

Prepare for Enterprise Data Fabric with Metadata Driven Data Quality

Marek Ovcacek, Ataccama



Enterprise Data Fabric

Data
Catalog

Customer



John Doe

PII

Data
Quality

87%

MDM

David Smith

FAMILY

Jo Smith

Reference
Data

CODE

SYMBOL

BHR



Data
Stories



Data
Integration



Make high quality data instantly available to

Data engineers

Data stewards

Value creators

Machines & algorithms

At the speed, quality and form required

How Data Fabric Solves This

Data
Sources

Enterprise Data Fabric

Data
Consumers



Decision Framework



AI

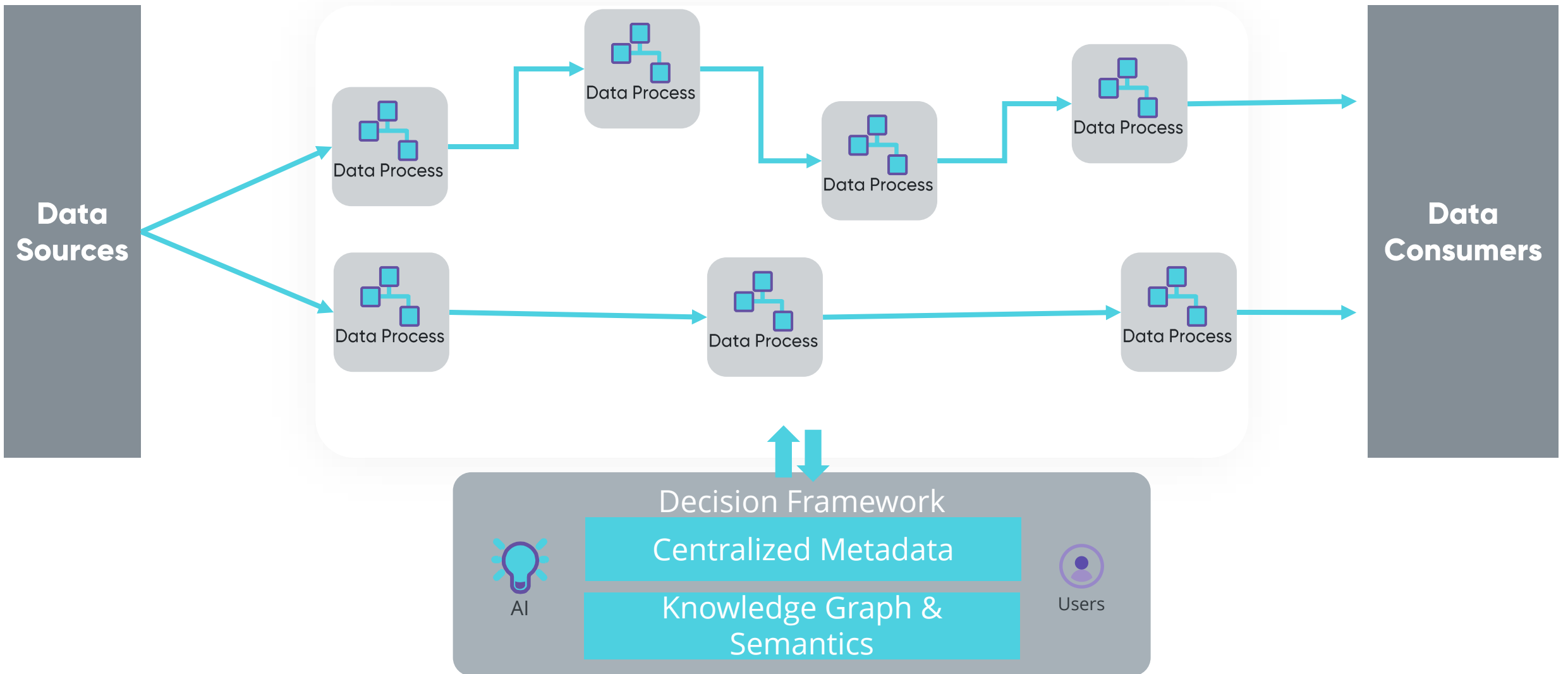
Centralized Metadata

Knowledge Graph &
Semantics



Users

How Data Fabric Solves This



Data Fabric Has to "Understand" State Of Data Quality

Of Entirety of your Data Landscape

How DQ is Done Right Now

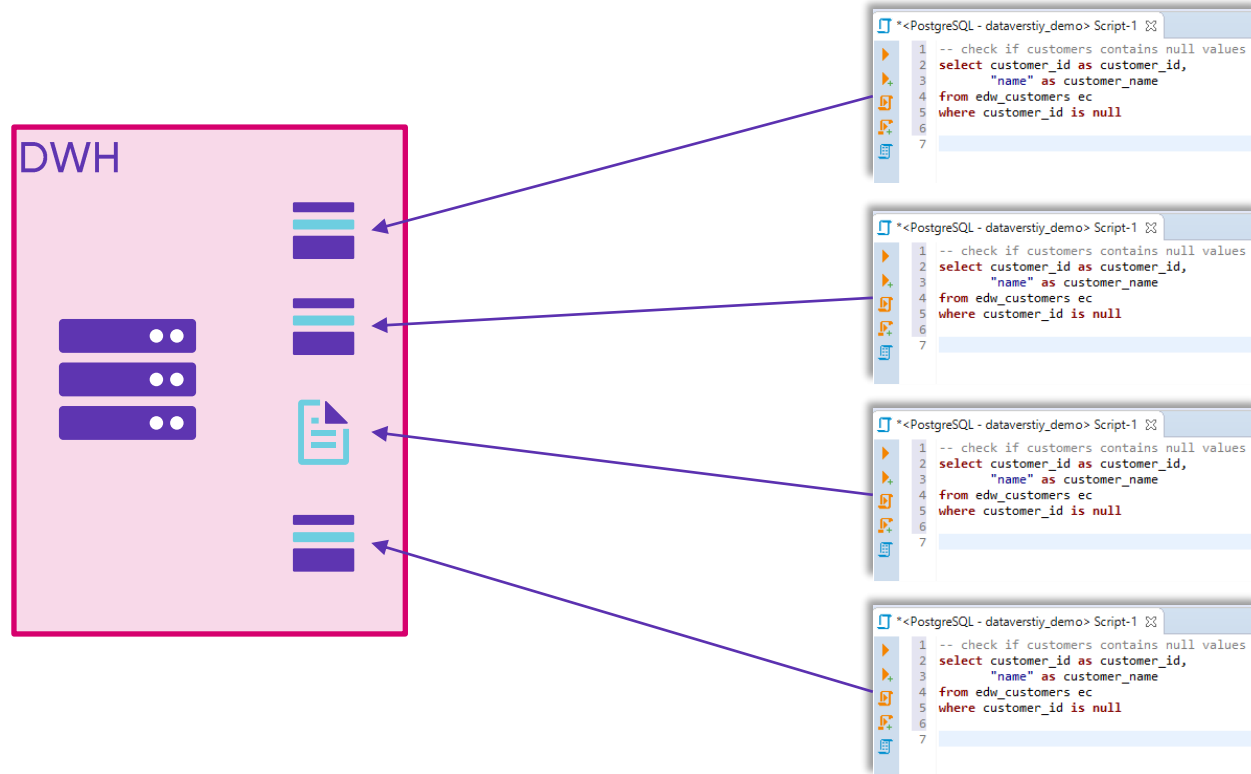
D	E	F	G	H	I
MANDATORY	DATA TYPE	REFERENCE/ CHECK TABLE	COMPLETENESS	VALIDITY	ACURACY
yes	CHAR	N/A	cannot be blank	1. Valid Format: alphanumeric; special character allowed;	Only values per Appendix are allowed
yes	INT	N/A	cannot be blank	1. Valid format: Numeric 2. Length: Must be 10 digit numeric (XXXXXXXXXX) Only	Only values per Appendix are allowed
yes	DEC	N/A	cannot be blank	1. Valid format: Numeric with 2 decimal	N/A

```
*<PostgreSQL - dataverstiy_demo> Script-1 ✕  
1 -- check if customers contains null values  
2 select customer_id as customer_id,  
3     "name" as customer_name  
4 from edw_customers ec
```

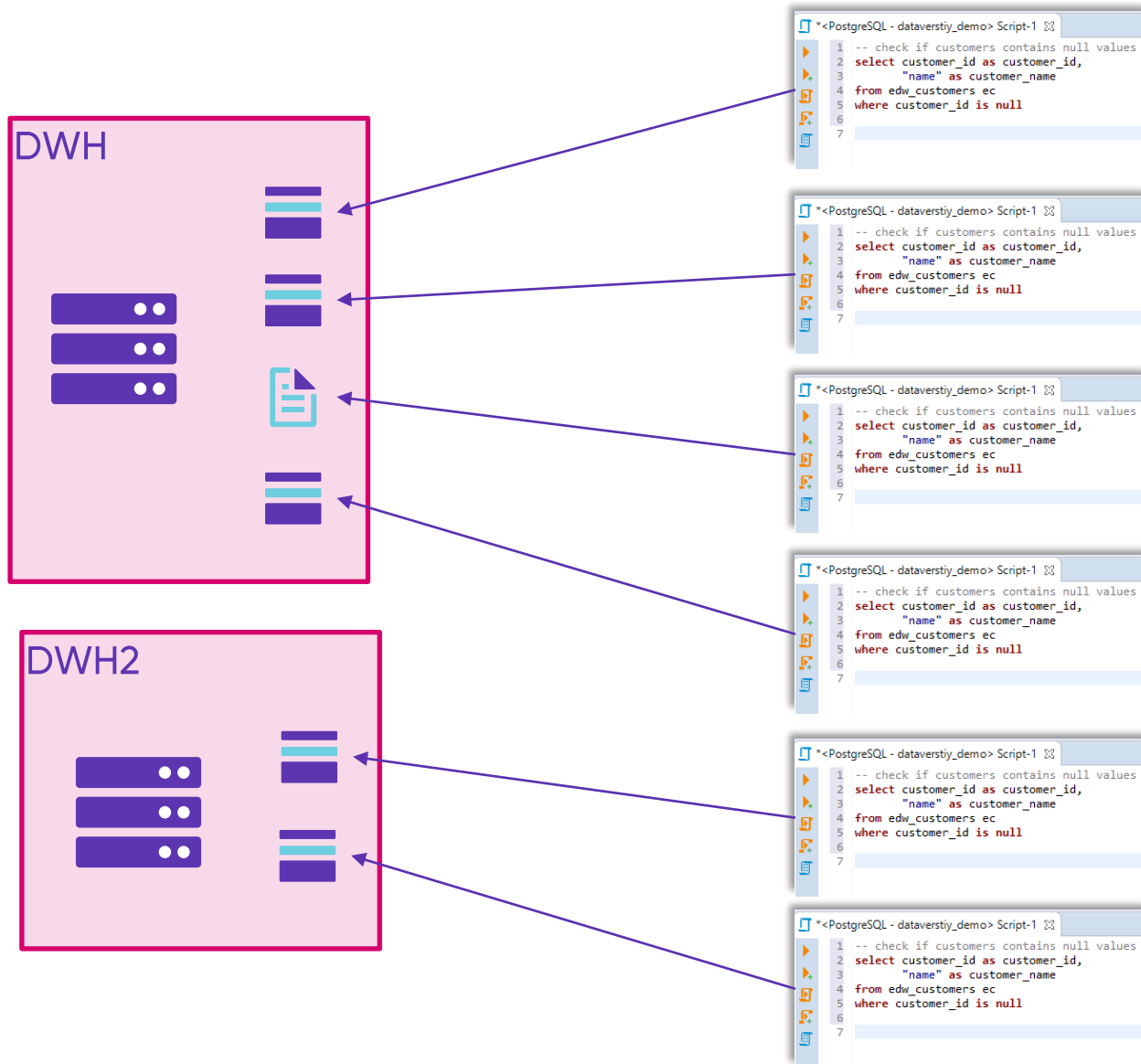
```
*<PostgreSQL - dataverstiy_demo> Script-1 ✕  
1 -- check if customers contains null values  
2 select customer_id as customer_id,
```

```
*<PostgreSQL - dataverstiy_demo> Script-1 ✕  
1 -- check if customers contains null values  
2 select customer_id as customer_id,  
3     "name" as customer_name  
4 from edw_customers ec  
5 where customer_id is null  
6  
7
```

