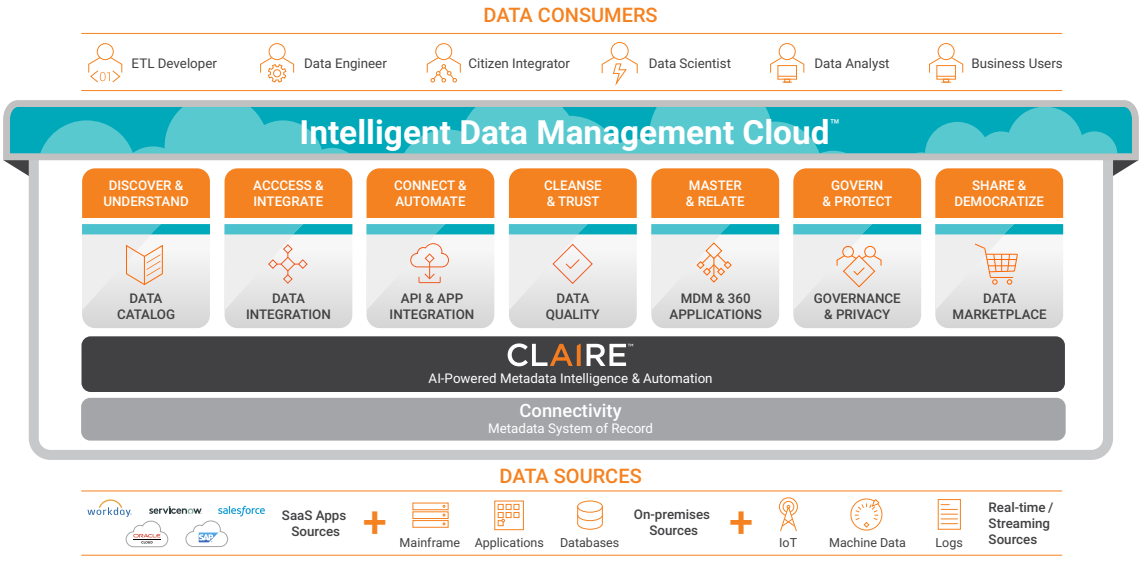



Figure 1: The Intelligent Data Management Cloud integrates data management capabilities with shared connectivity, operational insight and data and metadata intelligence.

3. Metadata: Informatica has long been known as a leader for its management of technical and business metadata. We have now increased our capabilities in this area by collecting a broader spectrum of metadata from across the enterprise, including:
   - Technical metadata, such as database tables, column information, data profile statistics, scripts and data lineage.
   - Business metadata, which captures context about data, including its meaning, relevance and criticality to various business processes and functions.
   - Operational metadata about systems and process execution to answer such questions as: When was the data last updated? When was the load process previously run? Which data was most accessed?
   - Usage metadata about user activity, including datasets accessed, search results clicked on and ratings or comments provided.

This broader collection of metadata is critical to ML. It provides datasets used to "train" the ML algorithms and enables them to adjust and produce better results.

4. Intelligence copilot: Informatica delivers an integrated combination of metadata and AI/ML with CLAIRE.

The metadata collected by the Intelligent Data Management Cloud provides a vast trove of information that the algorithms of CLAIRE can use to learn about an enterprise's data landscape. This knowledge helps CLAIRE make intelligent recommendations, automate the development and monitoring of data management projects and adapt to changes from within and outside the enterprise. CLAIRE is the copilot that helps drive the intelligence of virtually all the data management capabilities in the Intelligent Data Management Cloud.

CLAIRE acts as a copilot for a broad spectrum of users:

- Data engineers will find many implementation tasks partially or even fully automated
- Data analysts will find it easier to locate and prepare the data they need
- Business users will quickly identify data that should be subject to prescribed data governance and compliance controls
- Data scientists will gain an understanding of the data faster
- Data stewards will find it easier to visualize the quality of data
- Data security and privacy professionals will find it simpler to detect data misuse, protect sensitive data and demonstrate that appropriate controls are maintained
- Administrators and operators will have the power of predictive maintenance and performance optimization of data management processes

Here are some examples of how intelligence copiloted by CLAIRE is being used today.

## CLAIRE as Your Data Cataloging Copilot

Discovering and understanding your data is the first step in any data-driven initiative. CLAIRE copilots an ML-based discovery engine to scan and catalog data assets across the enterprise. An intelligent data catalog copiloted by CLAIRE can help data scientists, analysts and engineers find and recommend the necessary data, significantly reducing the time spent in data discovery and preparation.

### Advanced Relationship Discovery

One key data cataloging and modeling task is documenting relationships between datasets. CLAIRE uses machine-learning techniques to automatically identify primary and unique keys and joins across structured datasets. This reduces months of documentation effort to minutes. CLAIRE continuously improves its ability to identify relationships by including humans in the data-curation process. For example, users can accept or reject inferred relationships and learn from these actions with CLAIRE as their copilot.

For example, a data analyst at a bank creating a report about which customers are most likely to respond to a marketing campaign should be able to find existing products and loan information for all customers. However, given the siloed nature of data across the enterprise, finding such datasets across departments and data stores takes time and effort.

CLAIRE uses documented joins in the databases, joins performed in tools like BI and ETL, and statistics derived from data values to infer and recommend joins to the data analyst. This helps expand the user's analysis and uses virtually all the available information to create the right target audience for the campaign.

CLAIRE uses a variety of methods to discover keys and joins. To identify primary and unique keys, it combines profiling statistics such as null counts and uniqueness, along with column metadata that contains "ID" in the name, among other techniques. It then uses ML techniques like column signature analysis to discover joins and infer join keys at a large scale across numerous potential datasets.

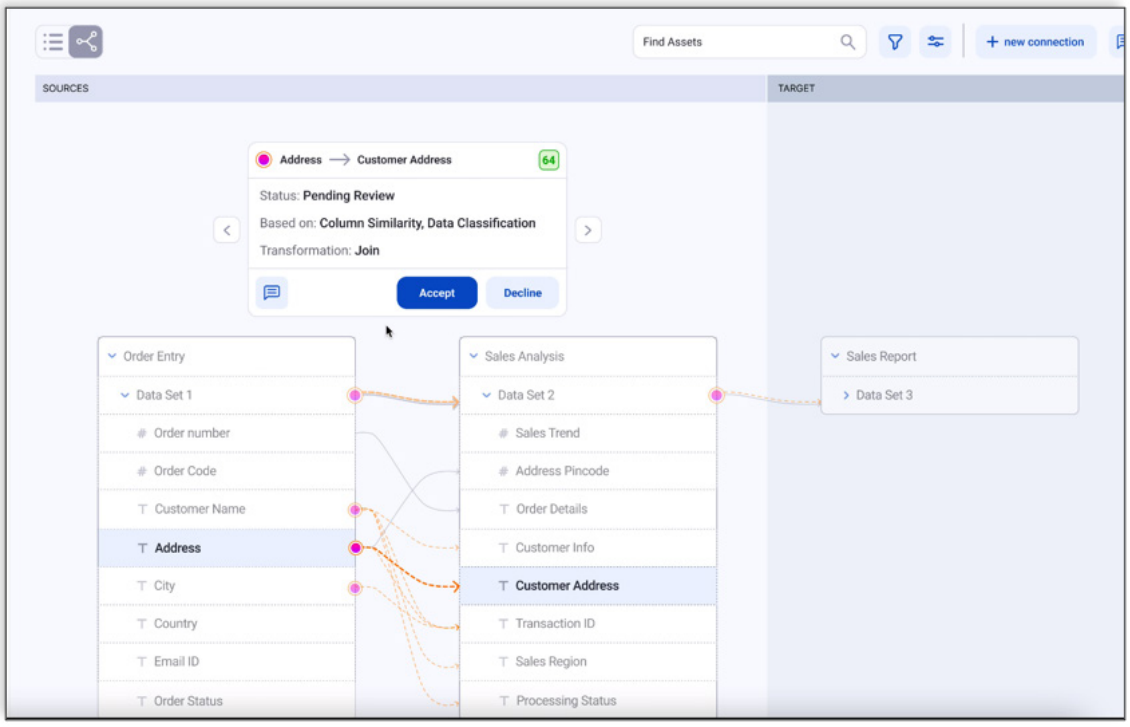



Figure 2: Discovering relationships through inference using machine-learning techniques.

## Intelligent Data Similarity

CLAIRE acts as a copilot using ML techniques like clustering to detect similar data across thousands of databases and file sets. Intelligent data similarity is one of the critical capabilities used for multiple purposes, including identifying data, detecting duplicates, combining individual data fields into business entities, propagating tags across datasets and recommending datasets as a trusted user copilot.

Data similarity computes the extent to which data in two columns is the same. A brute-force approach to try and compare all two-column pairs in an enterprise setting (say, across 100 million columns) would be computationally prohibitive. Instead, data similarity uses machine-learning techniques to cluster similar columns and identify likely matches.

The process works in multiple stages. First, columns are clustered based on column features. Then, data overlap is computed for unique values in each cluster. Finally, the most promising pairs are chosen for computing data similarity using the Bray-Curtis and Jaccard coefficients.

