Infosys topaz

WHITE PAPER

GENAI BASED DATA QUALITY, PROFILING AND HARMONIZATION

OVERVIEW

Data quality is a significant issue and stumbling block across all organizations, that use data analysis for deriving intelligent and informed business decisions. Unfortunately, this is across the industry and domains, and most data sources are riddled with various inaccuracies that make them unreliable, and worse with potential risks or perils.

About 79% of business leaders think data keeps people focused on the things that matter and that are relevant to the business, while, 76% think that data helps minimise the influence of personal opinions or egos in a business conversation.[1, 2]

Infosys®
Navigate your next

Infosys Solution Infosys Data Workbench (IDW) Description

IDW is the Infosys proprietary end-to-end solution that addresses enterprise level data quality challenges. It is bundled with Machine learning and AI capabilities offering functionalities as Data Profiling, Data Cleansing and Standardization.

IDW has been developed with Analytical MDM capabilities using traditional and ML based techniques. It has a comprehensive set of tools for data profiling, data standardization, address standardization using Google and Bing APIs customized for Insurance underwriting processes. IDW leverages supervised and unsupervised learning models to detect data anomalies and help in missing values correction. The MDM module has different matching techniques using deterministic, fuzzy, phonetic or hybrid approaches and ML based approaches to identify duplicates and generate a golden record using survivorship rules.

Key lifecycle process steps in sequence for IDW solution are depicted below for reference.

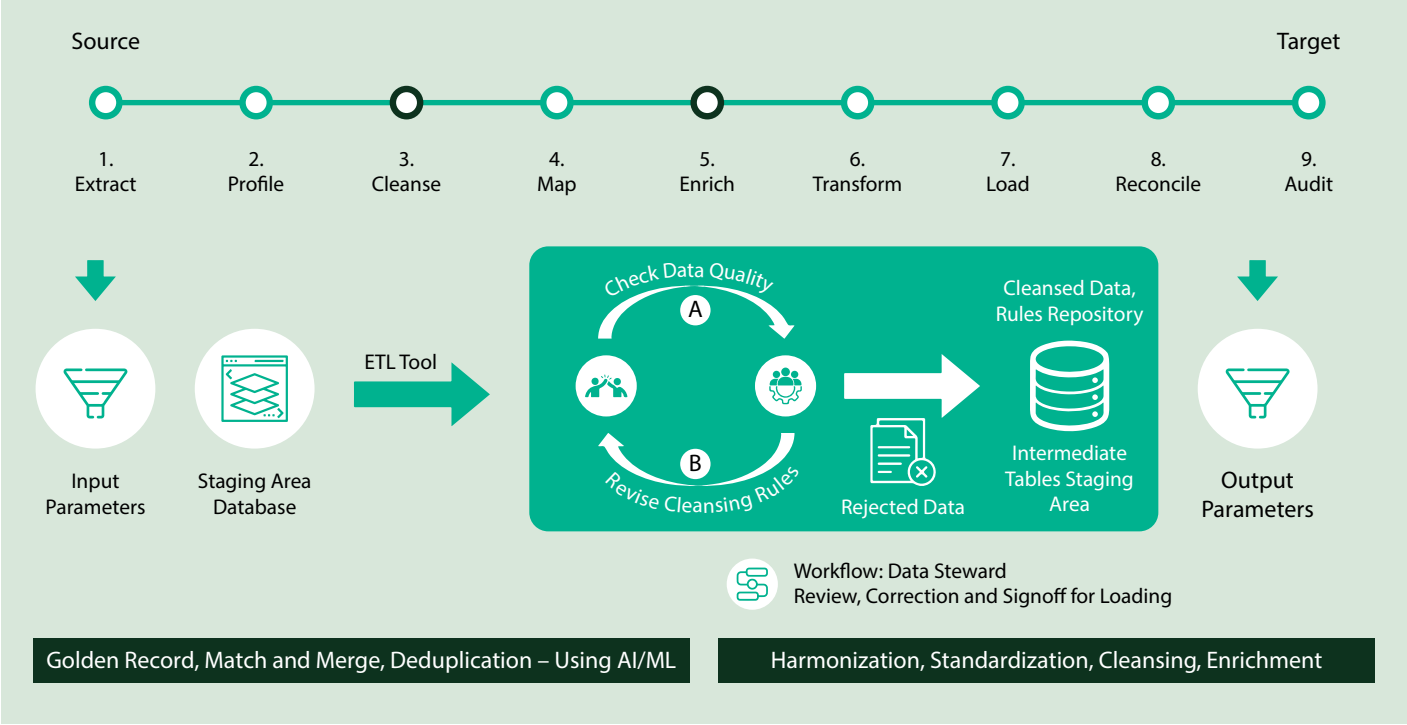


Figure 1: Key Data Quality Life Cycle Activities

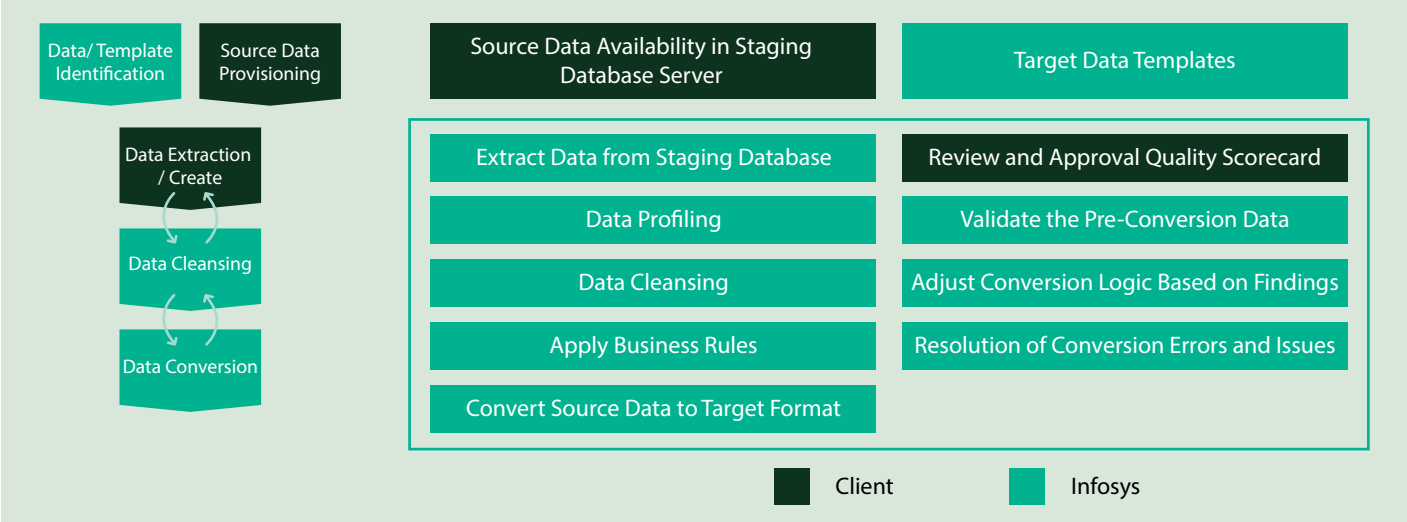


Figure 2: Key Data Quality Life Cycle Activities for Project Deployment

Once the data objects have been identified, the next steps will be to extract and cleanse the data, consolidate and transform the data in required format before uploading the same to the staging area.

