Informatica

Data Sheet

Informatica Enterprise Data Preparation

Benefits

- Rapidly discover, enrich, cleanse, and govern data pipelines for faster insights
- Prep data on cloud data lakes at scale
- Reduce data preparation efforts with advanced AI-powered automation
- Easily operationalize data preparation for reusability at enterprise scale

Simplify Self-Service Data Preparation Across Cloud and On-Premises Data Lakes

Data lakes are rapidly becoming the cornerstones of any digital transformation and big data initiative. They offer more flexible options to ingest, integrate, persist, and process massive volume and variety of datasets at high velocity across on-premises and cloud environments.

While data lakes offer a myriad of benefits, given the sheer complexity and diverse types of data that's ingested and stored, enterprises are challenged with building agile data engineering pipelines that can be operationalized at enterprise scale.

Industry research estimates data scientists are spending nearly 80% of their time on cumbersome data preparation tasks. DataOps teams, including data engineers, data scientists, and data analysts, must find, access, blend, cleanse, and transform the data they need into high-quality and governed datasets before it can be shared and consumed by the business to support a plethora of use cases including analytics and AI-enabled workloads.

Moreover, the use of a standalone solution approach coupled with a partially automated data preparation process often results in bottlenecks leading to higher inefficiencies, operational costs, and time-to-insight delays. Without scalable, repeatable, and intelligent mechanisms for discovering, cleansing, and curating data, all the opportunity that data lakes promise risks stagnation.

Powered by the Informatica® CLAIRE® engine, the industry's first metadata-driven AI, Informatica Enterprise Data Preparation enables raw big data to be systematically discovered, blended, cleansed, and transformed so that DataOps teams including data engineers, data scientists, and data analysts are empowered to turn datasets into trusted and governed information for use and analysis at enterprise scale.

Key Features

Machine Learning-Enabled Data Preparation and Data Cataloging

With Informatica Enterprise Data Preparation, DataOps teams can rapidly discover the data they have in a data lake using Google-like semantic search, including certified datasets along with key attributes about the data such as data domains, users, and usage as well as other related data assets. Users can easily visualize data sources, track datasets from source to destination, and enable effective data-driven business transformations with end-to-end data lineage and impact analysis capabilities.

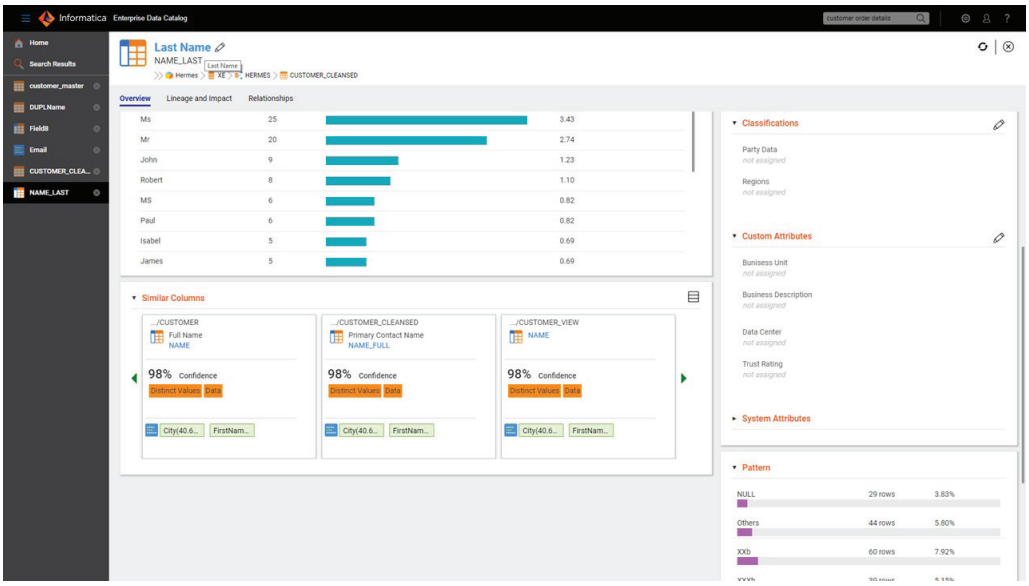

Figure 1: Rapid data discovery at enterprise scale with Google-like semantic search.

AI-Assisted Data Collaboration and Social Curation

With advanced data collaboration capabilities, users can harness the combined power of human expertise, social curation such as ratings and reviews, and AI-driven insights to automate data curation and enhance user experience.

Dataset and Recipe Recommendations

Data engineers, data scientists, and data analysts can easily collaborate with one another and share results in project workspaces. As they add datasets to their project workspace, machine learning algorithms powered by CLAIRE work in the background to recommend alternative datasets and recipe recommendations that simplify and accelerate the data preparation process.

