



FROM BLACK BOX TO GLASS BOX: A CONTEXT-AWARE APPROACH TO SDTM MAPPING AUTOMATION

Enhancing transparency and efficiency
in clinical data mapping

By:

Rostislav Markov, Amazon Web Services, Inc., New York, USA

Jacques Lanoue, Merck & Co., Inc., North Wales, USA

Ulf Nielsen, MSD Merck Sharp & Dohme AG, Zurich, Switzerland

GLASS-BOX APPROACH TO SDTM DATA TRANSFORMATION

Paper Overview

The paper presents an approach for automating SDTM mapping using interactive visual validation to enhance transparency and compliance.

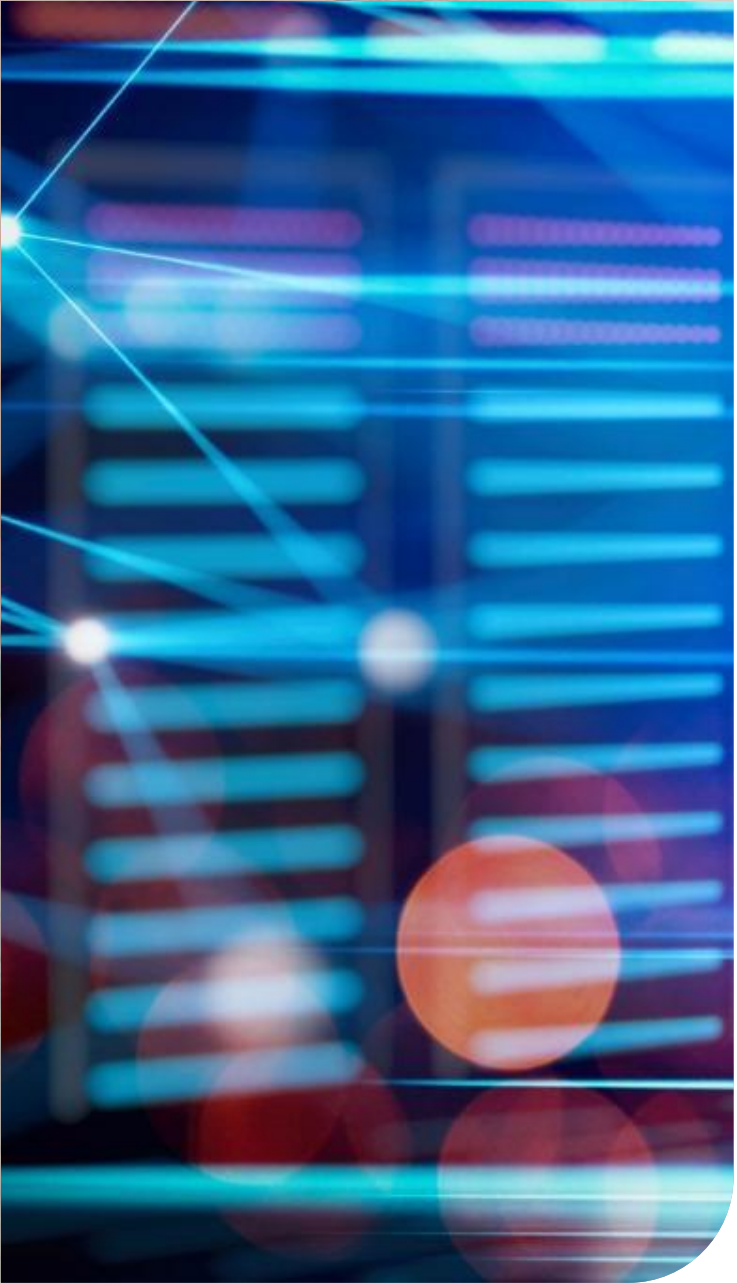
Author Affiliations

Authors come from leading organizations in cloud services and pharmaceutical industries across the USA and Switzerland.

Target Audience

The presentation targets peers interested in clinical data transformation, regulatory compliance, and AI integration in workflows.





ABSTRACT OVERVIEW

Specification-First Automation

AI generates machine-readable mapping specifications, avoiding opaque data transformations for transparency and control.

Layered Architecture Design

Data modeling, transformation logic, and specification generation are separated into distinct layers for modularity.

Reuse-First & AI Augmentation

The system leverages prior mapping specifications and applies AI to handle residual mapping needs efficiently.

Observability and Validation

An integrated dashboard provides UI-ready views of pipeline runs, metrics, and log correlation with strong validation controls.

INTRODUCTION TO SDTM MAPPING CHALLENGES



Challenges in SDTM Transformation

SDTM mapping is labor-intensive and error-prone due to protocol nuances and metadata variations in clinical studies.

