VIEW POINT

INTERPRETED DOMAIN-SPECIFIC MODELLING LANGUAGE - AI ADVISORY FOR DATA MANAGEMENT LIFECYCLE

Abstract

AI requires intelligent data management and wrangling capabilities to align to developed AI models and utilize the transformed organization data. This needs a robust end-to-end data lineage. The success of AI is dependent on the effectiveness of the training and prediction models designed by data scientists and guided techniques for understanding underlying data.

Data science and AI Advisory are key for deriving intelligence and insights from data.

Infosys®
Navigate your next

Need for Data AI Advisory

The overall volume of data traffic is expected to reach 20.6 zettabytes in 2023, while the number of connected devices and connections is projected to reach more than 25 billion by 2023. All this data needs to be processed and made usable and trustworthy while adhering to governance policies.

Due to inherent data volumes within organizations and limitations and/or restrictions on manual approaches, automated data cataloguing and re-certification are critical to ensure transparency, security, currency, and the final value. Governance is essential.

That is why AI is the preferred solution approach.

Data mining, machine learning, and artificial intelligence are all interrelated in making sense of data stacks in organizations. Some are overarching terms and involve a healthy mix of technical, functional, and application & system-related interplay as data meanders through enterprises.

Data Mining refers to the patterns, algorithms, and techniques leveraged in various tools to uncover insights from data and its mutations and variations. It aims to understand the correlation of data, primarily historic data, across systems and applications

Machine Learning is the process of training machines to sift through historical data and identify patterns and decisions without explicit programmatic rules and modules

Artificial Intelligence is the next level of machine learning, whereby patterns and algorithms are used, through sentiment analysis or similar techniques, to predict and suggest user actions for problem-solving, reasoning, planning, adapting, decisioning, and inferencing intelligently and independently. It is inspired by how the human brain learns and develops likes, preconceptions, and predilections to analyse and replicate human behaviour. It essentially enables a machine to emulate human behaviour.

To summarize, the resultant tool offering (IDSML from Infosys) needs to have capabilities for a complete machine learning life cycle from gathering data, wrangling, exploratory analysis, feature engineering, and modeling.

AI requires intelligent data management and wrangling capabilities to align to developed AI models and utilizes the transformed organization data. This needs a robust end-to-end data lineage. The success of AI is depends on the effectiveness of the training and prediction models designed by data scientists. This is heavily dependent on the continuous availability of clean data.

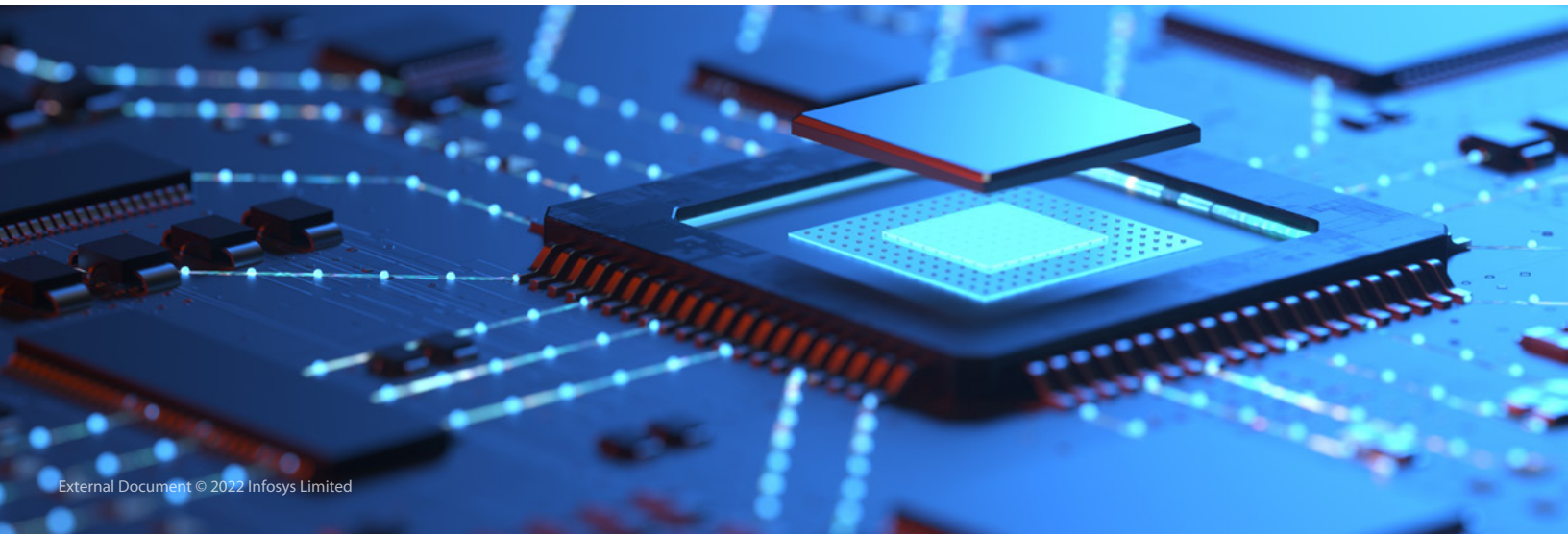
Major Requirements and Asks for AI Data Management

