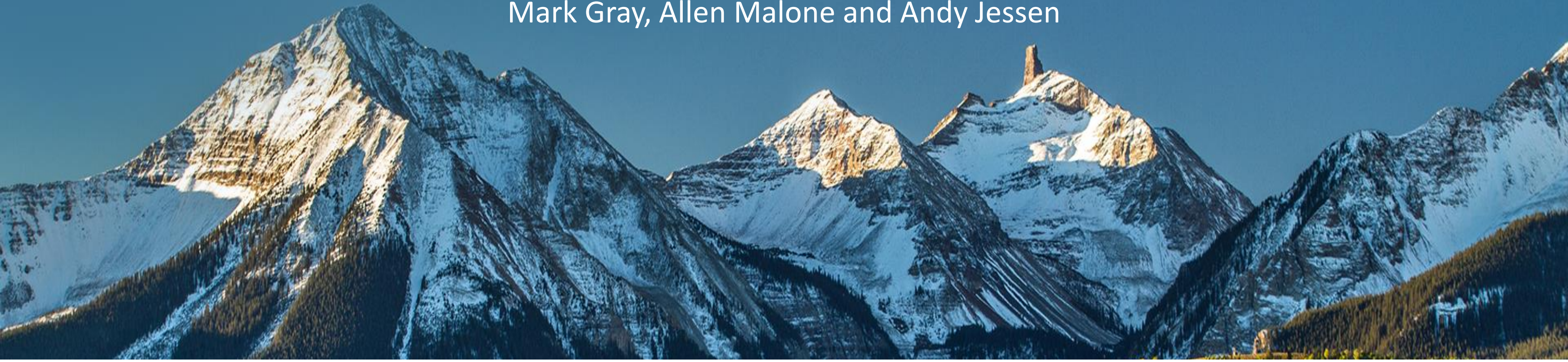
KAISER PERMANENTE®
Institute for Health Research

Data Quality Framework

HCSRN Scientific Data Resources Forum

September 9, 2025

Mark Gray, Allen Malone and Andy Jessen



KAISER PERMANENTE®
Institute for Health Research

VDW Infrastructure Team

Allen Malone

Andy Jessen

Brian Hixon

Mark Gray

Heather Tavel

Ravi Zalavadia

Mahesh Maiyani

Artie Runkle

James Lagrotteria

Rachel Zucker

Andy Sterrett

Diego Gomes

VDW Infrastructure Team

- **Vision and Mission of the VDW Team**
- **Our Vision**: *To be the research information partner of choice by providing dynamic, robust, and user-centered data collaborations aimed at deepening our understanding of factors that influence health outcomes and drive improvements in health care delivery.*
- **Our Mission**: *To provide high-quality, accessible, relevant, and timely data with the subject matter expertise for the research and evaluation communities so that information can be readily interpreted through technology and analytics.*

Justification for DQ Framework

- Pre-Dashboard State of QC
  - Monthly QCs are reviewed by the VDW team prior to each load
  - CESR/HCSRN conducts QCs for content areas as well as those initiated after a major change to assure adherence.
  - Projects that source their data from the VDW, such as Sentinel and Research Bank, also perform their own QCs
- Reasons for the DQ Framework
  - DQ programs had variation and minimal standardization. Same checks were done in different ways, programmed by different people (herd the cats), federated approach
  - Targeted currently to our Content Areas SMEs for "break/fixes". DQ dashboard would allow access to investigators and other downstream users of the VDW.
  - More timely QC. Instead of waiting for the CESR, HCSRN, and project related QC programs we are more proactive instead of reactive about fixes.

Challenges with V1.0

- **Delayed QC cycles**: Fixes were often reactive, waiting on external QC programs.
- **Lack of standardization**: Similar checks were implemented differently across teams (“herding cats”).
- **Limited Visibility/Access**: QC insights were mostly accessible to Content Area SMEs, not to broader stakeholders.

Objectives for New QC Framework (V2.0)

- **Proactive QC**: Enables earlier detection and resolution of data issues.
- **Standardized checks**: Centralized logic reduces duplication and inconsistency.
- **Greater accessibility**: A DQ dashboard empowers investigators and downstream users with real-time insights.

QC Model Inspiration

Model Referenced: *A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data*
Europe PMC Article Link

Why This Model?

- We didn't want to reinvent the wheel.
- The framework is well-vetted and aligns with our goals.
- It offers a structured approach to assessing data quality.

