CDPH California Department of Public Health

accenture

# Mastering Data at California Dept of Health

## An AI Enabled solution to Data Management and Person Identity

***Michael Powell**, Chief Registry & Assessment Section*
*Immunization Branch, California Department of Public Health*

***Ajali Sen**, Data and Analytics Leader Accenture*

*April 2025*

# Agenda

- A Story of Data
- Data Management
- Master Person Index

# A Story of Data

Lifecycle transformations and how to build confidence

# A Story of Data – Lifecycle Transformations

The data collected by CDPH undergoes multiple transformations throughout its lifecycle. Efficient Data Management and Person Identification processes are fundamental to maintaining and enhancing Data Quality.

At California Dept of Public Health, **individual immunization records are collected** through the California Immunization Registry (CAIR). These records help ensure individuals receive recommended vaccines on schedule, support providers in tracking patient history and aid public effort in managing outbreaks and improve vaccination coverage

The immunization insights provided to stakeholders, which include the Immunization Branch, Local Health Jurisdictions and CA residents, are the result of data that has **undergone numerous transformations** and modifications throughout its lifecycle

A patient may have used a **different information**, such as name or address than before, visited a new clinic, interacted with a new technician, or been logged into a new software system that generated an HL7 message.

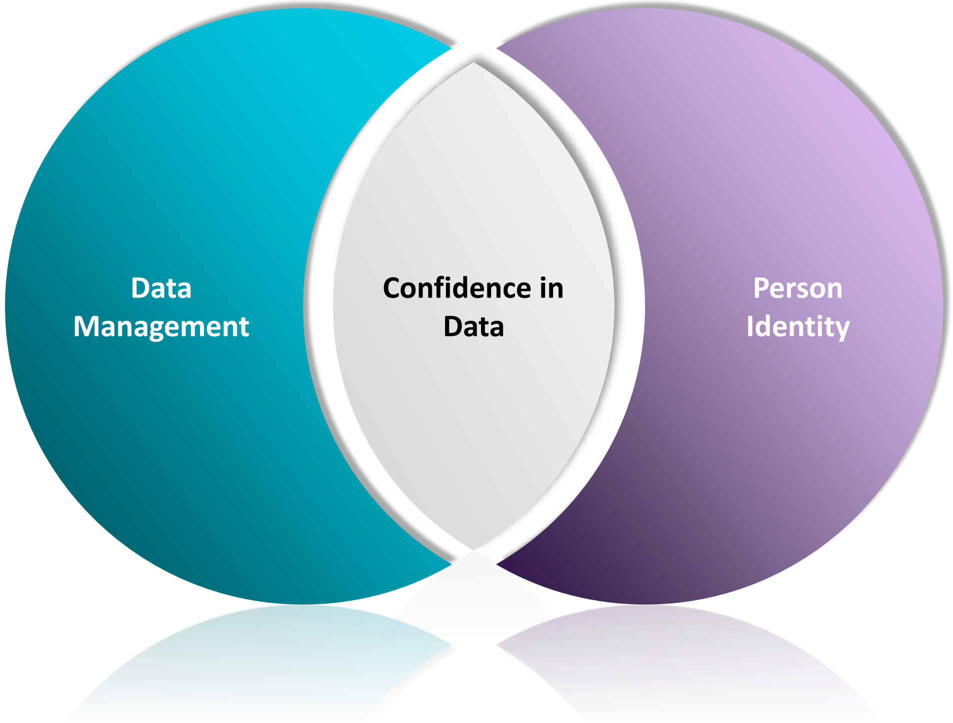
This message then passed through **multiple systems** before reaching an Immunization Information System (IIS), where it was broken down into various components stored across different tables, and later **reassembled to create a report**.

# A Story of Data – Building Confidence

To ensure confidence in data, it is crucial to have a clear understanding of what data has been collected and accurately identify who the data refers to, despite the complexities introduced during transformation processes