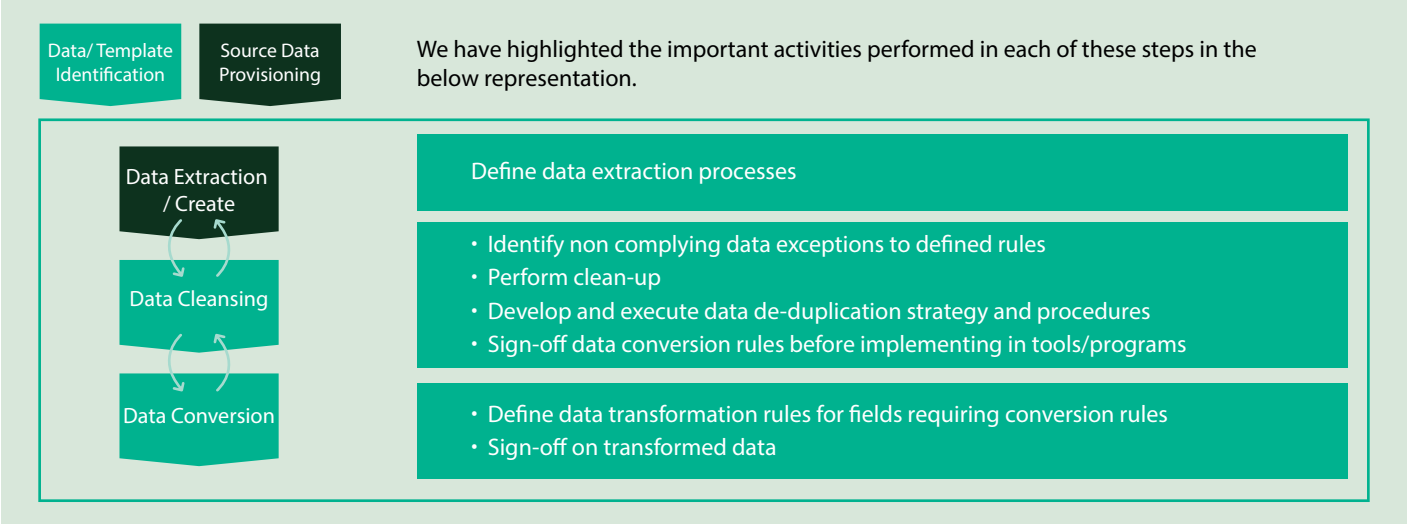


Figure 3: Sequential Data Quality Life Cycle Activities

1.1. Data Quality and Master Data Management

Profiling broadly examines parameters such as data values, value ranges, frequency distributions, metadata mismatches, various statistics, non-standard record formats, etc. The outcome from data profiling and data analysis will help to decide on the cleansing rules to not only standardize the data objects but also enrich the data objects to meet business regulatory requirements.

This is managed for multiple data categories:

- Master Data
- Transactional Data
- Historical Data

Business Rules for Master Data objects must be linked to transactional, and even historical data. Obsolete or incorrect data is determined by reviewing the linked transactional data and linked master data. For this reason, identification must occur in the source system(s).

There could be many contributors to incorrect / inconsistent data in legacy systems, for example:

- Entry errors
- Dummy values
- Multi-purpose fields
- Cryptic data fields
- Contradicting data values

1.2. Data Profiling

Data Profiling is the process of examining the data in an existing data source and collecting information about the data. Profiling is carried out on the legacy data extracts to gather relevant quality statistics from it. This activity feeds other data migration processes like data analysis, cleansing, data mapping, etc.

Data profiling is a technically led activity and requires extensive collaborations with Functional and Business SMEs /groups from client and domain for support and input.

- **Business Subject Matter Experts** - To identify known business process and data quality issues associated with the use of the source system and data to support the activities of the business area.
- **Technical Subject Matter Experts** - To identify technical and data quality issues associated with the source system.
- **Data Governance Board** - To guide the Data Cleansing / Migration team in the correct resolution of identified data quality issues.
- Data profiling offers a multitude of benefits, including improved data quality, faster project implementation, and enhanced user comprehension. A key advantage is its ability to unlock valuable business insights hidden within the data itself. This makes it a cornerstone technology for ensuring data accuracy in corporate databases.

1.2.1. Sample Data Quality Issues in Legacy Sources

#	Category	Examples	Fix for One-Time cleansing	Long Term Fix
1	Attributes with Blank or Invalid Values **(Completeness)**	• Mobile number, email address, country code, gender have blank or invalid values. • International prefix for mobile numbers has not been configured.	Cleansing rules can be defined to correct the values / derive them from other attributes if possible.	• Data capture mechanisms should be improved to have drop downs, checkboxes, and radio buttons where possible instead of Text boxes, auto population of few attributes. • validation/ verification mechanism to be implemented at source.
2	Lack of Accurate Information **(Accuracy)**	• Customer address • Different email address provided in different instances for the same customer.	• Business rules can be defined to get address details from reference databases based on zip code / national ID etc. • Survivorship rules can be defined to identify the most appropriate value for attribute.	Data capture mechanism to be improved by having important attributes as mandatory fields.
3	Lack of Data **(Completeness)**	Employment Information, PCI Information, company size related info, or third party involved in the accident are not documented.	Business rules can be defined to get these details from third party reference Databases based on other attributes if possible.	Data capture mechanism should be enhanced by designating important attributes as mandatory fields.
4	Inconsistent formats across Applications or lines of business **(Inconsistency)**	Gender	Data standardizations rules	Data standardization at source
5	Capturing More information than needed **(Accuracy)**		Source fields assessment exercise has to be carried out to identify irrelevant fields and decommission them.	Source fields Assessment exercise to be carried out to identify irrelevant fields and decommission them.
6	Duplicate Records **(Duplication)**	• Records with almost same name, mobile no, and email address. • If a customer has a company and is its CEO, then he will have two separate customer ids.	De-duplication rules can be defined to check for duplicate records based on a set of attributes.	Before creating a new customer, validation mechanism to check if a customer record with same name/ email / mobile no exists in the database needs to be implemented.
7	Lack of validation or controls at client creation - process to identify if the customer already exists. **(Duplication)**	Policies being issued with variants of the client names.	De-duplication rules can be defined to check for duplicate records based on a set of attributes.	Business process change to implement duplicate customer checks at the source itself.
8	Orphan Records **(Duplication)**	Orphan customers without policy.	Business rules can be defined to identify orphan customers and delete them after merging any important attribute data to the master record.	Business process change to delete orphan customers once it turns out to be a policy or after a certain time period.