Evolving and Right-sizing as per Data Evolution

Typically, data entities in an organization are called, massaged, and transformed millions of times from the master data tables. As data churn around, understanding the business context and standardization is a key evolutionary memory needed. Understanding the context of key mappings and inferring the consequent integration task is integral to a successful AI toolset.

Automated Governance

Associating business context to technical data tables and leveraging semantics to link key business processes is an invaluable capability for any governance use case. Automatically updating metadata of the ever-growing business catalogue not only ensures organic evolution but also helps reconstitute relationships, as customer data mining reveals fine-grained and more targeted business services introduction.



Major Features and Benefits of IDSML AI Data Management Advisory

Key Design Principles

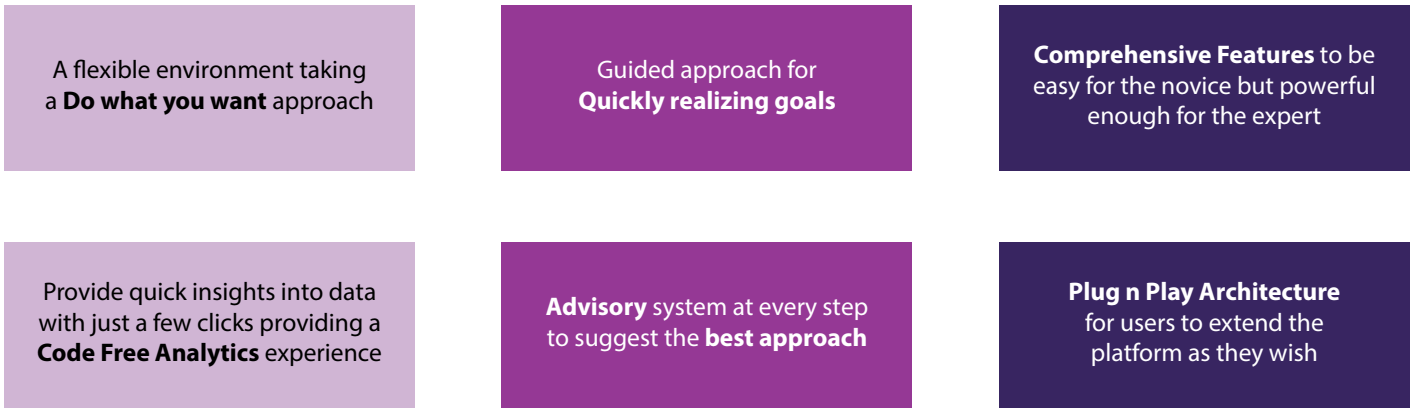


Figure 1: Key Business Asks and Features Offered

Integrated Data Management Platform

The AI platform needs to be a warehouse for all data types. For comprehensive data storage, it needs connectors to ingest and maintain an updated inventory and manage periodic scans of all data origination and transformation sources across enterprise environments, including legacy, mainframe systems, and custom-coded enterprise and ERP applications. It is critical to leverage metadata from ETL and discover lineage from applications, even without direct access and inside firewalls and gateway(s).