How DQ is Done Right Now



How DQ is Done Right Now





Data Growing Exponentially not only in Volume but also In Complexity

Traditional approaches don't scale anymore



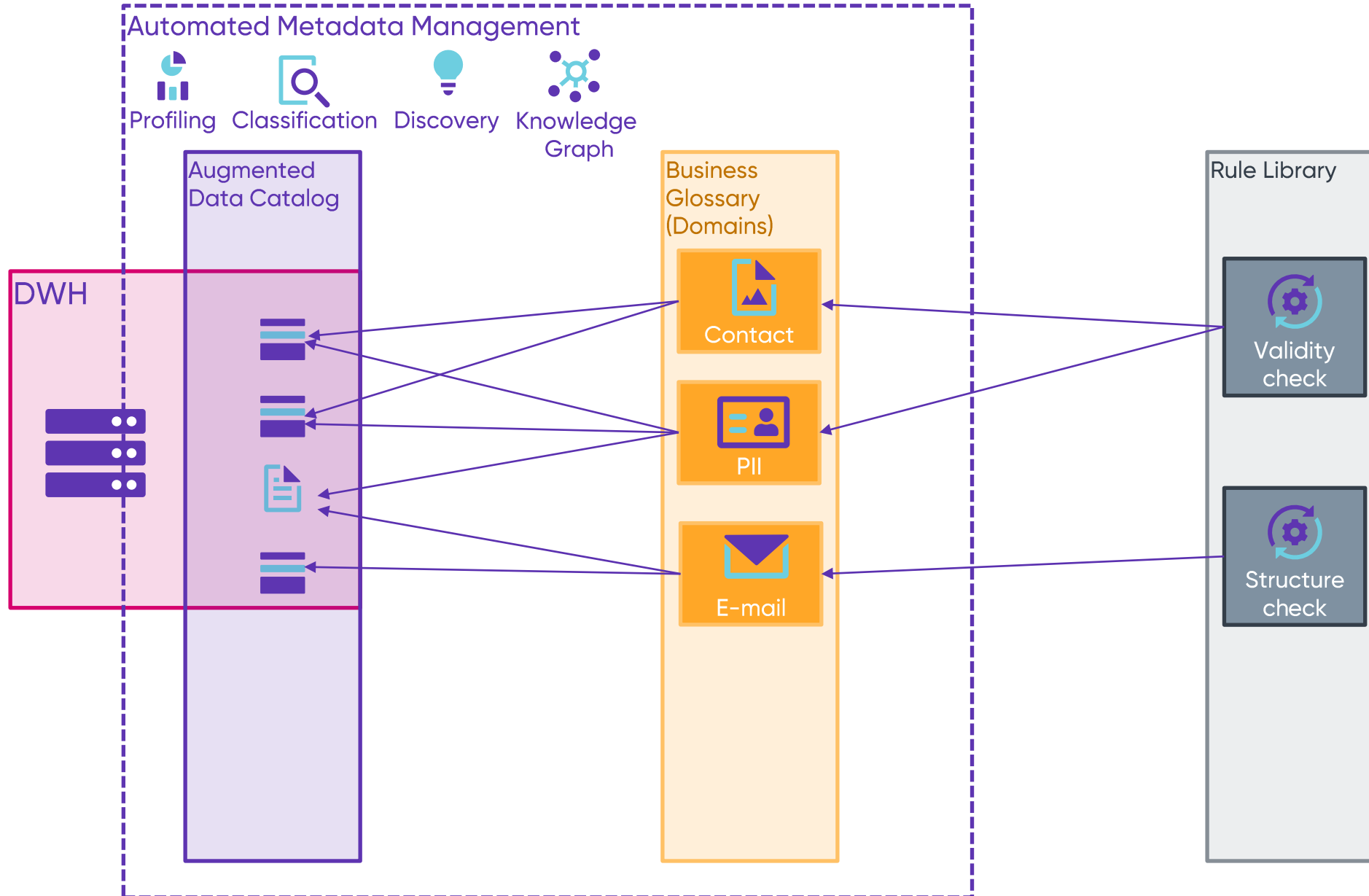




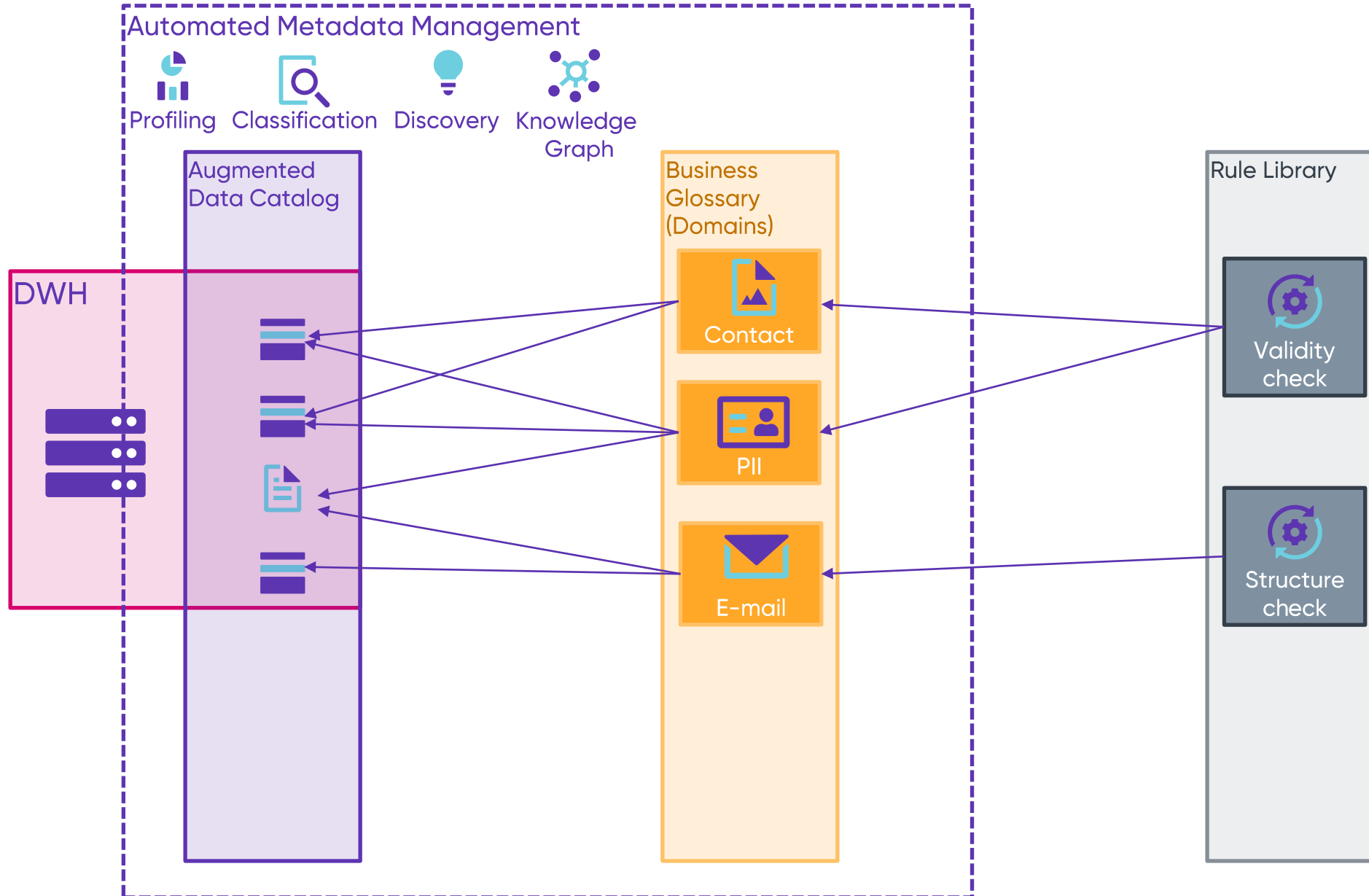
Automated Metadata Management + Metadata Driven Processing