Limitations of Traditional Approaches

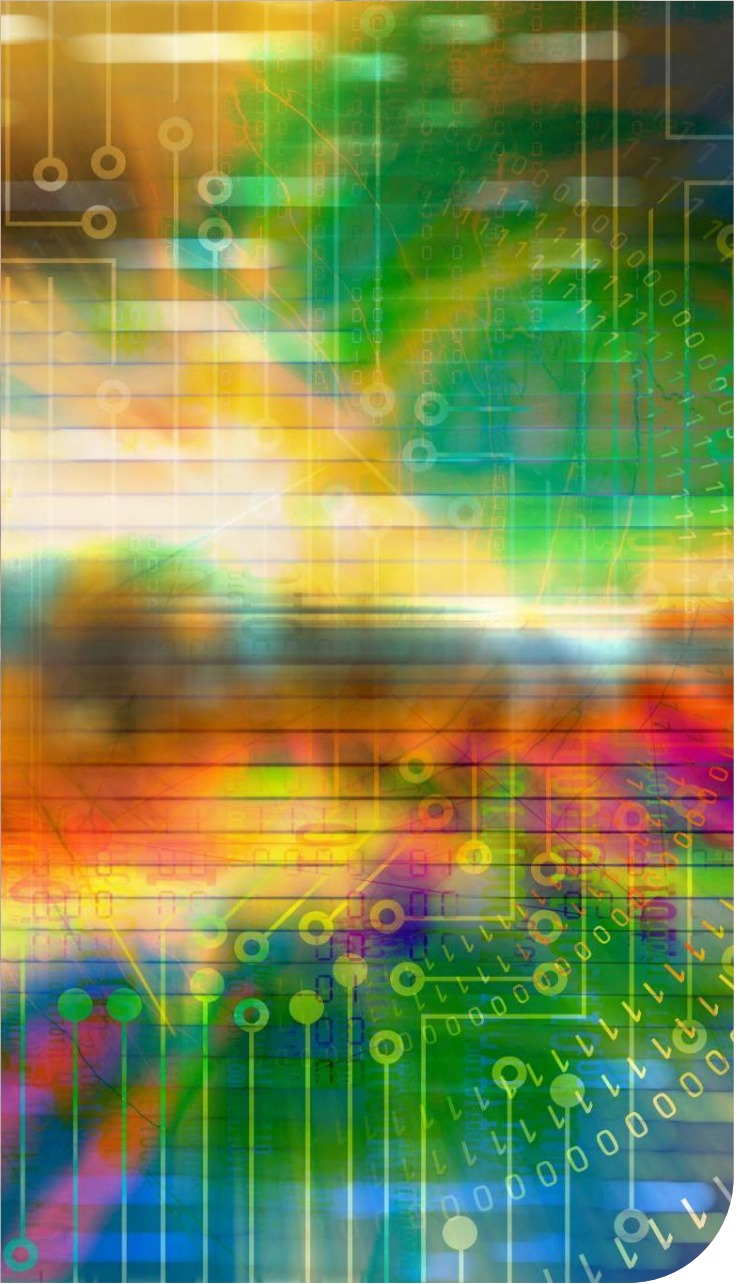
Rigid libraries and centralized governance limit flexibility and adaptability in clinical data transformations.

Risks of AI 'Black Box' Engines

AI-driven black box engines pose risks like hidden logic, limited traceability, and regulatory uncertainty.

Proposed Transparent Automation

Automating auditable specifications with deterministic engines ensures transparency, oversight, and compliance.



SPECIFICATION-FIRST AND REUSE BEFORE PREDICTION

Specification-First Principle

Transformation intent is declared before execution to ensure clear and controlled workflows.

Reuse Before Prediction

Prioritize reuse of mapping specifications and functions to improve efficiency across projects.

Reproducibility and Governance

Use version control, test-driven documentation, and validation to ensure reproducible and governed code.

Transparency and Human Oversight

Every mapping is traceable and AI suggestions remain visible and editable for regulation compliance.

SYSTEM ARCHITECTURE

Modular System Architecture

The architecture separates data modeling, mapping, governance, and execution to ensure flexibility and scalability.