Data Collection and Exploration

It is imperative to remove opacity and visualize data as it traverses across enterprise applications, including multiple data sources on-premises data sources, legacy databases, data warehouses, data lakes and mainframe systems, or SaaS applications and multi-cloud environments. Visibility should be extended to also includes ETL software, SQL scripts, programming languages, stored procedure codes, and "black box" applications

Metadata repository and Visualization

SQL ETL, databases, enterprise, and custom applications create data about data (metadata). This metadata and master data are key to mastering the data and data stewardship capabilities

Data Cleansing

As organizations grow organically, data entities within also grow like hydra-headed monsters. It is necessary to undertake periodic house cleansing of all entities that, at a minimum, include the below data capabilities -

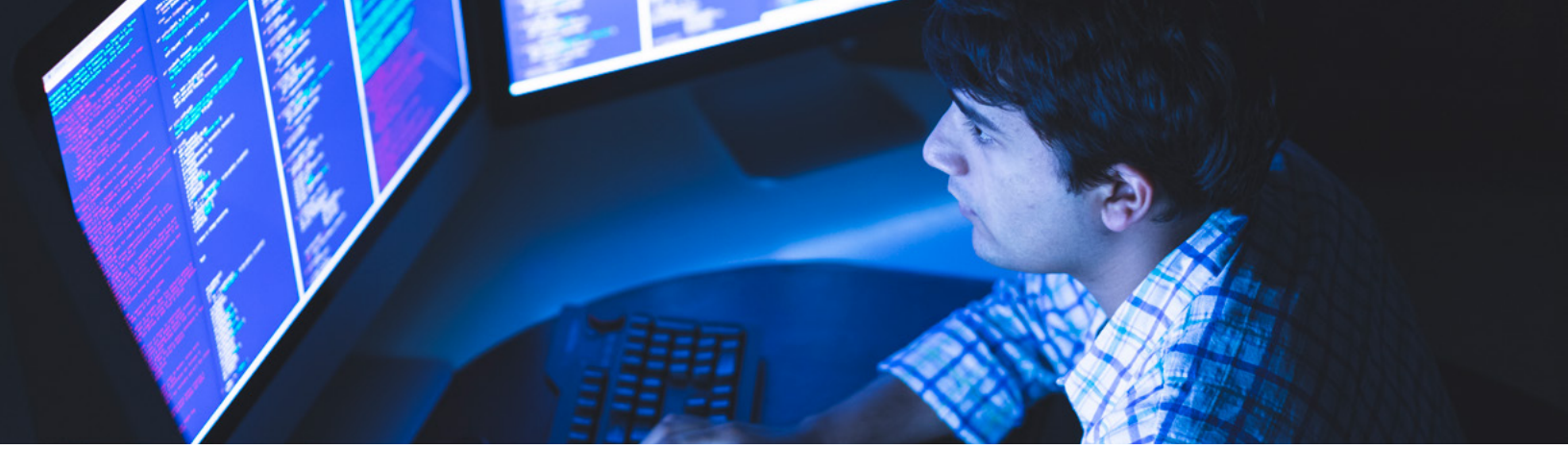
1. **Data Profiling** and **Business rule validation** against iterative re-usable business rule sets
2. Automated data and master data management using rules-driven profiling, cleansing de-duplication, and harmonizing dirty data
3. Automated data quality processing and regression testing for duplicate analysis and grouping strategy to enable **Match and Merge**
4. Use **contextual attributes**, **predictive** and **prescriptive rules** to discover and apply patterns and fuzzy logic algorithms

Entity Relationship Discovery and Maintenance

AI toolset progressively understands how disparate entities are integrated into a business context and learns from the processes to suggest automated classification and cataloguing for new entities. Tagging, data similarity, and semantic capabilities aid simple discovery and automated catalogue updation essential for the process.