#	Category	Examples	Fix for One-Time cleansing	Long Term Fix
9	Data Integration Issue **(Integrity)**	• Disconnect between member details and customer details across applications. • Independent applications for each LOB.	Linking & matching rules can be defined to a certain extent across the data sources.	MDM helps in identifying the golden customer record.
10	No reports to identify **Data Quality** issues	Periodic reports to Identify how many duplicate records are created in the last month.	DQ reports can be configured.	DQ reports can be configured.

Table 1: Common Data Quality Issues

1.2.2. IDW Enables the Above by the Following Capabilities

- Enhanced Data Profiling: IDW performs source data profiling and generates reports that helps in drawing out DQ Rules for further implementation.
- Enhanced Productivity: IDW offers a 40% or more productivity boost compared to manual methods of data profiling.

1.3. Data Cleansing

An iterative cleansing process starts as soon as the data quality has been assessed and data anomalies have been identified & signed off. Data cleansing includes the following iterative steps:

- Elimination of obsolete records.
- Removal of duplicate records.
- Correcting inaccurate records.
- Correcting incomplete records.

1.3.1. Identify and Resolving Missing Data

The recommended approaches to resolve such issue include:

- Using standard database tables or excel worksheets for data manipulation.
- Governance guidelines around master data during data population at source.
- Filling in missing data using data load programs, either through calculations or reference tables.

They help in achieving the following graduated Data Maturity for organizations.

1.3.2. IDW Helps in the Above Activities by the Following

- Enables solution to pre-fill unknown master attributes in transactional data by mapping to history/master data using Machine Learning models.
- Manual effort reduction by 80%+.
- Solution to de-duplicate master data from multiple sources and generate golden records.
- Low-cost solution for DQ assessment or one time data De-duplication.
- Data cleansing activities are performed to ensure availability of clean business data.
- Data needs to be standardized and correctly formatted to allow intelligence (e.g., Addresses should be derived from standard addresses as per Google / Bing, while legacy applications might allow a free text format).
- Mandatory data fields need to be populated with NOT NULL values to ensure all fields with NOT NULL attribute values hold correct data, else blanks are failures.
- Data De-duplication needs to be performed to master data to be populated across defined business applications. Strategy to identify, remove and rectify duplicate records can be agreed upon – either at source or during the extraction and conversion process.
- Data cleansing should pick up inaccurate data fields and apply business rules on data exclusion, or transformation and standardization to relevant entities.

Below operations are available out-of-the box. Custom rules can be easily built based on project's requirements

Operation	Description
Remove Special Chars	Removes special characters if present
Remove Spaces	Removes in-between or leading/trailing spaces
Remove Numbers	Removes numbers in data if present
Remove Alphabets	Removes alphabets in data if present
Value Conversions	Converts old value to new value. Ex: Male to M, Central Ave to Central Avenue
Pattern Standardization	Converts old pattern to new pattern. Ex: 124469 to 124-(469)

Table 2: Data Cleansing Tasks

1.4. Data Standardization

Data Standardization is achieved by Cross-system data standardization to achieve integrated data aligned with target system data, based on specific pre-defined rules. IDW also facilitates De-duplication of data within systems and across multiple systems.

1.4.1. IDW Helps in the Above Activities by the Following:

IDW enables discovery and consolidation of Golden records using patented algorithms, following the 3-step process of **Identifying Duplicates and Golden Records, De-duplication, and merging as per pre-defined Survivorship and Clustering rules.**

- Fuzzy match and phonetic match techniques are enabled for data deduplication.
- Partial matching / fuzzy logic feature is present in IDW which can be applied on relevant clusters
  (e.g., if name matches 80%, still it is considered in report as it could be a potential duplicate).

Clustering:

- When a group of records are identified to be potential duplicates they are grouped together and called a "Cluster".
- These clusters are then proposed to end users to confirm if they are actual duplicates or not.
- Once the duplicate vendors are identified, their other Information needs to be analyzed for duplicates, consolidate and merging.

Survivorship Rules

After 'grouping' the Customer's data, Survivorship rules are applied to **create** the **Golden Copy**.

Rule Scope	Rule Name	Description
Attribute	Most Complete	Longest String value is chosen
Attribute	Most Recent	Most recent value is picked
Attribute	Least Recent	Least recent value is picked
Attribute	Most Frequent	Most frequently occurring value is picked
Attribute	Trusted Source	Attribute value from the most trusted source is picked. e.g. Pick 'Address' where source_dept ='HR'
Attribute	Concatenated Value	Golden Record(s) will have concatenated value from all records for this attribute
Record Level	Trusted Source	The record from the most trusted source is picked
Record Level	Most Recent	Most recent record is picked
Record Level	Least Recent	Least recent record is picked

Table 3: Golden Record / De-duplication Parameters

Key Activities Details Performed (Including GenAI or ML Modules) Are:

1.5 Data Profiling ⇨ Data Cleansing ⇨ Data Enrichment (GenAI or ML Supported)

A workflow-based approach is used to enable checks and balances by different roles during the lifecycle.