Function Registry and Deployment

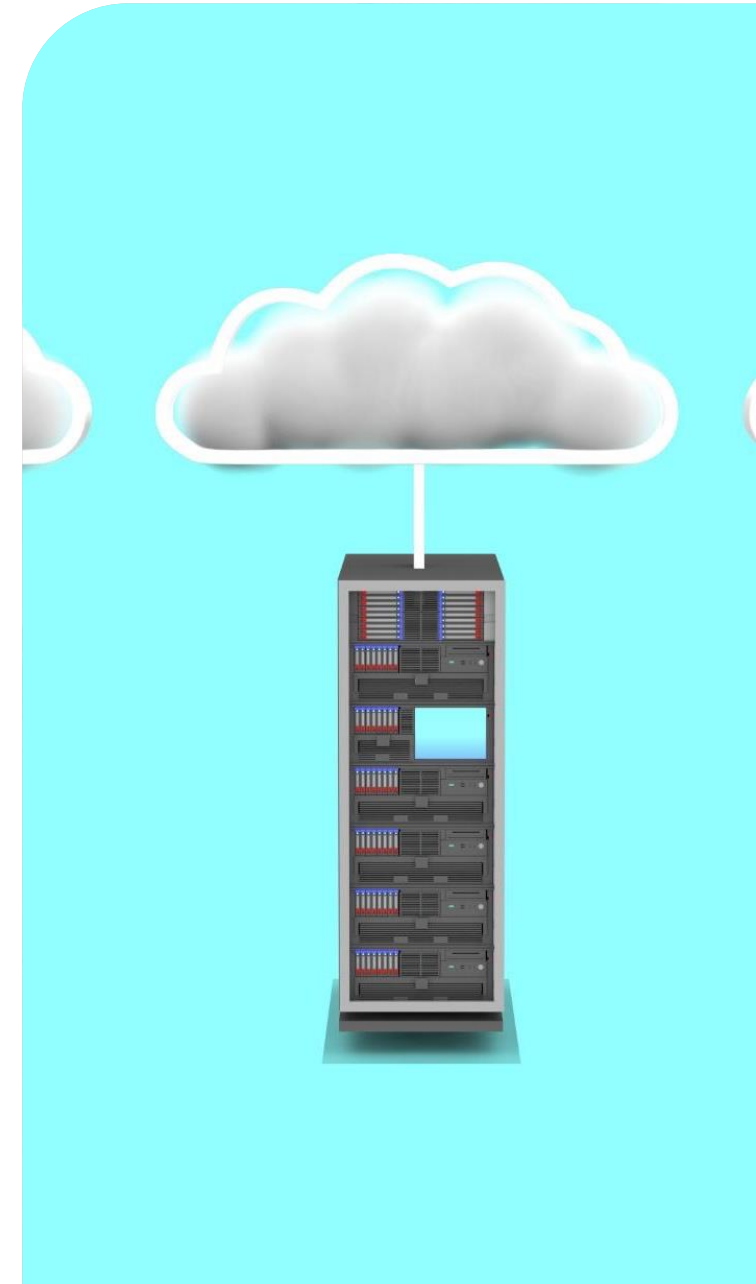
Functions are stored in a GitHub registry, implemented in Python, deployed via cloud resources, and released with JFrog.

