

# Find and Prepare Any Data for Self-Service Analytics

## Benefits

- Find and discover any data for self-service analytics with an easy-to-use interface
- Deliver high-quality and trusted data with an end-to-end data preparation pipeline
- Foster data preparation collaboration with data governance and data protection
- Operationalize data pipelines at scale

## Informatica Enterprise Data Preparation Provides End-to-End Data Discovery, Transformation, Enrichment, and Protection

### **Business Users Require On-Demand Access to Critical Business Data**

In today's data-driven economy, your ultimate business success depends on deriving better analytics insights from big data, faster. Business users are increasingly challenged to gain deeper insights into customers and products, optimize pricing, increase revenue, and reduce costs. Data scientists need more data in order to develop more accurate prediction models to help business users with forecasting and trend analysis.

### **Preparing Self-Service Data Can Be Time Consuming**

Both business users and data scientists need timely access to the wealth of business data in order to meet business imperatives. But these users can spend up to 80% of their time finding and preparing the data before they begin to use it. They need an easy-to-use self-service analytics solution to help them access the data faster, while ensuring compliance and protection of personal and sensitive data. Without self-service analytics, business users often manually wrangle their own data spending more time finding and acquiring data instead of focusing on the analysis.

### **Self-Service Analytics Needs Governance and Regulatory Compliance**

Democratizing data for self-service analytics requires delivery of more data to more users. Ensuring trust in the data and compliant use of the data requires policies and rules for quality, access and protection be in place. If governance isn't embedded into the activities data engineers and data analysts use to create a data pipeline you risk poor quality results, non-compliance with privacy regulations, and decreased customer trust.

The Informatica Enterprise Data Preparation solution provides critical self-service capabilities to empower data engineers and data analysts to quickly find, prepare, protect, and deliver data so data analysts and data scientists can create more value from the data. The solution leverages AI and machine learning to add further automations, intelligent search, and recommendations.

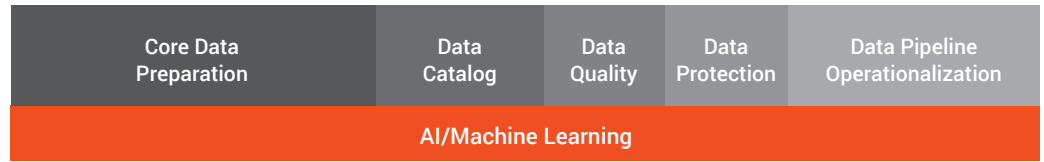


Figure 1: Enterprise Data Preparation pipeline leverages the data catalog for search, data preparation for transformation, data quality to apply business rules, data protection for masking, and is operationalized with the scalable Spark engine.

## Key Benefits

### Find and Access Any Data

Business analysts are able to easily access a wealth of data and find trusted data using Google-like semantic search and dynamic facets to filter and aggregate data assets. Metadata-driven AI, powered by the CLAIRE™ engine, assists in the data discovery and transformation process with relevant recommendations of new data assets that may augment the analytics. This greatly increases confidence in the data and reduces duplicate data which may affect the accuracy of the machine learning models.

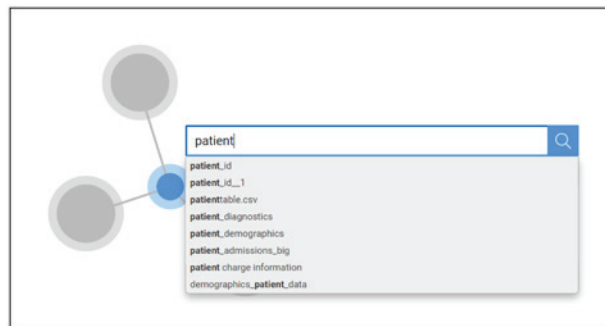


Figure 2: Data discovery using Google-like semantic search.