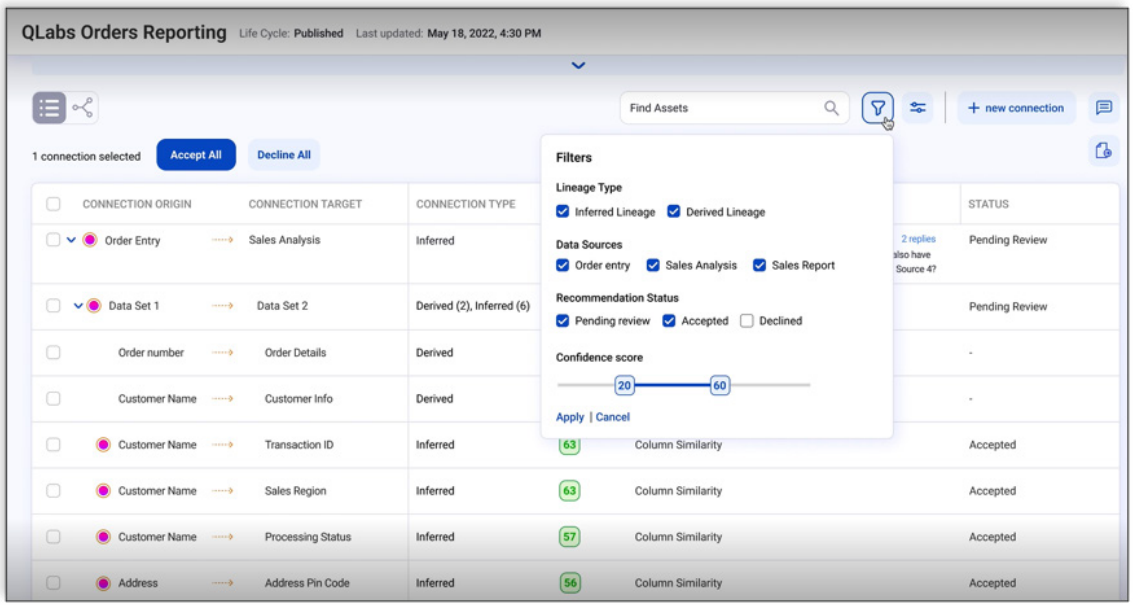



Figure 3: Identifying similar columns using clustering and the Bray-Curtis and Jaccard coefficients.

### Intelligent Domain Discovery with Tags

With CLAIRE as your copilot, you can automatically classify data fields by applying semantic labels to each column. These semantic labels are called data domains.

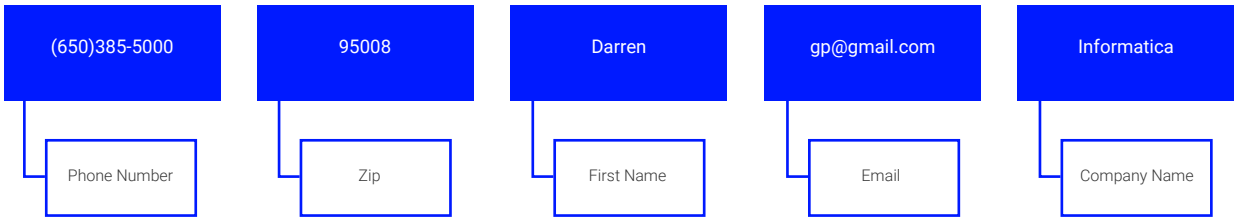



Figure 4: CLAIRE automatically classifies data fields and applies semantic labels called tags.

Usually, semantic labels are applied by evaluating rules based on regular expressions, reference tables or other complex hand-coded logic. It can be tedious to define and maintain thousands of such rules.

CLAIRE uses tags to simplify discovering and labeling data fields dramatically. For those columns not already classified, the user needs to provide a simple tag (say, “Claims Paid Date”) indicating the column content. The system learns by association and then auto-propagates this tag to all similar columns. The “facial recognition” for data technique is equivalent to tagging people in a Facebook photo, with the net effect that the same people are tagged in millions of other images.

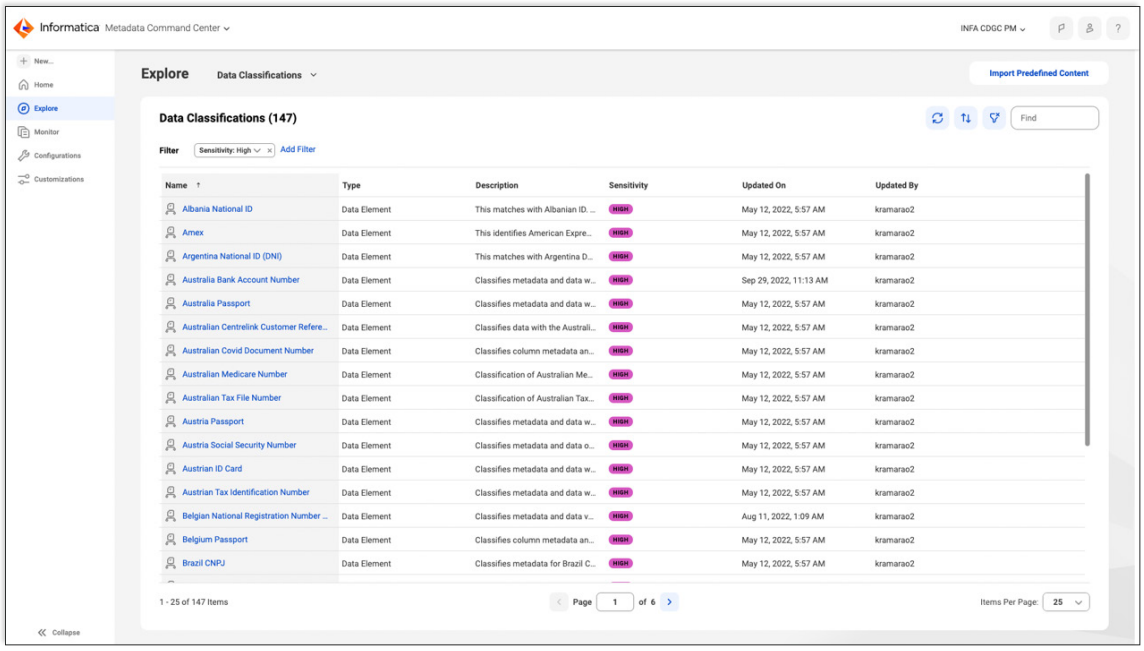



Figure 5: Automatic classification of data.

## Intelligent Entity Discovery

Once domains for columns have been identified, CLAIRE acts as a copilot by assembling these individual fields into higher-level business entities. The example below shows how a Purchase Order is created by combining fields identified as Customer and Product. Entity discovery learns how users have assembled disparate data fields in their analytics or data-integration processes and applies this learning to derive entities across the enterprise data landscape.

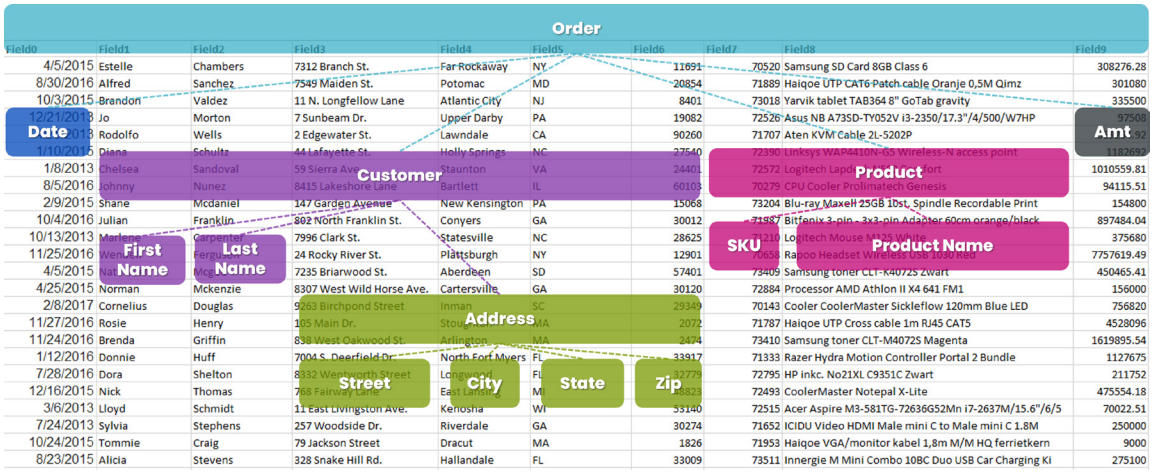



Figure 6: Combining data domains to detect entities from tables and files.

## CLAIRE as Your Analytics Copilot

CLAIRE-powered automation and intelligence acts as your copilot to help you significantly speed up analytic insights and processes, increase data availability and streamline data preparation for analytics. CLAIRE enhances data engineering productivity with data pipeline recommendations and the ability to parse complex multi-structured data automatically.