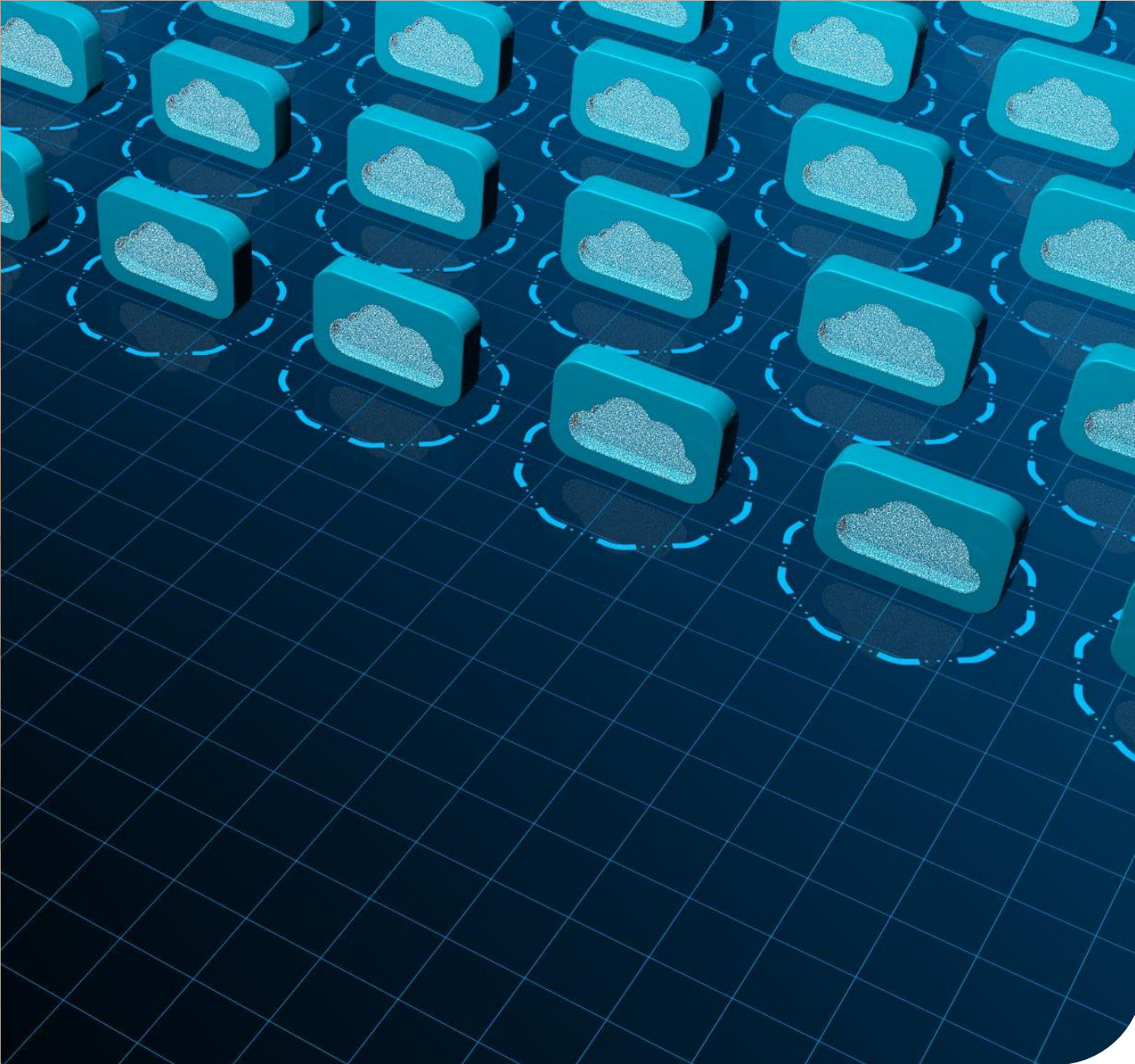
Metadata Indexing and Discovery

Metadata is indexed in OpenSearch search cluster, enabling function discovery in the web GUI and orchestrator.

Authoring Tool Features

The authoring tool supports dataset inspection, mapping design, AI predictions, version control, and production promotion.





SEPARATION OF DUTIES

Transformation Engine

The engine uses cloud functions to parse specs and create state machines for dataset transformations.

Web GUI Functionality

The GUI supports SQL modeling, query definition, dataset preview, and column validation for clear workflows.

Function Registry and Governance

A searchable registry enables users to find and execute authorized functions maintaining study isolation and governance.

MAPPING FEATURES AND OBSERVABILITY DASHBOARD

Variable-Level Mapping Features

Mapping includes grids for variables, completeness tracking, function selection, and tools for reuse and import.

Execution Monitoring and Trace Inspection

Dashboard enables monitoring of executions, error drilling, trace inspection, and comparison of specification versions.

Visual Pipeline with Metrics

Pipeline view shows processing steps color-coded by status with metrics for detailed troubleshooting.

Reduced Troubleshooting Effort

Detailed views and metrics decrease troubleshooting time, especially for partial failures or unexpected results.



Operations: DM_DSRND

Save

+ Add Operation

Select Table, and Columns ^ ▼

Remove

Table Name

DM ▼

Columns

Select columns ▼

Distinct



DM_AGE ✕

DM_AGE_TRANSLATED ✕

DM_AGE_UOM ✕

DM_AGE_UOM_TRANSLATED ✕

DM_BRTHYY ✕

DM_CRACE_ASIA ✕

DM_CRACE_ASIA_DECODE ✕

DM_CRACE_BLK ✕

DM_CRACE_BLK_DECODE ✕

DM_ETHNIC ✕

DM_ETHNIC_DECODE ✕

DM_O_SAFETHNIC_DRV ✕

DM_O_SAFETHNIC_DRV_DECODE ✕

DM_O_SAFSEX_DRV ✕

DM_O_SAFSEX_DRV_DECODE ✕

DM_O_SAFSITE_DRV ✕

DM_RACE01 ✕

DM_RACE02 ✕

DM_RACE03 ✕

DM_RACE04 ✕

DM_RACE05 ✕

DM_RACE06 ✕

DM_RACE07 ✕

DM_RACELBL ✕

Event.Date ✕

Event.Name ✕

Event.Status ✕

Event.VisitMethod ✕

EventGroup.Name ✕

EventGroup.SeqNbr ✕

Form.DataSource ✕

Form.LastModifiedDate ✕

Form.Name ✕

Form.SeqNbr ✕

ItemGroup.SeqNbr ✕

LBSEX ✕

LBSEX_DECODE ✕

Site.Number ✕

Study.Name ✕

Subject.Name ✕

Select Table, and Columns ^ ▼

Remove

Table Name

DSRND ▼

Columns

Select columns ▼

Distinct



DSRND_DMRNDNUM ✕

Join ^ ▼

Remove

Join Type

LEFT ▼

Table 1 Name

DM ▼

Table 1 Column Name

Subject.Name ▼

Table 2 Name

DSRND ▼

Table 2 Column Name

Subject.Name ▼

Operation

Select operation ▼

Mapping & Tracking

Functions

Dataset Operations

Deliverable

Package Configuration

Configuration

Logs

Import



▼

Pre-Processing Functions

Find function

	Functions	Version	Target Domains	Target Fields	Source Domains	Source Fields
	call_clean_invisible_characters	1	*	**	*	**
	call_source_with_header	1	*	**	*	**

▼

Transformation Functions

Find function

	Functions	Version	Target Domains	Target Fields	Source Domains	Source Fields
	get_usubjid	1	*	*.USUBJID, *.RSUBJID	*	*.Subject.Name, *.Study.Name
	get_tf_adv_zacomp			**	*	**
	direct_move				*	**
	suppqual_lookup			*.SUPPQUAL*	*	**

suppqual_lookup Description

×

The SUPPQLOOKUP function is designed to move the value of a variable in the source data to the Target variable QVAL in the SUPPQUAL (supplementary qualifiers) set in the target dataset and record the traceability information.