To solve the problem of scaling

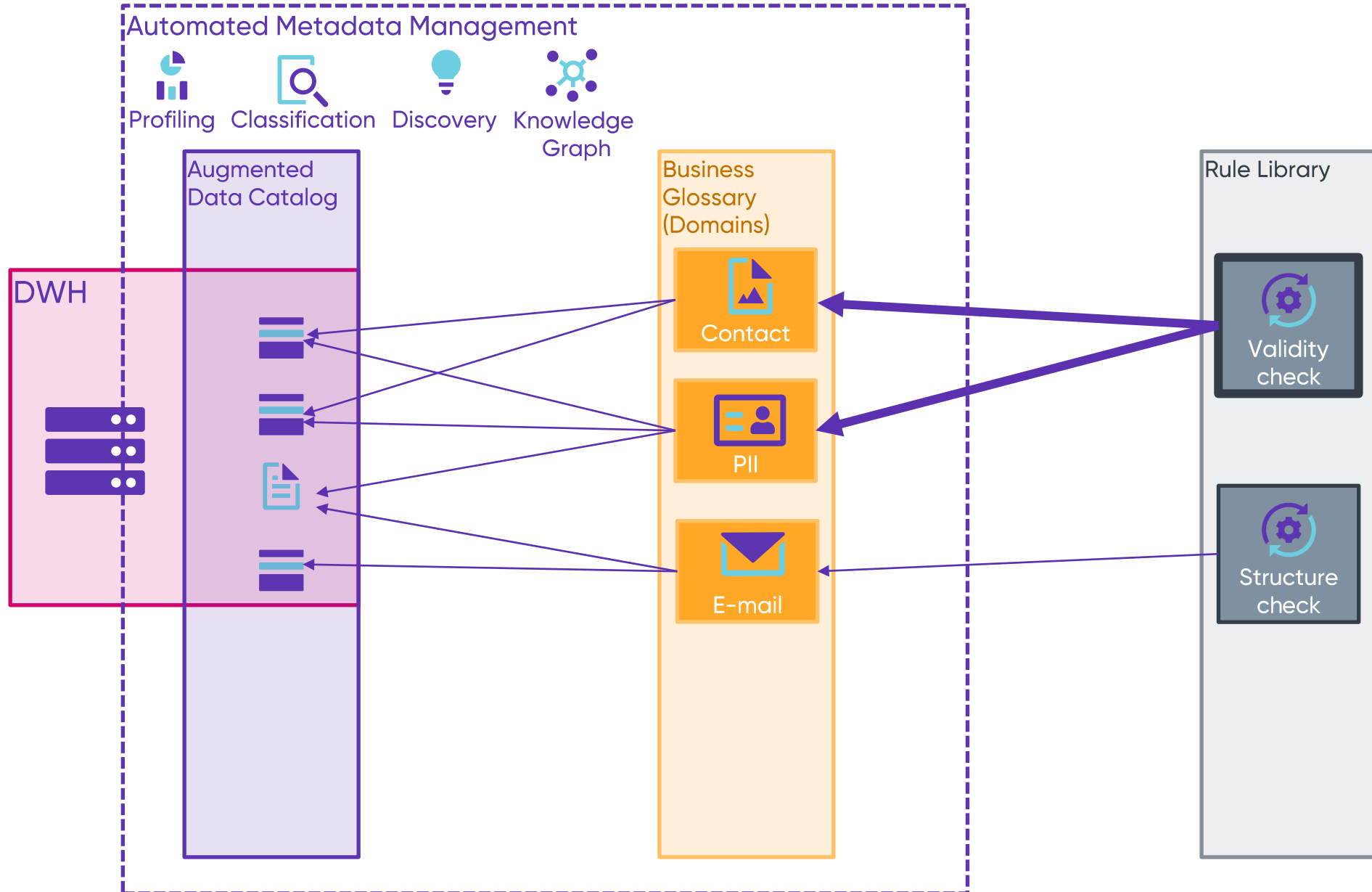
Metadata Driven DQ



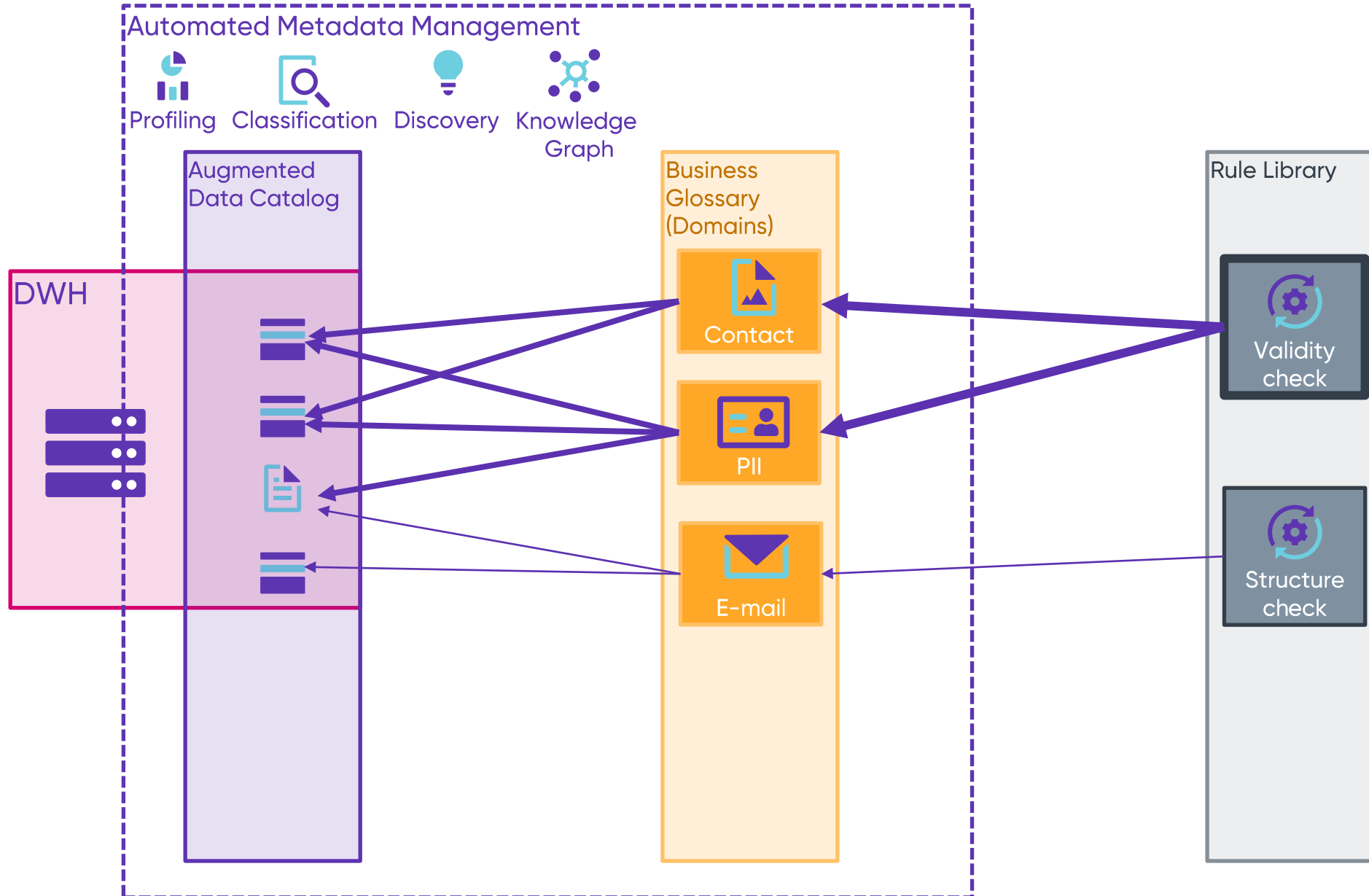
Metadata Driven DQ



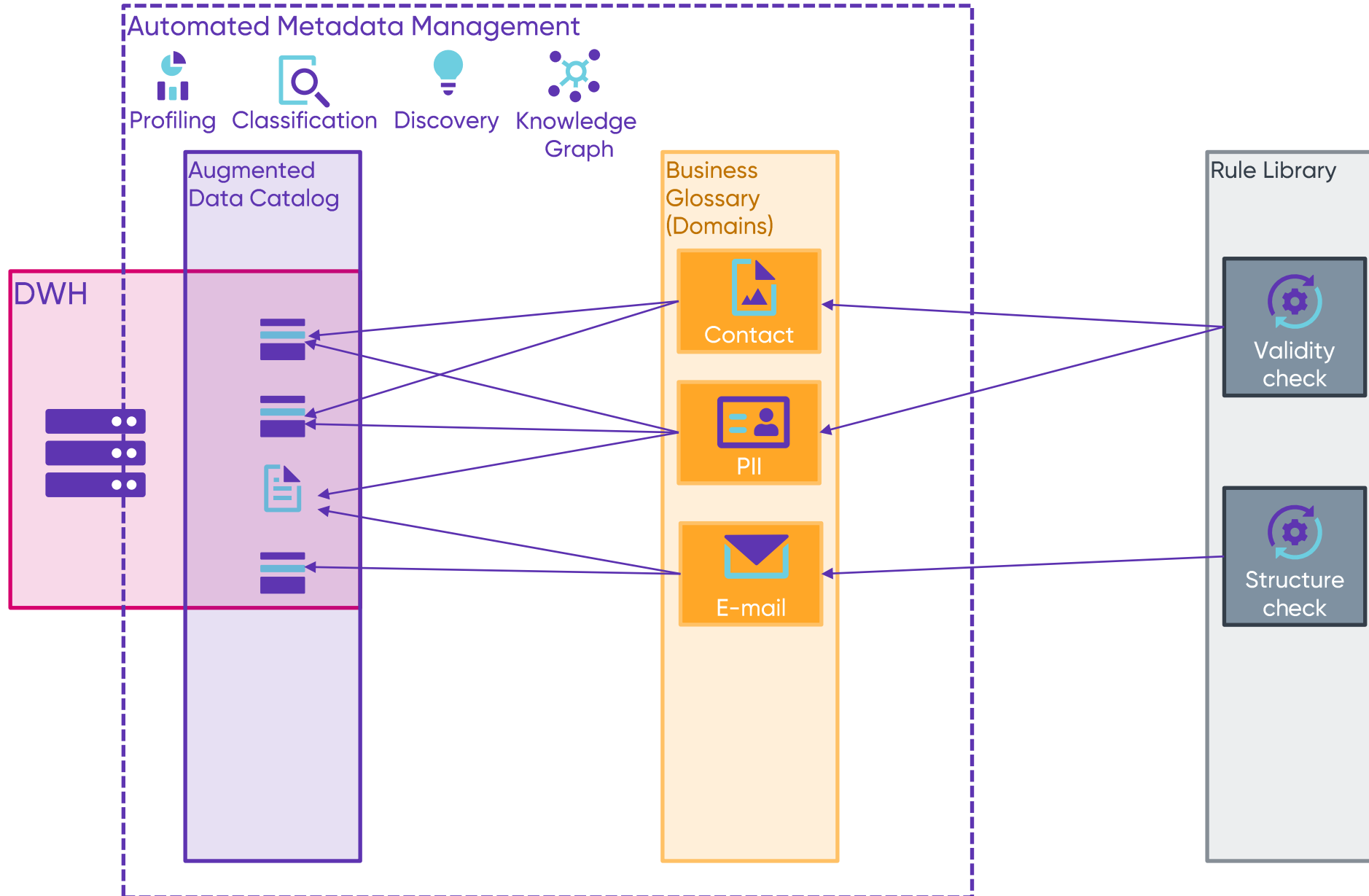
Metadata Driven DQ



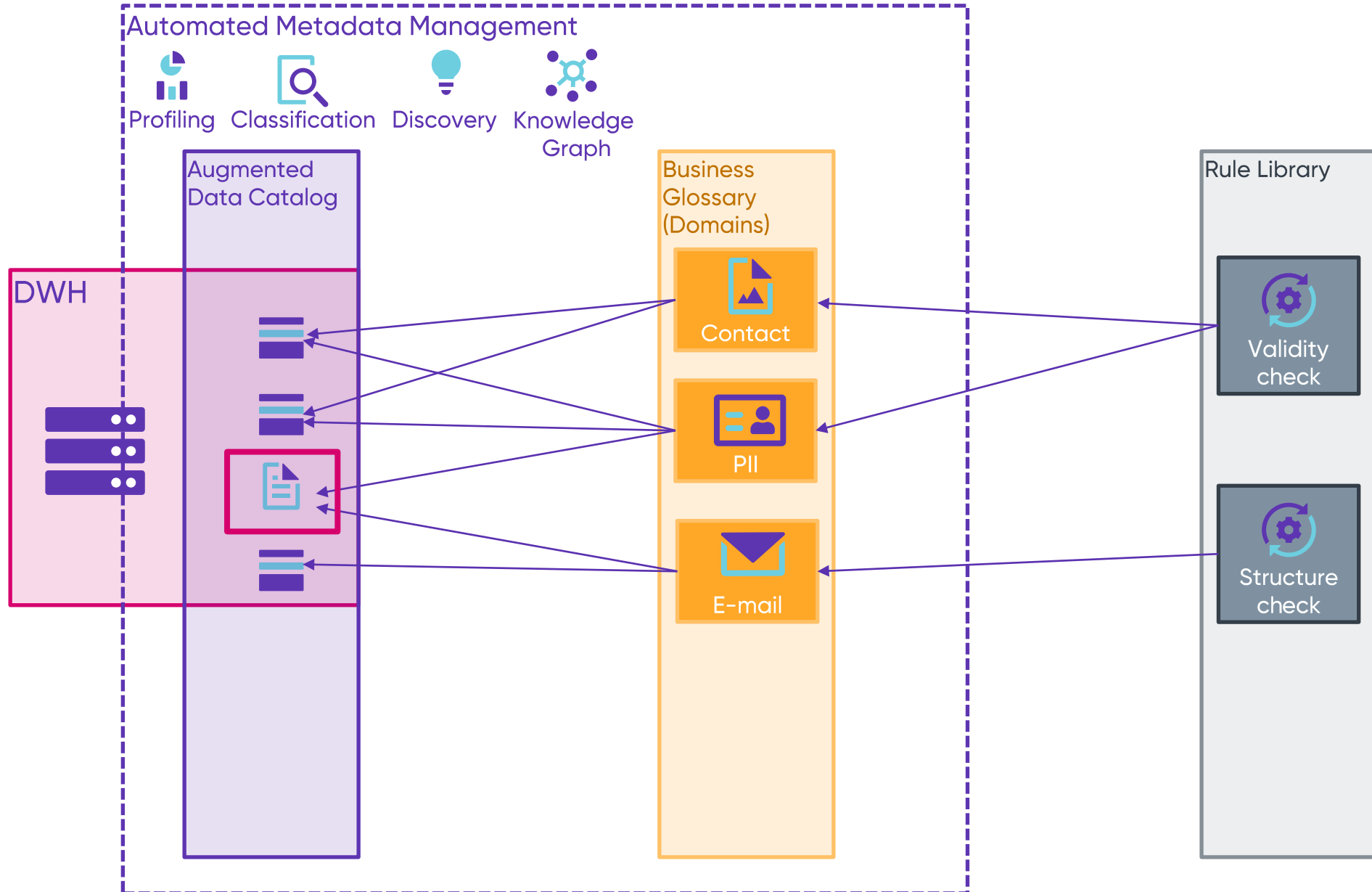
Metadata Driven DQ



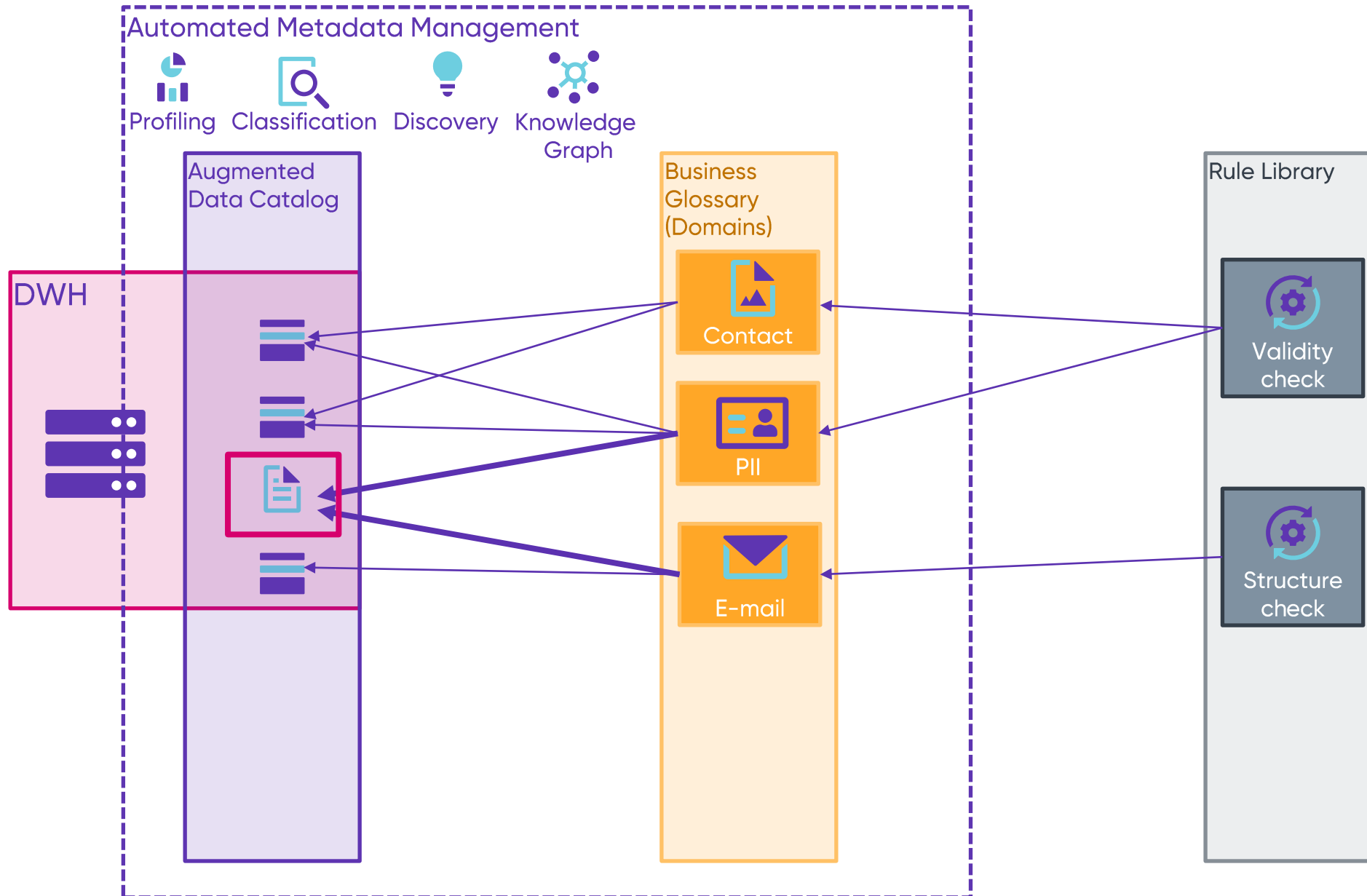
Metadata Driven DQ



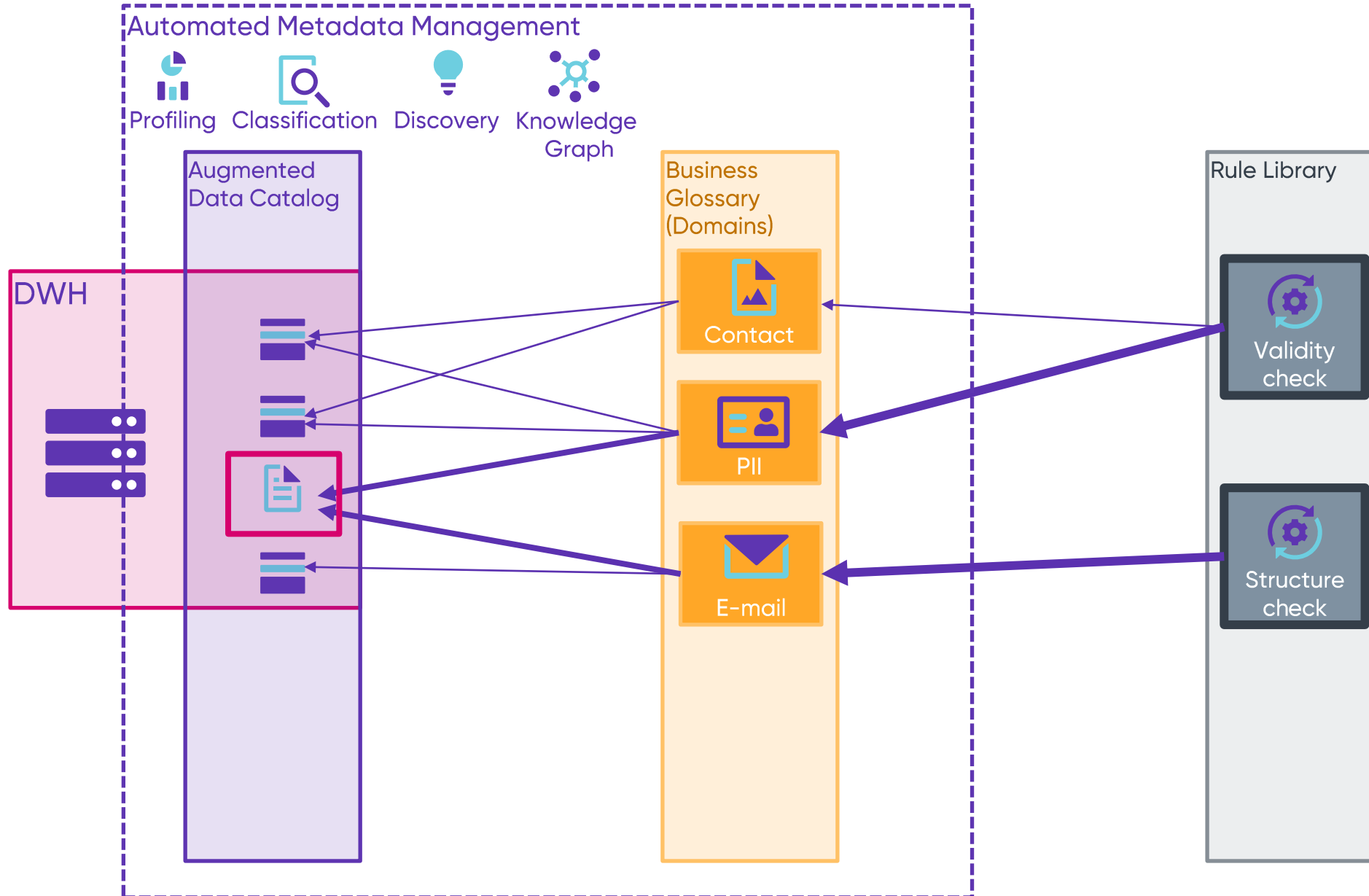
Metadata Driven DQ



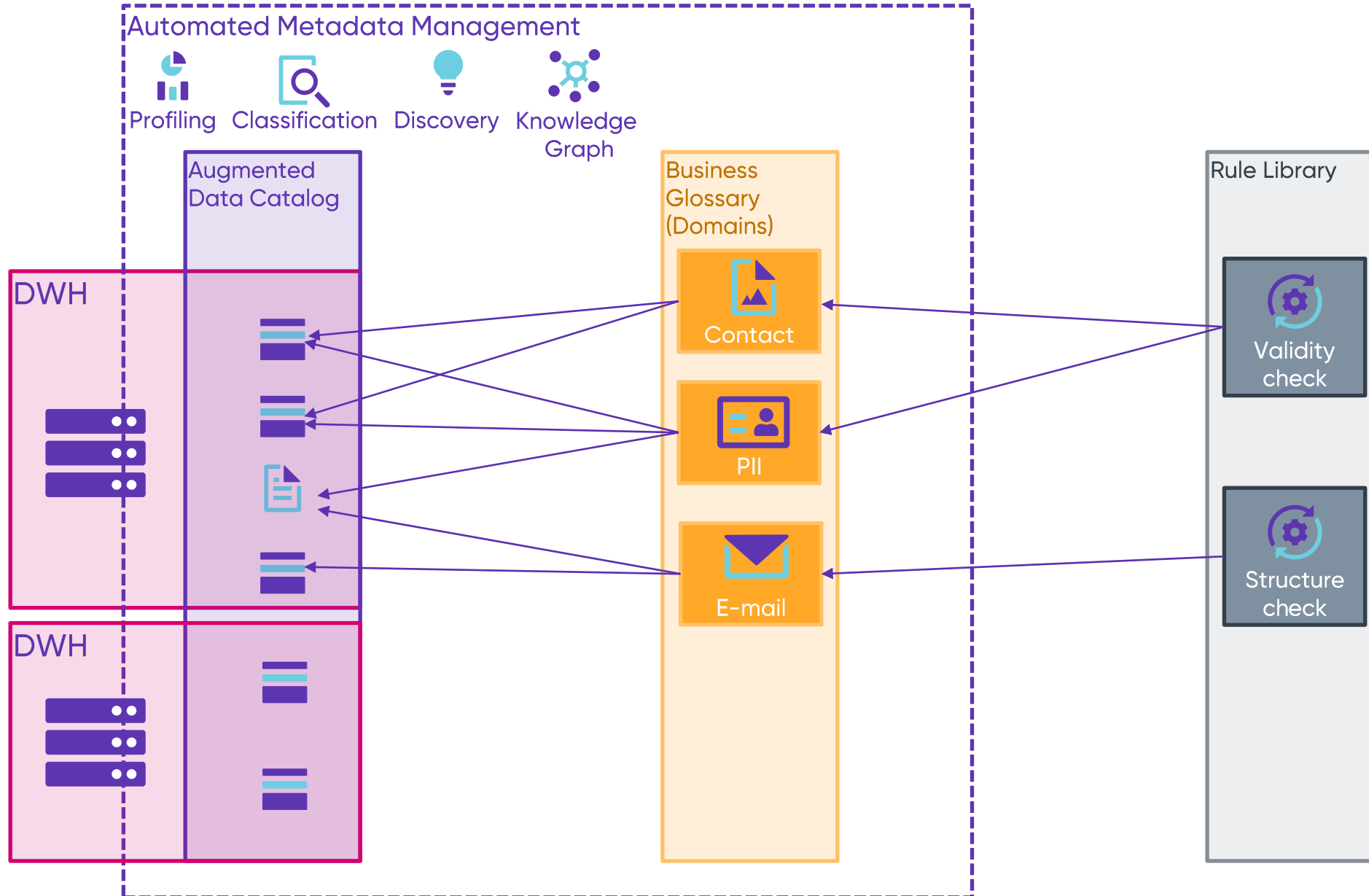
Metadata Driven DQ



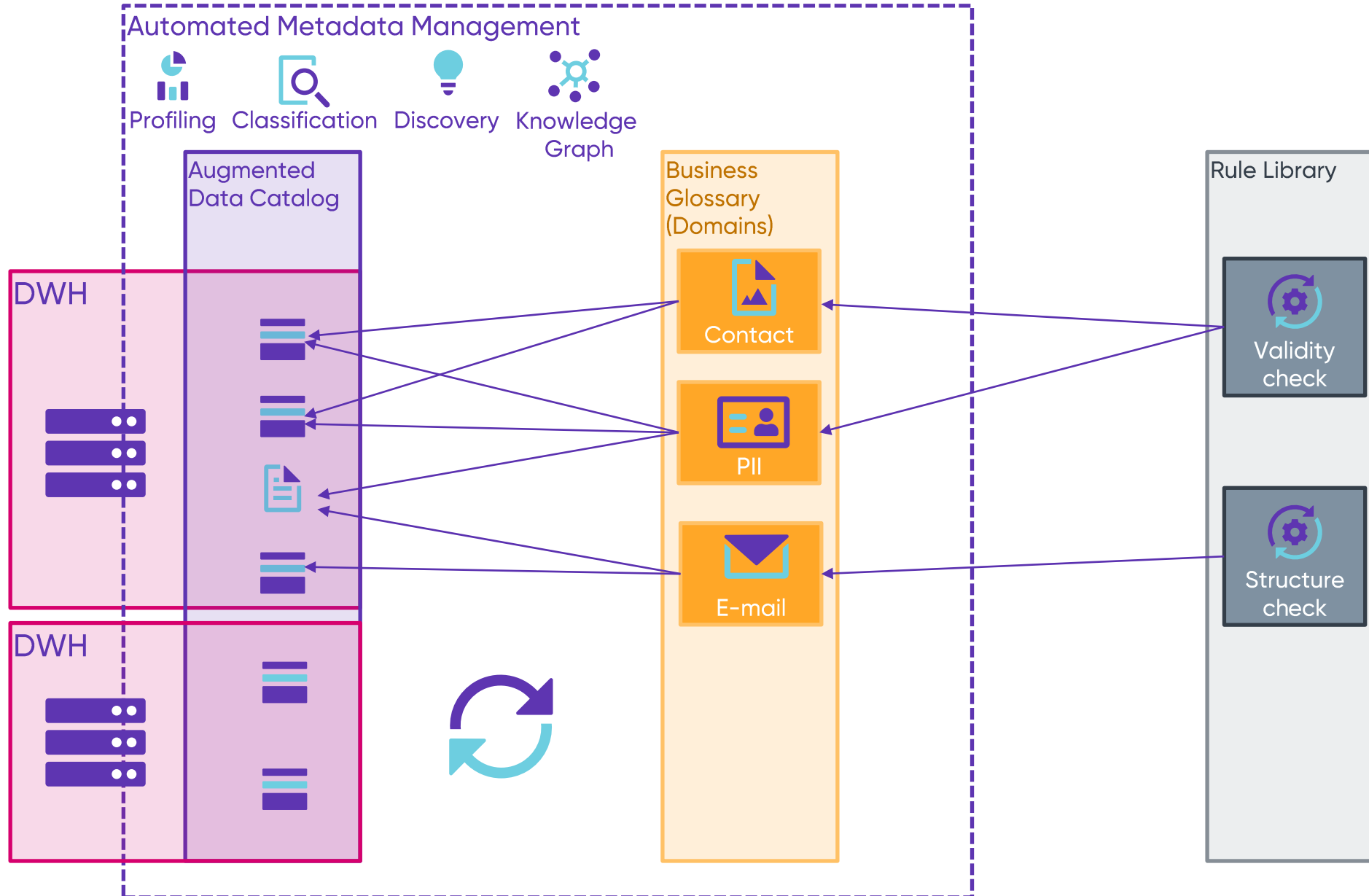
Metadata Driven DQ



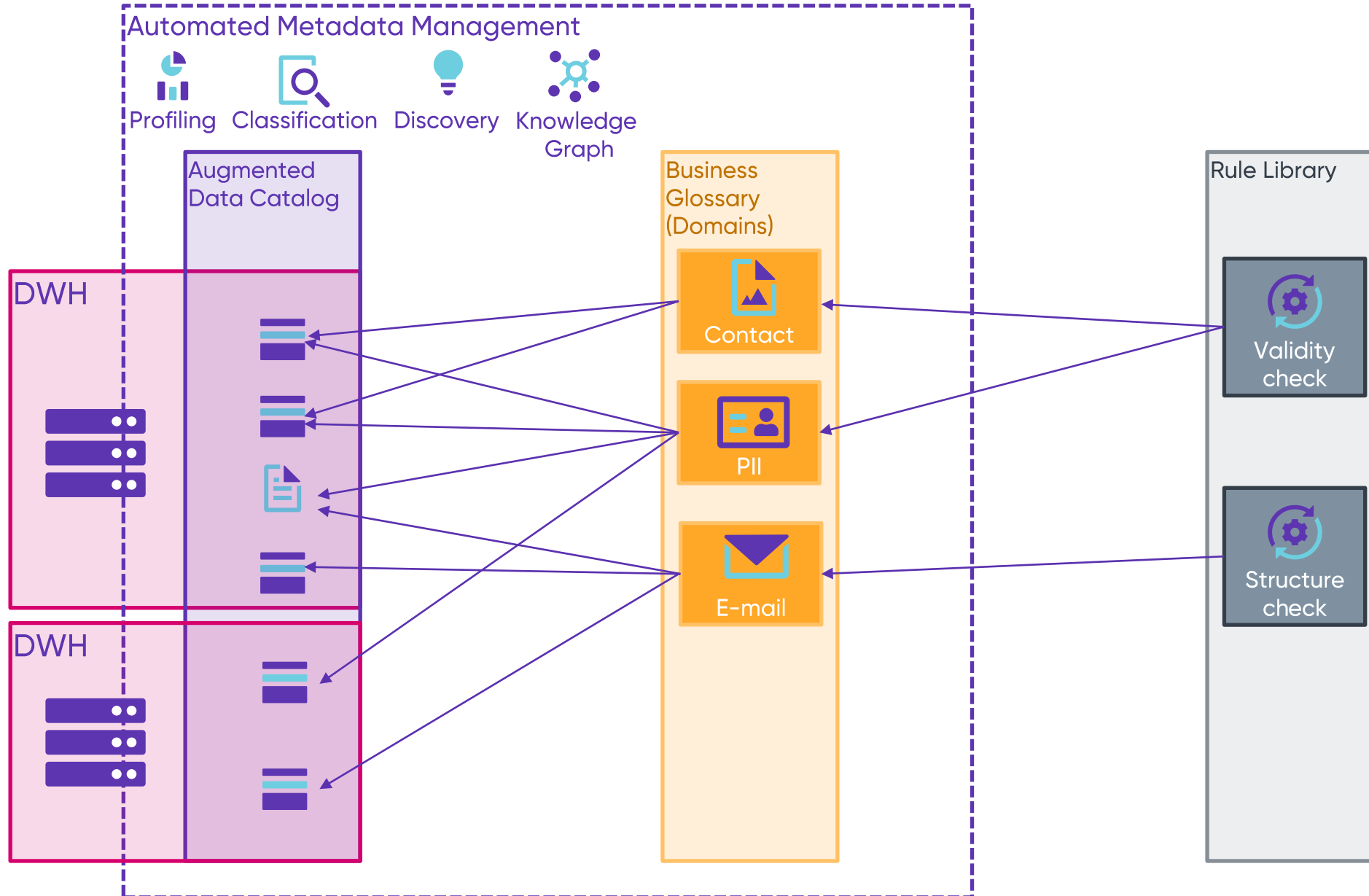