### Source Recommendation

Identifying the sources for datasets that require processing represents a significant challenge in analytic projects. With CLAIRE, you can provide a source table or combination, and it will recommend similar or related sources based on the relationship between datasets. It suggests joinable or unionable source tables that can be used for more contextual analysis, enabling you to gain greater insights.

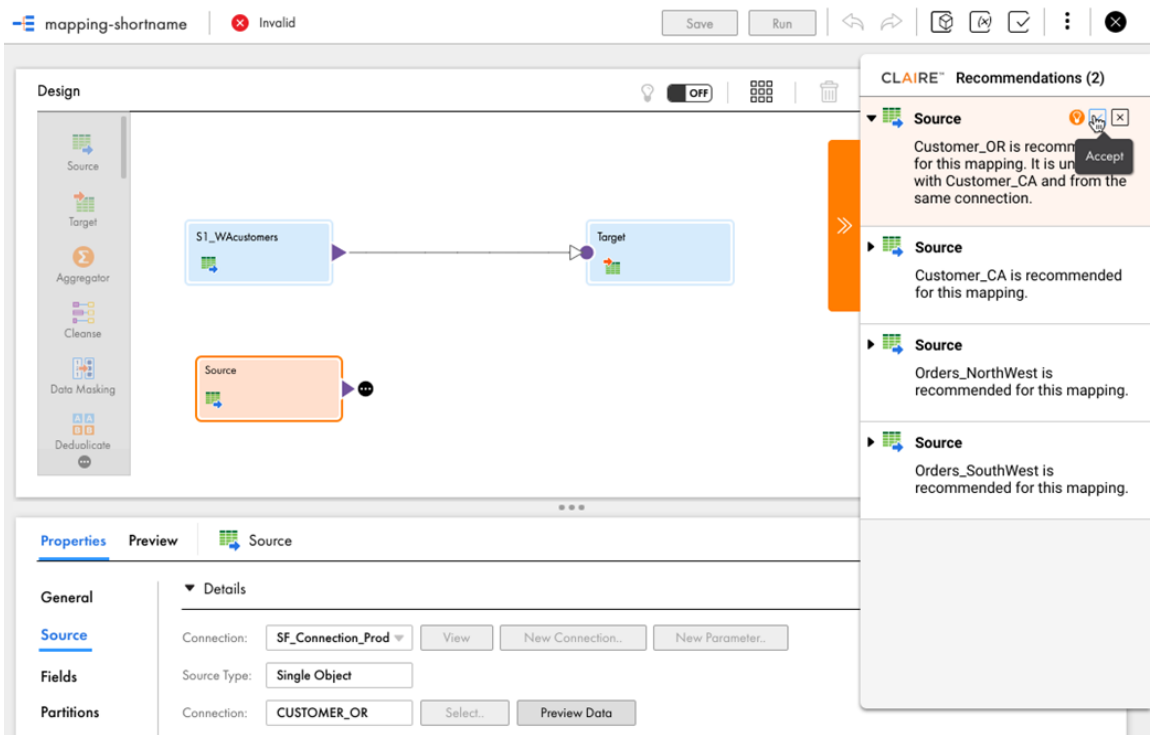



Figure 7. CLAIRE source recommendations.

## Transformation Recommendations

Close the design loop and enhance data engineer productivity with automated data integration mapping creation with predictions for the next transformation and expressions. When an organization opts in to receive CLAIRE-based recommendations, secure metadata from the organization's data pipelines is analyzed, and AI/ML is applied to act as a copilot providing design recommendations. This metadata is used to generate transformation and expression recommendations. CLAIRE becomes better with each utilization — acceptance or rejection of the recommendation. This accelerates development, automates repetitive tasks and enables more types of users to connect and integrate data quickly.

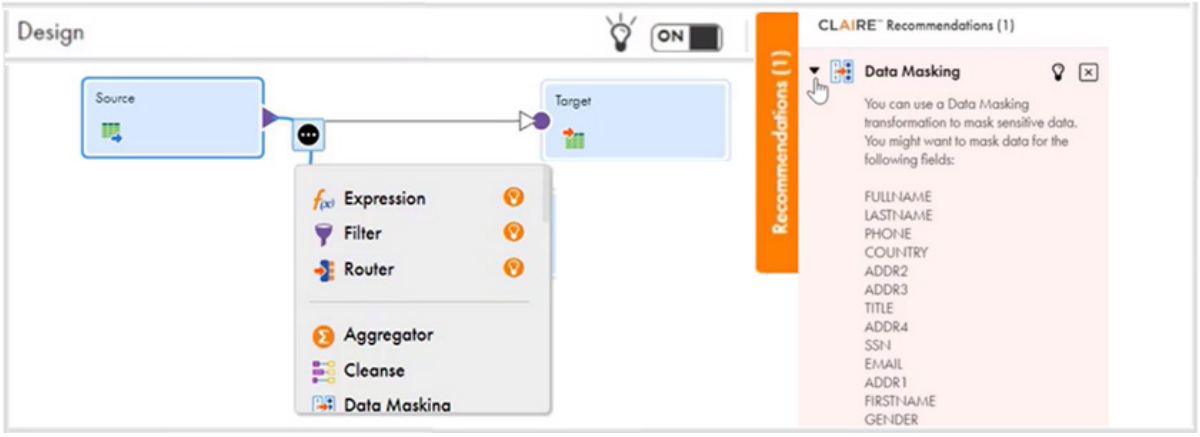



Figure 8: CLAIRE recommends next-best transformations when creating data pipelines.

### Automapping Recommendations

Automapping allows you to map data quickly and accurately. This feature can help reduce errors that come with manual mapping. It matches and aligns data integration elements between different sources and connections. Automapping suggests mappings based on data field names, structures and data types that match. This makes it easier for you to integrate data between different sources.

### Optimized At-Scale Process Execution

CLAIRE as a copilot uses various optimization methods to increase integration performance throughout the data pipeline. An intelligent optimizer decides on the best processing engine to run a big data workload based on performance characteristics; mapping recommendations are presented to data engineers based on past user activities; and a cost-based optimizer plus heuristics intelligently changes the join order in a data pipeline for optimum performance. These are just a few examples of where CLAIRE as your copilot optimizes data pipelines.

### Join-Column Recommendations

Use CLAIRE as your copilot to automatically suggest join columns (i.e., join keys) when a user chooses the action to combine two datasets. This saves data analysts hundreds of hours of manual effort to determine how best to merge datasets into a composite dataset for their analysis.

CLAIRE starts with the primary and foreign key relationships (i.e., Pk-Fk) defined in the source systems (e.g., relational databases such as Oracle) of the datasets imported into the data lake. If the same datasets are joined in other projects, this join column information will also be used for recommendations. This information is processed and ranked by CLAIRE and, as your copilot, it suggests the best join columns between two datasets. Moreover, the overlap percentage of data between the suggested columns is also shown based on sampling the datasets.

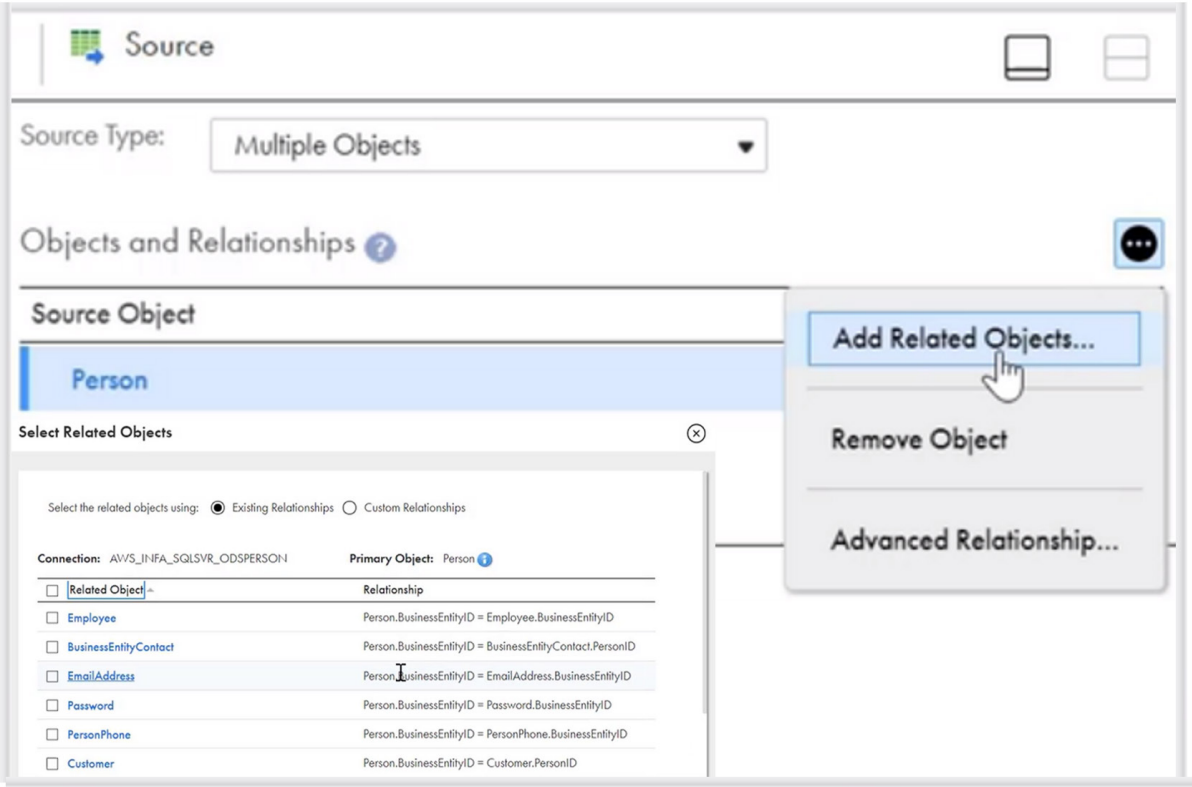



Figure 9: Automatic join-column suggestions when combining two datasets.