Specification Details [Info](#)

Mapping & Tracking

Functions

Dataset Operations

Deliverable

Package Configuration

Configuration

Logs

Import

☐ Prefilled Mapping

View Mapping

Sync Metadata

Save

Tools



☐	Source	SDTM Domains	Variables Mapped	Last modified by	TAUGs Used	Spec Author	Spec Reviewer	Review Status	QC End Date	Comments
☐	AE	AE, RS, SUPPAE, TR, TU	31/111							
☐	ALL_QUERIES	AE	14/40							
☐	ALL_QUERY_MESSAGES	AG	10/29							
☐	ARDENA_PK	CO, PC	26/42							
☐	All_Queries		0/28							NOT SUBMITTED
☐	All_Query_Messages		0/22							NOT SUBMITTED
☐	CM	AE, CM, SUPPCM	34/81							

Save

Find mappings

Dynamic Automapping

Actions	Source	Interim Source...	Source Variabl...	SDTM Domains...	SDTM Variable...	Functions	Function Para...	Group	Mapping Notes	Review Comm...
+  	AE		Subject.Name			Operational		0		
+  	AE		Subject.Name	AE	USUBJID			1		
+  	AE		Subject.Name	AE	USUBJID			1		
+  	AE		Subject.Name	HO	USUBJID			3		
+  	AE		Subject.Name	HO	USUBJID			3		
+  	AE		Subject.Name	AE	USUBJID	get_usubjid		1		
+  	AE		Subject.Name	TU	USUBJID	get_usubjid		1		
+  	AE		Subject.Name			Operational		0		
+  	AE		Subject.Name	FAAE	USUBJID			2		
+  	AE		Subject.Name	TR	USUBJID	get_usubjid		1		

Specification Details Info

[Mapping & Tracking](#) | [Functions](#) | [Dataset Operations](#) | [Deliverable](#) | [Package Configuration](#) | [Configuration](#) | [Logs](#) | [Import](#)[Log History](#) | [Log Details](#)

< 1 2 3 4 5 6 7 ... 11 >

Refresh



Job ID	User	Status	Issues	Source Forms	SDTM Domains (MUS)	SDTM Domains (Created)	Records Processed	Step Function ID	Execution ID	Trace ID
1-69848937-85eb197e0c6cfe34da58f39c		Completed	0 issues	44	32	60	168,249	1770293360	-1770293360:d8a51dcf-6a28-4dfa-a36d-8e77a51abbd	1-69848937-85eb197e0c6
1-698c9e50-3993c2e42f4bf09f63851719		Completed	0 issues	44	33	60	139,208	1770807200	-1770807200:925258d9-dbf8-4b7f-b975-7b8277bce033	1-698c9e50-3993c2e42f4t
1-6985dd09-7a88213625c71287710192e1		Completed	0 issues	44	33	30	78,247	1770380657	-1770380657:581cb82e-a368-4759-bc12-fde7b935b354	1-6985dd09-7a88213625c

Specification Details Info

[Mapping & Tracking](#) | [Functions](#) | [Dataset Operations](#) | [Deliverable](#) | [Package Configuration](#) | [Configuration](#) | [Logs](#) | [Import](#)[Log History](#) | [Log Details](#)

Job ID: 1-698e497e-3a0d

< 1 >

Refresh



States	SDTM Domains (MUS)	Status	Issues	Records Processed	Step Function ID	Execution ID	Trace ID
spec creation		Completed	0 issues	0			1-698e497e-3a0d
orchestration		Completed	0 issues	0	-orchestrator		1-698e497e-3a0d
▼ execution-1		Completed	29 issues	99,019	1770932730		1-698e497e-3a0d
▶ SDTM Domains(Created)		Completed	0 issues	88,321			
trigger engine		Completed	0 issues	0	engine-run		1-698e497e-3a0d
preprocessing		Completed	0 issues	0	preprocessing-		1-698e497e-3a0d
▶ sdtm minus		Completed	4 issues	10,698	abe7cb_sdtm_minus		1-698e497e-3a0d
derivations		Completed	25 issues	0	sdtm-derivations-		1-698e497e-3a0d

SPECIFICATION STRUCTURE AND AUTOMAPPING TOOLS

SECTION	DESCRIPTION
header	Study context, SDTMIG version, audit metadata
sdtmMinus	Source-to-target variable mappings
studyDatasetOperations	Declarative dataset operations (joins, filters, selections)
sdtmDerivations	Study-level derivation functions
preProcessingFunctions	Functions applied before mapping/derivation
transformationFunctions	List of all referenced transformation functions

```

1  {
2  "header": {
3    "Study ID": "STUDY-001",
4    "Study Name": "STUDY-001_internal",
5    "SDTM Version": "SDTMIG v3.4",
6    "Dictionary Versions": {
7      "Meddra": "",
8      "WHODD": "",
9      "LOINC": "",
10     "CT": ""
11   },
12   "Specification Name": "STUDY-001_internal_sdtmig3.4_v01"
13 },
14
15 "sdmMinus": [
16 {
17   "SourceTable": "AE",
18   "SourceVariable": "AEACN_RAW",
19   "InterimDataset": "",
20   "TargetList": [
21     {
22       "SDTMDomain": "AE",
23       "SDTMVariable": "AEACN",
24       "SDTMTransformationFunction": "uuid-direct-move",
25       "SDTMTransformationFunctionName": "direct_move (v1)",
26       "FunctionParameterValue": "",
27       "Group": "1",
28       "SDTMCodeList": "",
29       "SDTMType": "text"
30     }
31   ]
32 },
33 { ...
49 }
50 ],
51