Semantics-enabled Domain Association Discovery

Post creation of organization-level master data catalogue that (semi) automatically stitches and suggests lineage from all enterprise source(s) and application(s), plan to future proof the exercise with an 'impact analysis' and/or a 'what if' Monte Carlo tree view analysis leveraging the capabilities of integrated data platform. Data wrangling advisory options and model development are key enablers that can be effectively leveraged.



IDSML AI toolset offers the below key capabilities for data science and engineering –

Data Wrangling

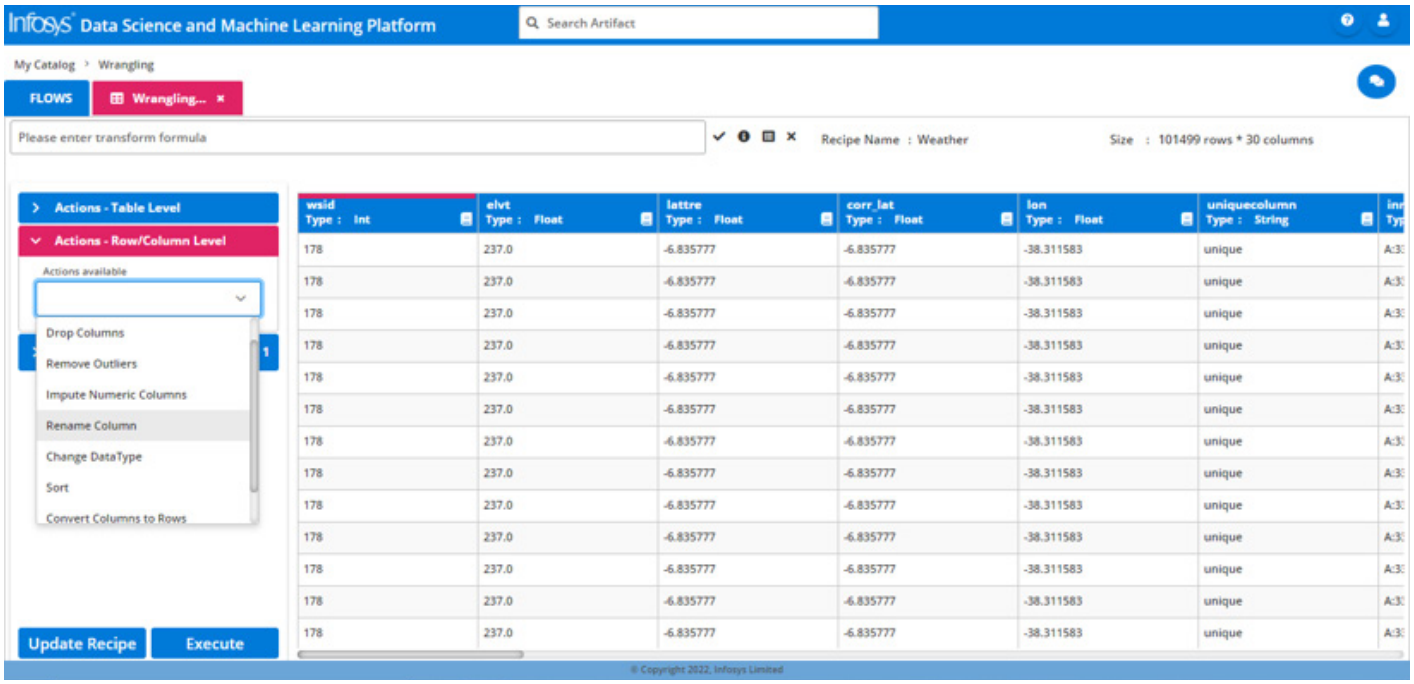


Figure 2: Data Wrangling

Bivariate and Multivariate Analysis

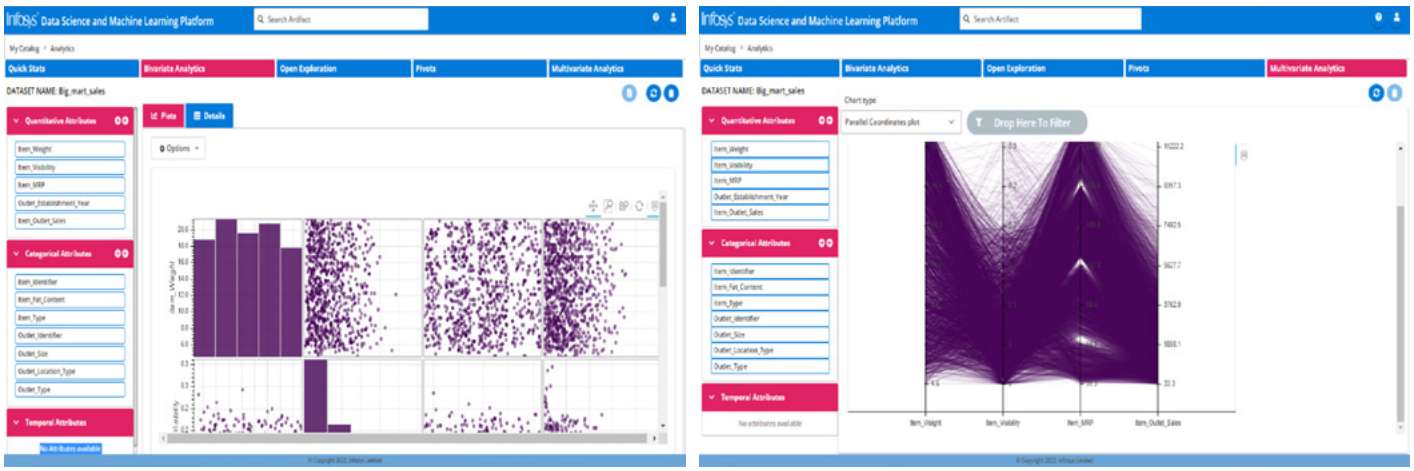


Figure 3: AI Tool Intelligent Data Analysis

Feature Engineering

Feature engineering capabilities enable the following curation possibilities and suggestions based on the historic data patterns, which may be accepted or manually improvised