Key Dimensions of Data Quality:

- Completeness
- Correctness
- Concordance
- Plausibility
- Currency

ME LOOKING FOR A SIMPLE SOLUTION
VENDOR QUOTING ME NEARLY 200K

QC Framework Timeline

March 2023

Start QC Framework Design

- Data harmonization (OHDSI)
- Dashboard – OMOP
- IHR Features

March 2024

Roll Out to SMEs

- VDW SMEs only
- Limited tables
- Training & Demos

May 2022

Internal VDW QC Audit

2022 goal to improve VDW data quality

July 2023

Develop

- Data modeling
- Coding
- Testing

SO YOU DON'T WANT TO INVEST IN DATA QUALITY
BUT YOU WANT TO BE DATA-DRIVEN?

Data Quality Feedback Loop

Solution Limitations

- Not covering all QC scenarios (e.g., composite measures)
- Error Thresholds are simple
- Implemented in SAS

There are two kinds of people.
1) Those who can extrapolate from incomplete data.

Application Architecture

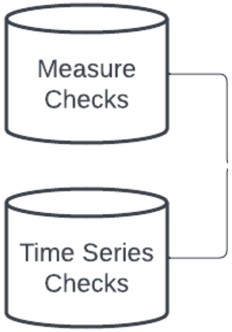
Microsoft SQL Server Database

The SAS System for ETL and data management

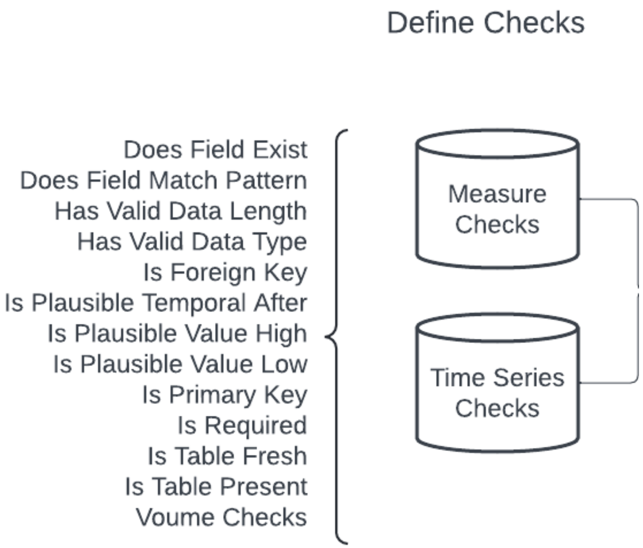
Tableau for Data Quality Reporting

Implementation

Define Checks



Implementation



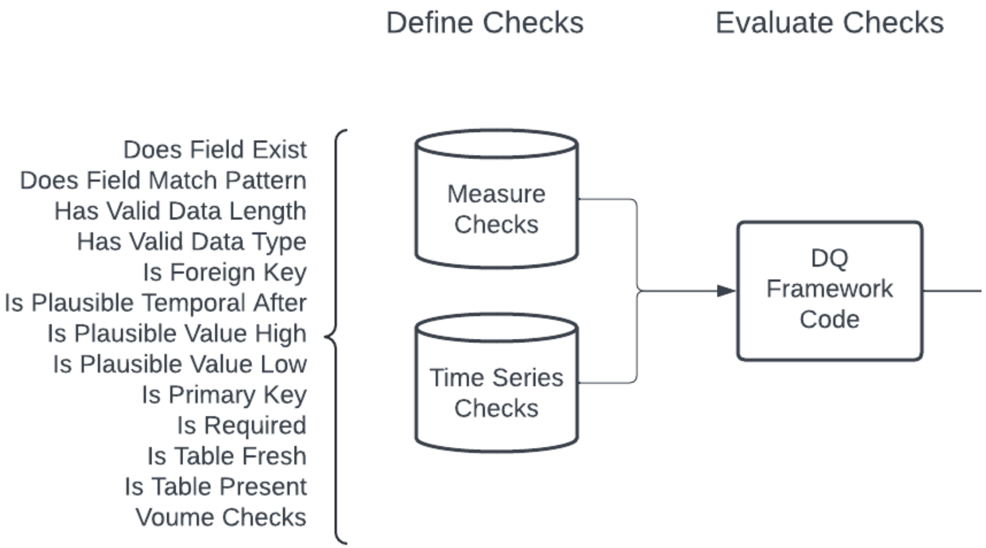
Define Checks

Does Field Exist
Does Field Match Pattern
Has Valid Data Length
Has Valid Data Type
Is Foreign Key
Is Plausible Temporal After
Is Plausible Value High
Is Plausible Value Low
Is Primary Key
Is Required
Is Table Fresh
Is Table Present
Voume Checks