## Mapplet/UDF Recommendation

AI-powered recommendations for mapplets and UDF can help users save time and effort. These objects contain a set of transformations that can be reused in multiple data maps, improving the accuracy and efficiency of the data mapping process.

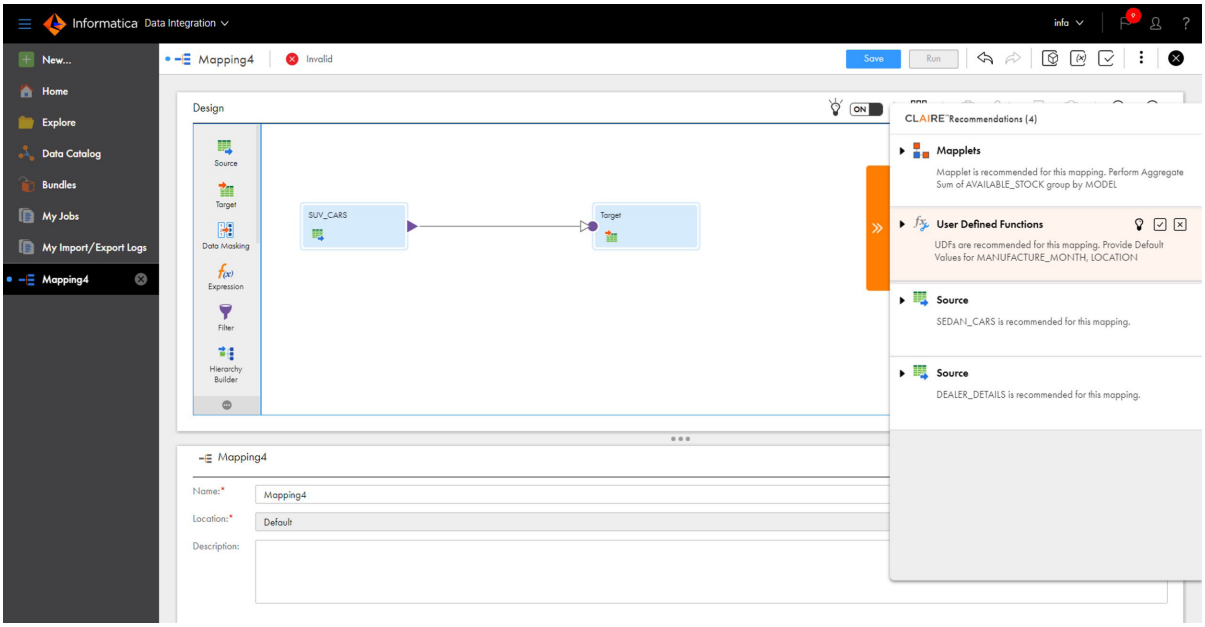



Figure 10: CLAIRE quickly provides mapplets and user-defined functions (UDF) to accelerate development time.

## Data Quality Rule Recommendation

AI-powered data quality rules can be implemented to help guarantee the accuracy and completeness of data. CLAIRE, acting as a copilot, simplifies identifying and applying data quality rules to columns in a dataset, ensuring that the data complies with business rules.

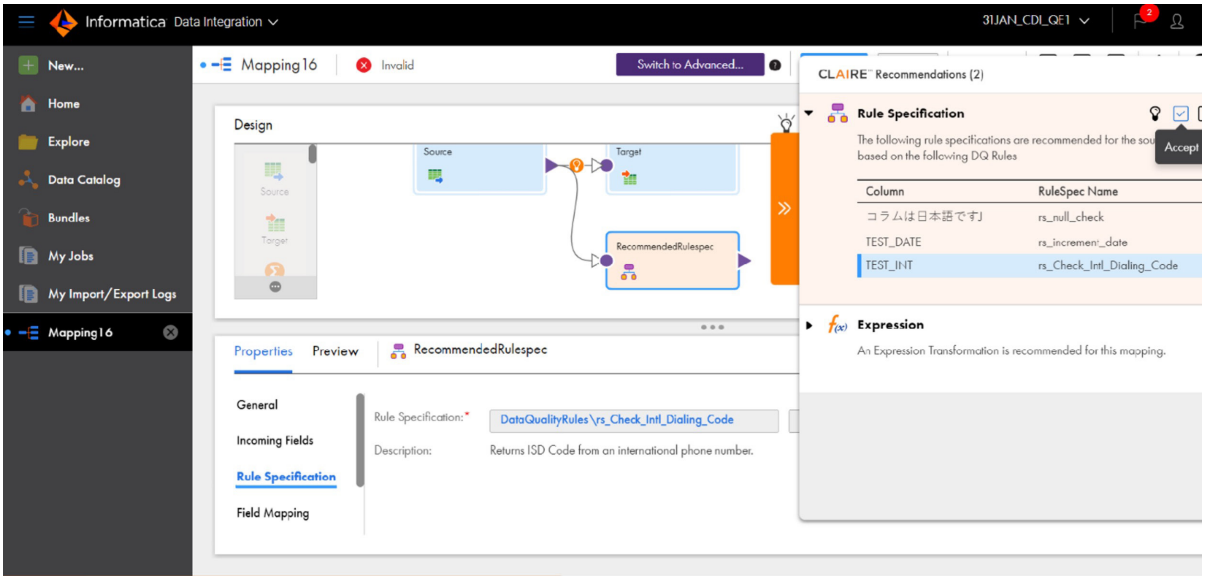



Figure 11: CLAIRE makes data quality rule specifications to help developers ensure their data is fit for use.

### Intelligent Data Recommendations

CLAIRE helps data analysts and scientists by suggesting suitable datasets for their projects. It analyzes the datasets that users have already chosen and recommends other similar, higher-ranked ones or other datasets that complement their choices. This helps users avoid repeating work that others have already done. CLAIRE's data recommendations are intelligent and save users time and effort. The recommendations include:

- A prepared version of the same data (substitutable data)
- Another table containing the same type of records (union-able data)
- A table that might be joined to enrich the data with additional attributes (join-able data)

Data recommendations use content-based filtering techniques to provide suggestions about additional datasets. The characteristics (terms) used for datasets include lineage information, user ranking and data similarity. Several similarity measures are used to score the equivalence of different datasets. Similar datasets are recommended based on users' scores and typical usage patterns obtained from querying the metadata graph.

### Intelligent Structure Discovery

An increasing amount of data is generated and collected across machines, enterprises and applications in unstructured or non-relational formats. These data types are characterized by the large volumes and their velocity, variety and variability. "Data drifting" is a term now commonly used to depict the fluctuation in the format, the pace and the content of data in these new data types.

Informatica Intelligent Structure Discovery (ISD) powered by CLAIRE is designed to automate the file ingestion and onboarding process so enterprises can discover and parse complex files. ISD provides out-of-the-box support for various data file formats, including clickstreams, IoT logs, CSV, text-delimited, XML, JSON, Excel, ORC, Parquet, Avro, PDF forms and Word table files.

With CLAIRE as your copilot, you can automatically derive the structure from these files, making them easier to understand and work with. Using a content-based approach to parsing files, it can adapt to frequent changes without affecting file processing.

ISD uses a genetic algorithm to automate the recognition of patterns in files. This approach uses the concept of "evolution" to improve results. Each candidate solution has a set of properties that can be automatically altered and then tested to determine if they provide a solution with a better fit. The resulting structures are then scored based on several factors, such as input coverage and derived domains. Top-scored structures enter a "mutation" phase where several changes are made to the structures, for example, combining substructures to see if the scores improve. It terminates the process when it determines the appropriate fitness of the structure to the data.

ISD also employs custom ML-powered named entity recognition (NER) and natural language understanding (NLU) mechanisms to identify fields and field types, simplifying following integrations and allowing external applications to use ISD as an underlying NER/NLU engine. For example, ISD detects PII information in incoming and outgoing API payloads, facilitating regulatory compliance and higher enterprise security. ISD is also used in data discovery, ingestion and streaming use cases.