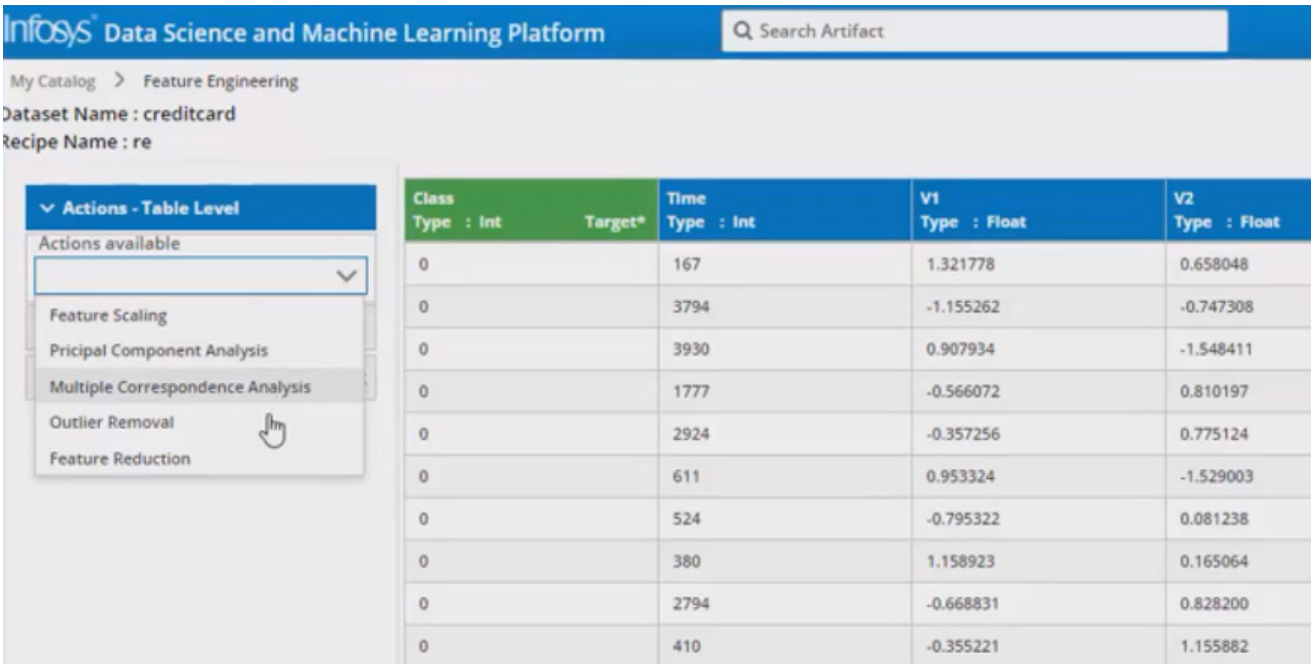


Figure 4: Leveraging Feature Engineering Capabilities

Intelligent Recommendations

Another interesting facet - due to its intelligence, the IDSML AI toolset can recommend transformation and expression recommendations by learning from and analysing the historic data. Feature engineering capabilities enable the following curation possibilities and suggestions based on the historic data patterns, which may be accepted or manually improvised.