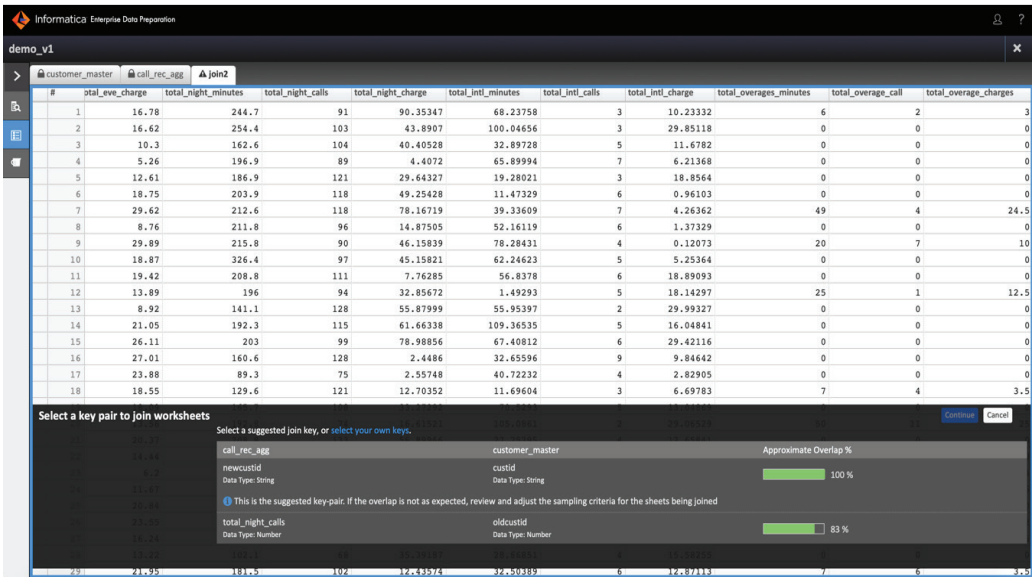

Figure 2: Advanced machine learning algorithms speed self-service data preparation.

Moreover, users can open datasets within the project workspace at any time via an easy-to-use and intuitive Excel-like user interface. They can visualize sheet-level and column-level descriptive statistic overviews, including value distributions, numeric, and data distributions. This metadata-driven approach to data preparation is the intelligent way to turn big data into trusted information assets that deliver sustainable business value across cloud and on-premises data lakes.

Optimized for Cloud Data Lakes

Data engineers, data scientists, and data analysts can easily import, upload, or publish files on Amazon S3 and Microsoft Azure ADLS. Enterprise Data Preparation comes with comprehensive support for ADLS Gen2, enabling users to build data pipelines with confidence.

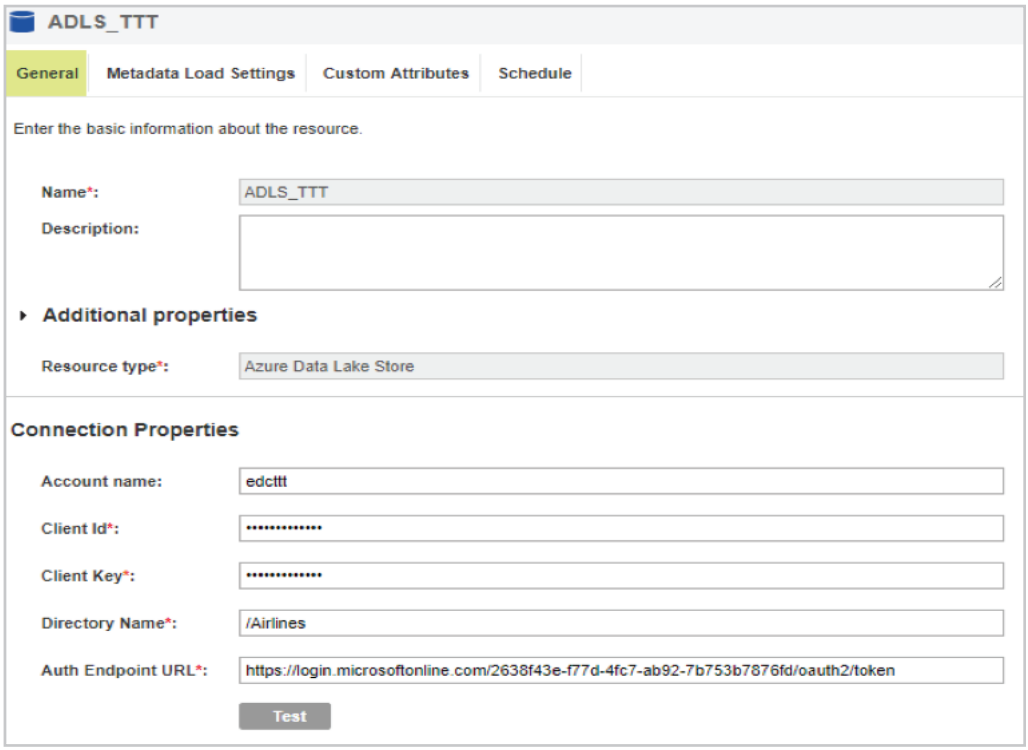

Figure 3: Prep data on cloud data lakes at scale.