### Deliver an End-to-End Data Preparation Pipeline

Informatica's self-service data preparation provides pre-integrated capabilities for catalog, preparation, data quality, data protection, and operationalization. The solution provides an Excel-like interface allowing business users to quickly combine, filter, and blend data into analytics insights and machine learning models. Crowdsource information about datasets provide business context and relevancy increasing data reusability.

## 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.

The screenshot displays the Informatica Enterprise Data Preparation interface. The main window shows a data table with columns: #, patient\_id, admission\_type\_id, discharge\_disposition\_id, and admission\_source\_id. The table contains 25 rows of patient data. On the left, a 'Recipe Steps' pane lists seven steps: 1. Join all rows from train\_adm\_demog and patient\_diagnostics using 'Full Outer' on key: patient\_id,patient\_id\_\_1; 2. Synchronize columns with source; 3. Edit all empty cells to 'NA' in admission\_type\_id; 4. Edit all empty cells to 'NA' in discharge\_disposition\_id; 5. Insert formula into new column: HighRisk - IF(num\_procedures>3, 1, 0); 6. Insert formula into new column: lab\_proc\_diff - 'num\_lab\_procedures - AVG(num\_lab\_procedures)'; 7. Edit all cells with value '7' to (blank) in medical\_specialty. Below the table, a 'Column Overview' pane for 'admission\_type\_id' shows a value frequency chart with categories: Emergency (4905), Urgent (1935), Elective (1786), Not Mapped (759), Not Available (614), and Newborn (1). A 'Suggestions' pane offers 'Uppercase' and 'Lowercase' options.

#	patient_id	admission_type_id	discharge_disposition_id	admission_source_id
1	1	Elective	Discharged to home	Physician Referral
2	2	Urgent	Discharged to home	Physician Referral
3	21	Elective	Discharged to home	Emergency Room
4	6	Elective	Discharged to home	Physician Referral
5	7	Elective	Expired	Physician Referral
6	10	Elective	Discharged to home	Physician Referral
7	11	Urgent	Discharged/transferred to a long term care hospital.	Physician Referral
8	13	Urgent	Discharged to home	Physician Referral
9	22	Elective	Discharged to home	Transfer from a hospital
10	15	Emergency	Discharged/transferred to SNF	Transfer from a Skilled N
11	4	Emergency	Discharged to home	Transfer from another hea
12	12	Emergency	Discharged/transferred to SNF	Transfer from another hea
13	3	Not Available	Discharged/transferred to home with home health service	
14	9	Not Mapped	Discharged to home	
15	20	Not Mapped	Not Mapped	
16	23	Emergency	Discharged to home	Emergency Room
17	24	Urgent	Discharged to home	Physician Referral
18	25	Elective	Discharged/transferred to another type of inpatient care institution	Physician Referral

Figure 3: Data preparation using an Excel-like interface where transformation steps are recorded for replay and automation.

## Foster Collaboration With Data Governance

The data preparation pipeline has many contributors including data engineers, data analysts, business analysts, and data scientists which requires a fair amount of ongoing collaboration. The recipe approach of recording the data preparation process fosters collaboration between the various users with annotations and sharable mappings enhancements. All users collaborate using the project workspaces and share datasets, data lineage, profile statistics, relationships, and transformations. And governance policies and rules ensures that only the authorized users have access to the data, and that it is used in a compliant and ethical manner.

## Operationalize Data Preparation at Scale

Operationalize the processing of data pipeline with advanced Spark support for performance and scalability to manage the lifecycle of the data pipeline at scale. The pipeline is also automated by scheduling the data ingestion, data preparation, and data delivery.

## Accelerating AI/ML Projects

Machine learning requires ongoing access to large volumes of data to continue to enhance and increase the accuracy of its models. Data scientists will leverage the dataset created by Enterprise Data Preparation and leverage an external machine learning platform for developing and executing the machine learning models. Some of the external platforms include Databricks Notebooks, DataRobot, AWS SageMaker, as well as programming languages such as Python and R.

## Next Steps

Learn more by visiting the [Informatica Enterprise Data Preparation](https://www.informatica.com/trademarks.html) product page.



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

IN17\_0619\_03710