Informatica

White Paper

Artificial Intelligence for Data-Driven Disruption

How the Machine Learning-Based Innovations in CLAIRE Are Driving a Big Leap in Data Productivity

CLAIRE™

This edition published May 2017

Table of Contents

Introduction

Digital transformation is real, and it is upon us. It's a matter of "disrupt or be disrupted." Organizations are driving transformative initiatives to improve financial performance and competitive position in their industries. Examples of these initiatives include deepening customer relationships, optimizing operations, personalizing healthcare, and preventing fraud.

The key factor driving the success of these initiatives is the ability to fuel them with trusted and timely data. It's pretty simple: Successful digital strategies are built on data. The competence you build around data management will determine how successful your digital strategy is. Or, put another way, your digital strategy will be only as effective as the data that informs it.

Odds are, however, that managing data "the way you always have" won't cut it. IT leaders are looking for ways to boost data management productivity to make better data available to all, faster.

Informatica's CLAIRE™ engine—or Cloud-scale AI-powered Real-time Engine—which uses artificial intelligence (AI) and machine-learning techniques powered by enterprise-wide data and metadata, will significantly boost the productivity of all managers and users of data across the organization.

Data Management Trends

It's time to think differently about data and data architecture. For decades, the focus has been on business systems and processes. While those are still important, the ability to deliver better, more timely, more complete data to your business initiatives will be what truly differentiates your organization in the marketplace. But, most IT budgets are growing slowly, so you also need to factor in doing more with your current resources.

The challenge of managing enterprise data has never been greater. To unleash the power of data, your IT organization must be able to manage:

1. **More Data:**
   - **Data volume:** 15.3 zettabytes per year in global data center traffic.
   - **Data complexity and variety:** There are many new sources and types of data, from within and outside the enterprise.
   - **Data velocity:** The rise of Internet of Things (IoT), with 20 billion connected devices, means always-on data streaming.
2. **More users:** With 325 million business data users and growing, everyone from business analysts to citizen data scientists to data stewards wants direct and timely access to data.
3. **More integration patterns:**
   - **Movement to cloud:** ERP suites are breaking up and moving to the cloud.
   - **Analytics technology:** The industry is moving to new technologies such as big data, NoSQL, and predictive analytics to complement data warehousing.
   - **Experimentation:** Users now want to use data to quickly form a hypothesis, try it out, either succeed or fail, and iterate quickly. It's all about speed over precision until they prove that a hypothesis is of value.

What This Means for IT leaders

All these trends combine to make the process of managing data that much more complex just as organizations are realizing that data is the fuel for their digital transformation.

This is an ideal opportunity to provide data-driven leadership to help your organization succeed. How will IT leaders meet the business need for better data faster without requiring an army of expensive developers, for instance?

With IT budgets growing slowly, if at all, there are three key ways to accomplish this:

- Increase automation and efficiency for data management tasks and projects
- Increase enablement of business self-service
- Increase collaboration to drive alignment between business and technical teams

What This Means for Business Leaders

Business leaders feel that they have the power to drive disruptive initiatives and ask questions they were never economically able to ask before. But, the results of their digital initiatives will be only as good as the data they run on.

Your #1 priority must be to build a plan to unleash the power of all your data.

It is important to build a data management competency as the foundation for all your digital initiatives. You need to manage data as an asset that is discoverable and usable by any user across your entire organization. And the data must be of a quality that is fit for purpose: high quality for important decisions and interactions and fair quality for rapid innovation and iteration. In terms of technology, hand coding or a collection of nonintegrated data management tools will not scale to meet the needs of the business.

What Is Machine Learning?

Machine learning is a technique where programs iteratively learn from data instead of being static. Machine-learning systems are used to build an input-based model that can be used to make predictions or decisions. These systems learn from the data and can adjust themselves accordingly to produce better results. The more data they have, the faster they learn and the more accurate their results.

Why machine learning for data management?

To scale up the speed of data delivery for critical business initiatives, you need to increase automation. That is where machine learning comes in. With enterprise-wide metadata visibility and machine learning, data management tools can be "taught" to make intelligent recommendations and to automate many data management tasks. Machine learning does not replace data analysts and other users; instead, it is key to increasing the productivity and effectiveness of the data management staff within an organization.