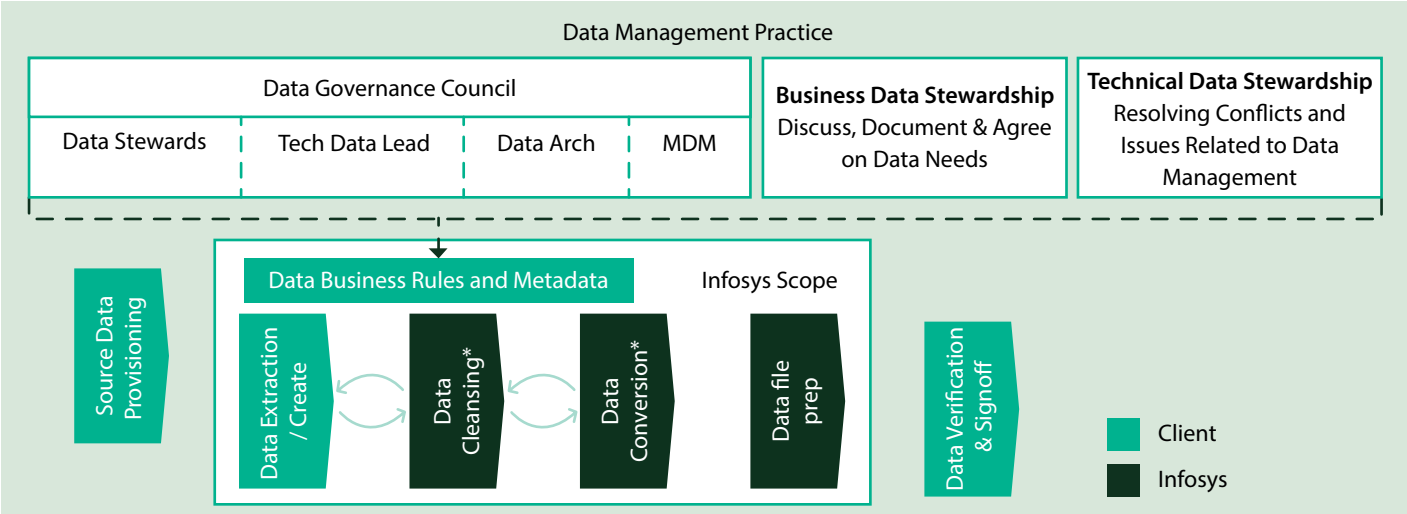


Figure 4 : Business View Representation of Data Quality Process

Modular Approach to Data Quality

Data Profiling Aids in the assessment of Source Data Quality	**Data Standardization** Helps in standardization and cleansing of data	**Business Rule validation** Implements business-specific rules to evaluate the quality of source data	**Data Reconciliation** Identifies missing/out-of-sync data across two data sets	**Outlier Detection using ML** Finds the outlier data in given dataset
Enrichment using ML Provides a solution to Pre-Fill unknown attribute values based on past data	**Data De-duplication** Provides solution to de-duplicate master data from multiple sources.	**Data Stewardship** Aids Business SMEs in the selection of matched records that are eligible for golden record generation.	**Golden Record Generation** Identifies golden records based on defined business rules and thresholds.	**Identity Resolution** Helps in linking data between multiple datasets using AI data matching techniques
Search and Lineage Aids in search of master data in large data sets using Solr.	**Match and Merge Editor** Enrich Golden records by editing and fixing data quality issues in the Master Data.	**Rule Extraction using ML** Provides data associations between elements of given data set	**Data Editor** Helps to edit the data records	**Dashboard and Reports** Interactive BI dashboard to derive Data Quality insights and generate customized reports.

Figure 5: Key Modules & Capabilities of IDW - Infosys Data Workbench Quality

2. GENAI Supported Data Quality

Infosys Data Workbench is a modular application and is coupled with GenAI, analytical MDM capabilities and Machine Learning (ML) based features. IDW is available as an On-Premises or Cloud (SaaS) offering, with options to host in different cloud environments.

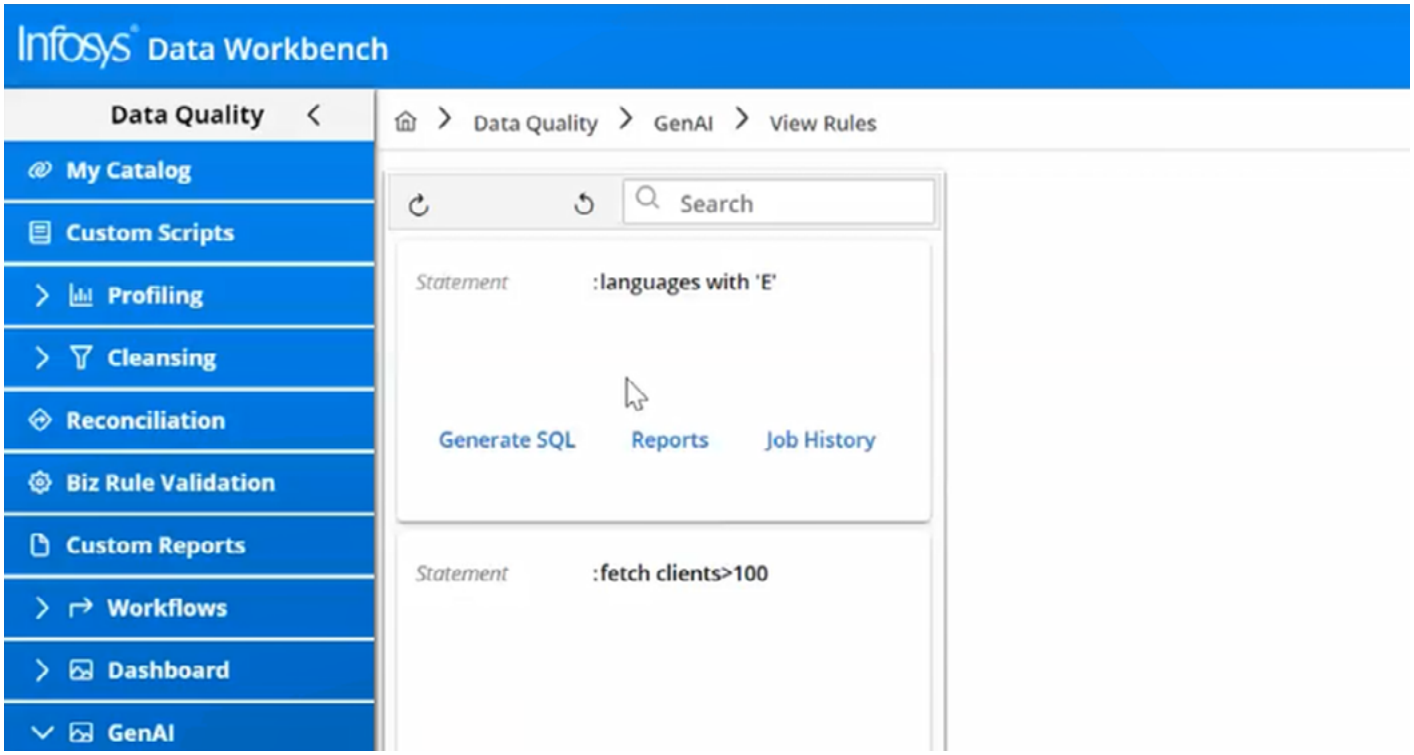
The following AI enabled modules are available out-of-the-box, complementing above mentioned programmatic data profiling, cleansing and enrichment features, and are detailed as below:

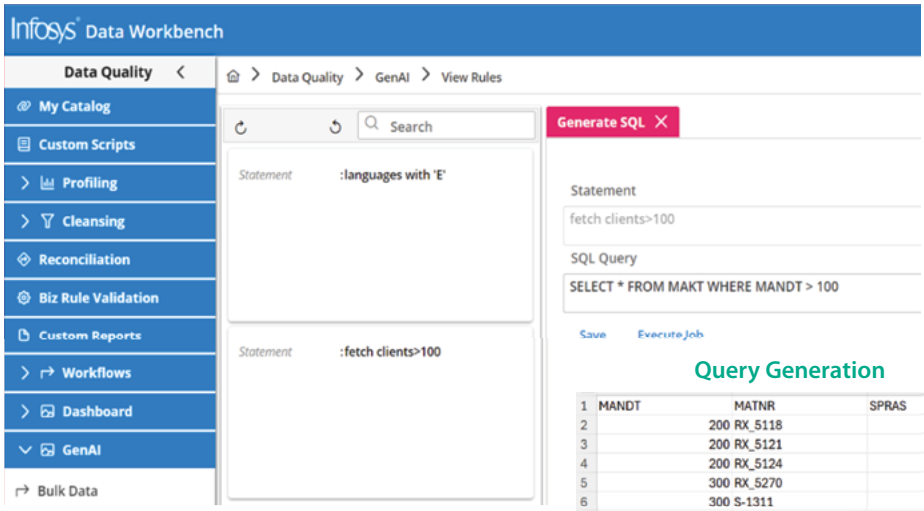
2.1. Generating SQL Queries for Profiling

Tasks needed for detecting and identification of issues in Table 1: Common Data Quality issues can be created in IDW using GenAI queries. A Business user can update a .csv template in NLP with business details e.g. table and fields details in the attached table.

Source Table	Business Rule	DBtype
MAKT	languages with 'E'	MYSQL
MAKT	fetch clients>100	MYSQL

Once populated, compatible SQL queries can be generated by IDW based on natural language queries posed by business Users, even in bulk mode. These can be a combination of filter based or flowchart algorithm queries and be executed in a workflow mode to enable sequential data profiling.





The NLP Input statements are converted into Queries for performing Data Quality Checks on the Input Data.

Query Generation

Statement
fetch clients>100

SQL Query
SELECT * FROM MAKT WHERE MANDT > 100

Save Execute Job

Profiling Report

	MANDT	MATNR	SPRAS	MAKTX	MAKTG	ProfilingStatus
2	200	RX_5118		1 Counterbalancing Syste	COUNTERBALANCING S	Y
3	200	RX_5121		1 Hand	HAND	Y
4	200	RX_5124		1 Centre counter balance	CENTRE COUNTER BAL	Y
5	300	RX_5270		1 Drive	DRIVE	Y
6	300	S-1311		1 Internal services	INTERNAL SERVICES	Y
7	100	T-ZS206		1 Desktop Repair Service(	DESKTOP REPAIR SERVI	N
8	100	SPPSJ_CONS		1 Consulting Services	CONSULTING SERVICES	N
9	100	ROW(B2671)	E	1-Feb-21	1-Feb-21	N
10	100	2969		1 247391-BTW_SHFT_1_2	247391-BTW_SHFT_1_2	N
11	100	T-FF200		1 Fertilizer, liquid, form B	FERTILIZER, LIQUID, FO	N
12	100	R47201-RETRNG472_4		1 R47201-RETRNG472_42	R47201-RETRNG472_42	N
13	100	ROW(B2086)	E	1-Feb-21	1-Feb-21	N
14	100	SPPSJ_ENG-JE		1 Engineering-Junior(Exte	ENGINEERING-JUNIOR(	N
15	100	790057-INTK_CAMSHF		1 790057-INTK_CAMSHFT	790057-INTK_CAMSHFT	N
16	100	3049		1 G1200-GSKT_EXH_PIPE	G1200-GSKT_EXH_PIPE	N

Data is Profiled based on the Query and Data Quality can be assessed through the reports