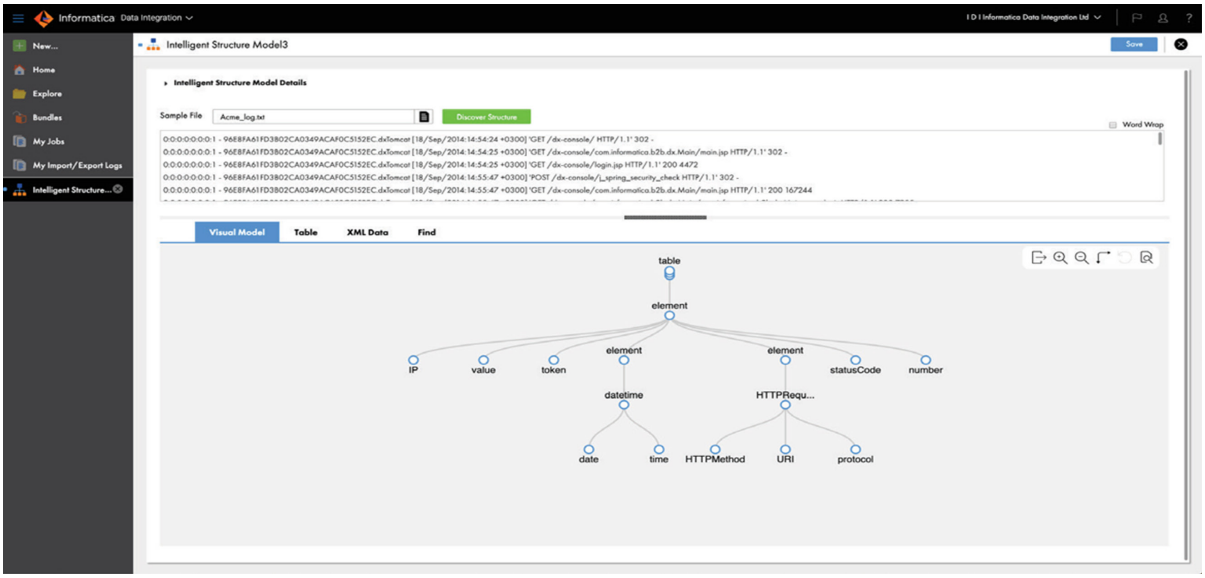



Figure 12: Intelligently finding structure in unstructured data files.

## CLAIRE as Your Auto Tuner

CLAIRE Tuning helps users fine-tune their data integration jobs effortlessly through a simple, user-friendly process that eliminates the need for expert knowledge. This transformative capability makes informed decisions on the optimal property values to deploy. The process begins by initiating the jobs while using specific job properties as a baseline. As the runtime unfolds, insights are gleaned that guide the next value to be tested. This iterative cycle continues, refining the parameters based on the outcomes deserved. After a limited number of iterations, the best-suited values emerge.

CLAIRE employs methodologies like Latin Hypercube Sampling to select the most appropriate values. This technique ensures a comprehensive exploration of potential values, resulting in a well-distributed range of possibilities.

The Smart Hill Climb approach further augments the process, which strikes a balance between exploration and exploitation. This equilibrium permits in-depth exploration of new avenues while maximizing the utilization of valuable data points.

The marriage of data-driven insights with sophisticated techniques like Latin Hypercube Sampling and Smart Hill Climb empowers users to make informed decisions regardless of their technical expertise. This innovation accelerates the optimization process and ensures that the chosen values are well-suited to your specific context.

## CLAIRE as Your Auto Scaler

CLAIRE harnesses the power of advanced ML algorithms, predictively scaling up or down the computing resources required for efficient data processing. The autonomic capability minimizes the risk of over-provisioning or underutilization, which can result in wasted costs and performance bottlenecks. By constantly analyzing historical usage patterns and real-time data flows, CLAIRE learns the intricacies of the system's behavior and understands the correlation between varying workloads and infrastructure requirements.

One of CLAIRE's distinguishing features is its ability to adapt and respond in real time. As incoming data processing tasks surge or diminish, CLAIRE dynamically provisions or de-provisions computing resources, ensuring that the infrastructure matches the immediate workload. This can help not only to guarantee optimal performance but also maximize cost efficiency. For instance, during peak demand periods, such as Black Friday sales or large-scale data ingestion, CLAIRE intelligently allocates additional resources to prevent performance degradation while eliminating the need for manual intervention.

Furthermore, CLAIRE's functionality extends beyond mere scaling. It has an inherent understanding of

the intricacies of different data processing tasks. It evaluates data volume, complexity and processing requirements to make informed decisions. This leads to a more nuanced and effective scaling strategy. CLAIRE's insights also empower organizations to plan resource allocation more effectively, leading to a reduction in operational costs and a boost in overall productivity.

In essence, CLAIRE emerges as an autonomous and intelligent orchestrator of computational resources, eliminating the complexities associated with manual scaling and adapting infrastructure to meet data processing needs. With its machine learning prowess, real-time responsiveness and keen understanding of diverse workloads, CLAIRE presents a transformative leap forward in infrastructure management, aligning technological resources with business goals seamlessly and efficiently.

## CLAIRE as Your Master Data Management Copilot

Copiloting with CLAIRE-powered automation and intelligence using advanced AI/ML enriches and improves the accuracy of 360 views for customers, products, suppliers and other domains. Various blended AI/ML techniques ranging from deterministic, heuristic and probabilistic algorithms to contextual synthesis matching and active learning entity matching are employed to provide rapid, scalable and highly accurate record matching and enrichment of master data.

### Identity Matching

CLAIRE's NAME3 identity matching encapsulates 30-plus years of training and tuning using a variety of techniques such as intelligent key generation for indexing and blocking, semantic text stabilization and comparison of party and location data, edit lists and text stabilization rules for 80 populations and intelligent weighting of feature importance for different purposes. These powerful techniques enable indexing and blocking on multiple fields, client-defined match and anti-match rules given requirements and implementation-defined match and anti-match rules to complement other AI rules.

### Entity Matching

Entity matching finds data records that refer to the same real-world entity (e.g., customers, products, etc.). Data records can be unstructured (e.g., customer information hidden in a web chat) and structured. Match classification compares a matching pair and determines whether there is a match, maybe match or non-match, along with a confidence level. Some techniques use human-configured rules (i.e., declarative rules) or AI rules (i.e., a machine-learned configuration). The best matching results are achieved when these two techniques are blended.

CLAIRE employs declarative rules created by subject-matter experts and powerful AI rules in the form of

a learned random forest classifier. To accelerate the AI training process, CLAIRE uses supervised active learning, where a user is presented with micro batches of 10 or 20 match pairs for labeling (i.e., match, maybe-match, no-match). Once labeled, CLAIRE retrains the random forest classifier and determines the next-best match pairs for further labeling.

This iterative process enables CLAIRE to infer blocking rules, remove obvious non-matches, perform blocking, train a model and perform entity matching. It is important to note that CLAIRE uses supervised active learning rather than crowd-sourced or multi-user learning.

CLAIRE as your copilot uses a combination of string comparisons/similarities such as Jaccard and cosine similarity, declarative rules derived from data profiling and stabilized datasets (population files, nicknames, semantic comparisons, etc.). These declarative rules address gaps and exceptions and help accelerate the active-learning training process (i.e., reduce the number of match pairs required for learning), accelerate AI rule feature building and increase match accuracy. For example, when a name, birthdate and social security number compare strongly, the rule classifies that as a match. This blending of declarative and AI rules accelerates training and improves the matching accuracy.

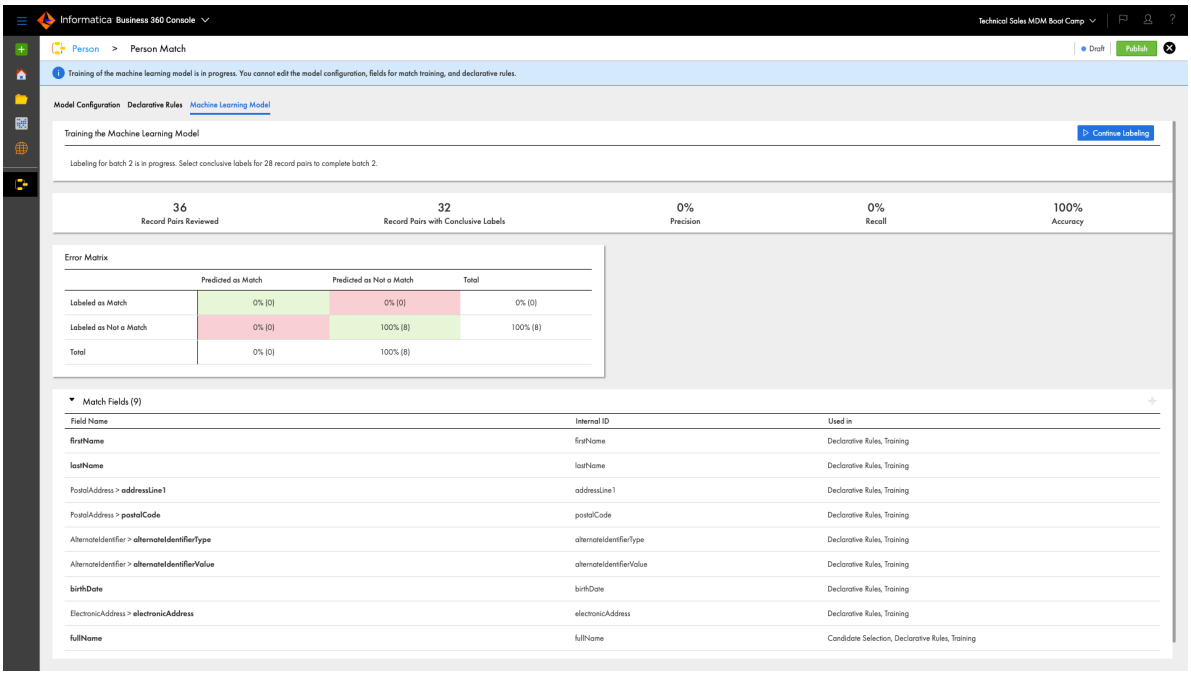



Figure 13: CLAIRE blends declarative and AI rules to accelerate training and improve matching accuracy.