Measure Checks

Time Series Checks

Implementation



Define Checks

Evaluate Checks

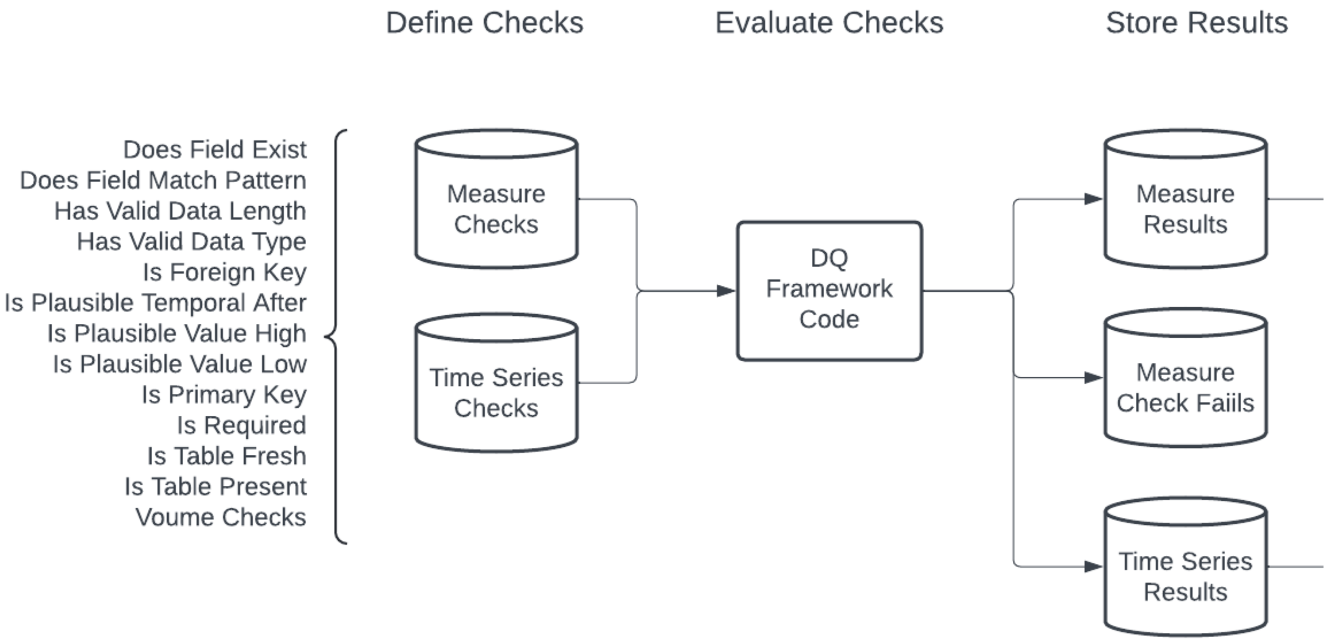
Does Field Exist
Does Field Match Pattern
Has Valid Data Length
Has Valid Data Type
Is Foreign Key
Is Plausible Temporal After
Is Plausible Value High
Is Plausible Value Low
Is Primary Key
Is Required
Is Table Fresh
Is Table Present
Voume Checks

Measure Checks

Time Series Checks

DQ Framework Code

Implementation



Define Checks

Evaluate Checks

Store Results

Does Field Exist
Does Field Match Pattern
Has Valid Data Length
Has Valid Data Type
Is Foreign Key
Is Plausible Temporal After
Is Plausible Value High
Is Plausible Value Low
Is Primary Key
Is Required
Is Table Fresh
Is Table Present
Voume Checks