Figure 6 : IDW Module for NLP Supported SQL Generation

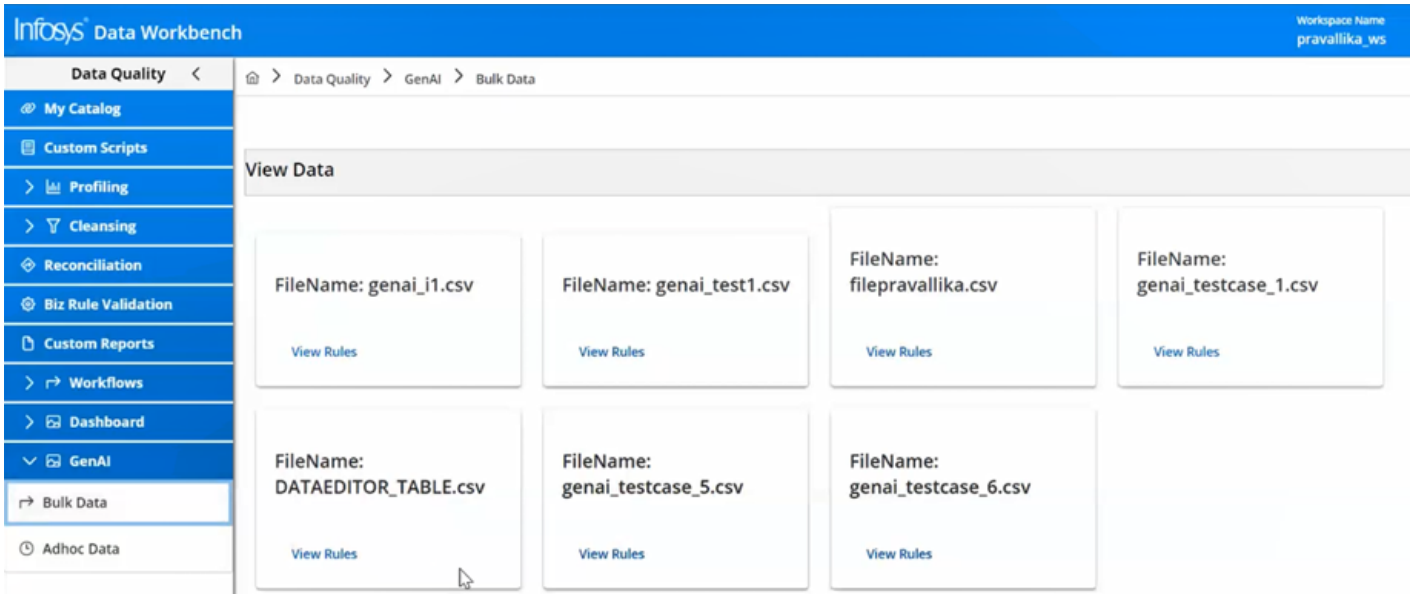


Figure 7: Bulk SQL Query Generation for Workflow Execution

MANDT	MATNR	SPRAS	MAKTX	MAKTG	ProfilingStatus	ProfilingRemarks	ROW_NUMBER	DATA_REMEDIA
200	RX_5118	1	Counterbalancing Syste	COUNTERBALANCING S	Y		1	N
200	RX_5121	1	Hand	HAND	Y		2	N
200	RX_5124	1	Centre counter balance	CENTRE COUNTER BAL	Y		3	N
300	RX_5270	1	Drive	DRIVE	Y		4	N
300	S-1311	1	Internal services	INTERNAL SERVICES	Y		5	N
100	T-ZS206	1	Desktop Repair Service(	DESKTOP REPAIR SERVI	N		6	N
100	SPPSJ_CONS	1	Consulting Services	CONSULTING SERVICES	N		7	N
100	ROW(B2671)	E	1-Feb-21	1-Feb-21	N		8	N
100	2989	1	247391-BTW_SHFT_1_2	247391-BTW_SHFT_1_2	N		9	N
100	T-FF200	1	Fertilizer, liquid, form B	FERTILIZER, LIQUID, FO	N		10	N
100	R47201-RETRNG472_4	1	R47201-RETRNG472_42	R47201-RETRNG472_42	N		11	N
100	ROW(B2086)	E	1-Feb-21	1-Feb-21	N		12	N
100	SPPSJ_ENG-JE	1	Engineering-Junior(Exte	ENGINEERING-JUNIOR(	N		13	N
100	790057-INTK_CAMSHF	1	790057-INTK_CAMSHFT	790057-INTK_CAMSHFT	N		14	N
100	3049	1	G1200-GSKT_EXH_PIPE	G1200-GSKT_EXH_PIPE	N		15	N

Figure 8: Sample Data Profiling Output Based on SQL Query Execution

2.2. Generating SQL Queries for Cleansing & Enrichment

Similar to Data Profiling, IDW leverages GenAI to generate equivalent SQL queries for cleansing, enrichment and/or standardization of the identified data with anomalies. IDW provides ability for business user to approve / reject / re-work on the tool generated output. The following are referenced:

Figure 9: IDW Leveraging GenAI for Cleansing

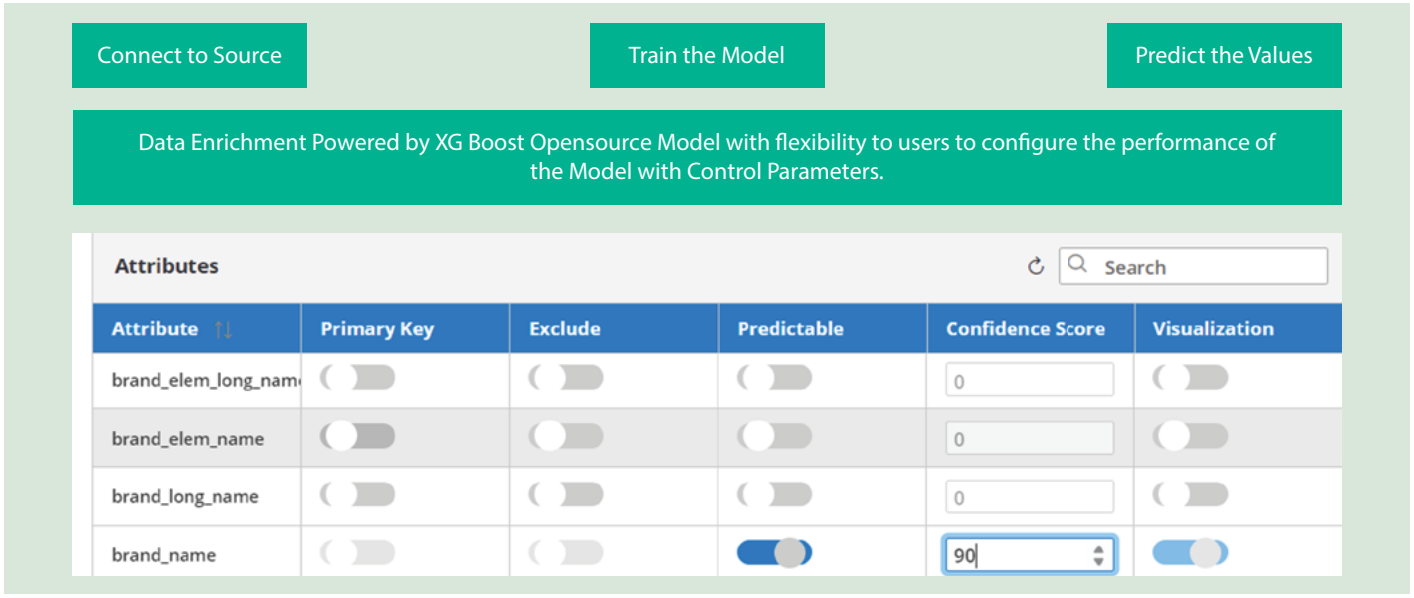


Figure 10: IDW Leveraging GenAI for Data Enrichment

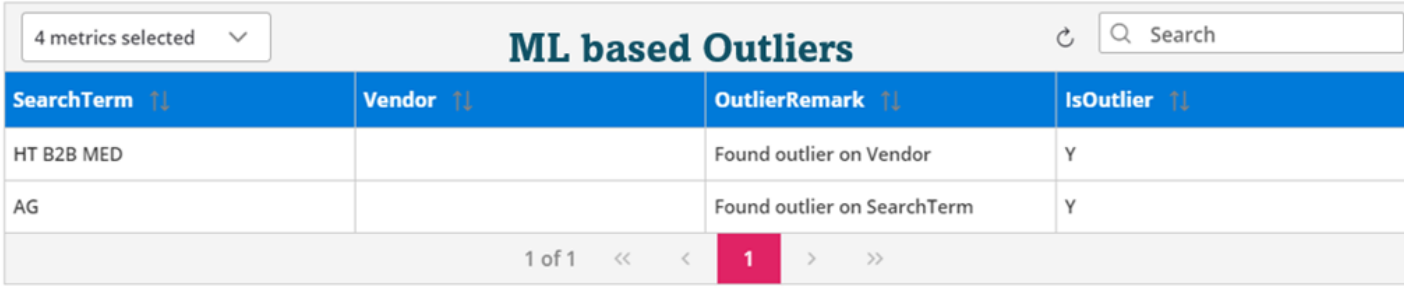


Figure 11: IDW Leveraging GenAI for Outlier Detection Based on Statistical Data Analysis

Infosys Data Workbench brings modular Machine-learning Capabilities for Outlier Detection, Associative Rule Mining, and Data Enrichment.

Outlier Detection

Uni-variate analysis to find Outliers in Numerical and Categorical data with the help of Machine Learning techniques.

Data Enrichment

Leverage Powerful Supervised Machine learning algorithms to enrich the data and fill the missing fields.