## CLAIRE as Your Data Governance and Data Quality Copilot

AI and ML are essential to automating today's most challenging data governance tasks: finding data, measuring its quality and enabling collaboration to help govern it. CLAIRE automatically generates policy rules (e.g., data quality), ties business semantics to technical metadata and helps copilot users with the most relevant and trusted data for their business needs.

### Insights From Profiling Results

Insights provided by our copilot CLAIRE offer a method for automatically discovering data quality issues. These can range from anomalous data values to complex inconsistencies. Insights in data profiling automate the process of detecting data quality issues.

The CLAIRE engine is designed to provide insights and recommendations on your data. You have the power to approve or reject these recommendations. CLAIRE can automatically run on profiles, allowing you to detect data quality issues. If you approve the recommendations, the data quality and profiling services will automatically create and apply data quality rules to your profile. This will allow you to monitor quality issues and act as needed.

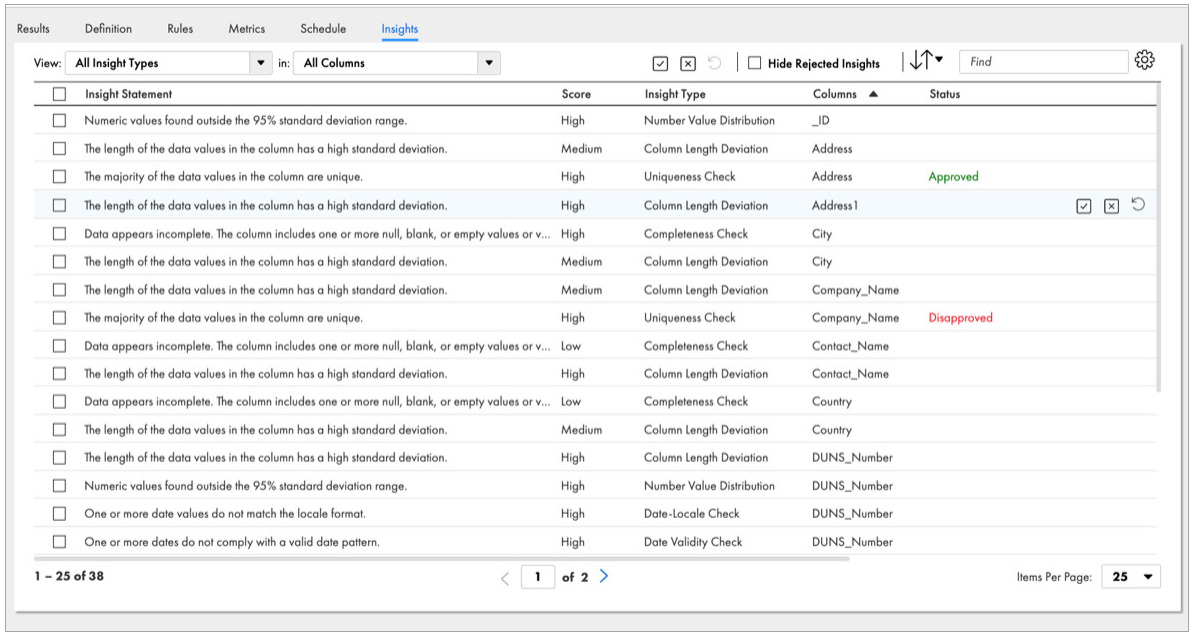



Figure 14: Automated data quality insights for approval and automation of applicable rules.

## Automatically Associate Business Terms With Physical Datasets

Data governance requires the documentation of business artifacts, definitions, stakeholders, processes, policies and more. To enable a truly aligned view, users must be able to associate definitions and business views to the underlying technical implementations in their data estate. Typically, this task is slow, laborious and error-prone — relying on key people to communicate and manually line up technical manifestations one by one — a task that can take days, weeks or even months to complete.

**Informatica Data Governance and Catalog** can shortcut this process. With CLAIRE as your copilot, users are provided recommendations of relevant and appropriate data elements to be linked as metadata scans are completed. This cuts down the task of searching for, validating and linking data elements, allowing data stewards and the data governance office to focus on their critical tasks. As implementations progress, the process can be completely automated.

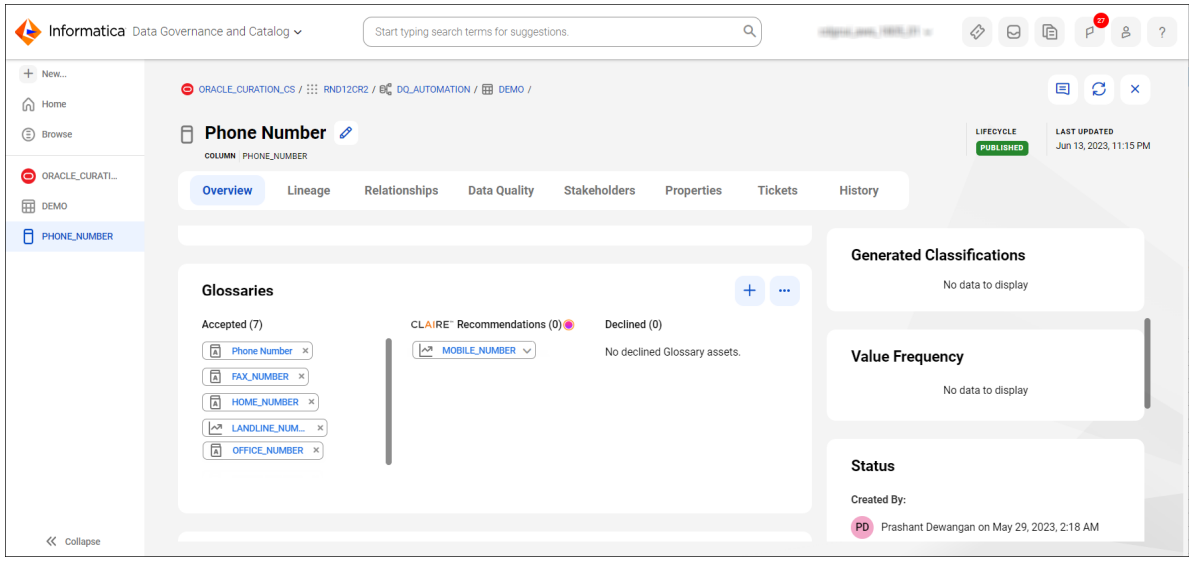



Figure 15: Automatic association of business terms with physical datasets.

### Automatically Assess Data Quality

A key performance indicator (KPI) in data governance is data quality throughout a system that supports a process, underpins policies and so on. The data governance office needs to ensure data is complete, accurate, consistent, valid and more. In short, it must be trustworthy and good enough to support the business operations. As data governance implementations grow, assessing quality for an increasing number of systems and fields across the data landscape, from databases to data lakes, becomes increasingly time-consuming.

With the help of your CLAIRE AI copilot, Cloud Data Governance and Catalog — in coordination with **Informatica Cloud Data Quality** — can automate the application of data quality measurements across the enterprise, saving thousands of hours of work. The data governance team relates data quality rules for various data quality dimensions to business terms and critical data elements, and the underlying system then generates the required data quality checks on the different systems. It reports the metrics back to the governance office.

This automation is enabled by combining three key pieces of information:

1. Knowledge of critical business elements and data quality rules required
2. Portable and executable data quality rules and a flexible execution engine from Cloud Data Quality
3. Metadata details from physical data assets from the data catalog

CLAIRE combines this information to generate data quality rule execution jobs in Cloud Data Quality against the physical data assets from Cloud Data Governance and Catalog. CLAIRE also maintains the business user context from Cloud Data Governance and Catalog to ensure the results are displayed in the correct dashboards and aggregated views for consumption by the governance office.

The automation enables governance programs to scale faster than ever, removing thousands of hours of manual labor associated with creating data quality assessments and linking them to the governance context individually. Using CLAIRE as a copilot also helps ensure that when new physical assets are identified, they can be automatically assessed for quality. In addition, new domains are discovered using named entity extraction or classifier in data quality rules.

