Informatica

White Paper

Artificial Intelligence for the Data-Driven Intelligent Enterprise

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

About Informatica

Digital transformation changes expectations: better service, faster delivery, with less cost. Businesses must transform to stay relevant and data holds the answers.

As the world's leader in Enterprise Cloud Data Management, we're prepared to help you intelligently lead—in any sector, category or niche. Informatica provides you with the foresight to become more agile, realize new growth opportunities or create new inventions. With 100% focus on everything data, we offer the versatility needed to succeed.

We invite you to explore all that Informatica has to offer—and unleash the power of data to drive your next intelligent disruption.

Table of Contents

"Data and analytics leaders face complexity in their data landscape. Our data management solution predictions acknowledge key developments and growing demand for cloud capabilities, connected data architectures, metadata and the automation of routine and nonroutine tasks through application of AI."[1]

— Gartner

The Importance of AI

Artificial intelligence (AI) and machine learning (ML) are powering the digital transformations happening in every industry around the world. AI is top of mind for boardroom executives as a strategy to transform their businesses. And it has become pervasive in enhancing our daily life, from the movies we watch to the cars we drive. AI/ML is critical in discovering new therapies in life sciences, reducing fraud and risk in financial services, and delivering truly 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—be it for next-best actions, customer satisfaction tracking, efficient operations, and innovative products. Machine learning in general, and especially deep learning, is data hungry. To get the accuracy required, ML needs vast amounts of data for training. This data must be an accurate reflection of the current state of business. AI trained with bad or limited data will have a terrible impact on business initiatives, to the point where it has a reverse impact on the desired outcome.

For effective AI, in which the right features are used and trained for, we need to tap into a wide variety of data from inside and outside the organization. This data must be brought together in a way an ML model can be built and trained. This needs data management. Not only is it a question of dealing with the scale and complexity, it is also about trust. Is the data being used to train the model coming from the right systems? Have we removed personally identifiable information (PII) and adhered to all regulations? Are we transparent, and can we prove the lineage of the data that the model is using? Can we document and be ready to show regulators or investigators that there is no bias in the data? All this requires good control and a basis of data management. Without a solid foundation of data management, AI is incomprehensible and unreliable—in other words, without data management, AI can be a black box that has unintended consequences.

AI Needs Data

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

Why do data scientists tasked with building AI/ML models need high-quality data? Take, for example, a prediction model 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 both 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—all the more reason to get it right. We should all expect "explainable AI" to become a regulated mandate, not just an option. Without metadata-driven lineage and traceability, AI-powered applications and insights cannot be deployed into production.

[1] Gartner, Predicts 2020: Data Management Solutions, Rick Greenwald, Donald Feinberg, Mark Beyer, Adam Ronthal, Melody Chien, 5 December 2019.

AI needs intelligent data management to quickly find all the features for the model; automatically transform data to meet the needs of the AI model (feature scaling, standardization, etc.); deduplicate data and provide trusted master data about customers, patients, partners, and products; and provide end-to-end lineage of the data, including within the model and its operations. The success of AI is dependent on the effectiveness of the models designed by data scientists to train and scale it. And the success of those models is dependent on the availability of trusted and timely data.

Data Needs AI

AI/ML also plays a critical role in scaling the practices of data management. Due to the massive volumes of data needed for digital transformation, 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: use poor-quality data, deliver untrustworthy insights.

According to recent research, the overall volume of data center traffic is expected to reach 20.6 zettabytes in 2021, while the number of connected devices and connections is projected to reach more than 25 billion by 2022[2]. All this data needs to be processed and made usable and trustworthy while adhering to governance policies. Adding to all this is the requirement to move quickly and respond to changes in businesses strategy and processes. The effort involved in preparing the data for digital transformation initiatives has gone up in complexity, in step with the amount of data growth. According to LinkedIn, the position of data scientist is one of the most promising jobs in the U.S.[3] And the number of data engineers sought by companies has recently seen a 96% year-over-year change.[4] But hiring alone is not enough to manage the increase in data volume.

Don't Take a Linear Approach to an Exponential Challenge

We cannot solve these challenges by simply throwing more engineers and developers at the problem—this is not an issue that can be solved at 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 learning, processing is expensive, and governance and privacy steps are repeated over and over again. So how can organizations move at the speed of business, enable self-service, better serve their customers, increase operational efficiency, and rapidly innovate?

[2] Cisco, Global Cloud Index Forecast and Complete Visual Networking Index Forecast.
[3] LinkedIn, "LinkedIn's Most Promising Jobs of 2019."
[4] Datanmi, "Data Engineering Continues to Move the Employment Needle."

This is where AI shines. AI can automate and simplify tasks related to data management—across discovery, integration, cleansing, governance, and mastering of data. Machine learning methods 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 to developers, analysts, stewards, and business users, speeding up tasks through automation and augmentation with recommendations and next-best actions.

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. CLAIRE leverages all enterprise unified metadata to automate and scale routine data management and stewardship tasks.

Four Major Benefits of AI for Data Management

In general, AI benefits data management teams in four major ways: improving the productivity of data professionals, improving the efficiency of operations, providing a more intelligently guided data experience and deeper 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 every day. The monitoring of these processes is largely 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 and proactively flag outlier values, as well as predict issues that may occur if not handled ahead of time.

Data experience: 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 documentation on primary-key/foreign-key relationships 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: A common but tedious step in data governance is to associate business terms to 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 a combination of natural language processing (NLP) techniques and business-type identification. This can dramatically reduce the drudgery of this error-prone task. In this era of cloud, it is important to note that this approach works for SaaS applications as well. 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 a banking example.

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 cost of wealth advisors
- Reduce the cost of project-related regulatory compliance

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

AI, as it relates to data management, 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 channels, whether online, mobile, or phone. A 360-degree view of POIs helps banks discover patterns and networks of fraudulent activity much faster, potentially saving millions.

And 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 analytic models and algorithms. More and better data, combined with AI/ML and advanced analytics, has been proven to yield significant results, such as improving next-best offers and identifying fraud.

AI also powers data governance that ensures policies are not just documented but actually 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 "Intelligence" in the Intelligent Data Management Cloud

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

1. The Intelligent Data Management Cloud™: We have delivered an integrated, end-to-end cloud-native data management platform for maximum productivity. By providing unified connectivity, metadata, and operations management, the unified platform accelerates the development and deployment of new data management projects. The platform provides a powerful and consistent set of capabilities for managing data across on-premises, cloud, multi-cloud, and multi-hybrid sources. We call this unified data management platform the Intelligent Data Management Cloud.

This platform is modular: Start with any single tool and grow at your own pace:

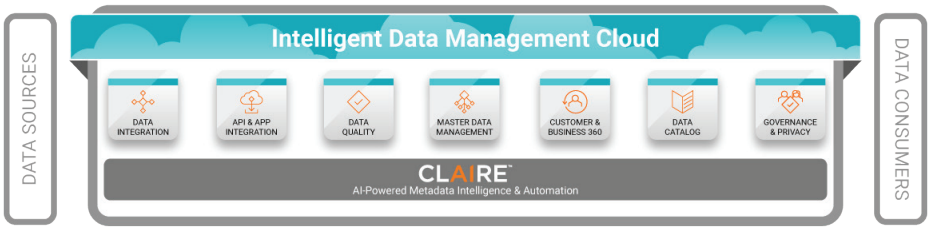

Figure 1: The Intelligent Data Management Cloud integrates data management capabilities with shared connectivity, operational insight, and data and metadata intelligence.

2. Metadata: Informatica has long been known as a leader for its management of technical and business metadata. Informatica now has increased its 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 last 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 machine learning. It provides datasets that are used to "train" the machine-learning algorithms and enables them to adjust and produce better results.

3. Intelligence: Informatica is delivering an integrated combination of metadata and AI/machine learning 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 what drives the intelligence of all the data management capabilities in the Intelligent Data Management Cloud.

CLAIRE helps a wide 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 delivered by CLAIRE is being used today.

CLAIRE for Data Cataloging

Discovering and understanding the data you have is the first step on any data-driven initiative. CLAIRE provides a machine-learning based discovery engine to scan and catalog data assets across the enterprise. An intelligent data catalog powered by CLAIRE can help data scientists, analysts, and data engineers find and recommend the data they need, significantly reducing the time spent in data discovery and preparation.