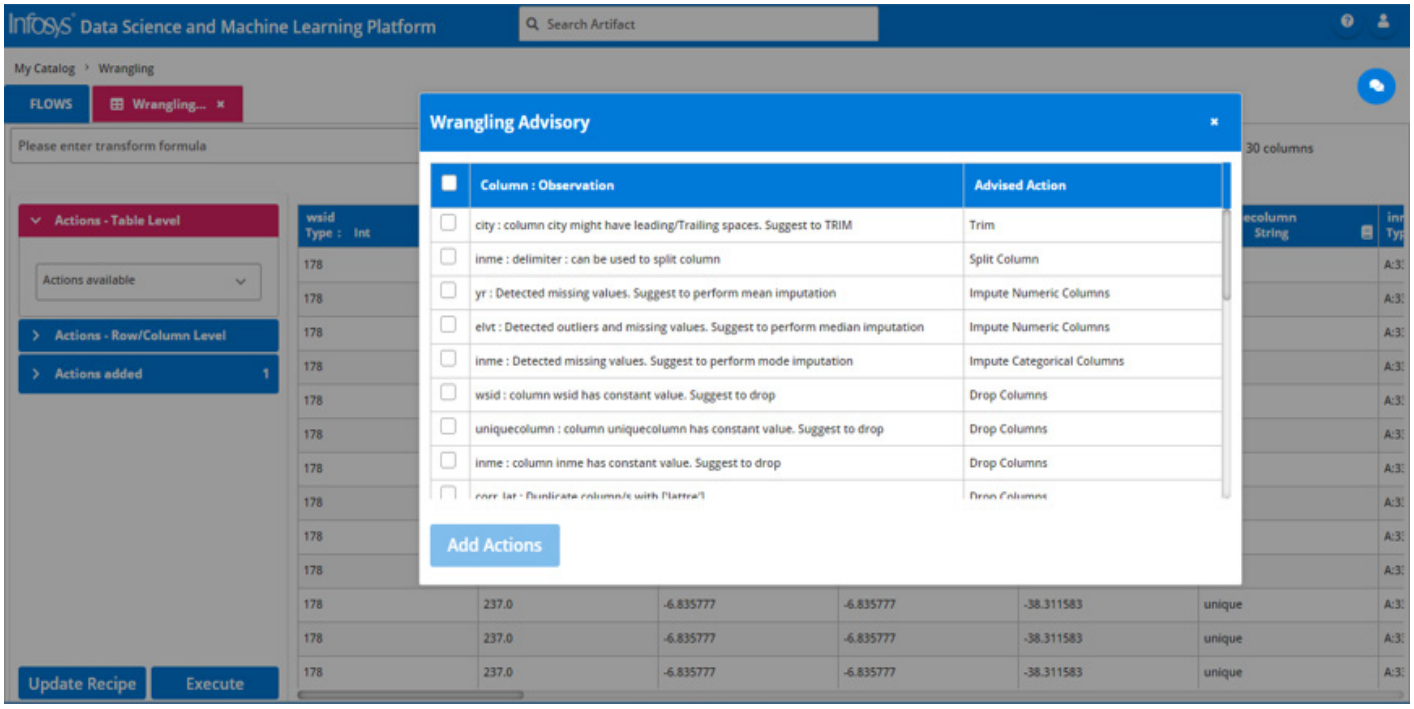


Figure 5: Leveraging Intelligent Recommendation Capabilities

Pipelining

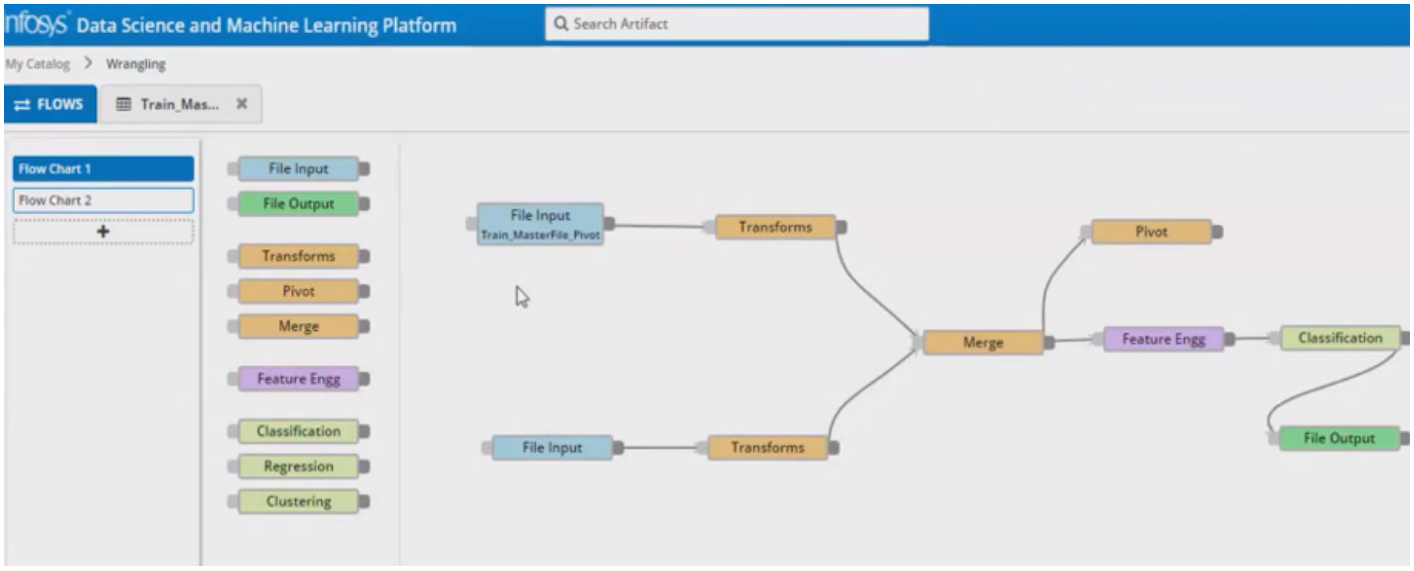


Figure 6: Leveraging AI Toolset for a Data Intelligent Organization

The data science pipeline refers to the process and tools used to gather raw data from multiple sources, clean the gathered data, analysis, feature engineering, model building, and prediction. IDSML offers automated pipelines for the machine learning life cycle.

IDSML offers a wide variety of guided techniques for understanding underlying data - ranging from deterministic, heuristic, and probabilistic algorithms to contextual synthesis matching and active learning entity matching. These are deployed in a plug-and-play architecture construct to provide rapid and scalable data insight techniques that have the ability to manipulate data matches and enrich data and/or master data. Added features are contextual attributes' intelligence capabilities to suggest joins - Artificial Intelligence (AI) powered lineage discovery brings insights into "control" relationships, as joins and logical-to-physical models. E.g., deleting a column used in a join can impact a report that depends on that join.

All these techniques enable a comprehensive discovery and impact analysis, especially when the relationships are undocumented, undiscovered, or dynamic.

A summary of the end-to-end data journey is presented below for quick validation

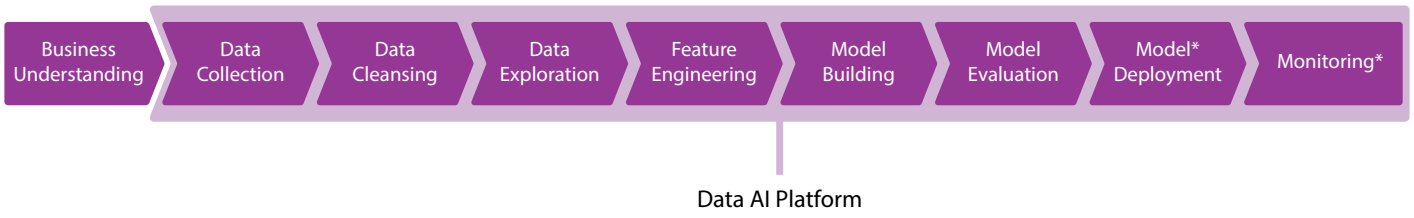


Figure 7: End-to-End AI Lifecycle Capabilities of Infosys IDSML

And the tool offers an almost entirely Low code or Code free analytics experience for the newbie as well as the experienced data scientist.