Metadata Driven DQ



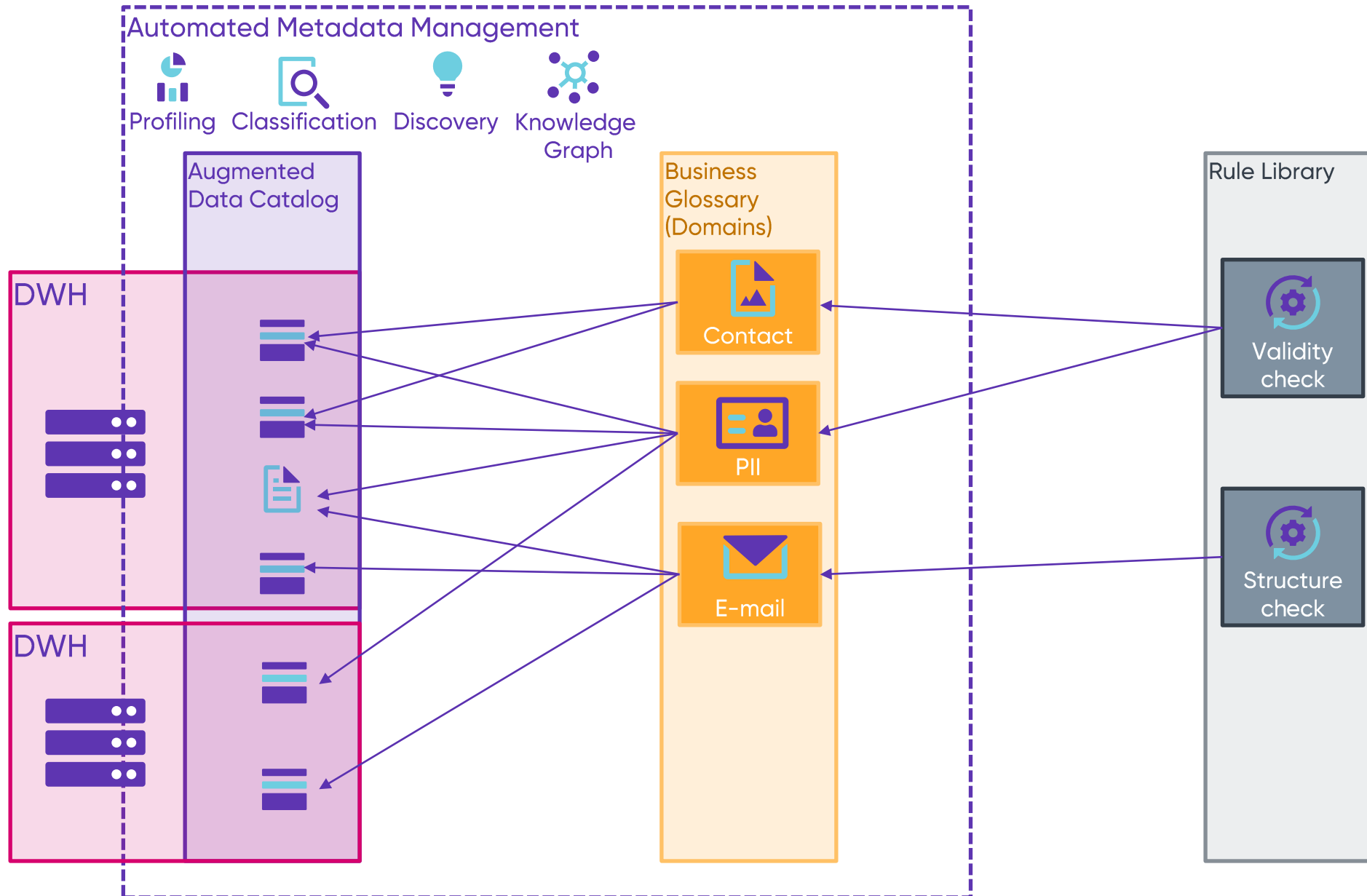
Metadata Driven DQ



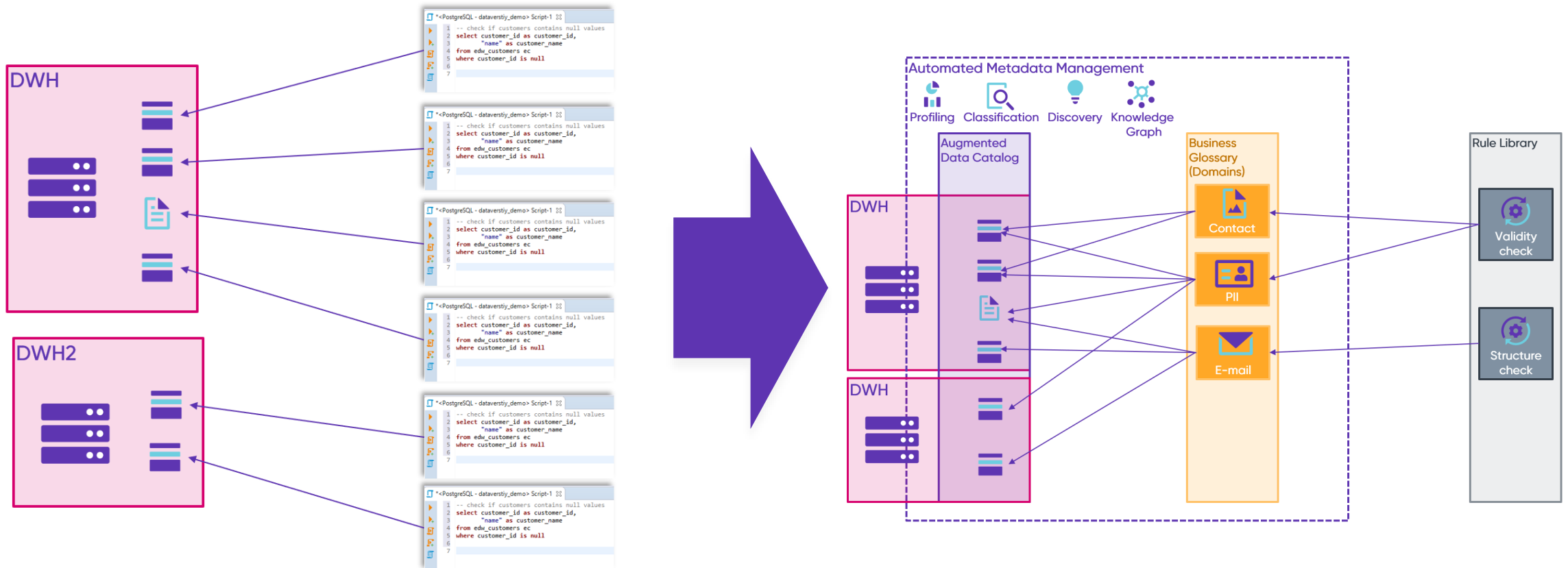
Metadata Driven DQ



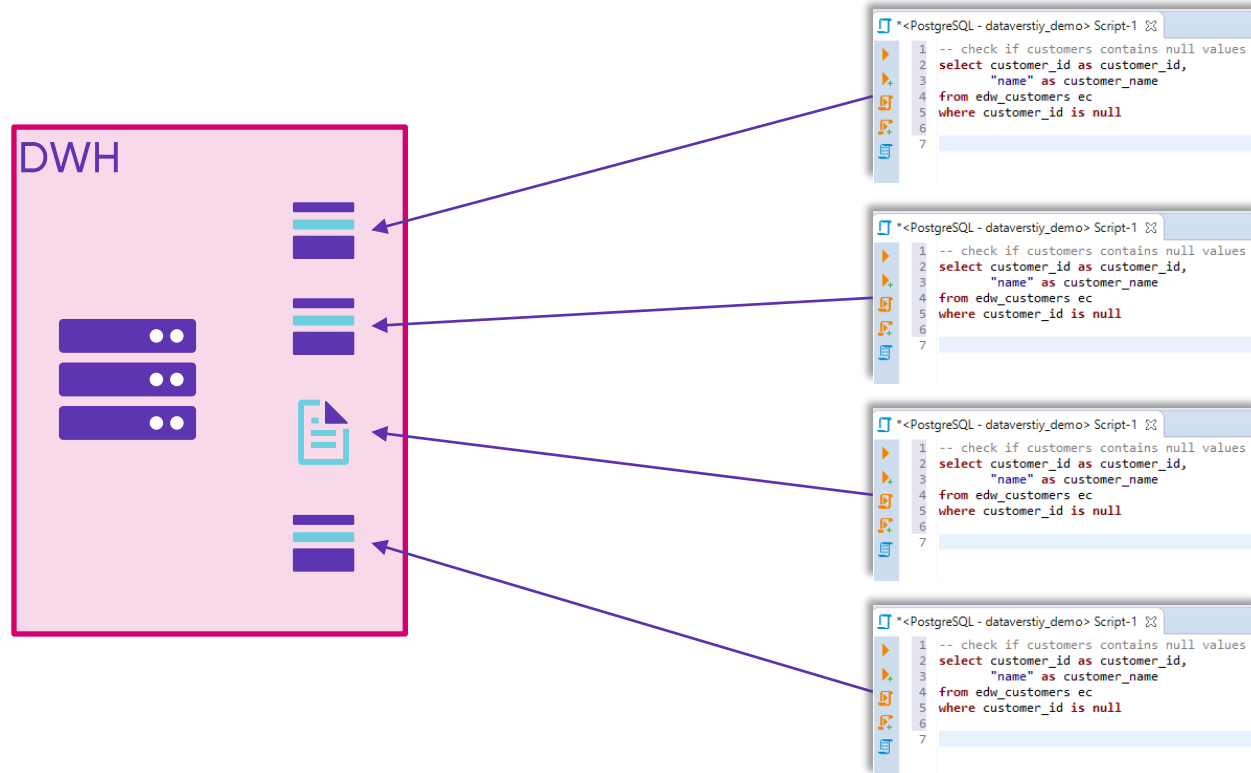
Metadata Driven DQ



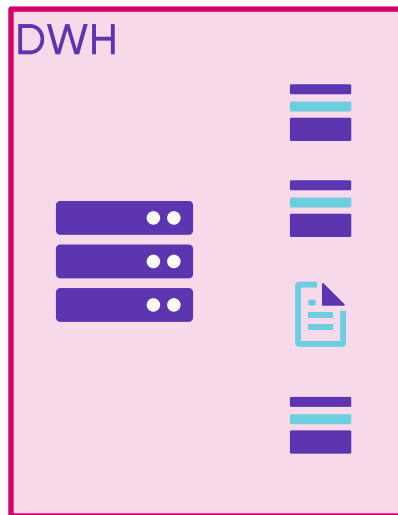
How Do you get There?



Re-Usable Rules



Re-Usable Rules



1 VALID_US_STATE Condition Builder

WHEN

is not empty

AND

is from reference data

+

1 COMPLETENESS Condition Builder

WHEN

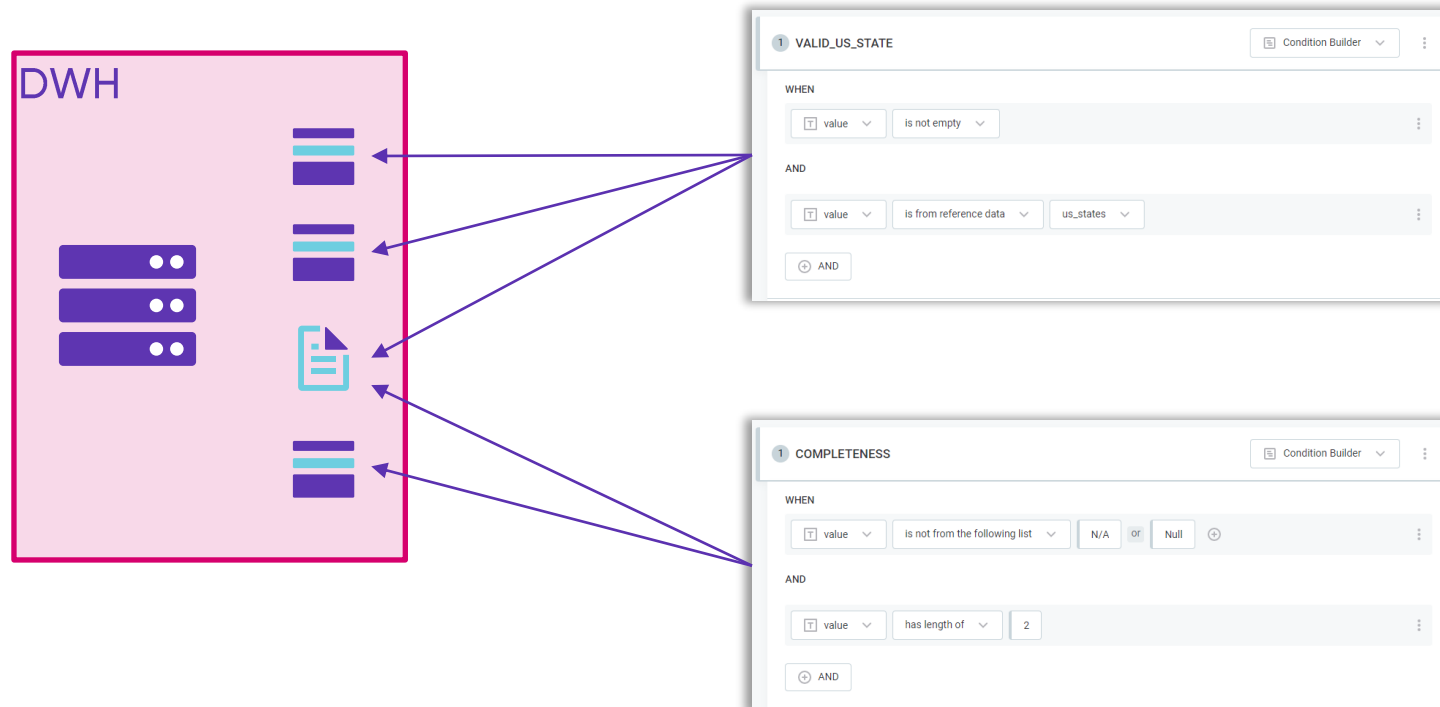
is not from the following list or

AND

has length of

+ AND

Re-Usable Rules