Advanced Relationship Discovery

One key data cataloging and data modeling task is to document relationships between datasets. CLAIRE uses machine-learning techniques to automatically identify primary keys, 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—e.g., users can accept or reject inferred relationships and CLAIRE learns from these actions.

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, it is difficult to find such datasets across departments and data stores. CLAIRE uses documented joins in the

databases, joins performed in other 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 all the available information to come up with the right target audience for the campaign.

CLAIRE combines multiple techniques for key and join discovery. For keys, profiling statistics like uniqueness, null counts, column metadata (e.g., column names containing "ID") and others are combined to discover primary and unique keys. Joins and join key inference then use a combination of machine-learning techniques like column signature analysis to discover joins at scale across many potential datasets.

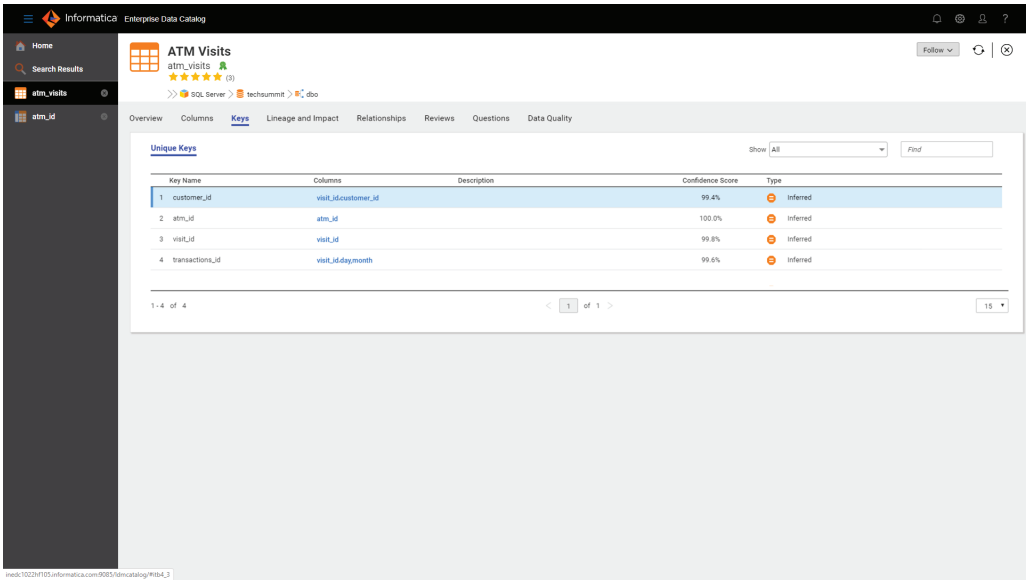

Figure 2: Discovering unique keys through inference using machine-learning techniques.

Intelligent Data Similarity

CLAIRE uses machine-learning techniques like clustering to detect similar data across thousands of databases and file sets. Intelligent data similarity is one of the key capabilities used for multiple purposes, including identifying data, detecting duplicates, combining individual data fields into business entities, propagating tags across datasets, and recommending datasets to users.

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 on the basis of column features. Then, data overlap is computed for unique values in each of these clusters. Finally, the most promising pairs are chosen for computing data similarity using the Bray-Curtis and Jaccard coefficients.

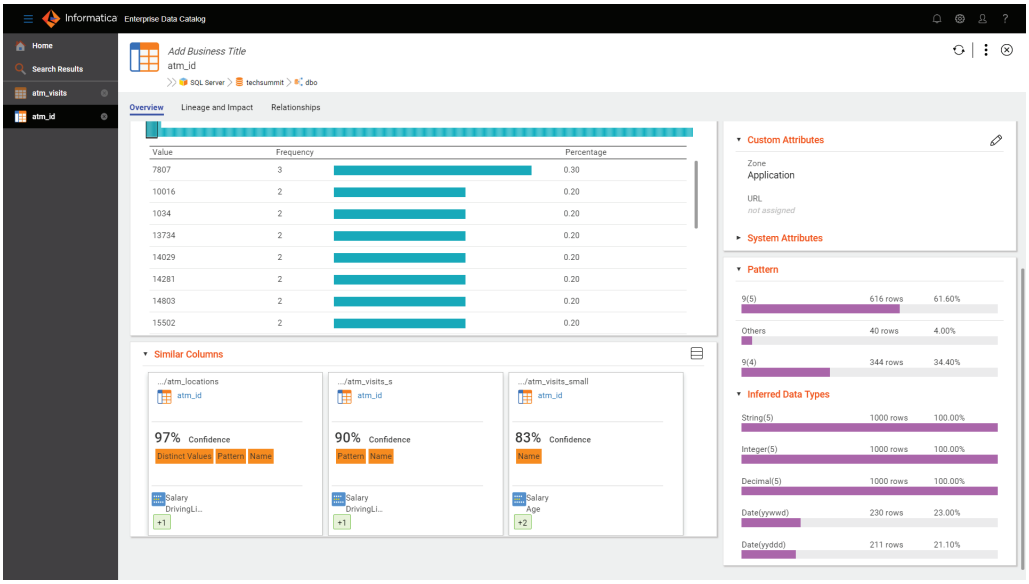

Figure 3: Identifying similar columns using clustering and the Bray-Curtis and Jaccard coefficients.

Intelligent Domain Discovery With Tags

CLAIRE is capable of classifying data fields by applying semantic labels to each column. These semantic labels are called data domains.

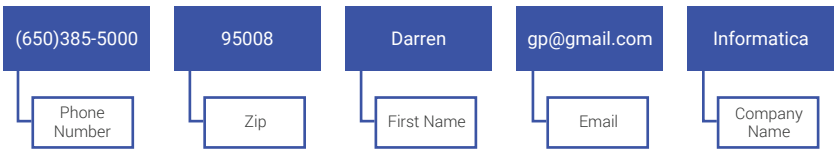

Figure 4: CLAIRE automatically classifies data fields and applies semantic labels called tags.

Usually semantic labels are applied by evaluating rules based on regular expressions, reference tables, or other complex hand-coded logic. Defining and maintaining thousands of such rules is tedious.

CLAIRE instead uses the concept of tags to dramatically simplify the process of discovering and labeling the data fields. For those columns not already classified, the user just 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 photos.

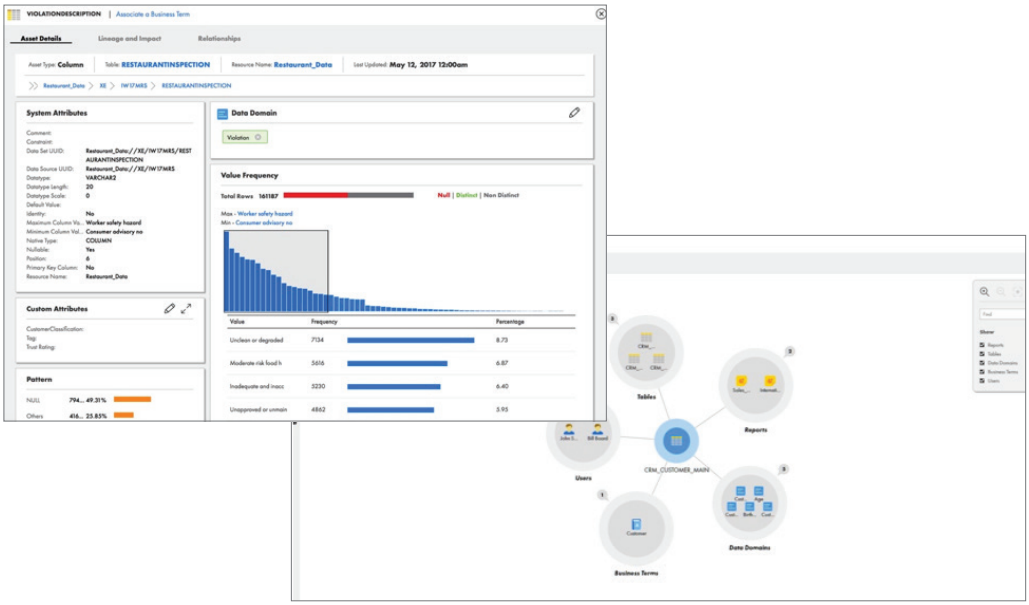

Figure 5: Automatic classification of data.