Scalable and Flexible Options

Users can prepare master datasets from various structured, semi-structured, or unstructured data in CSV, Excel, JSON, Parquet, Avro, or text delimited file formats, and integrate them seamlessly with structured relational tables in the data lake without writing a single line of code. Final master datasets can be published as relational tables or files which can be consumed for further analytics, reporting, and data science projects.

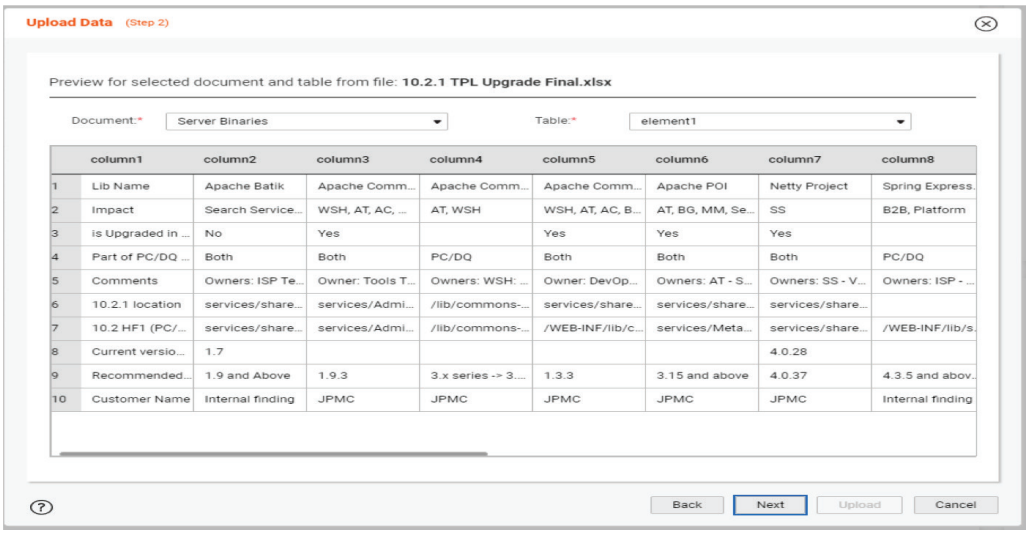

Figure 4: Import, upload, or publish as CSV, Excel, Parquet, Avro, JSON, or text delimited file formats.

Automated Data Cleansing, Blending, Error Handling, and Transformation

Users can iteratively prepare data for analysis with prebuilt data cleansing, blending, and transformation to filter, aggregate, merge, lookup, shape, and join data. They can leverage advanced automation capabilities such as fuzzy matching and consolidation, automatic structure discovery, and automatic error detection. They can perform column-level data cleansing and data transformation using string, math, date, and logical operations. Metadata capabilities allow users to wrangle and mashup datasets. Guided intelligence helps prepare datasets with recommendations such as join keys when blending datasets.