**Data Management Tools**

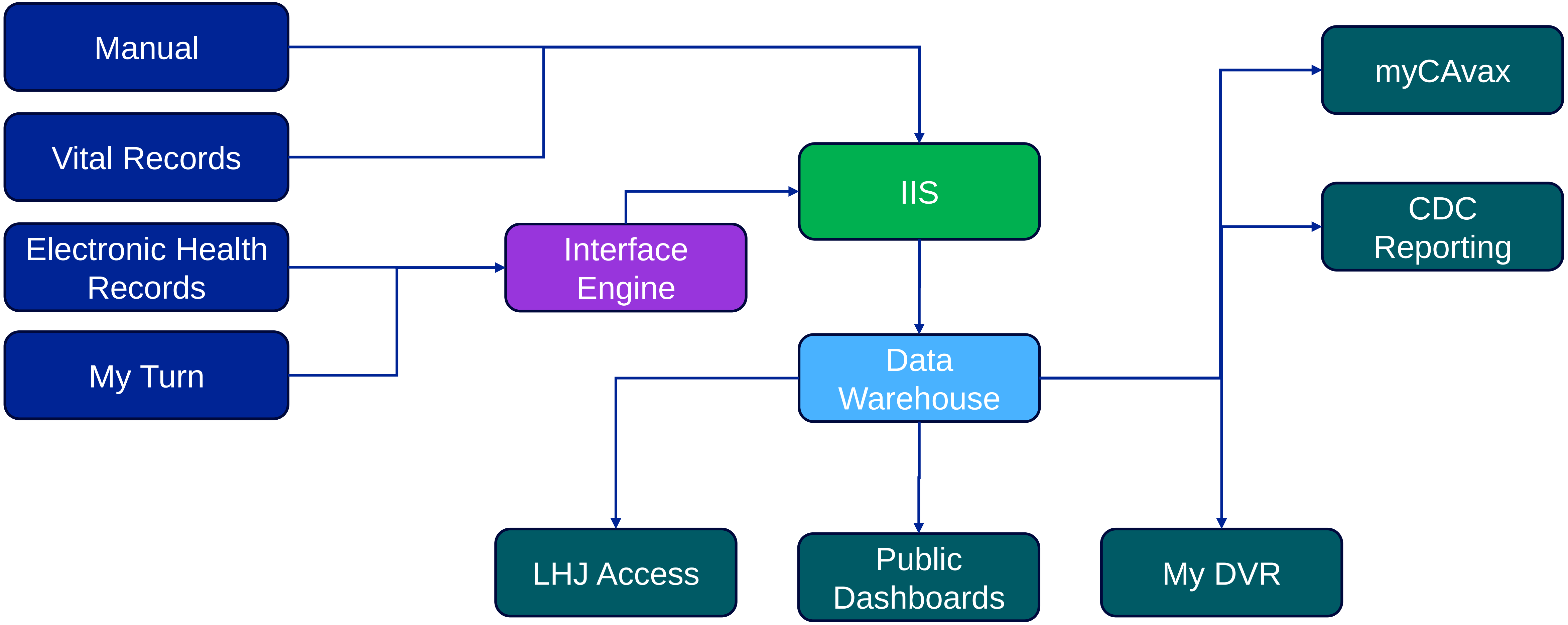
Master Data Management (MDM), Data Classification and Data Lineage promote accurate, secure and accessible immunization records



**Mater Person Index (MPI)**

An MPI uniquely identifies individuals across systems, even when direct identifiers like social security numbers are unavailable, promoting accurate mapping of immunization records to the correct individuals