Intelligent Entity Discovery

Once domains for columns have been identified, CLAIRE can assemble these individual fields into higher-level business entities. The example below shows how an entity called Purchase Order is created by combining fields identified as Customer and as Product. Entity discovery learns from 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.

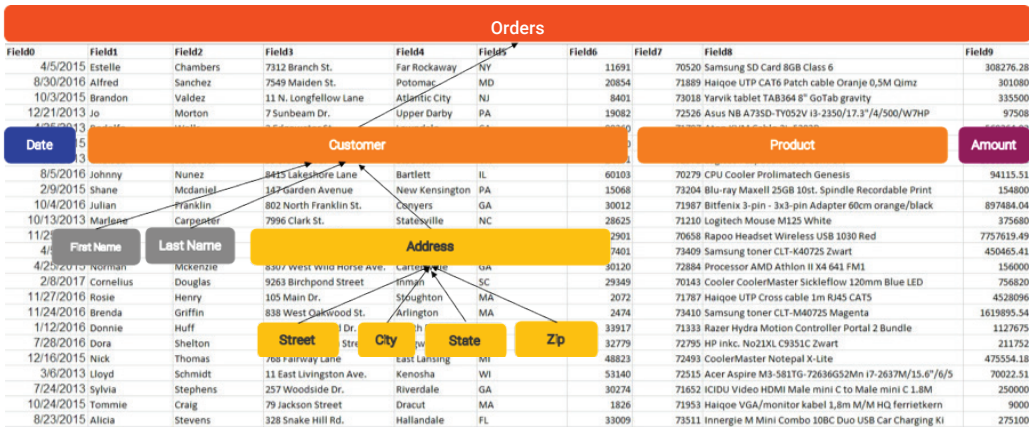

Figure 6: Combining data domains to detect entities from tables and files.

CLAIRE for Analytics

CLAIRE-powered automation and intelligence significantly speeds up analytic insights and processes, increases data availability, and streamlines data preparation for analytics. CLAIRE enhances data engineering productivity with data pipeline recommendations and the ability to parse complex, multi-structured data automatically.

Transformation Recommendations

Close the design loop and enhance data engineer productivity with automated data integration mapping creation with predictions for next transformation and expressions. When an organization opts in to receive CLAIRE-based recommendations, anonymous metadata from the organization's data pipelines is analyzed and AI/ML is applied to offer 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 quickly connect and integrate data.

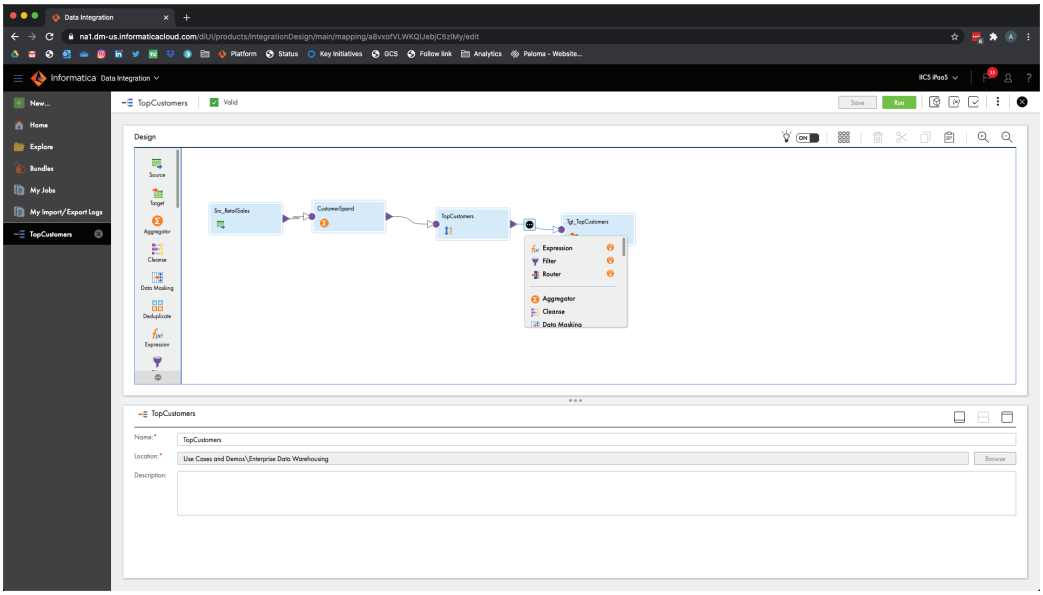

Figure 7: CLAIRE recommends next-best transformations when creating data pipelines.

Optimized At-Scale Process Execution

CLAIRE uses a variety of optimization methods to increase performance of integration throughout the data pipeline. A smart 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 change the join order in a data pipeline for optimum performance. These are just a few examples of how CLAIRE optimizes data pipelines.

Join-Column Recommendations

CLAIRE automatically suggests 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 in trying 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 original source systems (e.g., relational databases such as Oracle) of the datasets which were imported into the data lake. If the same datasets are joined in other projects, this join column information will also be used for recommendations. All of this information is processed and ranked by CLAIRE to suggest the best join columns between two datasets. Moreover, based on sampling the datasets, the overlap percentage of data between the suggested columns is also shown.

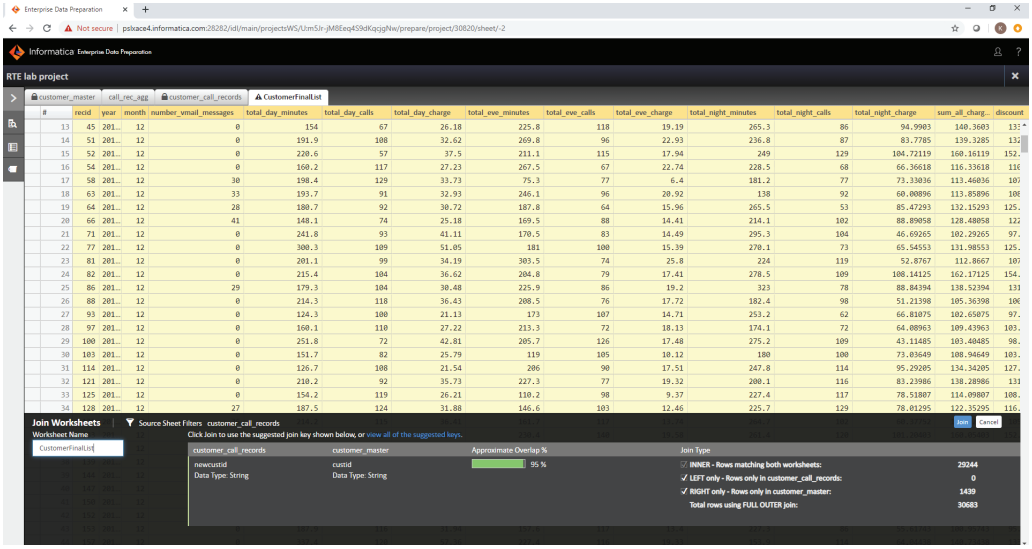

Figure 8: Automatic join-column suggestions when combining two datasets.

Apache Zeppelin Visualization Recommendations

Informatica Enterprise Data Preparation uses Apache Zeppelin to view the worksheets in the form of a notebook that contains graphs and charts. When the user opens the notebook of a publication, the user can see CLAIRE visualization recommendations. When the user opens the notebook for the first time following its publication, the user sees histogram visualizations of derived numeric columns. If the publication does not contain any derived numerical columns, the user sees a “SELECT * FROM” table query in the first paragraph of the notebook.