```

```

52 "studyDatasetOperations": {
53   "datasetOperations": [
54     {
55       "DM_EXAMPLE_JOIN": [
56         {
57           "type": "TableAndColumns",
58           "sequence": 1,
59           "tableName": "DM",
60           "tableColumns": "Study.Name,Subject.Name,DM_AGE,DM_BRTHYR",
61           "distinct": false,
62           "filterText": ""
63         },
64         {
65           "type": "TableAndColumns",
66           "sequence": 2,
67           "tableName": "RAND",
68           "tableColumns": "RANDNUM",
69           "distinct": true,
70           "filterText": ""
71         },
72         {
73           "type": "Join",
74           "sequence": 3,
75           "joinType": "LEFT",
76           "table1Name": "DM",
77           "table1ColumnName": "Subject.Name",
78           "table2Name": "RAND",
79           "table2ColumnName": "Subject.Name",
80           "operation": "",
81           "additionalJoinConditions": []
82         }
83       ]
84     }
85   ]
86 },
87

```

```

88 "sdmDerivations": [
89 {
90   "Function": "uuid-get-informed-consent",
91   "Version": 1,
92   "Name": "std_fn_get_informed_consent",
93   "Sequence": 0
94 },
95 {
96   "Function": "uuid-get-first-dose",
97   "Version": 1,
98   "Name": "std_fn_get_first_dose",
99   "Sequence": 1
100 }
101 ],
102
103 "preProcessingFunctions": [
104 {
105   "Function": "uuid-clean-invisible-chars",
106   "Version": 1,
107   "Name": "std_fn_clean_invisible_chars"
108 }
109 ],
110
111 "transformationFunctions": [
112 {
113   "Function": "uuid-direct-move",
114   "Version": 1,
115   "Name": "std_fn_direct_move"
116 },
117 {
118   "Function": "uuid-iso8601",
119   "Version": 1,
120   "Name": "std_fn_iso8601"
121 }
122 ]
123 }
124

```

Shop For Mapping [Info](#)

< 1 2 >

✓ Review Selection(s)



☐	Source ▾	Study ▾	Spec ▾	Percent Match ▾	Target(s) ▾	Promoted Date ▾	Last Processing Time ▾
☐	AE		_sdtmig3.4_v14	100%	AE, AG, CM, FAAE, HQ, NOT SUBMITTED	2025-09-25 13:28:20:926182	2025-10-07 04:31:16
☐	AE		sdtmig3.4_v9	100%	AE, FAAE, HQ, NOT SUBMITTED	2025-10-16 15:34:23:141835	2025-10-30 10:56:23
☐	AE		sdtmig3.4_v8	100%	AE, FAAE, HQ, NOT SUBMITTED	2025-10-16 13:38:15:355738	2025-10-30 10:56:23
☐	AE		sdtmig3.4_v7	100%	AE, FAAE, HQ, NOT SUBMITTED	2025-10-15 22:17:21:092070	2025-10-17 11:25:27
☐	CM		sdtmig3.4_v9	100%	CM	2025-10-16 15:34:23:141835	2025-10-30 10:56:23
☐	CM		sdtmig3.4_v8	100%	CM	2025-10-16 13:38:15:355738	2025-10-30 10:56:23
☐	CM		sdtmig3.4_v7	100%	CM	2025-10-15 22:17:21:092070	2025-10-17 11:25:29
☐	CM		sdtmig3.4_v6	100%	CM	2025-10-13 13:59:37:280894	2025-10-17 11:25:29
☐	CM		sdtmig3.4_v2	100%	CM	2025-12-18 17:32:16:669340	2026-01-09 04:30:58



AI-ASSISTED MAPPING AND RAG PIPELINE

Contextual Prompting Pipeline

The pipeline uses study metadata and transformation functions to provide structured prompts for AI model responses.