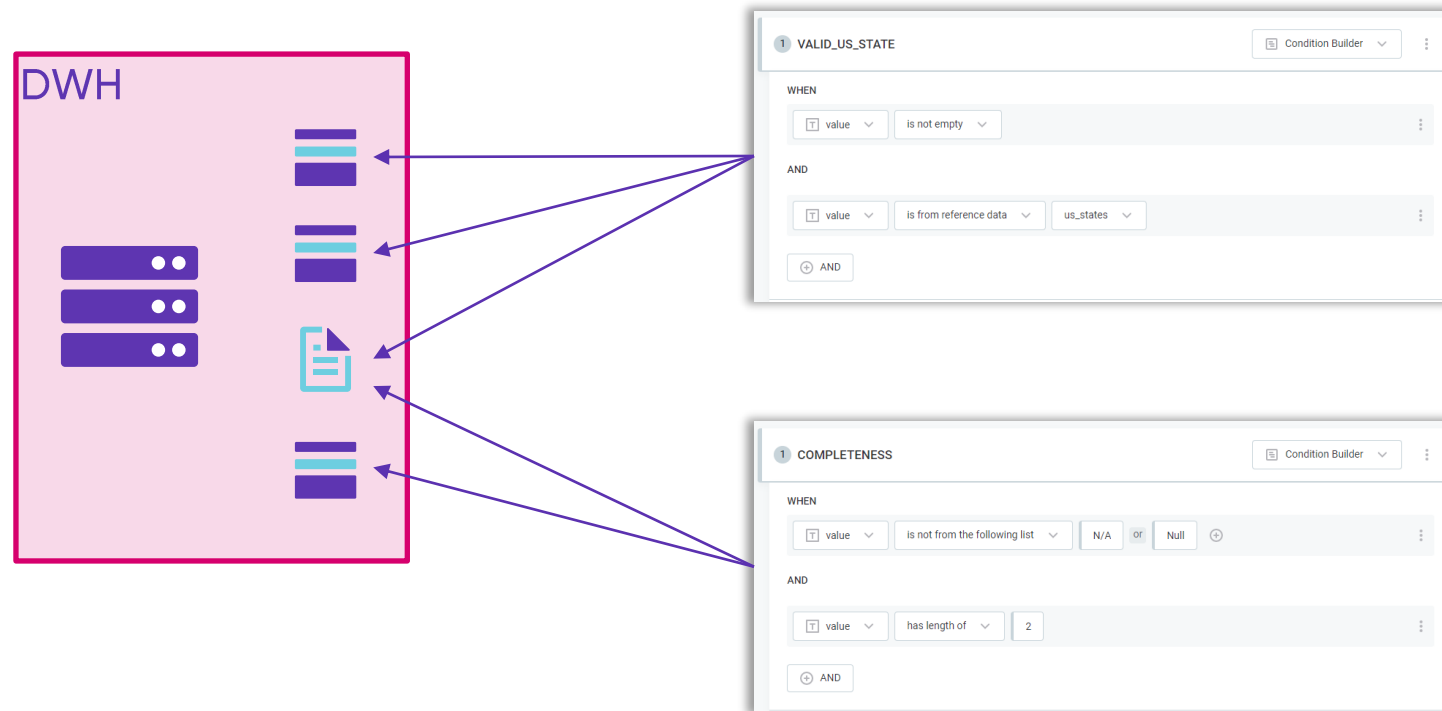
Rule management & Re-usable Rules

The screenshot shows a web interface for managing data quality rules. On the left is a sidebar with navigation links: KNOWLEDGE CATALOG, OVERVIEW, DATA CATALOG (with sub-links for Data Assets, Business Systems, Data Sources, Data Marts, and Infrastructure), Rules (selected), Components, BUSINESS GLOSSARY, BUSINESS PROCESSES, and PROJECTS. The main content area is titled 'Customer' and contains a table of rules. The table has columns for Name, Purpose, Usage, Input Attributes, and Description. It lists various rules such as '[CAN] ZIP code', 'ISIN', 'E-mail simple', 'URL', 'IPv6', '[CAN] SIN validation', '[USA] Zip code', '[CAN] Social Insurance Number', '[North America] Phone Number', and 'IPv4'. Each rule entry includes a checkbox, a name, a purpose (Detection or DQ Evaluation), usage statistics (number of rules, data items, and tests), input attributes (like Zip Code, SIN, E-mail, URL, Website, IP Address, Phone), and a description. A search bar at the top of the table allows filtering by name, owner, or date. A 'Create' button is visible in the top right corner of the main area.