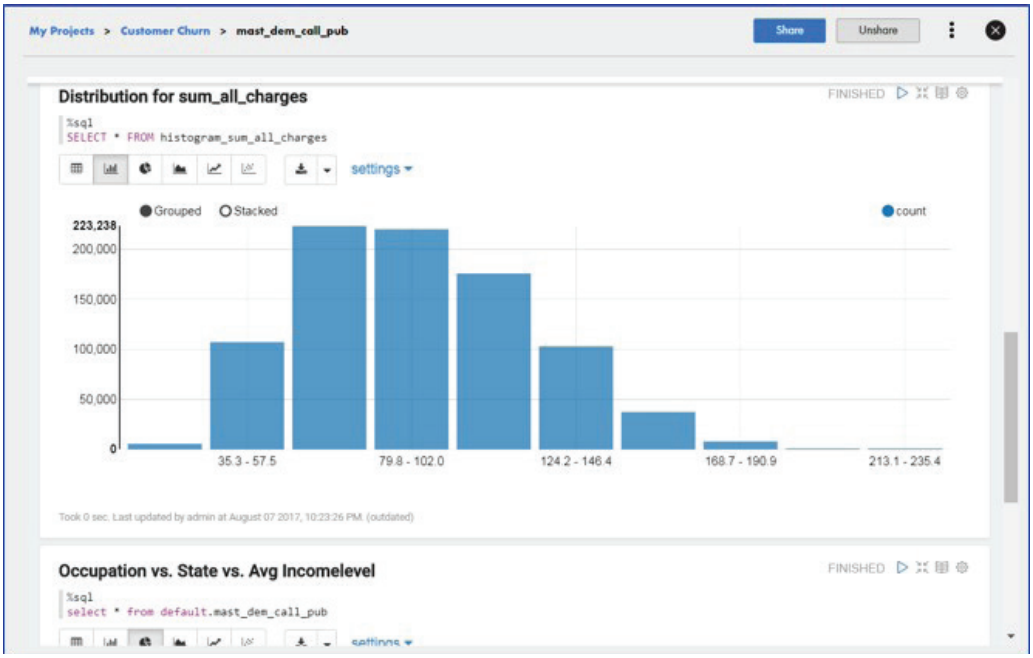

Figure 9: Recommended visualization in Apache Zeppelin notebook.

Intelligent Data Recommendations

CLAIRE provides data analysts and data scientists with suggestions on which datasets to use for their projects. It observes the datasets the users have selected and suggests other similar and better-ranked ones or additional datasets that may complement the ones they are using. Intelligent data recommendations help users avoid repeating the same work that many of their colleagues may already have performed. 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. These scores are then used to recommend datasets with similar properties. Complementary items are recommended by querying the metadata graph to find datasets commonly used together by different users.

Intelligent Structure Discovery

An increasing amount of data is generated and collected across machines, enterprises, and applications in unstructured or non-relational format. These data types are characterized not just by the large volumes, but also by their velocity, variety, and variability. “Data drifting” is a term that is 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 a variety of data file formats, including clickstreams, IoT log, CSV, text-delimited, XML, JSON, Excel, ORC, Parquet, Avro, PDF forms, and Word table files. CLAIRE 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 file 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 on the basis of 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 appropriate fitness of the structure to the data.

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

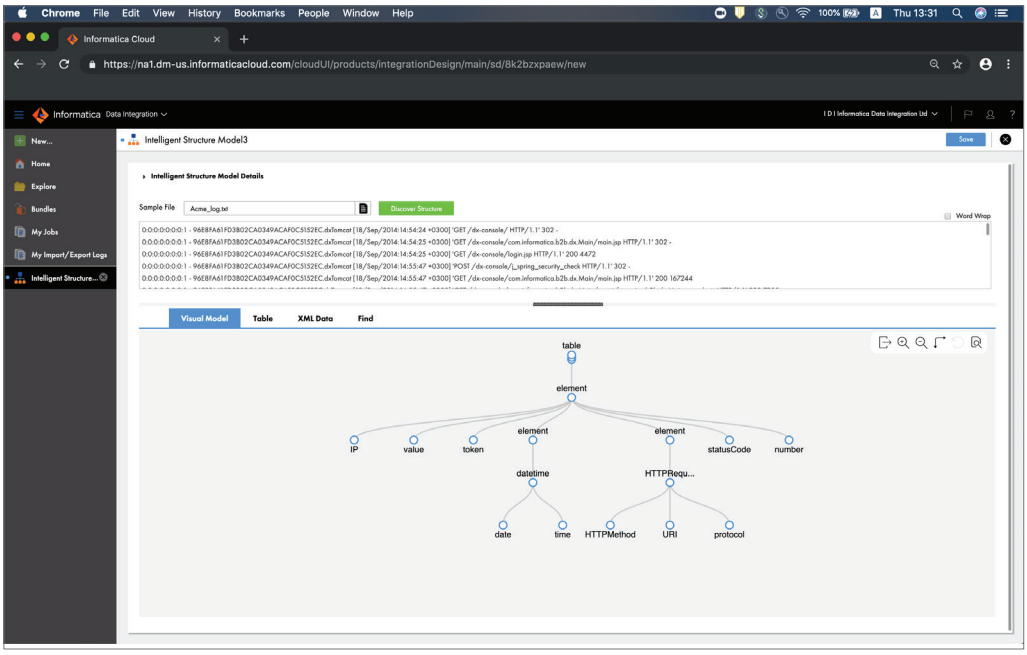

Figure 10: Intelligently finding structure in unstructured data files.

CLAIRE for Master Data Management

CLAIRE-powered automation and intelligence using advanced AI and machine learning enriches and improves the accuracy of business 360 views for customers, products, suppliers, and other domains. A variety of 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.

Synthesis Matching

Synthesis is a next-generation matching technique that addresses, for example, matching prospects to customers, matching interactions and unstructured data to customers, and discovers non-obvious relationships. It uses "contextual attributes," machine learning, NLP, and a combination of probabilistic matching with declarative rules to accomplish this.

Demographic attributes (e.g., name, address, and phone number), interaction attributes (e.g., email, webchats), and contextual attributes (e.g., when, what, where, who) are powerful in matching customer-related data with a given confidence level. NLP can pull out the "contextual attributes" from unstructured text to provide many more data points for use in the matching process. An ML algorithm can be very effective in matching when using a supervised training approach where data stewards and subject-matter experts label a properly selected set of match pairs as either matches or non-matches. These labeled match pairs form a training set that is used to produce a matching algorithm.

Synthesis will stitch together a full 360-degree customer view consisting of demographic, account, transaction, interaction, and unstructured data. Traditional matching algorithms merge records together to form a single customer view, whereas synthesis matching manages all customer data in a graph. Data is related together with confidence levels, where it is then possible to provide multiple views, or "perspectives," of a customer.

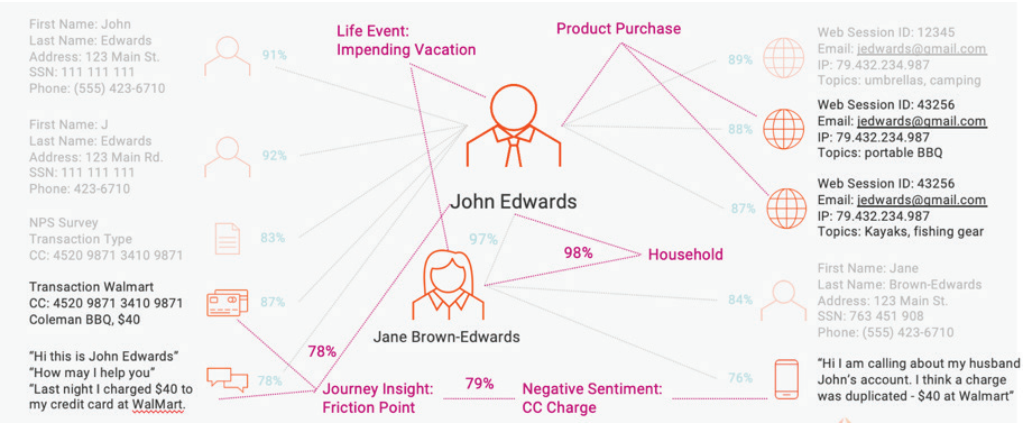

Figure 11: Synthesis matching and reasoning infers intelligence that is then stored as part of the Customer 360.

Identity Matching

CLAIRE's NAME3 identity matching encapsulates 30-plus years of training and tuning using a variety of techniques such as smart 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 match pair and determines whether there is a match, maybe-match or non-match along with a confidence level. There are techniques that 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 together.

Declarative rules, created by subject-matter experts, complement powerful AI rules that CLAIRE employs in the form of a learned random forest classifier. CLAIRE uses supervised active learning (as opposed to crowd-sourced or multi-user learning) to accelerate the AI training process where micro batches of 10 or 20 match pairs are presented for labeling to a user (i.e., match, maybe-match, no-match). Once they are labeled, CLAIRE retrains the random forest classifier and determines the next-best match pairs to present to a user in this iterative labeling process. CLAIRE uses the labeled pairs to infer blocking rules (i.e., remove obvious non-matches), perform blocking, train a model, and perform entity matching.