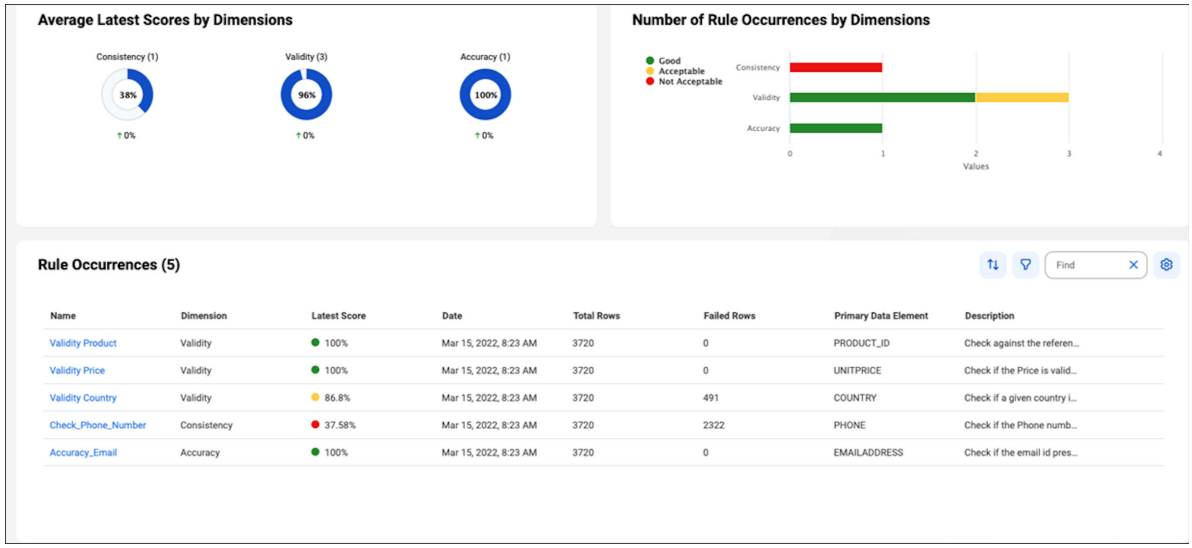



Figure 16: Automatic data quality assessments across the entire data estate save thousands of hours of manual labor.

## ML/NLP-Assisted Data Quality Rule Creation

Data quality is a crucial imperative for a data governance program, and in larger implementations, there can be many data quality rules. To help data stewards identify the correct rules to use, your AI copilot CLAIRE can help identify rules and generate any missing ones.

A Cloud Data Governance and Catalog user can specify their rule requirement in plain text (for example: “Customer Identifiers must have eight characters and start with C”) and invoke CLAIRE to help. CLAIRE will analyze the user requirement through ML and NLP techniques and translate it into a technical representation. Based on this representation and associated metadata (for example, Glossary Term name), CLAIRE will automatically generate a new data quality rule to satisfy the Informatica Data Quality repository requirement and link it back to the Cloud Data Governance context.

In addition, CLAIRE automatically associates data quality rules to cloud profiles based on the column and source object name match. By default, Data Profiling associates rules with columns of Oracle, flat file, ODBC and Amazon S3 V2 connections. As users create new profiles against core objects from one of these sources, CLAIRE will automatically suggest best-practice data quality rules that should be applied to the measurement.

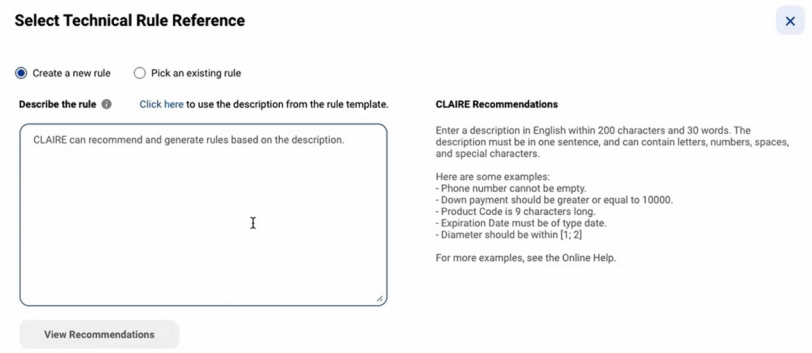



Figure 17: Automatic data quality rule creation using NLP.

A large global healthcare company had a full-time employee mapping 21,000 technical assets with 6,000 business terms, a process that took two months. With data governance and data catalog, CLAIRE automated the mapping of 18,000 technical assets with 99% accuracy in 8 minutes.

## CLAIRE as Your DataOps Copilot

With CLAIRE assisting you, organizations can accelerate data processing pipelines, automating many aspects of data management for continuous integration (CI) and continuous delivery (CD) related to DataOps.

### Insightful and Predictive Analytics for Data Management Environments

Operational analytics helps understand the current usage of existing projects and resources and plan for future capacity. It offers parameters for building charge-back models while supporting multiple lines of business (LOBs) on a single data management platform. Based on continuous observation of resource utilization trends, data-volume processing projections are offered to help with capacity planning. CLAIRE takes this to the next step by auto-scaling data management runtime resources.

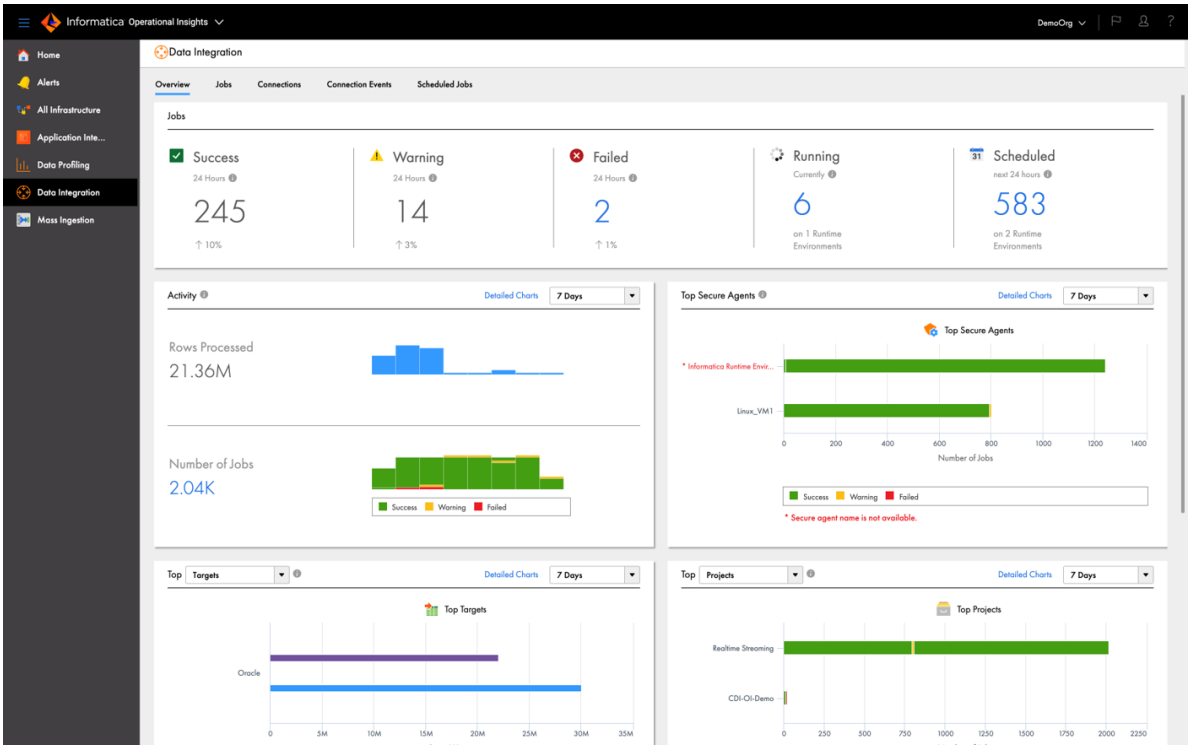



Figure 18 Operational Insights resource utilization for Informatica domain processes.

## Anomaly Detection in Job Runs

CLAIRE automatically detects anomalies related to job run times, data processed, data loaded, resources consumed, throughput and more. Automatically detecting these anomalies helps IT proactively fix issues with data integration jobs before impacting downstream business processes. The Seasonal Hybrid ESD algorithm is used to detect anomalies in job-run behavior. This algorithm considers seasonality (month-end peak load, holiday season, etc.) and weeds out jobs with expected aberrations induced by business cycles.