Machine learning can be used to improve tasks that are tedious or impossible to do at human scale. Some examples include:

1. Discovery and identification

- Data quality rules, and business entity discovery
- Semantic search, pattern identification, and data classification,
- Anomaly detection and notification

2. Predictive operations

- Burst to handle data spikes
- Prioritize operational issue investigations
- Self-heal to handle changes to environments

3. Next-best-action & recommendations

- Suggest data sets, transforms, and rules
- Auto-map, cleanse, and standardize from sources to target
- Self-integrate new sources of data

The basis for machine learning in data management

Effective machine learning requires large training data sets. In a data management context, the best source of data is an enterprise-wide data catalog. Most enterprises have thousands of databases, data files, applications, and analytics systems. By harvesting the metadata across these data repositories, enterprises can build a richly populated catalog. The combination of machine learning and a data catalog with enterprise-wide visibility into metadata would provide the basis for intelligence that would make a truly meaningful and positive impact on data management productivity.

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.

Informatica CLAIRE: The "Intelligence" in the Intelligent Data Platform

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

1. The Intelligent Data Platform (IDP): We have delivered an integrated end-to-end 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, and big data sources. We call this unified data management platform the Intelligent Data Platform.

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

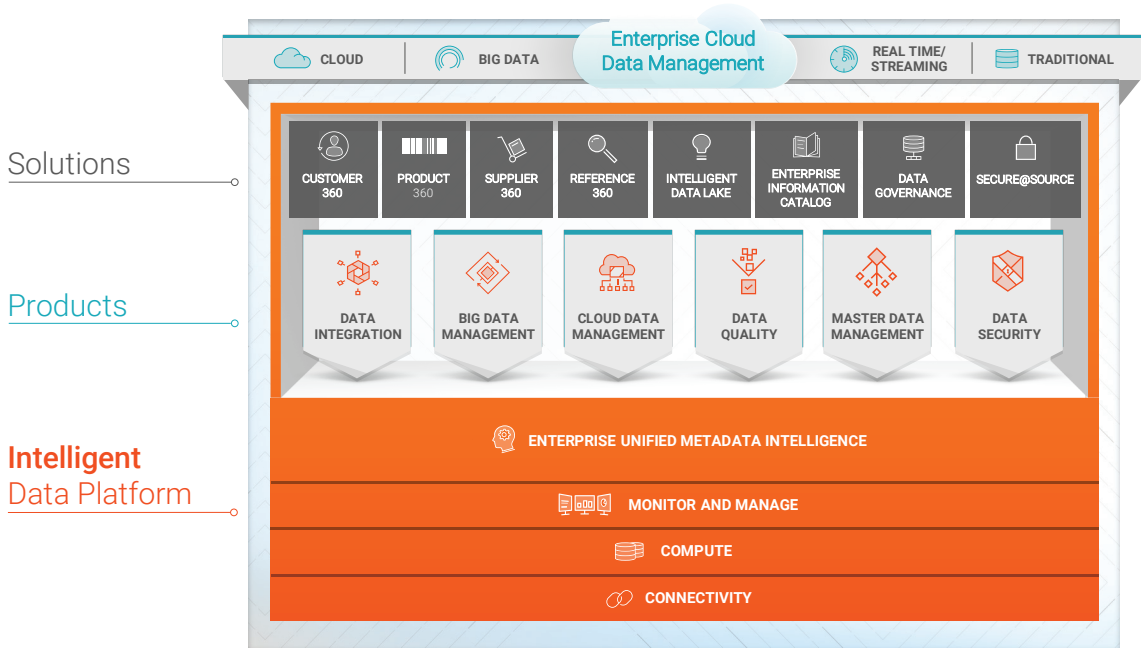

Figure 1: The Intelligent Data Platform integrates data management capabilities with shared connectivity, operational insight, and data and metadata intelligence.

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, and data profile statistics
   - Business metadata, which captures context about data, its meaning, relevance, and criticality to various business processes and functions
   - Operational metadata, about systems and process execution, such 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 data sets accessed, search results clicked on, ratings or comments provided

This broader collection of metadata is critical to machine learning. It provides data sets that are used to "train" the machine learning algorithms and enables them to enables them to adjust to 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 Platform 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 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 Platform.

CLAIRE in Action

CLAIRE helps a wide spectrum of users:

- Data developers 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 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.

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 data sets, and recommending data sets to users.

Data similarity computes the extent to which data in two columns are 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.

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.

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 Facebook photo, with the net affect that the same people are tagged in millions of other photos.