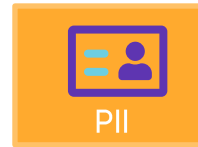
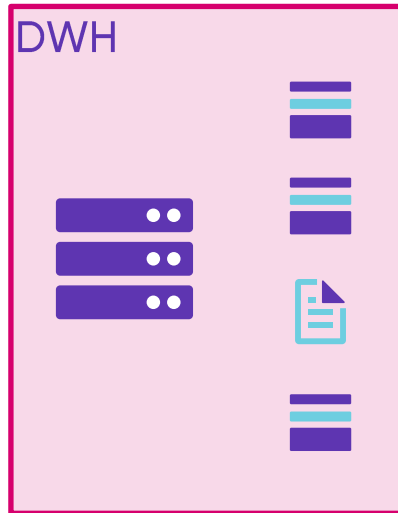
☐	Name	Purpose	Usage	Input Attributes	Description
☐	EQ [CAN] ZIP code	Detection	3 368 1,789	Zip Code	Canadian postal code
☐	EQ ISIN	Detection	2 123 1,434	SIN	This domain validates ISIN and by doing so te
☐	EQ E-mail simple	DQ Evaluation	2 87 1,100	E-mail	Simple validation of E-mail structure
☐	EQ URL	Detection	3 56 123	URL Website	URL (Uniform Resource Locator)
☐	EQ IPv6	Detection	5 40 231	IP Address	An Internet Protocol Version 6 address
☐	EQ [CAN] SIN validation	DQ Evaluation	2 70 321	SIN	Validates format and checksum of canadian S
☐	EQ [USA] Zip code	Detection	3 100 2,789	Zip Code	Postal code used by the United States Postal
☐	EQ [CAN] Social Insurance Number	Detection	2 102 2,329	SIN	A social insurance number SIN is a number is
☐	EQ [North America] Phone Number	Detection	3 50 1,812	Phone	Phone numbers data in the North American fo
☐	EQ IPv4	Detection	5 20 567	IP Address	An Internet Protocol Version 4 address

- > **Reusable Data Source Agnostic** rules
- > Easy-to-use **collaborative rule management**
- > Full rule **history** and rule **version comparison** available
- > Central rule library, **shared** in various projects

Map rules to metadata Instead of data



Map rules to metadata Instead of data



1 VALID_US_STATE Condition Builder

WHEN

[T] value is not empty

AND

[T] value is from reference data us_states

AND

1 COMPLETENESS Condition Builder

WHEN

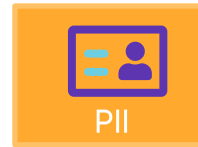
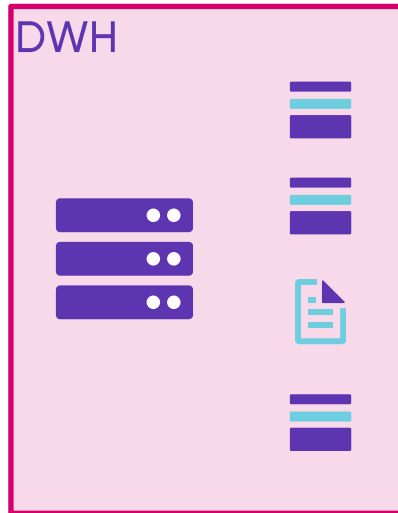
[T] value is not from the following list N/A or Null

AND

[T] value has length of 2

AND

Map rules to metadata Instead of data



1 VALID_US_STATE Condition Builder

WHEN

[T] value is not empty

AND

[T] value is from reference data us_states

AND

1 COMPLETENESS Condition Builder

WHEN

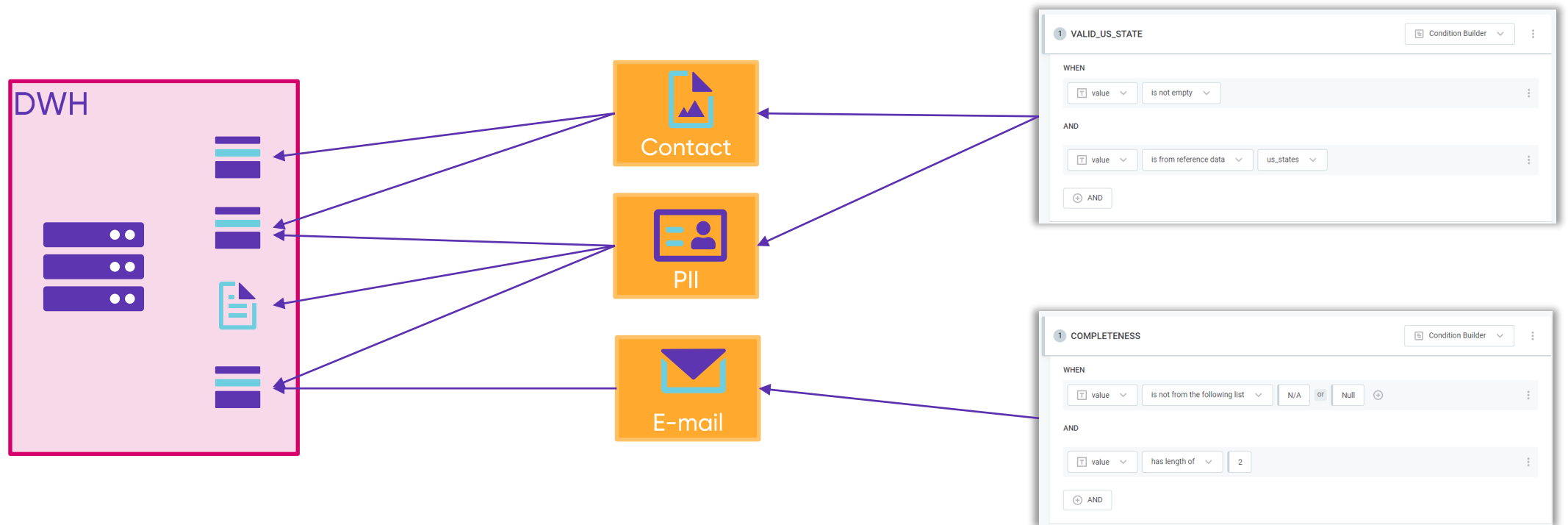
[T] value is not from the following list N/A or Null

AND

[T] value has length of 2

AND

Automatically Generate Metadata Mapping



Knowledge catalog

The screenshot displays a 'Knowledge Catalog' interface. On the left is a sidebar with navigation options: OVERVIEW, DATA CATALOG (selected), BUSINESS GLOSSARY, BUSINESS PROCESSES, and PROJECTS. The main area is titled 'Catalog Items' and features a search bar with the placeholder 'Filter by name, owner, creation date...'. Below the search bar is a table listing various data assets. Each row includes a checkbox, a name, terms, data quality status, number of records, number of attributes, system, and insights. Data quality is visualized with green and red progress bars. Some items have a blue 'V+6' icon, and others have a yellow warning triangle.

	Name	Terms	Data Quality	# Records	# Attributes	System	Insights
☐	src_person	PII Employee Enum		1 345 987	12	CRM	
☐	Master customer	PII Customer		576	57	EMEA	V+6
☐	Customers 2019	PII Customer		987 879	6	BRIC	
☐	comp	Account		98	23	CRM	
☐	Customer campaigns	Customer Campaign		7 768	2	EMEA	
☐	invoice-20200829.pdf	PII Account		none	none	BRIC	
☐	employees_2020	PII Employee		1 345 987	12	CRM	
☐	Master address	Address		576	57	EMEA	V+6
☐	customers_2019_ext	PII Customer		987 879	6	BRIC	
☐	account_list	PII Account		98	23	CRM	
☐	Reports summary					EMEA	
☐	invoice-20200531.pdf	Account		none	none	BRIC	
☐	src_person	PII Employee Enum		1 345 987	12	CRM	
☐	Master customer	PII Customer		576	57	EMEA	V+6

- > Locate, capture and store metadata
- > Capture data relations and lineage
- > Share metadata assets across the organization
- > Automatically calculate DQ metrics

Business glossary

Home > Knowledge Catalog > Business Glossary

Create

Filter by name, owner, creation date...

View: Category

☐	Name	Abbreviation	Type	Data Quality
☐	Account	-	Category	View
☐	Campaign	-	Category	View
☐	Employee	-	Category	View
☐	Invoice	-	Category	View
☐	Order	-	Category	View
☐	Activity	-	Sub-category	View
☐	Date	-	Sub-category	View
☐	Item	-	Sub-category	View
☐	Order item	-	Element	View
☐	Order item ID	-	Element	View
☐	Unit price	-	Element	View
☐	Order	-	Sub-category	View
☐	Person	-	Sub-category	View
☐	Product	-	Sub-category	View

- › Central component of Ataccama ONE managing all business terms
- › Relationships among business terms are possible and their hierarchy is available
- › Occurrence statistics across all connected data sources
- › Advanced conditions or rules as part of the business term definition
- › Link with relevant data assets, links to DQ rules

Data profiling

The screenshot shows a 'Knowledge Catalog' interface. On the left is a sidebar with navigation options: Overview, Data Catalog, Data Assets, Business Systems, Data Sources, Data Marts, Infrastructure, Rules, Components, Business Glossary, Business Processes, and Projects. The main area displays a profile for 'party_full' with 2999 records and 7 attributes. The 'Profile' tab is active, showing a table of attributes with their respective data quality insights, top values, and mask analysis.

Name	Terms	Insights	Top 3 Values	Mask Analysis
src_primary_key		3 Duplicates	3% NaN 8% 145 8% 146	3% LLL 47% DDD 58% DDDD
src_name	Last Name	3 Duplicates	24% Null 3% Green 2% Kazner	8% LLLL 8% LLLLL Show All +29
src_sin	Social Insurance Number	NULL: 24%	24% Null 8% 183792775 8% SIN: 999678852	24% LLL: 000000000 18% 000000000 Show All +22
src_card	Credit Card Number	7 Exceptions	2% ##### 1% ##### 8% #####	98% 0000000000000000 2% LLLL
src_email		Email ✓ 80%	3% jane.doe@gmail.com 2% noname@mail.com 8% agent87@gmail.com	19% LLLLLL0LLLL.LLL 8% LLLL0LLL.LL Show All +16
src_birth_date	Birth Date	Outlier Value Detected in given attribute	38% 1988-01-01 6% Null 1% 1989-09-01	74% 0000-00-00 8% 00.00.0000 Show All +16
src_address		Address ✓ 80%	58% Null 3% N/A 1% 25 Linden Str	58% LLL 12% 0000-0000-0000-0000 Show All +16

- › Automated out of the box profiling
- › Data domain discovery
- › AI-driven classification
- › SME-driven process
- › Automated detection of outliers and exceptions
- › Sharing of profile reports