The focus has been to create a single platform that can help alleviate all these issues and be flexible enough for savvy scientists to tinker with and change it for their specific needs.

About the Authors

Eggonu Vengal Reddy

Eggonu Vengal Reddy is a Principal Product Architect with over 20 years of experience in Data Management, specifically Data Warehousing, Data Modeling, Big Data, and Data Science. He has provided architecture and design to develop tools and solutions to handle enterprise-wide database migrations, master data management, data quality and wrangling, explorative analysis, and feature engineering in the Machine Learning life cycle.

Gopinadh Bapatla

Gopinadh Bapatla is a Technology Architect with over 20 years of experience in Application Development, Data Analysis, Data Science, Machine Learning, and Big Data. He has provided architecture and design to develop tools and solutions to handle data quality by data wrangling, explorative analysis, feature engineering, and building Machine Learning models.

Tushar Subhra Das

Tushar Subhra Das is a Senior Business Data Analyst with over 10 years of experience in Data Migration and Governance. He has worked with Europe and Australia-based insurance and logistics clients for Data migration, MDM and Data Quality, and process governance. In his current role, Tushar is responsible for APAC and EMEA data migration deployments and enhancements, including product developments for iDSS as the next-generation industry-standard data management platform.

For more information, contact askus@infosys.com

Infosys®
Navigate your next