CLAIRE uses a combination of string comparisons/similarities such as Jaccard, declarative rules derived from data profiling, stabilized datasets (population files, nicknames, semantic comparisons, etc.), and user-defined rules that handle exceptions. 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, whenever name, birthdate, and SSN compare strongly then the rule classifies that as a match. This blending of declarative rules and AI rules accelerates training and improves the matching accuracy.

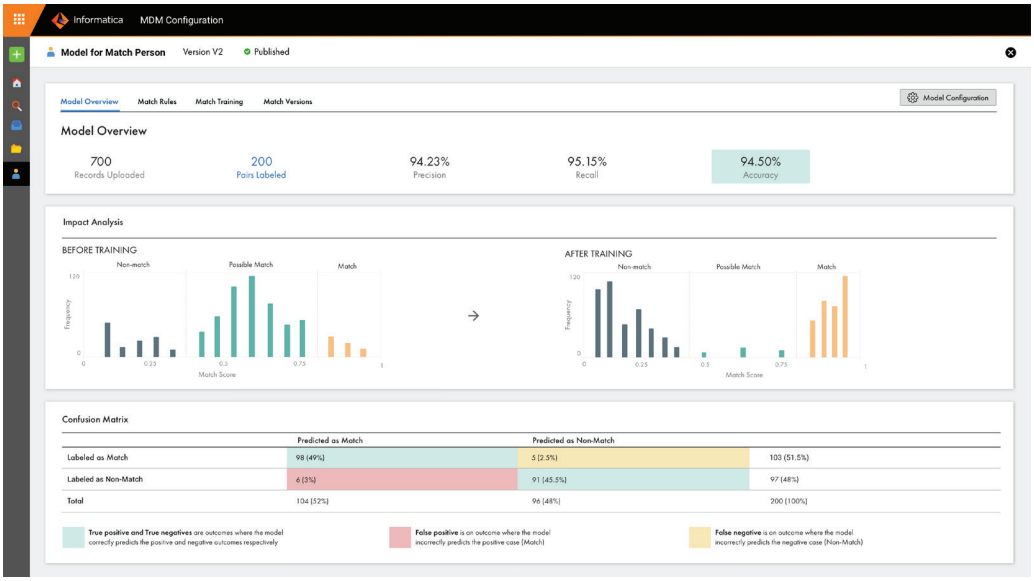

Figure 12: Entity Matching

CLAIRE for Data Governance and Compliance

AI and machine learning 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) and ties business semantics to technical metadata and helps guide users to the most relevant and trusted data for their business needs.

Automatic Data Quality Enrichment

CLAIRE uses an NLP approach based on Stanford NER to parse and extract entities from unstructured text. Typically, to extract entities from strings (e.g., product code), users end up writing parsing rules using reference tables and regular expressions. The amount of data, complexity, and patterns are constantly increasing; writing all possible rules to match every input is not practical or scalable. Instead CLAIRE uses pre-trained models to identify and extract entities based on Stanford NER.

CLAIRE uses machine learning to classify incoming text, for example: Language, Product Type, and Tech Support Issue. The machine-learning methodology used is referred to as supervised learning with Naïve-Bayes and Max Entropy (multinomial logistic regression). Supervised learning is used to train models and assign labels. Subsequently the trained model can be deployed during data processing to label, route, and process different classes of input—e.g., deal with "engine problems" separately from "configuration" ones with similar meanings and distinguish between uses of words with multiple meanings. CLAIRE automates image tagging and classification by leveraging NLP and ML models for product classification and extracting image meta-tags.

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 Axon Data Governance and Enterprise Data Catalog, CLAIRE automated the mapping of 18,000 technical assets with 99% accuracy in 8 minutes.

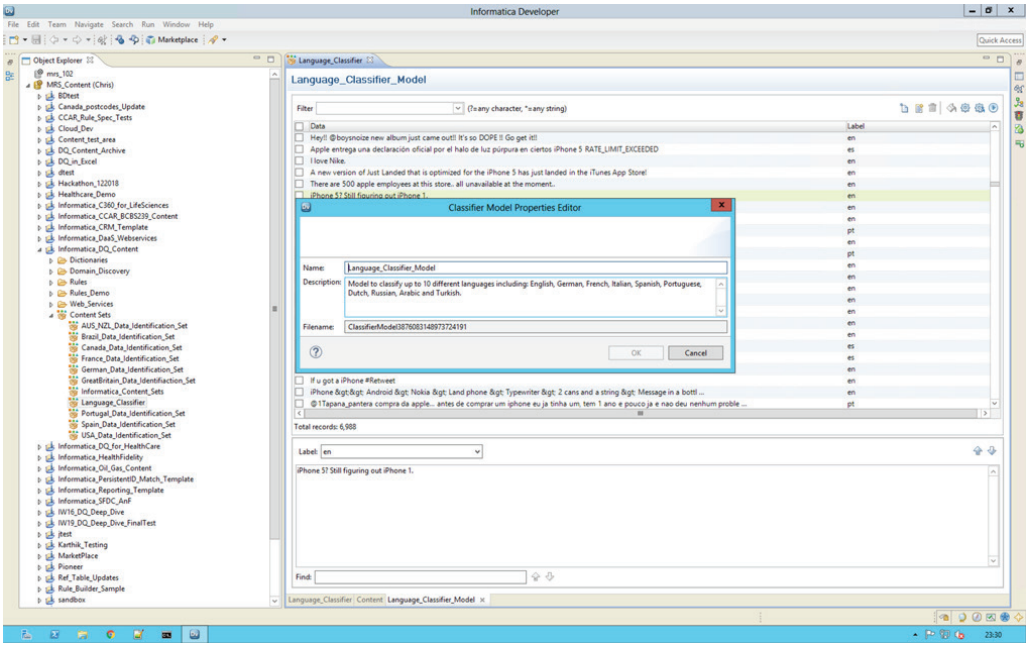

Figure 13: Machine learning NLP classifies text and extracts entities.

Automatically Associate Business Terms With Physical Datasets

Data governance requires the documentation of business artifacts, definitions, stakeholders, process, policies, and more. To enable a truly aligned view, it is key for users to 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 Axon Data Governance, through tight integration with the Informatica Enterprise Data Catalog, can shortcut this process. CLAIRE provides users 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.

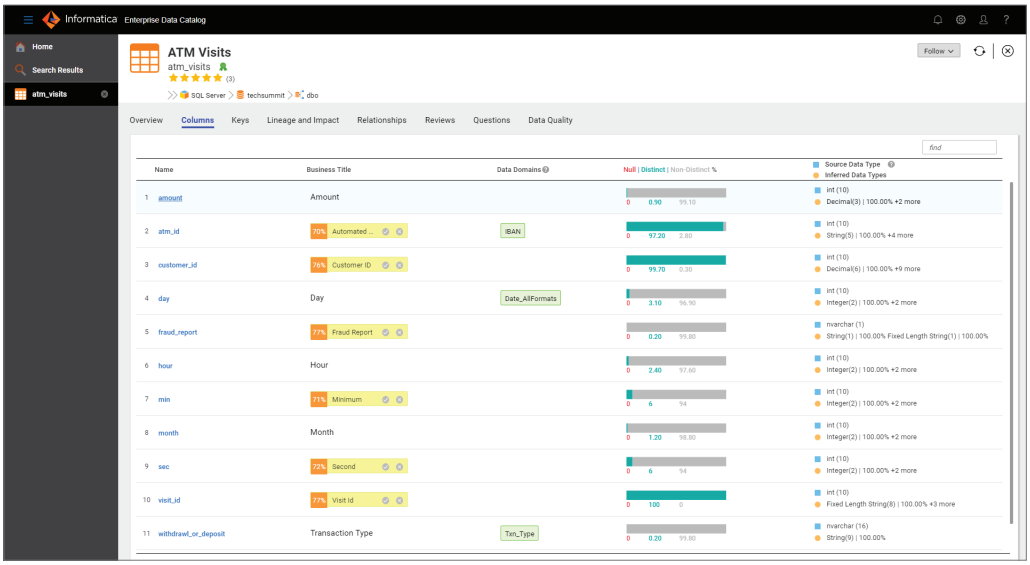

Figure 14: Automatic association of business terms with physical datasets.