Embedding-based Study Profiles

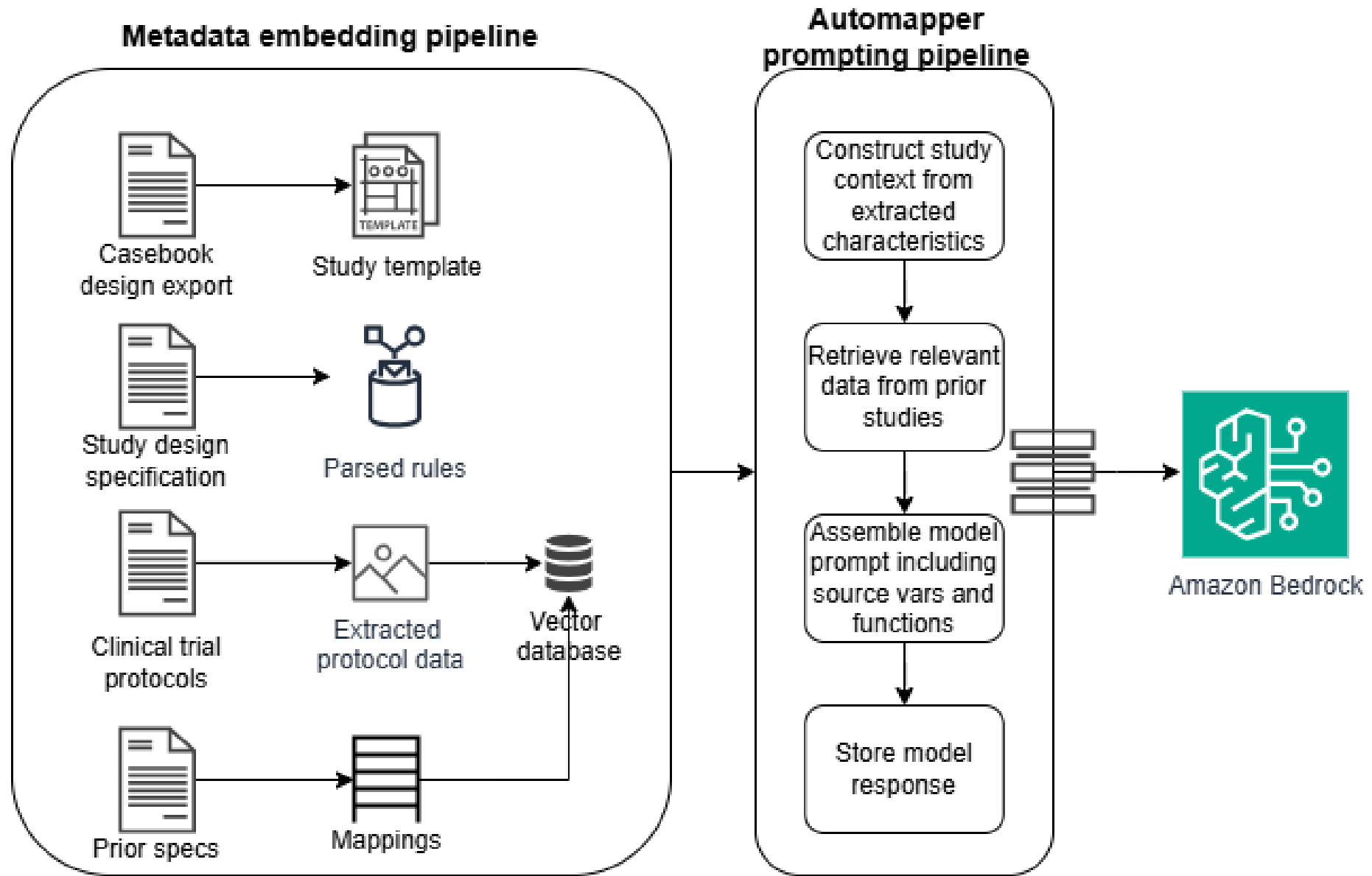
Embedding pipelines generate study profiles from EDC exports, grounding responses using RAG techniques.

Human Review & Compliance

AI-generated code is reviewed by programmers and tested in preview environments to ensure compliance and accuracy.

Balanced Automation Approach

The system balances AI automation with human oversight to reduce manual effort while maintaining transparency and compliance.



GOVERNANCE, COMPLIANCE, AND OBSERVED BENEFITS

Multi-layered Governance Enforcement

Governance is applied through function registry, schema validation, restricted operations, and execution transparency to ensure compliance.