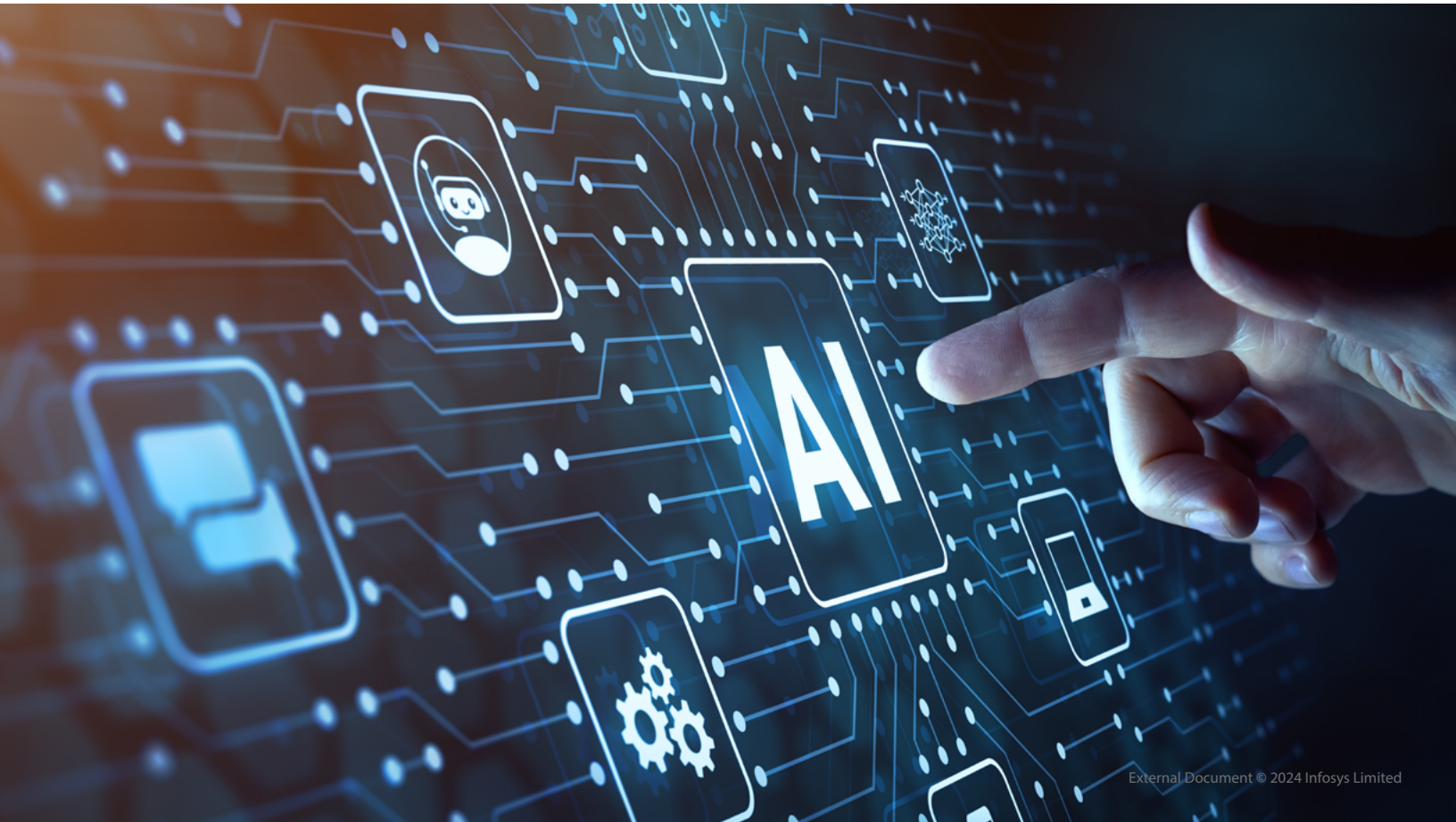
Rule Mining

Powerful associative rule mining engine to discover patterns and rules within data and also cleanse data simultaneously.

Figure 12: IDW USPs and Approach

Outlier Detection	Data Enrichment	Rule Mining	GEN - AI Data Profiler
Solution to Identify Outliers in a dataset.	A solution to pre-fill unknown attributes within the datasets by training the model with historical data	Provides data associations between elements of a given data set to discover patterns and Rules.	Accelerating Data profiling capabilities aiding business SMEs to understand data easily and intuitively.
Intuitive outlier reporting dashboards provides users to spot outliers	Identify the right confidence score and ML model Parameters of the predicted value before enriching the data	Cleanse data simultaneously and validate using stewardship based on the logic and the patterns discovered.	GEN AI-powered Query generator with NLP as Input. This contributes to the efforts of functional SMEs to query data for rules and visualizations.

Figure 13: IDW GenAI Modules



Value Proposition – Infosys Data Workbench

Infosys Data Workbench is a Data Quality platform, coupled with light weight Analytical MDM module and AI-ML based Data quality features. IDW leverages powerful Apache spark core to perform data management activities at scale