Measure Checks

Time Series Checks

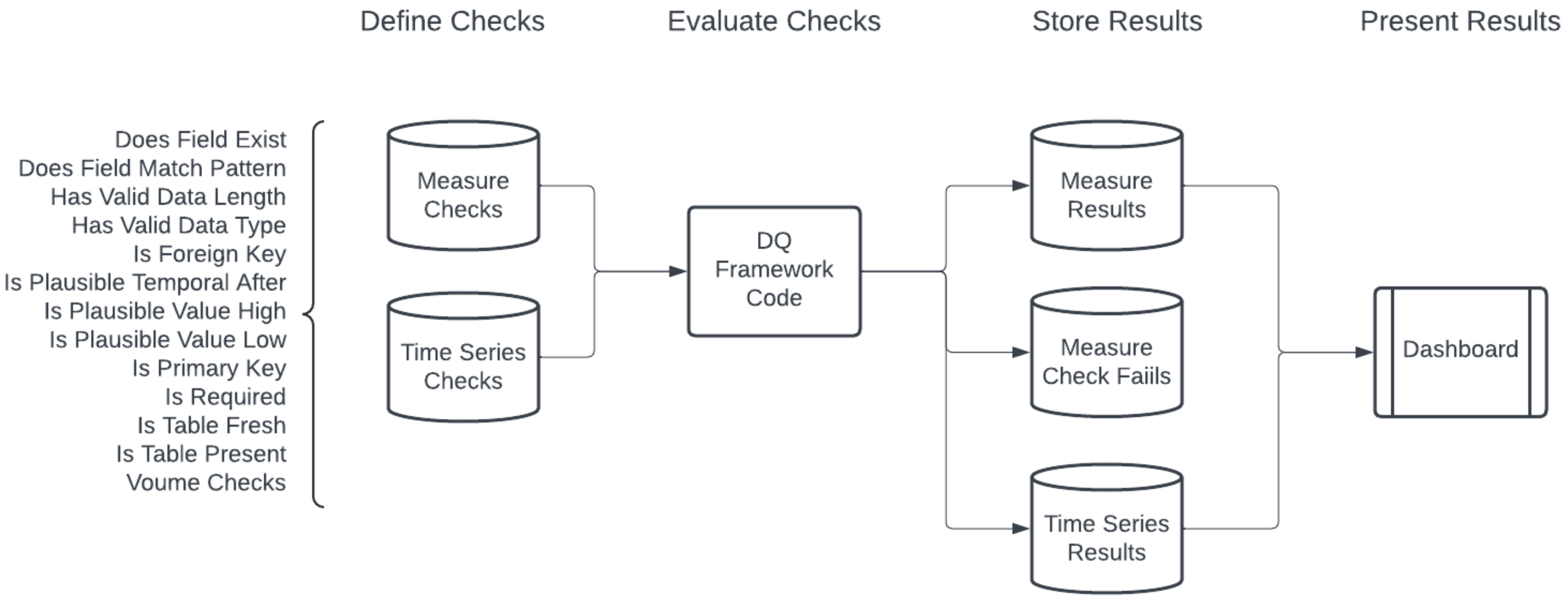
DQ Framework Code

Measure Results

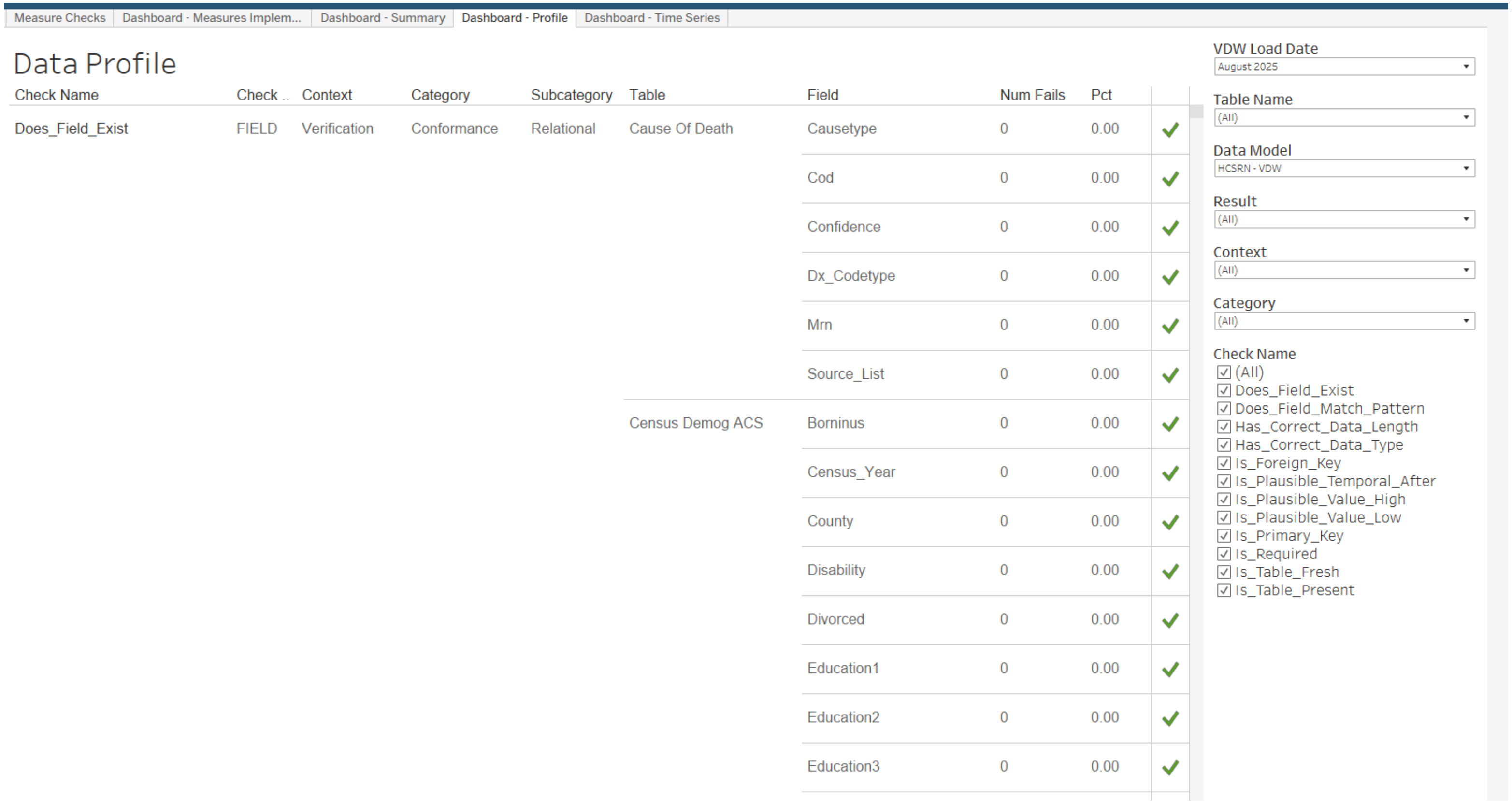
Measure Check Faiils

Time Series Results

Implementation



Tour of the Dashboard



Data Profile

Check Name	Check ..	Context	Category	Subcategory	Table	Field	Num Fails	Pct	
Does_Field_Exist	FIELD	Verification	Conformance	Relational	Cause Of Death	Causetype	0	0.00	✓
						Cod	0	0.00	✓
						Confidence	0	0.00	✓
						Dx_Codetype	0	0.00	✓
						Mrn	0	0.00	✓
						Source_List	0	0.00	✓
					Census Demog ACS	Borninus	0	0.00	✓
						Census_Year	0	0.00	✓
						County	0	0.00	✓
						Disability	0	0.00	✓
						Divorced	0	0.00	✓
						Education1	0	0.00	✓
						Education2	0	0.00	✓
						Education3	0	0.00	✓

VDW Load Date
August 2025

Table Name
(All)