Compliance Through Specifications

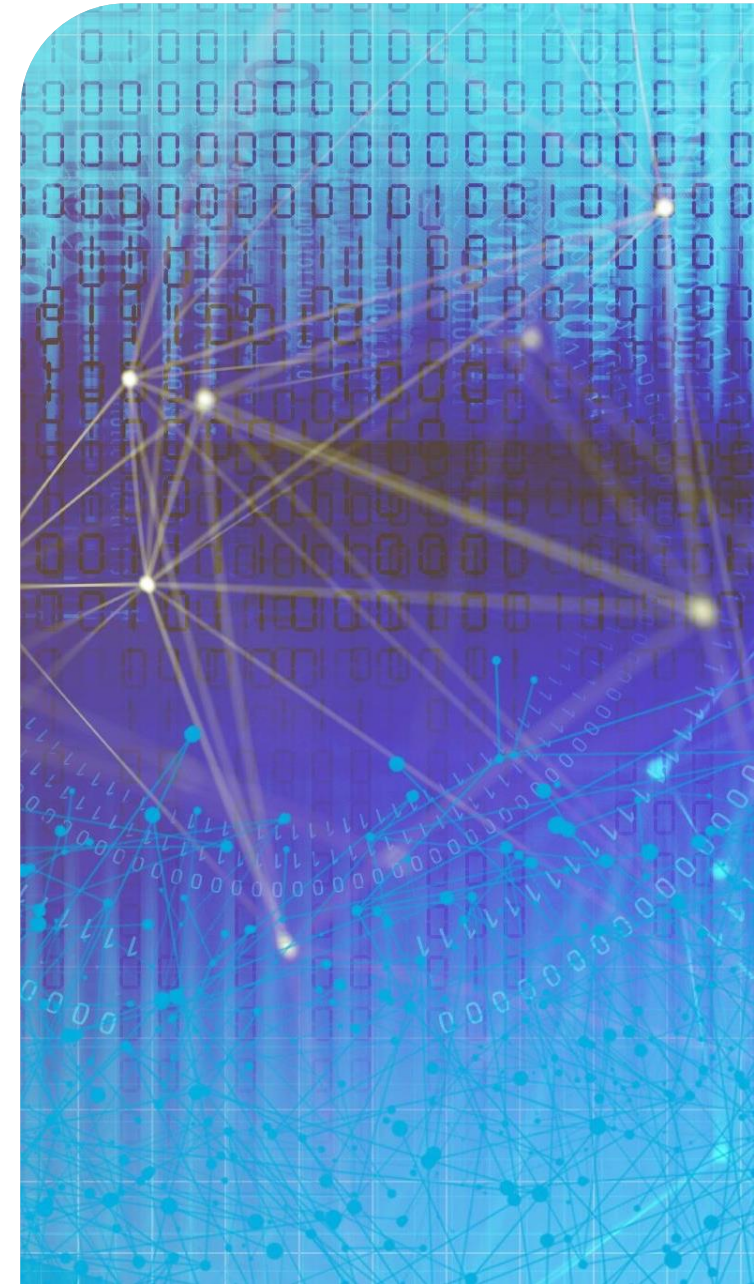
Specifications act as configuration artifacts serving as lineage anchors, supporting compliance without executable code.

Efficiency and Automation Metrics

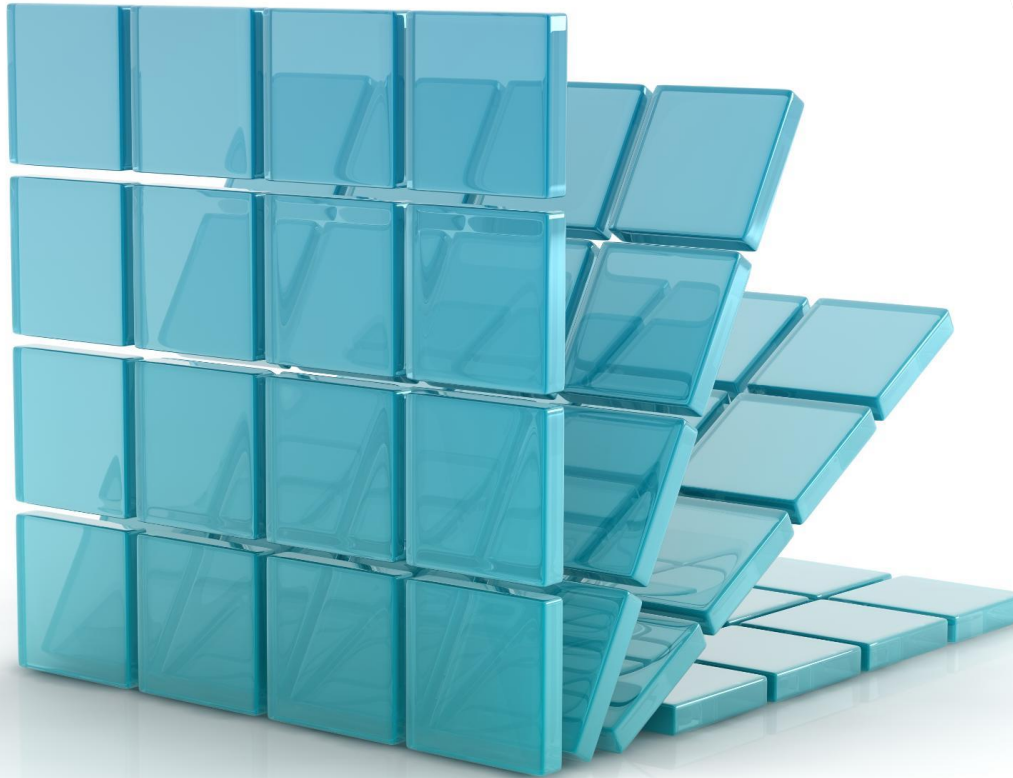
The platform achieves 60–80% reuse and nearly 90% automation, projecting 32,000 hours saved annually at scale.

AI-assisted Mapping Benefits

AI, enhanced by RAG-based grounding, is expected to reduce manual effort and accelerate data mapping processes.



SUMMARY AND FINAL THOUGHTS



Transparent Automation Architecture

The architecture ensures transparency and scalability while automating SDTM dataset operations efficiently.

Governed AI Augmentation

AI is used within clear boundaries, governed by code, version control, and test-backed execution.

Reproducible Clinical Data Delivery

The specification acts as an intelligence layer ensuring consistent and reproducible SDTM delivery across studies.