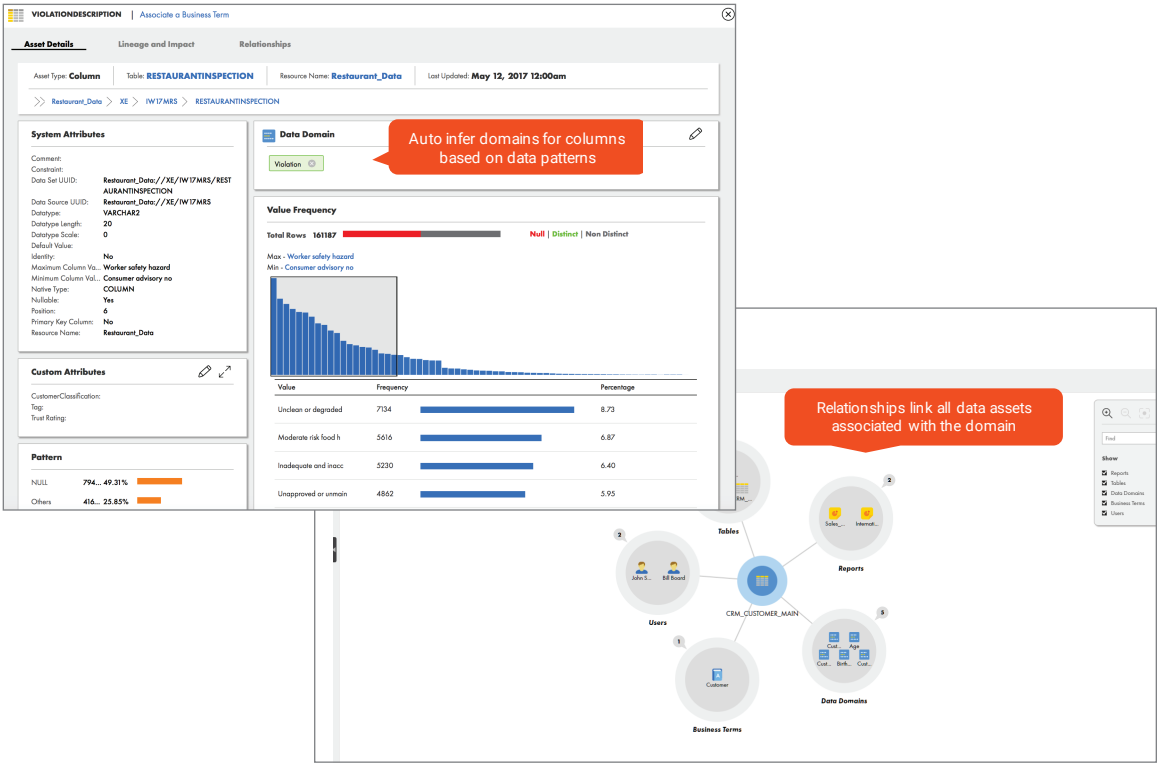

Figure 3: Automatic classification of data.

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.

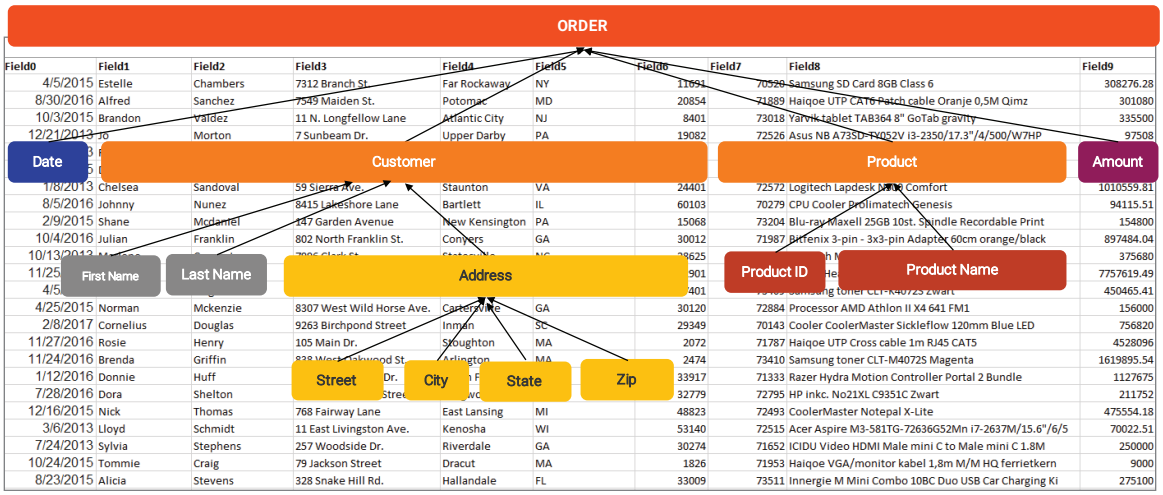

Figure 4: Combining data domains to detect entities from tables and files.