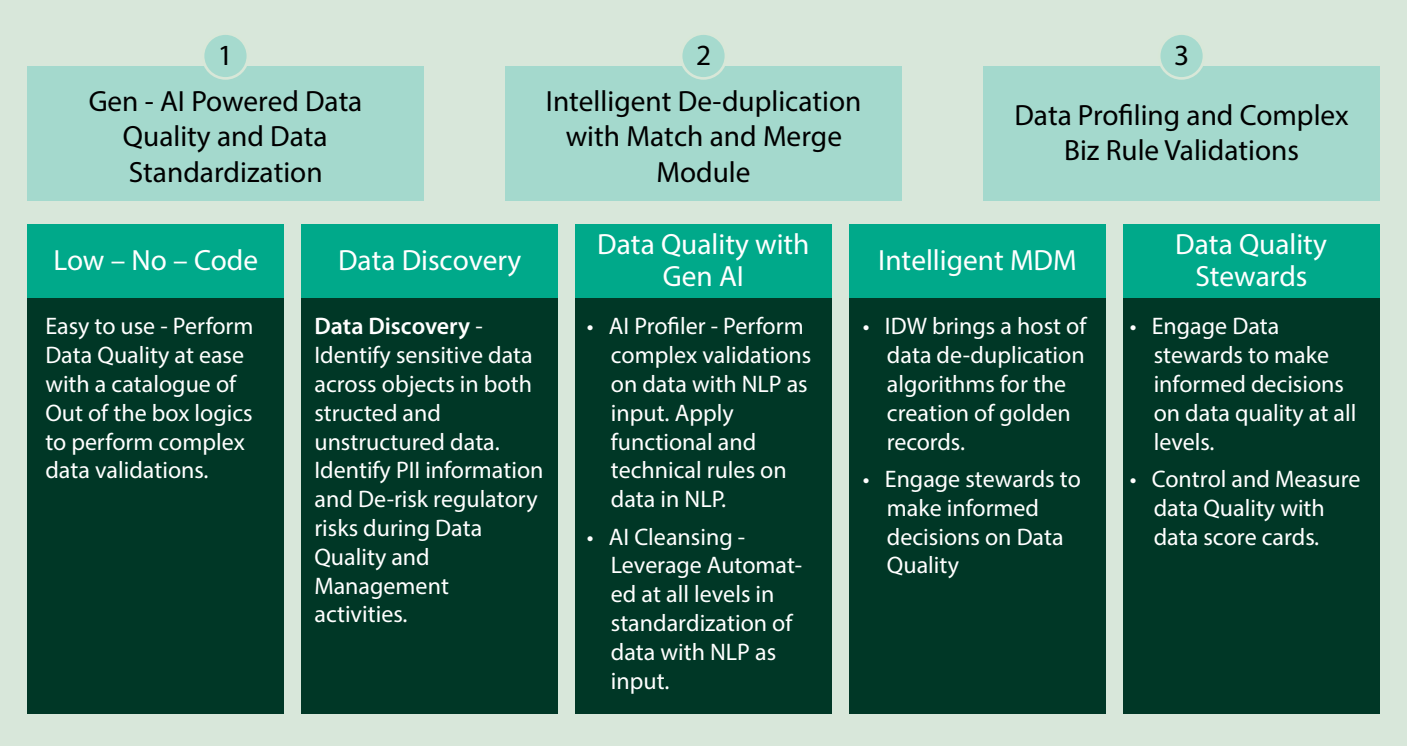


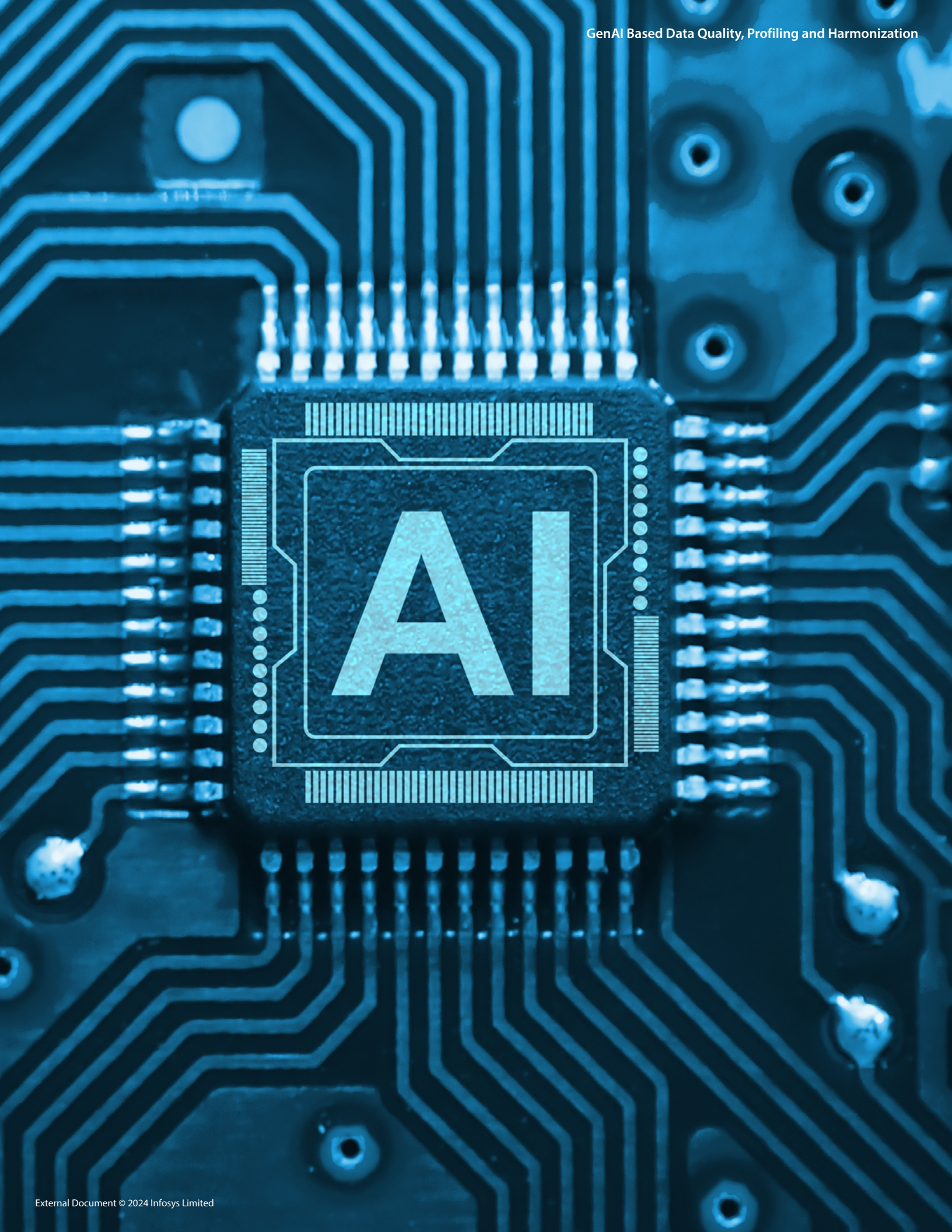
Figure 14: IDW Value Proposition

IDW Positioning

Unique Features of IDW Tool compared to other Data Quality tools

1. Complex rule builder to perform logical and transformational validations.
2. Built-in accelerators (SAP Auto validation rules) to generate automated data quality rules during migration.
3. Analytical MDM capabilities powered by intelligent algorithms to achieve deduplication and for the creation of golden records.
4. Data comparators for data gap assessments.
5. ML based features for Data Quality, Outliers and Rule Mining.

The primary focus of IDW has been to create a single platform that can help enable rules-based alleviation of all data issues for data profiling and cleansing process and help derive maximum value of data across the organization.



References

- https://www.expresscomputer.in/news/78-of-indian-business-leaders-believe-data-reduces-uncertainty-and-drives-better-decisions-salesforce-research/97266/
- https://economictimes.indiatimes.com/tech/technology/80-indian-business-leaders-say-data-crucial-in-decision-making-report/articleshow/99845304.cms?from=mdr

About the Authors

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.

Tushar Subhra Das is a Senior Business Data Analyst with over 15 years of experience in Data and Governance. He has worked with Europe and Australia-based insurance and logistics clients for Data management, MDM and Data Quality, and process governance. In his current role as product manager, he is responsible for commercialization of data management tools, deployments and enhancements, including product development for Generative AI supported data management platforms.

Sridhar Sivakoti is Senior Technology Manager with over 20 years of experience in Data Management, Data quality and Business Intelligence areas. He involved in architecture and design to develop applications and solutions to handle enterprise-wide BI projects. He managed large scale of Data projects with various ETL and Reporting tools. Handled end to end life cycle of Data like Migration, Modernization, Upgradation and moving the data from on premise to various cloud.

For more information, contact askus@infosys.com

Infosys®
Navigate your next