Data Model
HCSRN - VDW

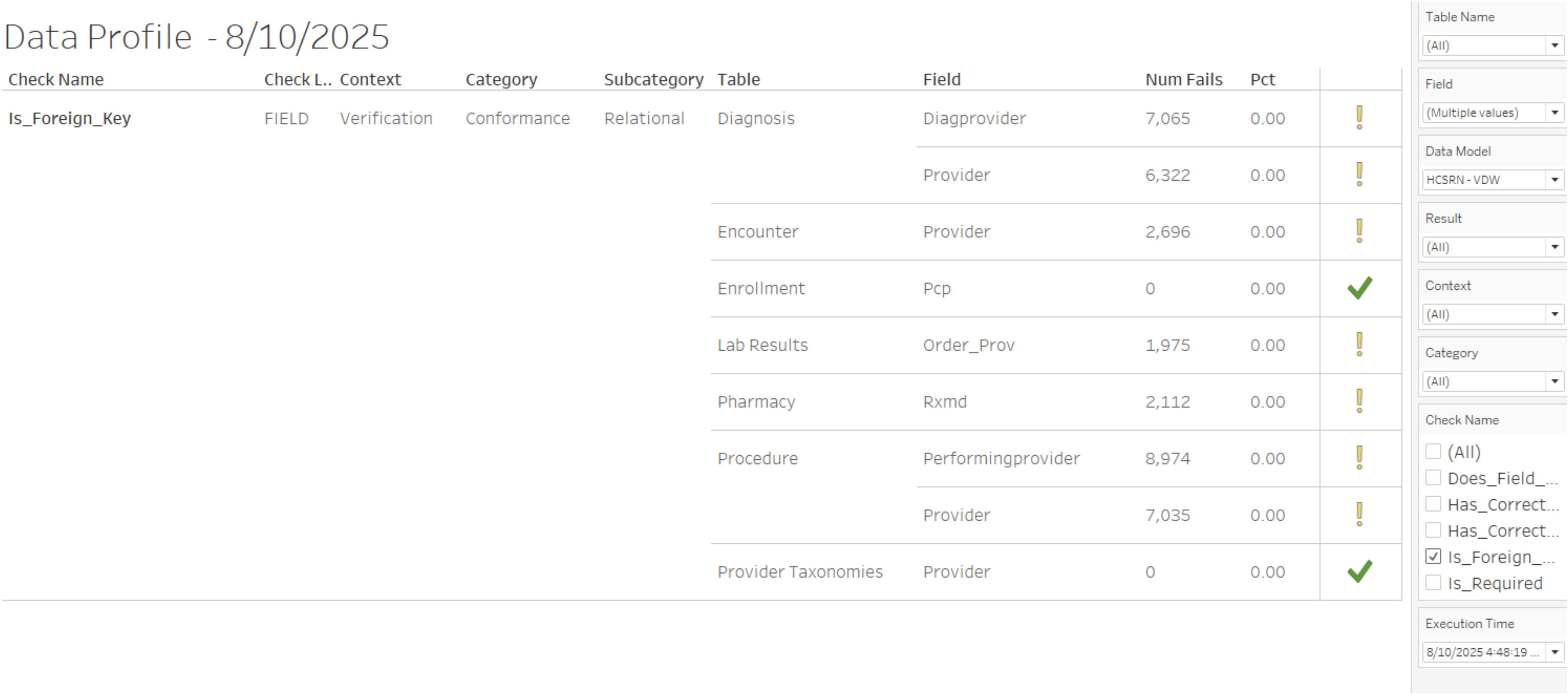
Result
(All)

Context
(All)

Category
(All)

Check Name
- ☑ (All)
- ☑ Does_Field_Exist
- ☑ Does_Field_Match_Pattern
- ☑ Has_Correct_Data_Length
- ☑ Has_Correct_Data_Type
- ☑ Is_Foreign_Key
- ☑ Is_Plausible_Temporal_After
- ☑ Is_Plausible_Value_High
- ☑ Is_Plausible_Value_Low
- ☑ Is_Primary_Key
- ☑ Is_Required
- ☑ Is_Table_Fresh
- ☑ Is_Table_Present

Provider Example (Before)



Data Profile - 8/10/2025

Check Name	Check L..	Context	Category	Subcategory	Table	Field	Num Fails	Pct	
Is_Foreign_Key	FIELD	Verification	Conformance	Relational	Diagnosis	Diagprovider	7,065	0.00	!
						Provider	6,322	0.00	!
					Encounter	Provider	2,696	0.00	!
					Enrollment	Pcp	0	0.00	✓
					Lab Results	Order_Prov	1,975	0.00	!
					Pharmacy	Rxmd	2,112	0.00	!
					Procedure	Performingprovider	8,974	0.00	!
						Provider	7,035	0.00	!
					Provider Taxonomies	Provider	0	0.00	✓

Table Name
(All)

Field
(Multiple values)

Data Model
HCSRN - VDW

Result
(All)