Solution



Capture

Automatically discover
Metadata

Import metadata

Manually create
metadata



Manage

Detect and manage
business domains

All data assets in
a knowledge catalog

Curate and update



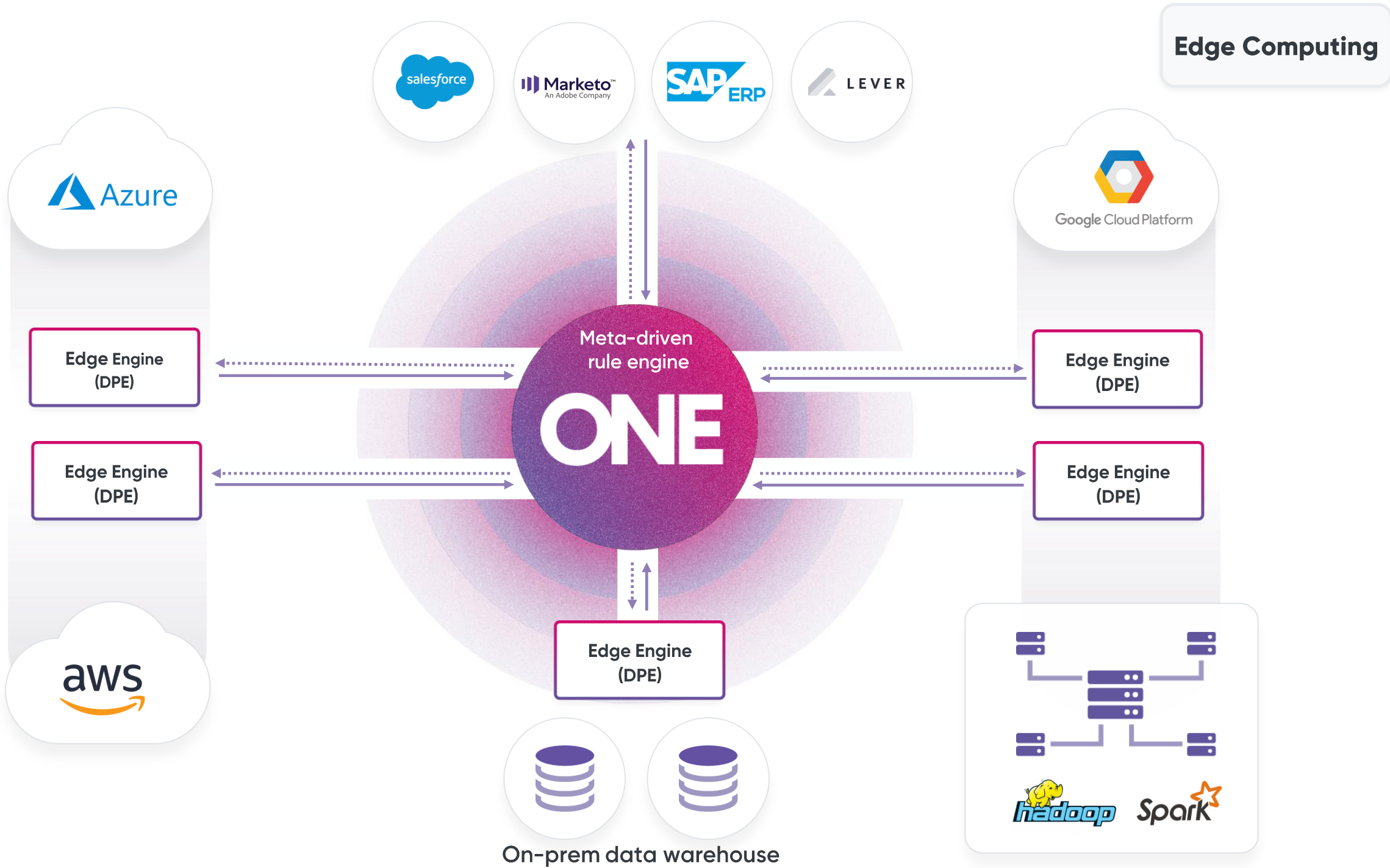
Execute

Automatically Generate Data
Pipelines

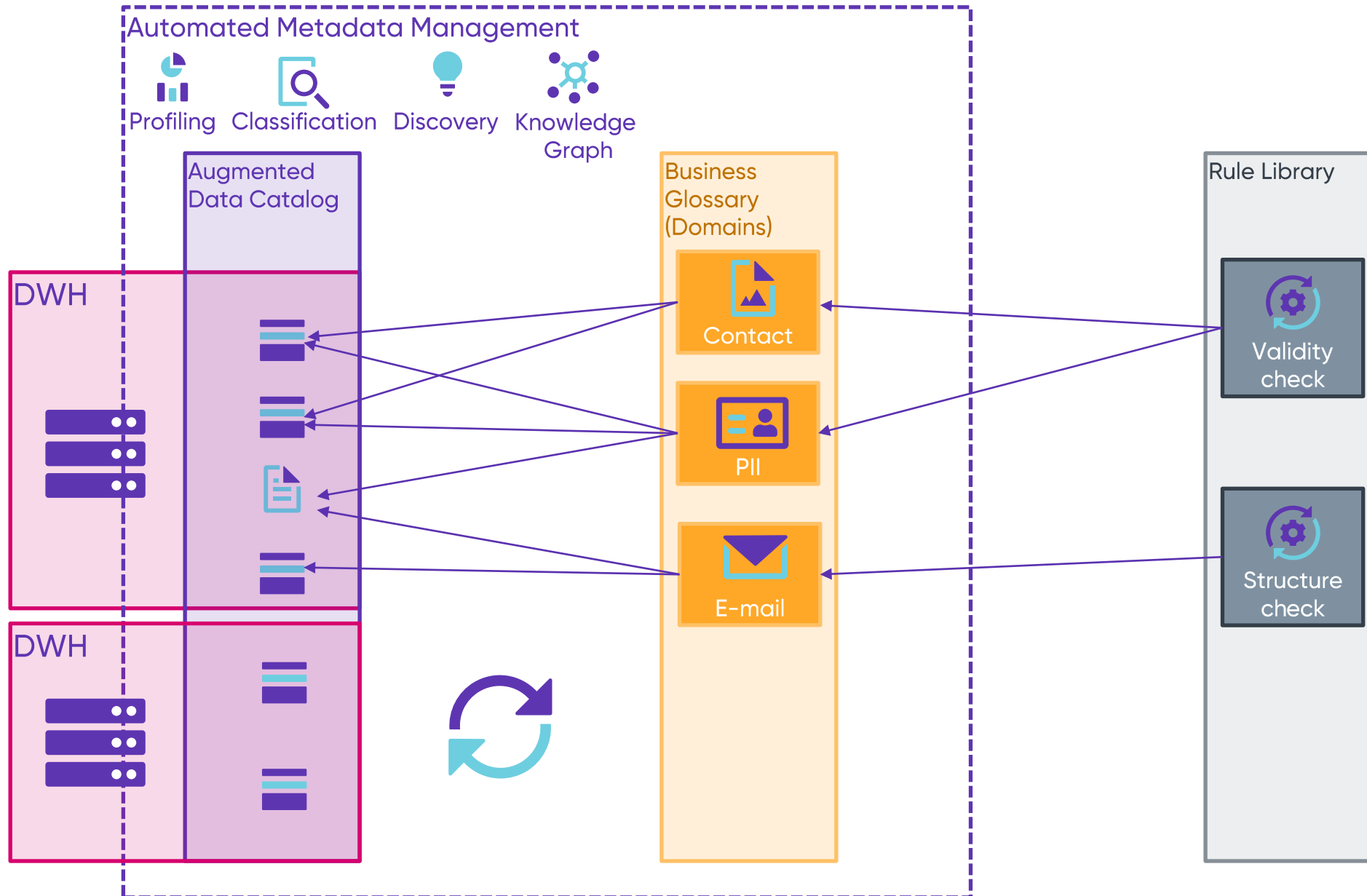
Data Source Agnostic Rules

Horizontally Scalable

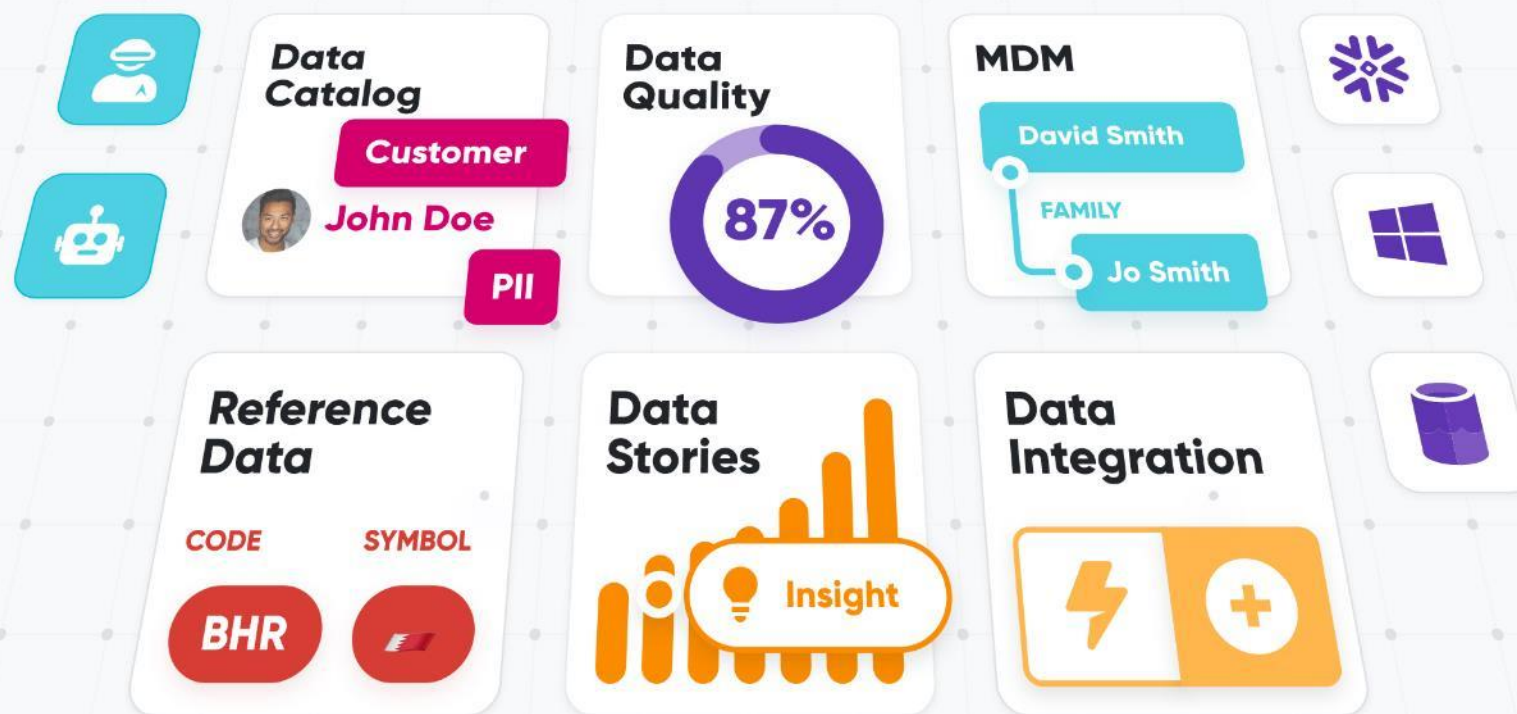
Multi-Cloud Edge processing



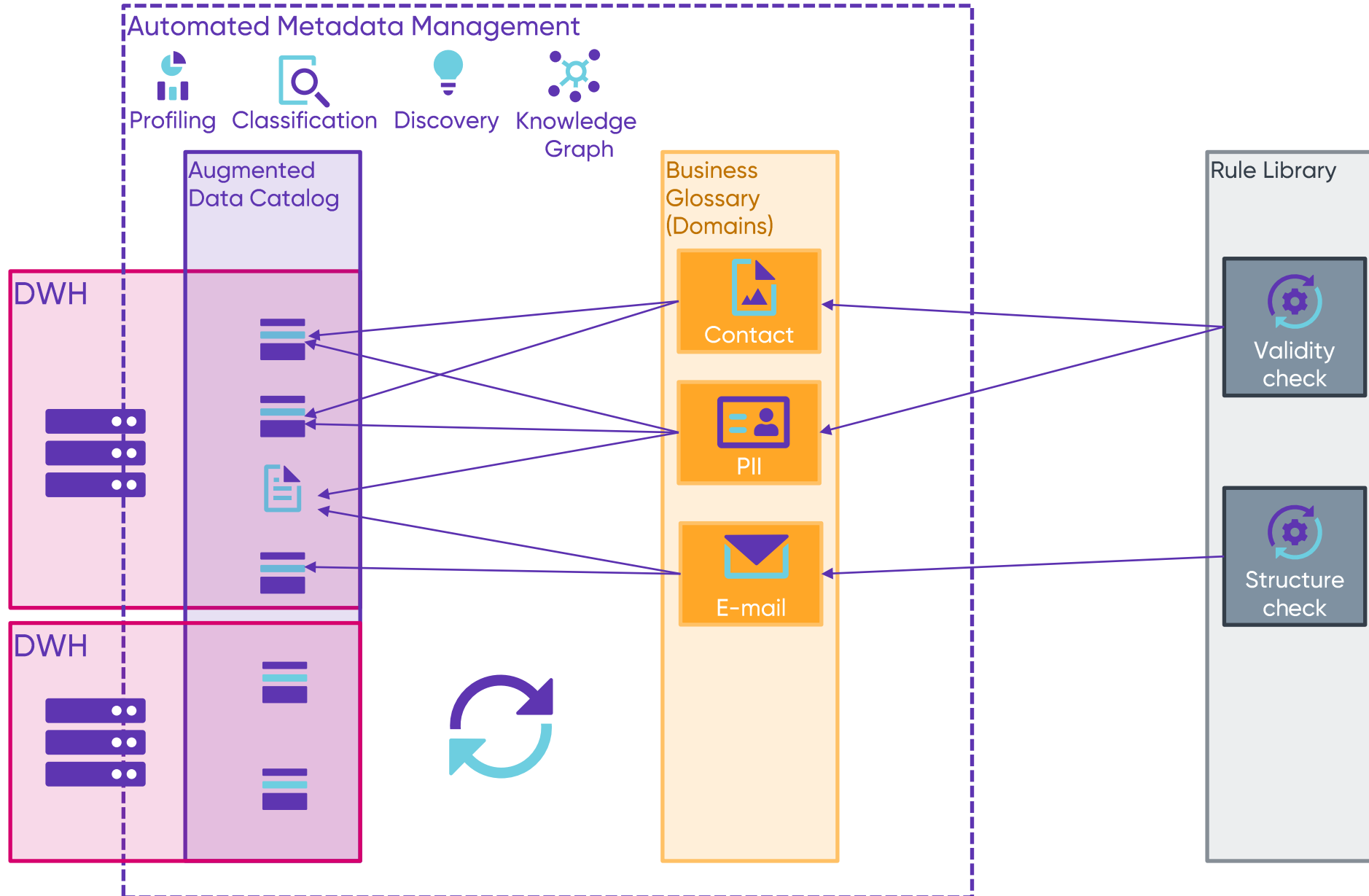
Metadata Driven DQ



Data Quality Fabric



Metadata Driven DQ



Q & A

Learn more and schedule a demo at Ataccama.com