Automatically Assess Data Quality

A key performance indicator (KPI) in data governance is the quality of data 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.

Through CLAIRE, Axon Data Governance—in coordination with Informatica Data Quality and Informatica Enterprise Data Catalog—can automate the application of data quality measurements across the enterprise, saving thousands of hours of work. The data governance team associates 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 various systems and 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 from Axon
2. Portable and executable data quality rules and a flexible execution engine from Informatica Data Quality
3. Metadata details from physical data assets from Enterprise Data Catalog

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

The automation enables governance programs to scale faster than ever before, removing thousands of hours of manual labor associated with creating data quality assessments and linking them back to governance context one by one. CLAIRE also ensures any new physical assets identified are automatically assessed for their quality. In addition, new domains are discovered using Named Entity Extraction or Classifier in data quality rules.

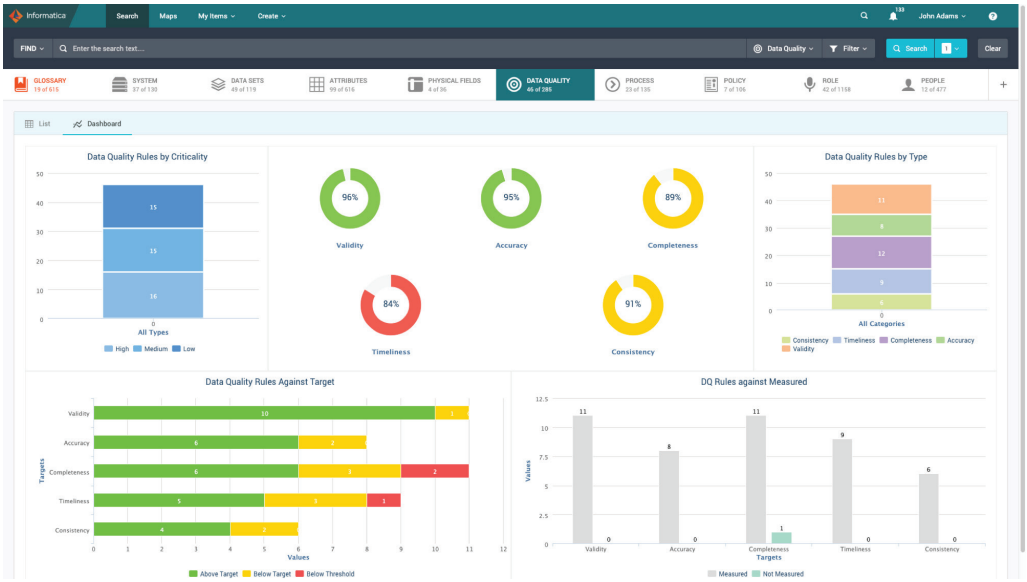

Figure 15: Automatic data quality assessments across the entire data estate save thousands of hours of manual labor.

ML/NLP-Assisted Data Quality Rule and Identification

Data quality is a key 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, CLAIRE can help not only identify rules, but also generate missing rules.

An Axon Data Governance 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. Through ML and NLP techniques, CLAIRE will analyze the user requirement and translate it into a technical representation. Based on this representation, as well as associated metadata (for example: Glossary Term name), CLAIRE will search the Informatica Data Quality Rules and identify any potential candidates. The user can then either choose from a matched existing rule or (if none are applicable) request CLAIRE to generate a new data quality rule.

If no applicable rule has been found, CLAIRE will automatically generate a new data quality rule to satisfy the requirement in the Informatica Data Quality repository and link it back to the Axon Data Governance context. In addition, CLAIRE automatically associates data quality rules to cloud profiles based on Microsoft Common Data Model (CDM) and Salesforce sources. 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.

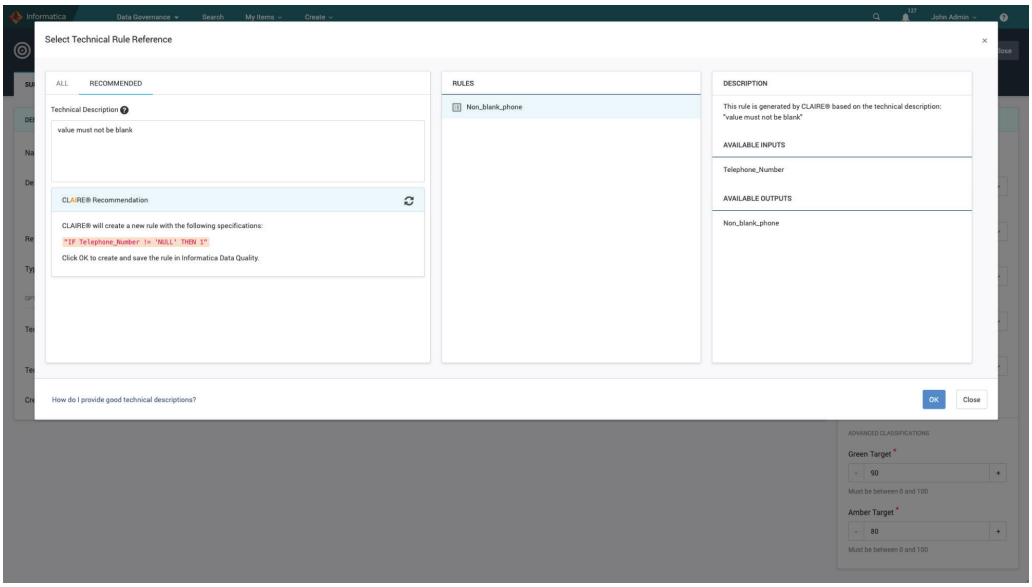

Figure 16: Automatic data quality rule identification using NLP.

CLAIRE for Data Privacy and Protection

With intelligent data privacy solutions powered by CLAIRE, organizations can gain an enterprise-wide view and analysis of personally identifiable information (PII) within data assets. AI-driven automation enables you to discover personal and sensitive data, understand data movement, link identities, analyze risk, and remediate problems.

Subject Registry Identity Mapping

CLAIRE determines identity correlation to sensitive data that provides data mapping for privacy compliance and data-subject access reporting. CLAIRE evaluates and scores data that in combination can identify data subjects. In addition to exact matching, various advanced techniques, including named-entity recognition (NER), are used to improve results commonly found when data is combined from different sources.

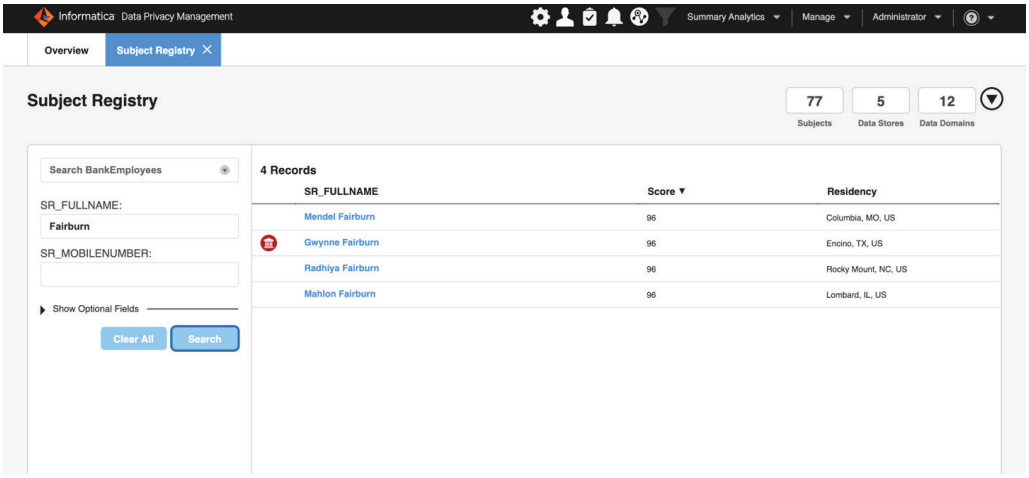

Figure 17: Subject registry identity mapping for privacy compliance and data subject access reporting.

Sensitive Data Mapping and Movement

CLAIRE leverages and extends the lineage capabilities mentioned above to also identify how sensitive data proliferates across repositories to support security and privacy compliance requirements. CLAIRE determines both upstream and downstream movement as well as related metadata, such as the specific type of data, process, protection status, and location of the data, to evaluate if violations have occurred. For example, a violation might occur if personal data is moving from a source to a target across geographic boundaries, or if data onboarded for billing processes is now being proliferated to other departments or locations for marketing processes that may be in violation of privacy regulations. Policy or process stakeholders can then be notified for remediation.