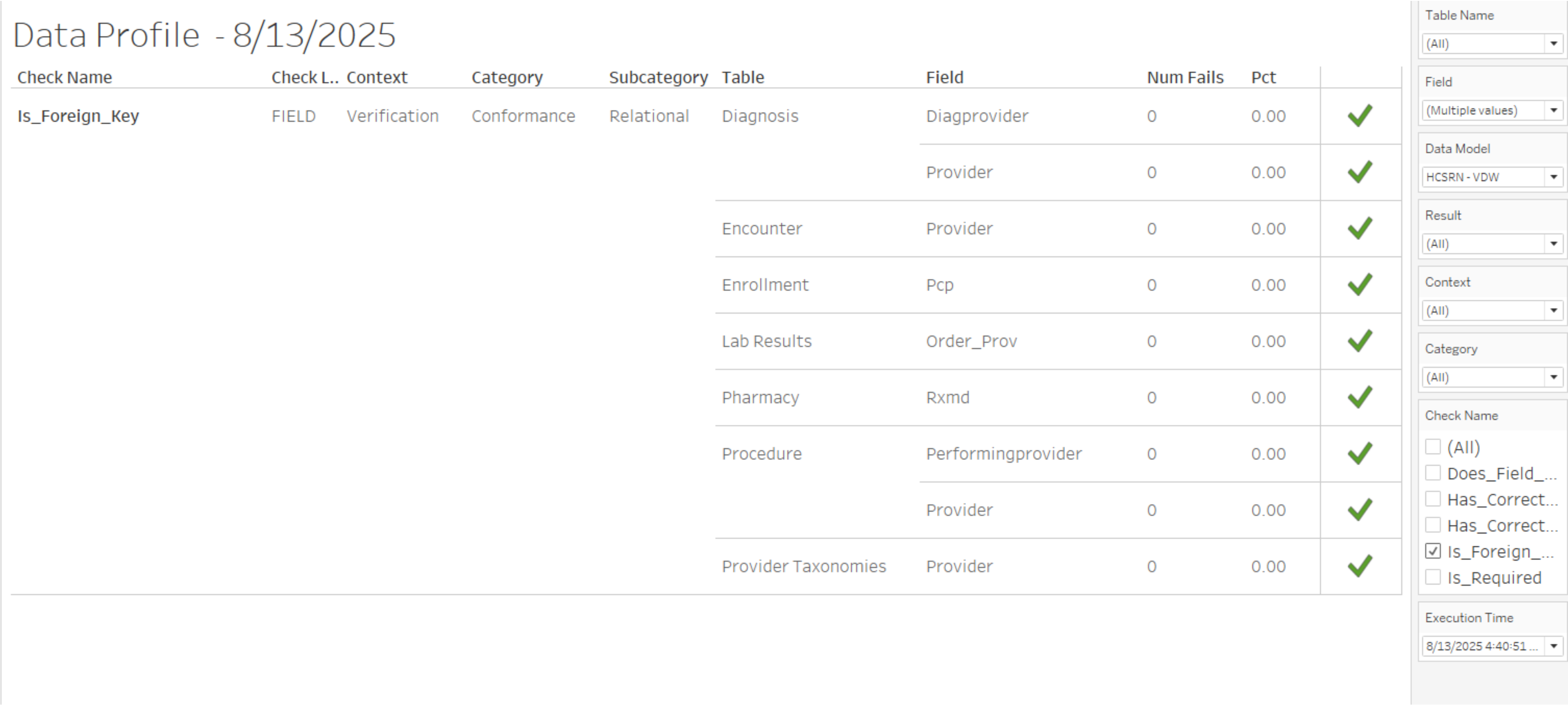
Context
(All)

Category
(All)

Check Name
- □ (All)
- □ Does_Field_...
- □ Has_Correct...
- □ Has_Correct...
- ☑ Is_Foreign_...
- □ Is_Required

Execution Time
8/10/2025 4:48:19 ...

Provider Example (After)



Data Profile - 8/13/2025

Check Name	Check L..	Context	Category	Subcategory	Table	Field	Num Fails	Pct	
Is_Foreign_Key	FIELD	Verification	Conformance	Relational	Diagnosis	Diagprovider	0	0.00	✓
						Provider	0	0.00	✓
					Encounter	Provider	0	0.00	✓
					Enrollment	Pcp	0	0.00	✓
					Lab Results	Order_Prov	0	0.00	✓
					Pharmacy	Rxmd	0	0.00	✓
					Procedure	Performingprovider	0	0.00	✓
						Provider	0	0.00	✓
					Provider Taxonomies	Provider	0	0.00	✓

Table Name: (All)

Field: (Multiple values)

Data Model: HCSRN - VDW

Result: (All)

Context: (All)

Category: (All)

Check Name
- □ (All)
- □ Does_Field_...
- □ Has_Correct...
- □ Has_Correct...
- ☑ Is_Foreign_...
- □ Is_Required

Execution Time: 8/13/2025 4:40:51 ...

Challenges

- Applying the Table Driven Design
- Extending OMOP DQ Dashboard
- Improving Performance

Questions