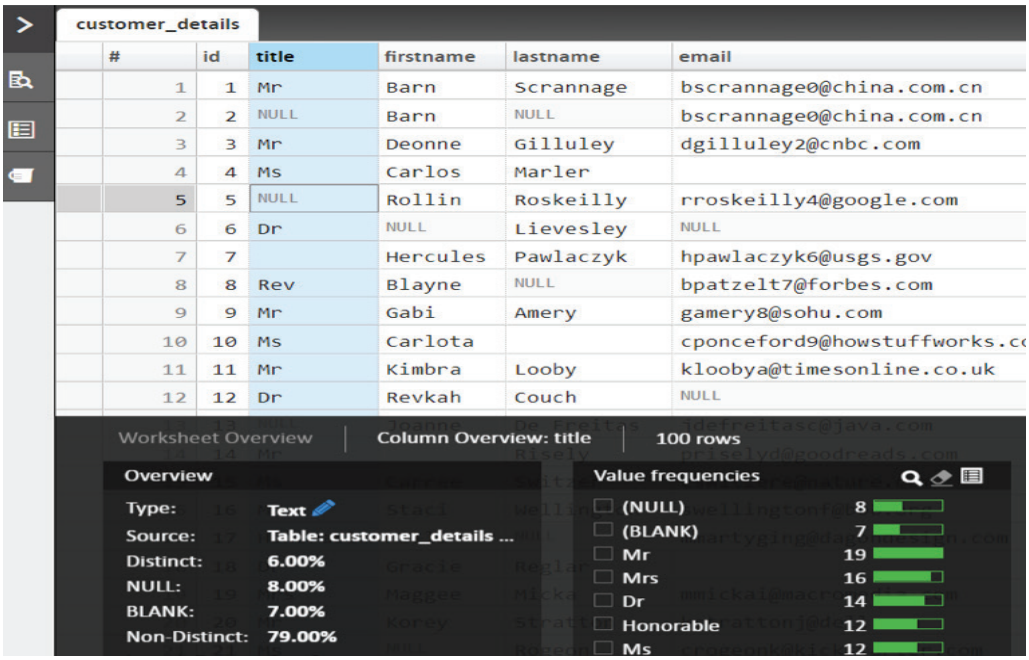

Figure 5: Empower users with intelligent error handling, prebuilt data cleansing, and transformation.

Intelligent Data Quality and Data Masking

Informatica Enterprise Data Preparation provides a holistic and agile approach to managing data quality while enriching and standardizing any data at scale. Users can easily manage data quality tasks, visualize data quality levels, and react quickly to any potential problems. Powerful masking and encryption capabilities to identify and protect sensitive and personal data enables users to enforce adherence to security and privacy policies. With trusted, high-quality, and protected datasets, DataOps teams can enable more accurate analysis.

Operationalized Data Preparation With Reusable Workflows

DataOps teams often have to repeat data preparation activities on new sets of data, which squanders any gains from ongoing scale and reusability. With Informatica Enterprise Data Preparation, all steps are recorded in recipes enabling users to automatically generate data flows that can be scheduled on a repeatable basis to operationalize analytical insights.

Fully Governed User Privilege Control

Governance is critical to any data preparation initiative, especially in self-service environments. Informatica Enterprise Data Preparation provides comprehensive IT-governed user activity control for import, upload, publish, export, or download activities on various files and relational resources in the data lake.

Key Benefits

Easily Discover and Collaborate on All Your Data in On-Premises and Cloud Data Lakes

Data cataloging is the foundational first step for any modern data preparation initiative. With petabytes of data residing in on-premises and cloud data lakes, DataOps teams including data scientists, data engineers, and data analysts can use the AI-powered Informatica Enterprise Data Catalog to easily find the data they have with Google-like semantic search. With Informatica Enterprise Data Catalog, teams can understand key attributes about the datasets, including business context and relevancy, and determine whether the data is certified, whether it comes from trusted sources, who owns the datasets, and obtain a holistic relationship view with end-to-end data lineage and impact analysis. Additionally, an automated machine learning-based discovery process transforms related data assets into intelligent recommendations that may be of interest to users. This greatly increases confidence and reduces duplicate datasets from being created for similar projects. Moreover, to drive greater efficiencies and overcome data silos, advanced social curation and data collaboration capabilities enable DataOps teams to easily share, collaborate on, review, rate, and follow the datasets that are of interest to them. This "wisdom of crowds" helps to enrich and curate data, making it even more valuable when prepping and operationalizing data for use and for self-service analytics at enterprise scale.

Empower DataOps Teams With AI-Powered, End-to-End Data Preparation and Intelligent Automation

The sheer complexity of the data that resides in data lakes requires AI-powered automation for modern data preparation initiatives at enterprise scale. Informatica Enterprise Data Preparation leverages the power of the CLAIRE engine and the use of advanced machine learning algorithms to automate various tasks in the data preparation pipeline such as data discovery, inferring relationships about datasets, pattern recognition, recommendation of alternative datasets, and recipe recommendations. With an intuitive, easy-to-use Excel-like user interface coupled with various prebuilt data cleansing, blending, transformation, data quality, and masking capabilities, users can iteratively explore, combine, clean, enrich, and transform raw data into curated datasets for self-service data integration and BI- or analytics-driven use cases as well as for building robust and accurate machine learning models in conjunction with DataRobot.

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.

Operationalize Data Preparation at Enterprise Scale

To truly derive value from the data that resides across on-premises and cloud data lakes, DataOps teams must think of operationalization of curated and governed datasets at enterprise scale. With Informatica Enterprise Data Preparation, all steps in the data preparation pipeline are recorded in recipes enabling users to automatically generate data flows that can be scheduled on a repeatable basis to operationalize machine learning models and analytical insights. Users can build, manage, and deploy the lifecycle of the data preparation pipeline at scale across Apache Spark, Kafka, Hive, Amazon S3, and Microsoft ADLS data lakes. The end-to-end data preparation capabilities empower DataOps teams to operationalize the process of data pipelines with comprehensive support for governance, performance, and scalability.

For more information, visit the Informatica Enterprise Data Preparation product page.

Informatica

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

IN06_1120_03237

© Copyright Informatica LLC 2019. Informatica, the Informatica logo, and CLAIRE 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.