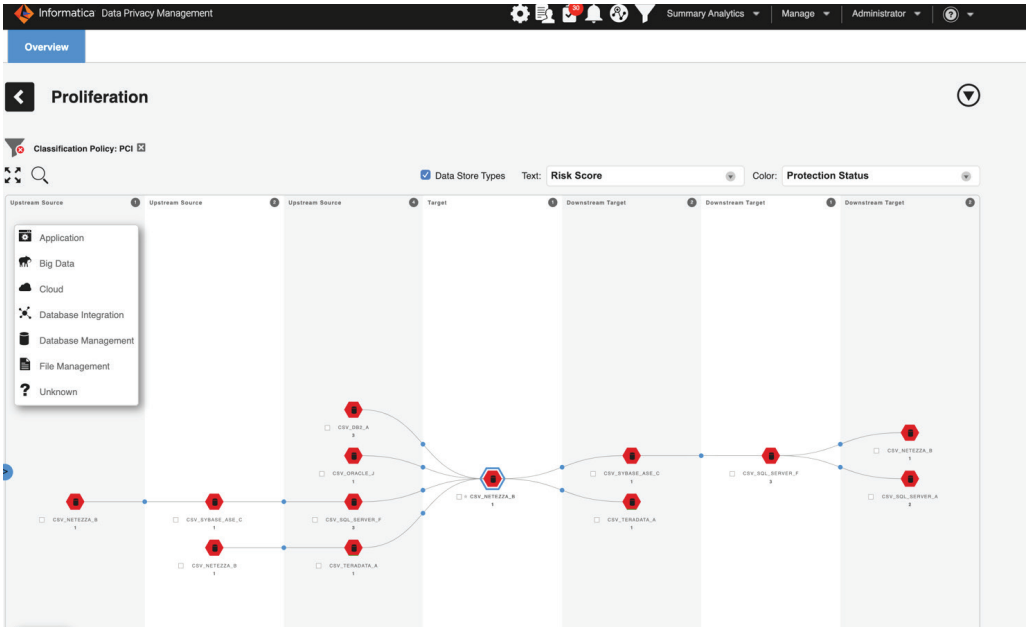

Figure 18: Identify and track the movement of sensitive data across repositories.

Risk Simulation Plans

Privacy regulations increasingly require organizations to have data-protection plans. CLAIRE can help organizations simulate the impacts of these protection plans to ensure greater return on investment and facilitate budget processes. CLAIRE evaluates the protection techniques applied to one or more data domains and then calculates the change in risk score, exposure of sensitive data, and residual risk cost for each of the selected data stores and the aggregated impact for the organization using an expected utility model.

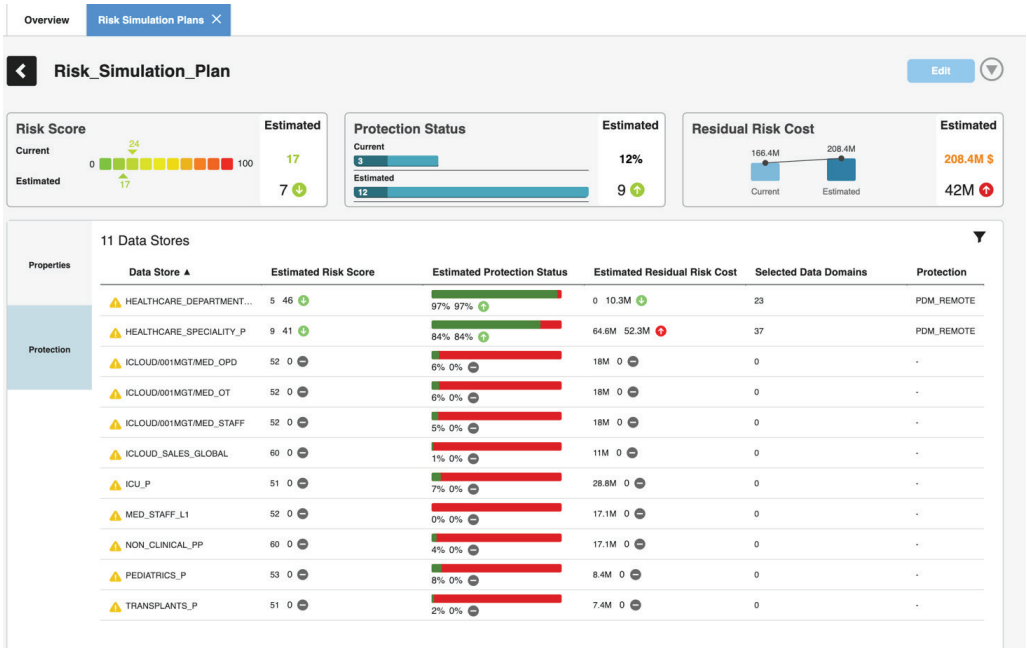

Figure 19: CLAIRE evaluates protection techniques applied to data domains to determine risk.

Intelligent Anomaly Detection

CLAIRE uses statistical and machine learning approaches to detect data outliers and anomalies. The user behavior analytics (UBA) capability detects patterns of user behavior that might be risky and expose an organization to data misuse. UBA is capable of detecting impersonation, credential hijacking, and privilege escalation attacks.

UBA applies unsupervised machine learning to a multidimensional model of user activities, which include the number of data stores accessed by the user, the number of requests made, and the number of affected records across different systems. Principal component analysis is applied to this model for dimensionality reduction. The BIRCH technique is applied for unsupervised hierarchical clustering to find users whose behavior was different during a given period. To validate the anomalous behavior, distance-and density-based outlier detection methods are employed and the statistical Grubbs' test for outliers is performed to confirm that objects indicated by the first two methods are indeed outliers in the cluster system.

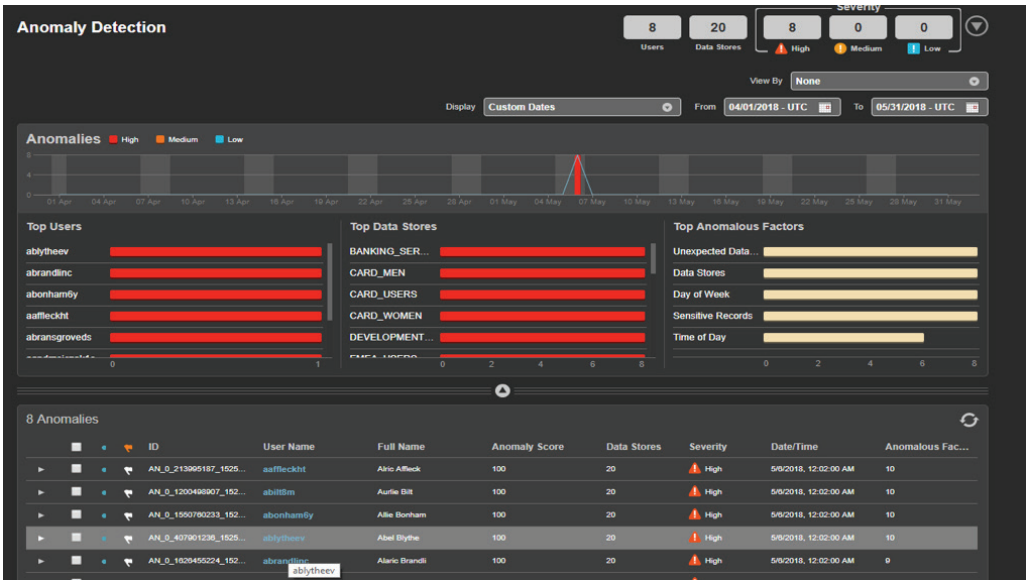

Figure 20: User behavior analytics to automatically detect user anomalies that may indicate data misuse.

Real-Time API Data Protection

Protect sensitive data (e.g., PII) in real time by identifying personal data leakage in APIs, blocking and masking data. Informatica API Management incorporates data protection libraries to block sensitive data on incoming and outgoing API calls, minimizing the risk of exposing sensitive data.