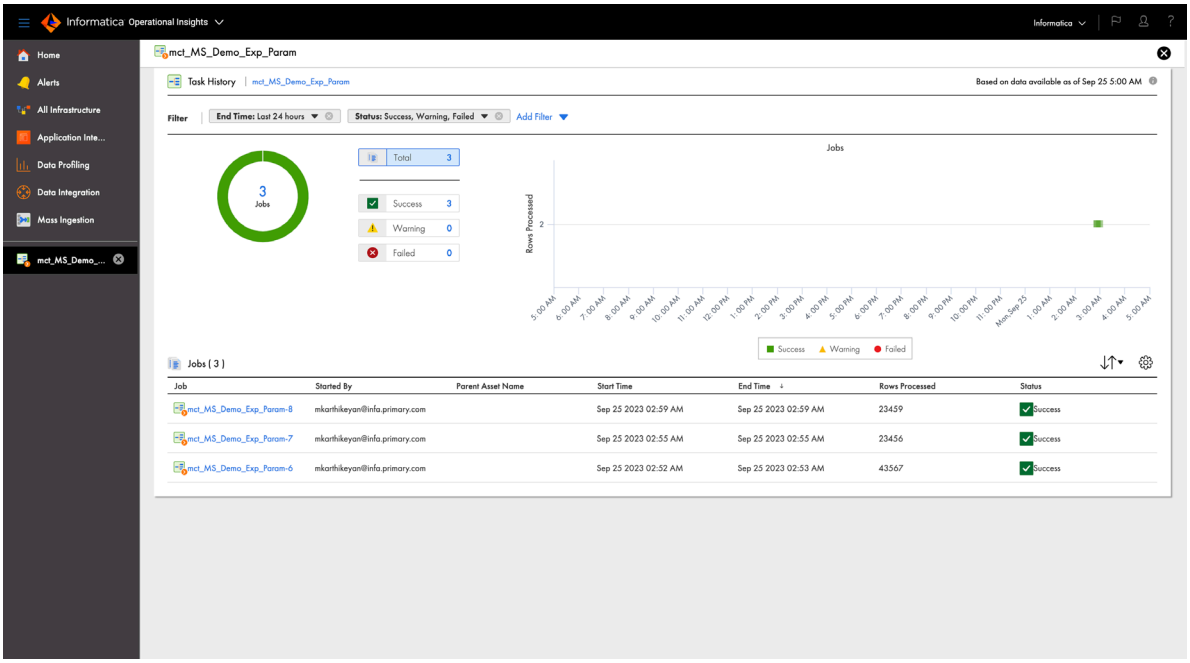



Figure 19: CLAIRE automatically detects anomalies related to Informatica jobs and data processing.

## CLAIRE GPT

The next-gen version of CLAIRE is delivering the industry's first generative AI-powered data management solution that will transform and redefine the users' data management and data consumption experience. With CLAIRE GPT as the primary interface for data management, users can interact with and manage their data through a natural language interface and create rapid first drafts of mappings, quality rules and governance artifacts so that the data team can refine and implement across the enterprise.

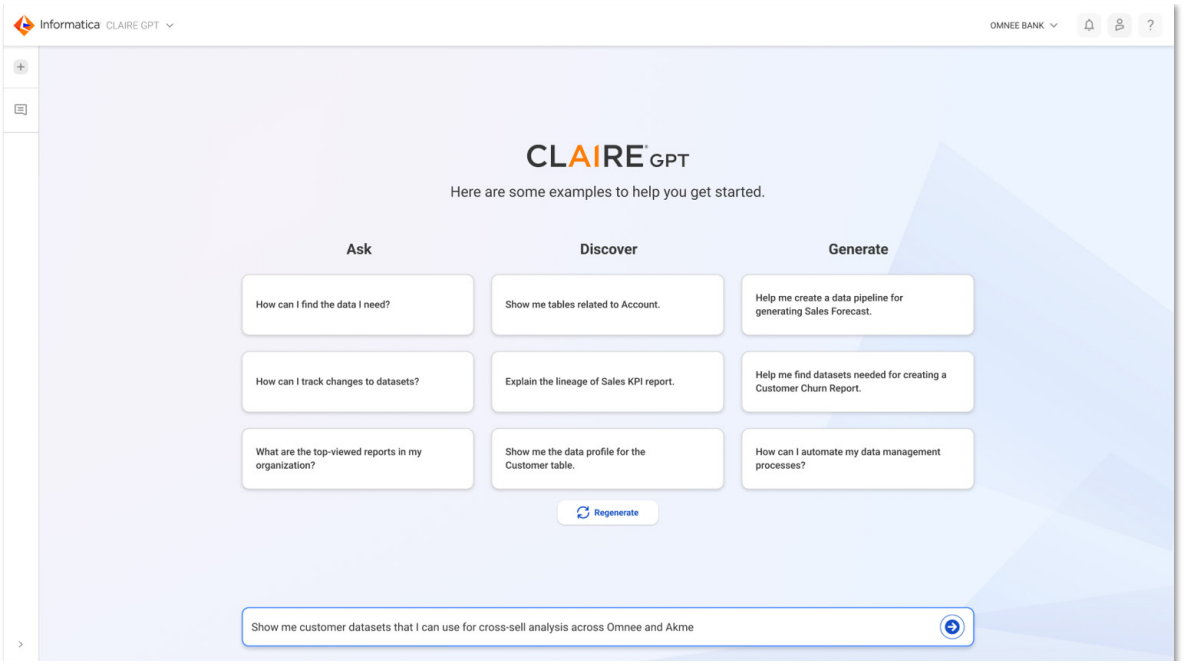



Figure 20: CLAIRE automatically detects anomalies related to Informatica jobs and data processing.

### Simplify Data Access and Insights

With CLAIRE GPT, data consumers can quickly discover data assets, explore metadata, find profiling stats, ask data questions and craft initial data pipelines through a simplified Natural Language interface. Interacting with and managing data through a text-to-data-management interface democratizes access to data across the enterprise, extending to a broader data consumer community, thus accelerating a data-driven culture.

### Automate Repetitive Tasks

Improve quality and transparency of data by automating repetitive tasks such as debugging, testing, refactoring and documentation. CLAIRE GPT helps data engineers quickly explore the datasets and find data quality issues. It also helps document and test the pipeline.

### Reduce Data Management Costs

Use CLAIRE GPT for knowledge that normally requires input from a subject matter expert (SME). CLAIRE GPT serves as an assistant SME helping data engineers understand the business context. CLAIRE GPT helps simplify, accelerate and optimize data management operations, driving enormous gains in productivity for data teams.

## CLAIRE in the Future

As CLAIRE develops, it will continue to increase productivity and efficiency, enabling data leaders to leverage intelligent automation for faster, better insights and more effective data management. Future capabilities include:

1. Self-integration. Automatically integrate newly arriving data into the data integration processes. Identify data, locate integration patterns that process similar data and automatically transform and move data with learnings from millions of existing mappings and user actions.

2. Development assistance. Provide recommendations to users and suggest next-best-actions during the development process, including:
   - Transformation auto-completion
   - Template recommendations
   - Masking-type suggestions for sensitive data
   - Data quality suggestions for cleansing and standardization
   - Automatic performance optimizations

3. Automapping. Detect master data entities across the enterprise and automatically map them to the master data model, applying the requisite transformations and quality rules.

4. Auto-enrichment: Add context to data assets automatically for relevant discovery and wider data understanding and utilization.

5. Auto-data quality (DQ). Automate the discovery of data quality issues as well as the application of solutions for these issues based on CLAIRE-generated recommendations.

6. Self-heal. Handle external system issues such as low memory or compute power gracefully. For example, add additional compute ('burst to cloud') to deal with spikes in data

7. Self-tune. Based on historical information, current data volumes and available system resources, predict and adjust schedules or compute resources to meet performance criteria

8. Self-secure. Automatically detect sensitive data and mask it before it leaves a secure region

## Conclusion

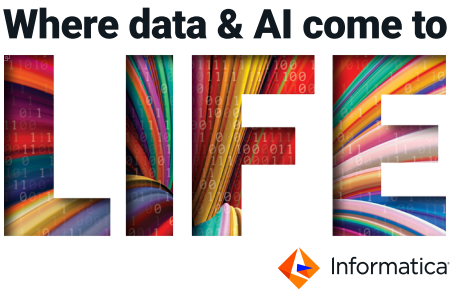
Today's data-centric business strategies are built on a foundation of data. Winning requires building a competence in data management to unleash the power of data. With all the challenges that data management presents under ordinary circumstances, traditional approaches can't scale to meet today's requirements — to say nothing of tomorrow's.

One way to leverage your data to drive disruption is to standardize on an end-to-end data management platform that uses the power of data, metadata and CLAIRE as your ML/AI copilot. The result will be increased productivity for virtually all users of the platform, including technical, operational, business and business self-service users.

Contact us to learn more about how you can use CLAIRE and the Intelligent Data Management Cloud to harness the power of your data.

# About Us

Informatica (NYSE: INFA) brings data and AI to life by empowering businesses to realize the transformative power of their most critical assets. When properly unlocked, data becomes a living and trusted resource that is democratized across your organization, turning chaos into clarity. Through the Informatica Intelligent Data Management Cloud™, companies are breathing life into their data to drive bigger ideas, create improved processes, and reduce costs. Powered by CLAIRE®, our AI engine, it's the only cloud dedicated to managing data of any type, pattern, complexity, or workload across any location — all on a single platform.




Worldwide Headquarters
2100 Seaport Blvd,
Redwood City, CA 94063, USA
Phone: 650.385.5000
Toll-free in the US: 1.800.653.3871

**informatica.com**
**linkedin.com/company/informatica**
**twitter.com/Informatica**

CONTACT US

IN09-3328-1023

© Copyright Informatica LLC 2023. Informatica and the Informatica logo are trademarks or registered trademarks of Informatica LLC in the United States and other countries. A current list of Informatica trademarks is available on the web at **https://www.informatica.com/trademarks.html**. Other company and product names may be trade names or trademarks of their respective owners. The information in this documentation is subject to change without notice and provided "AS IS" without warranty of any kind, express or implied.