Intelligent data recommendations

CLAIRE provides data analysts and data scientists with suggestions on which data sets to use for their projects. It observes the data sets the users have selected and suggests other similar and better-ranked ones or additional data sets 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:

1. A prepared version of the same data (substitutable data)

2. Another table containing same type of records (union-able data)

3. 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 data sets. The characteristics (terms) used for data sets include lineage information, user ranking, and data similarity. Several similarity measures are used to score the equivalence of different data sets. These scores are then used to recommend data sets with similar properties. Complementary items are recommended by querying the metadata graph to find data sets commonly used together by different users.

Intelligent structure discovery

CLAIRE can derive structure from messy device and log files, making them easier to understand and work with. By using a content-based approach to parsing files, it can adapt to frequent file changes without affecting file processing.

Intelligent structure discovery uses a genetic algorithm to automate the recognition of patterns in the files. In this approach, it uses the concept of "evolution" to improve results. Each candidate solution has a set of properties that can be altered and then tested to determine if they provide a solution with a better fit. It neither requires user input to define the structure of the file nor is specific to a set of industry file formats. Initial structures of the file are derived based on basic delimiter-based parsing. These 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.

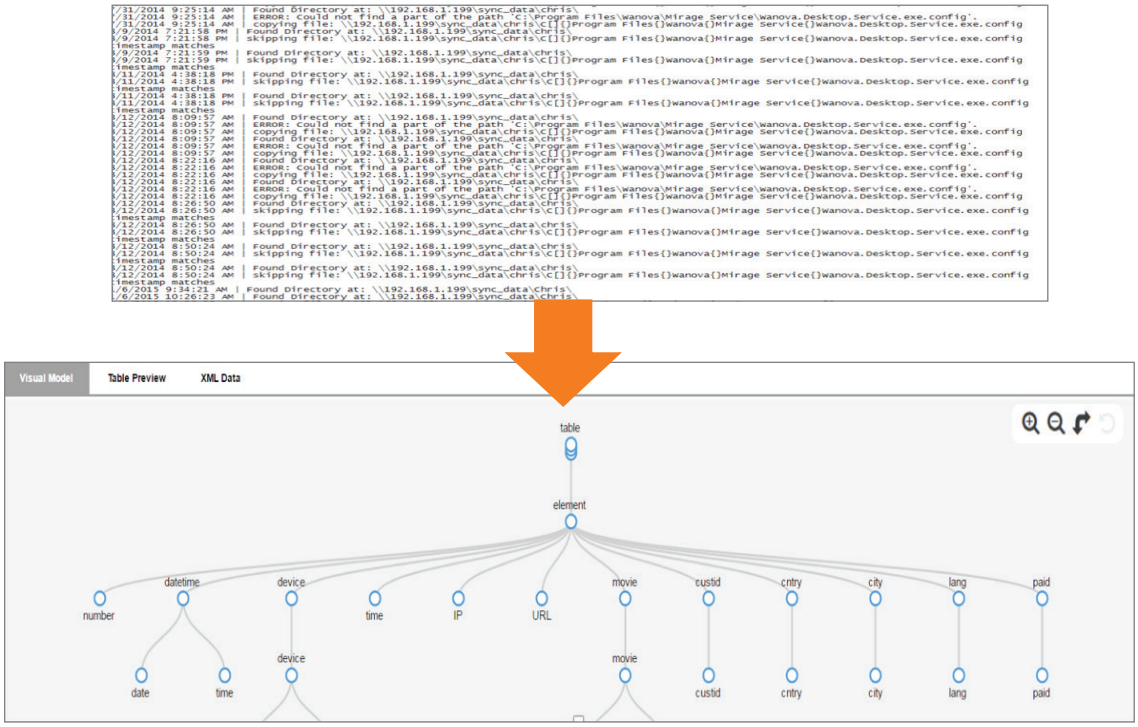

Figure 5: Intelligently finding structure in unstructured data files

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 multi-dimensional 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.

Here are some examples of CLAIRE capabilities launching in the future:

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

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

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

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

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

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

About Informatica

Digital transformation is changing our world. As the leader in enterprise cloud data management, we're prepared to help you intelligently lead the way. To provide you with the foresight to become more agile, realize new growth opportunities or even invent new things. We invite you to explore all that Informatica has to offer—and unleash the power of data to drive your next intelligent disruption. Not just once, but again and again.

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 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 Platform to harness the power of your data.

Informatica

Worldwide Headquarters, 2100 Seaport Blvd, Redwood City, CA 94063, USA Phone: 650.385.5000 Fax: 650.385.5500 Toll-free in the US: 1.800.653.3871 informatica.com linkedin.com/company/informatica twitter.com/Informatica

IN09_0517_3328