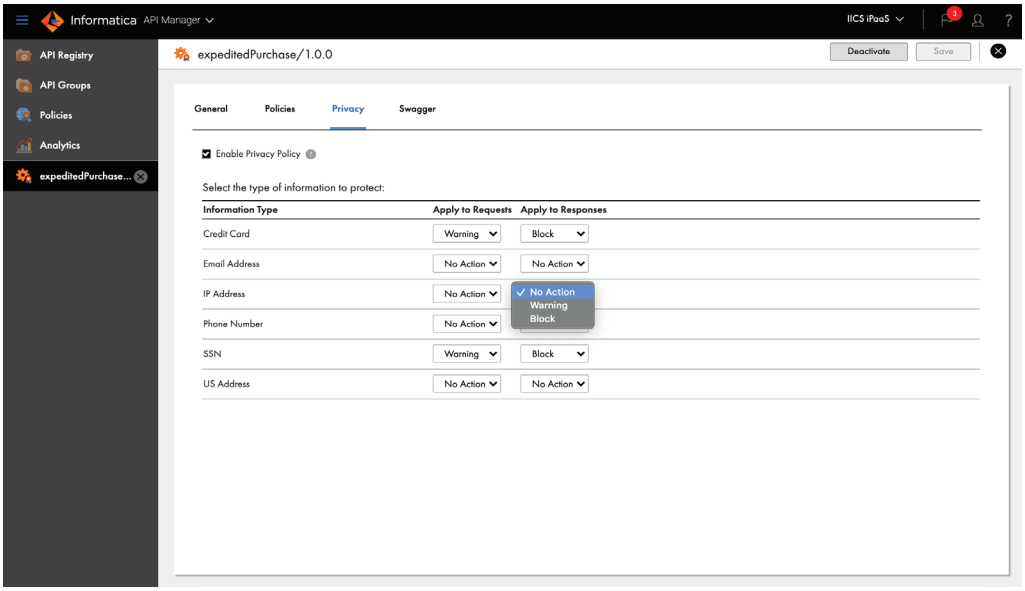

Figure 21: Block access to sensitive data on incoming and outgoing API calls.

CLAIRE for DataOps

With CLAIRE, 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 in understanding the current usage of existing projects and resources and in planning for future capacity. It offers parameters for building charge-back models while supporting multiple 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 offering auto-scaling of data management runtime resources.

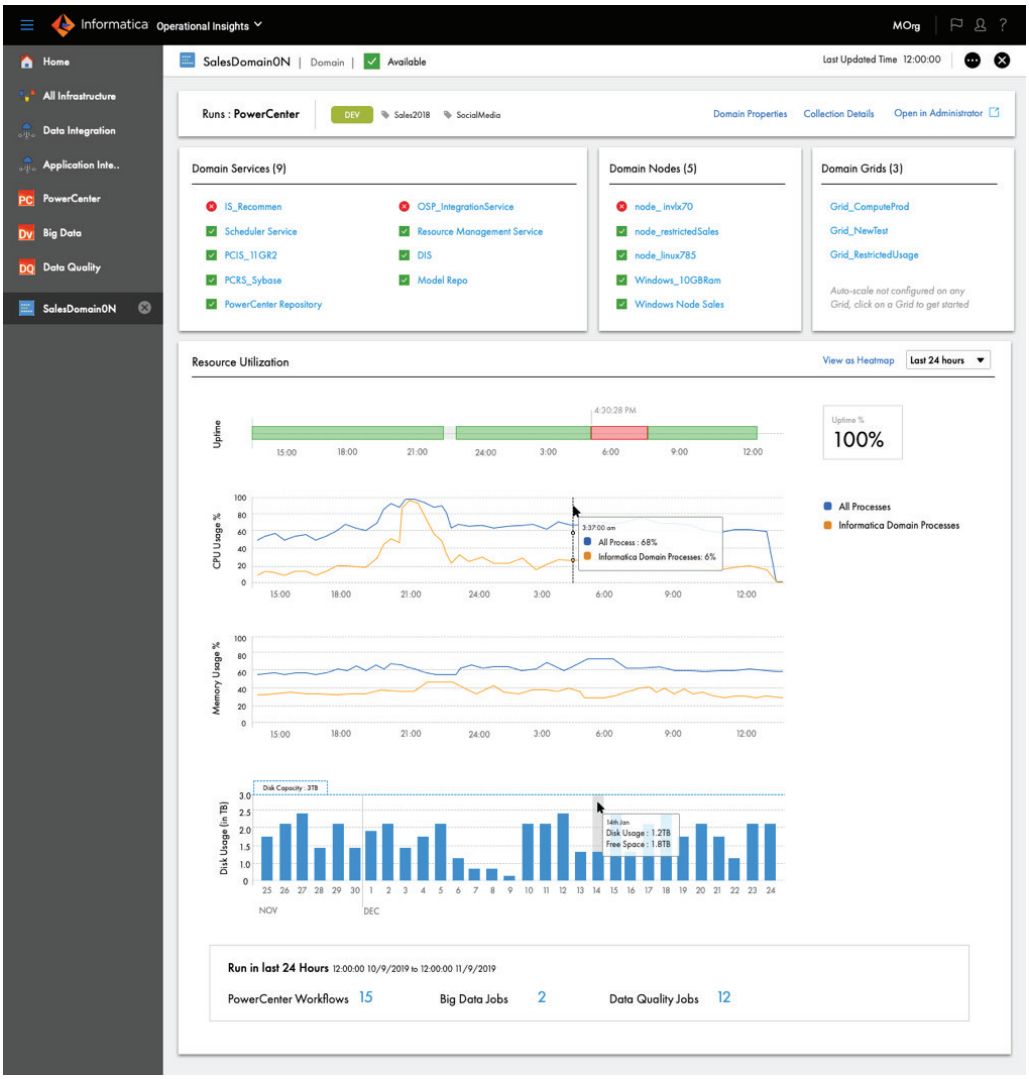

Figure 22: Operational Insights resource utilization for Informatica domain processes.

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 takes seasonality (month-end peak load, holiday season, etc.) into consideration and weeds out jobs with expected aberrations induced by business cycles.

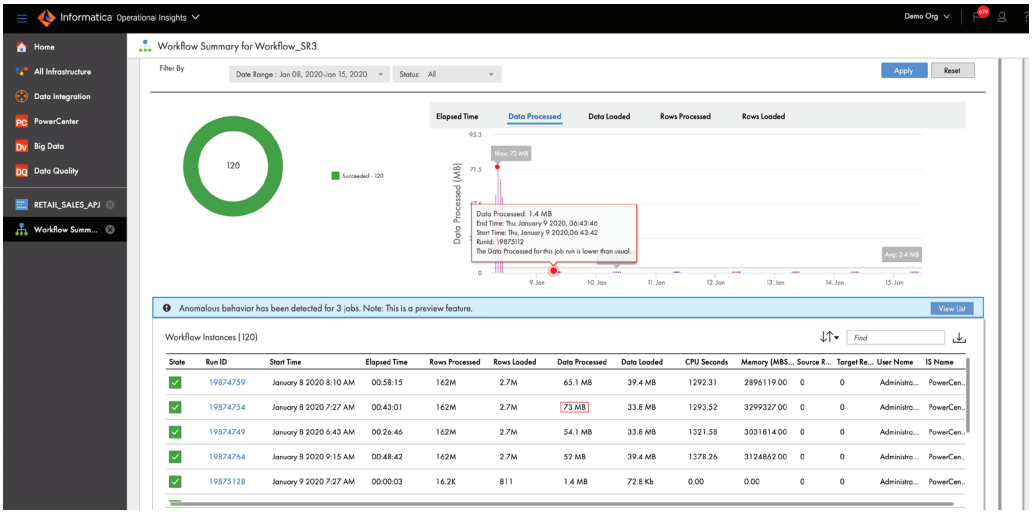

Figure 23: CLAIRE automatically detects anomalies related to Informatica jobs and data processing.

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. **Auto-mapping**: Detect master data entities across the enterprise and automatically map them to the master data model applying the requisite transformations and quality rules
4. **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
5. **Self-tune**: Based on historical information, current data volumes, and available system resources, predict and adjust schedules or compute resources to meet performance criteria
6. **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 machine learning/AI to enhance the productivity of all users of the platform: technical, operational, business, and particularly business self-service.

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

Informatica

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

IN09_0621_03328