# A Story of Data – Lifecycle Transformations



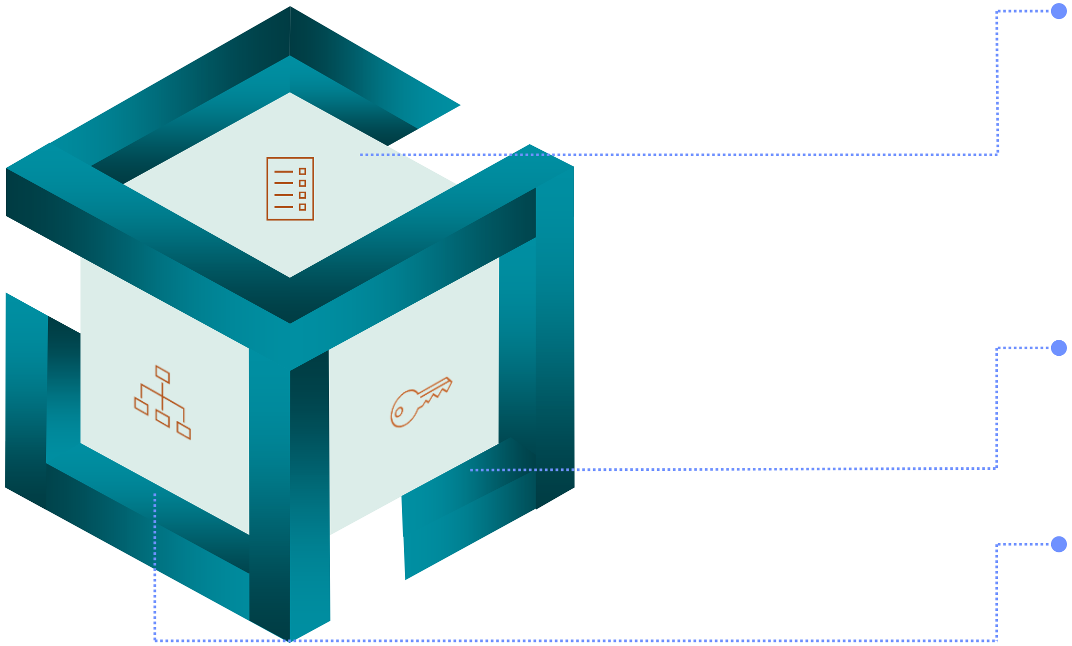
CDPH

# Data Management

The What, Where and Who of Data

# Data Management -Capabilities

The California Dept of Public Health employs Microsoft Purview’s for managing immunization data in a secure and transparent way that promotes precision and compliance

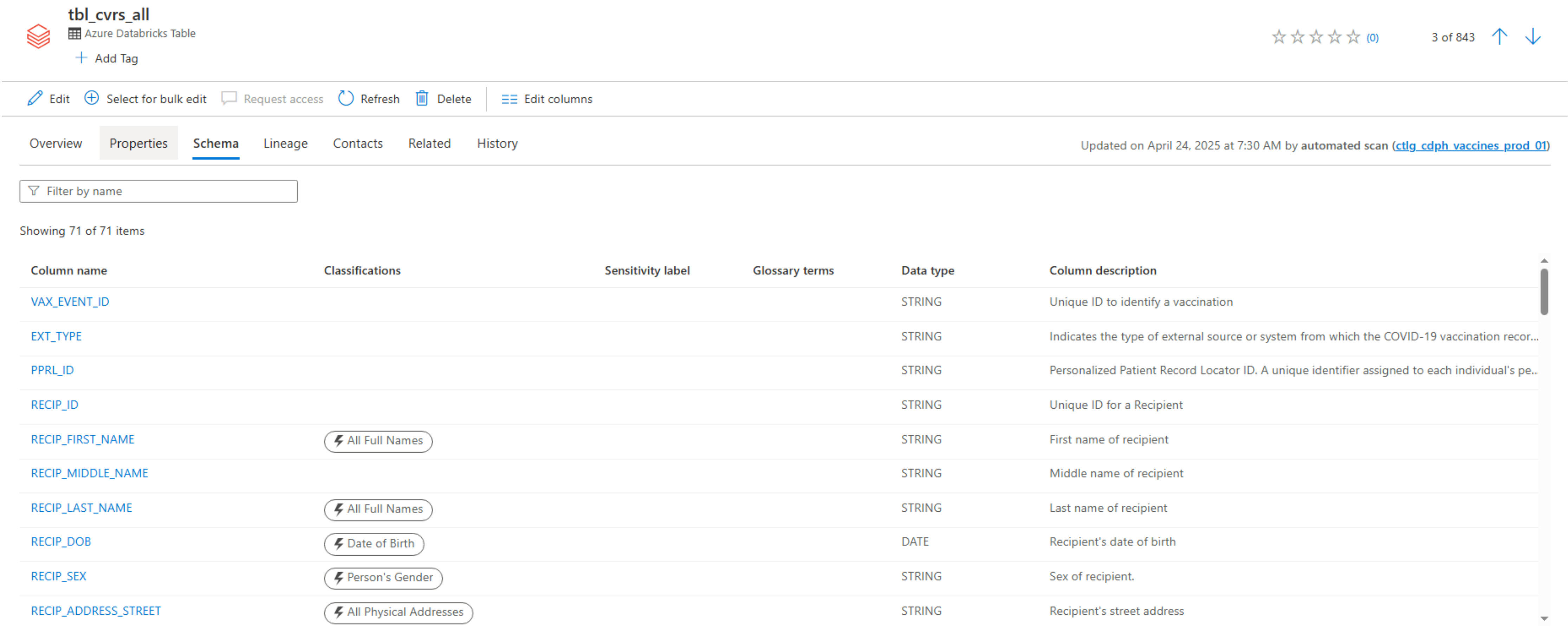


- **Data Discovery & Cataloging**: Auto-scans IRIS, DVR and Public Reporting systems to catalog all data assets
- **Glossary & Metadata**: Common definitions to ensure consistency across immunization branch stakeholders

**Data Classification**
Tags sensitive data and enforces access rules

**Data Lineage**
Visual maps from source to reporting layer for full transparency

# Data Catalog– Example



tbl_cvrs_all

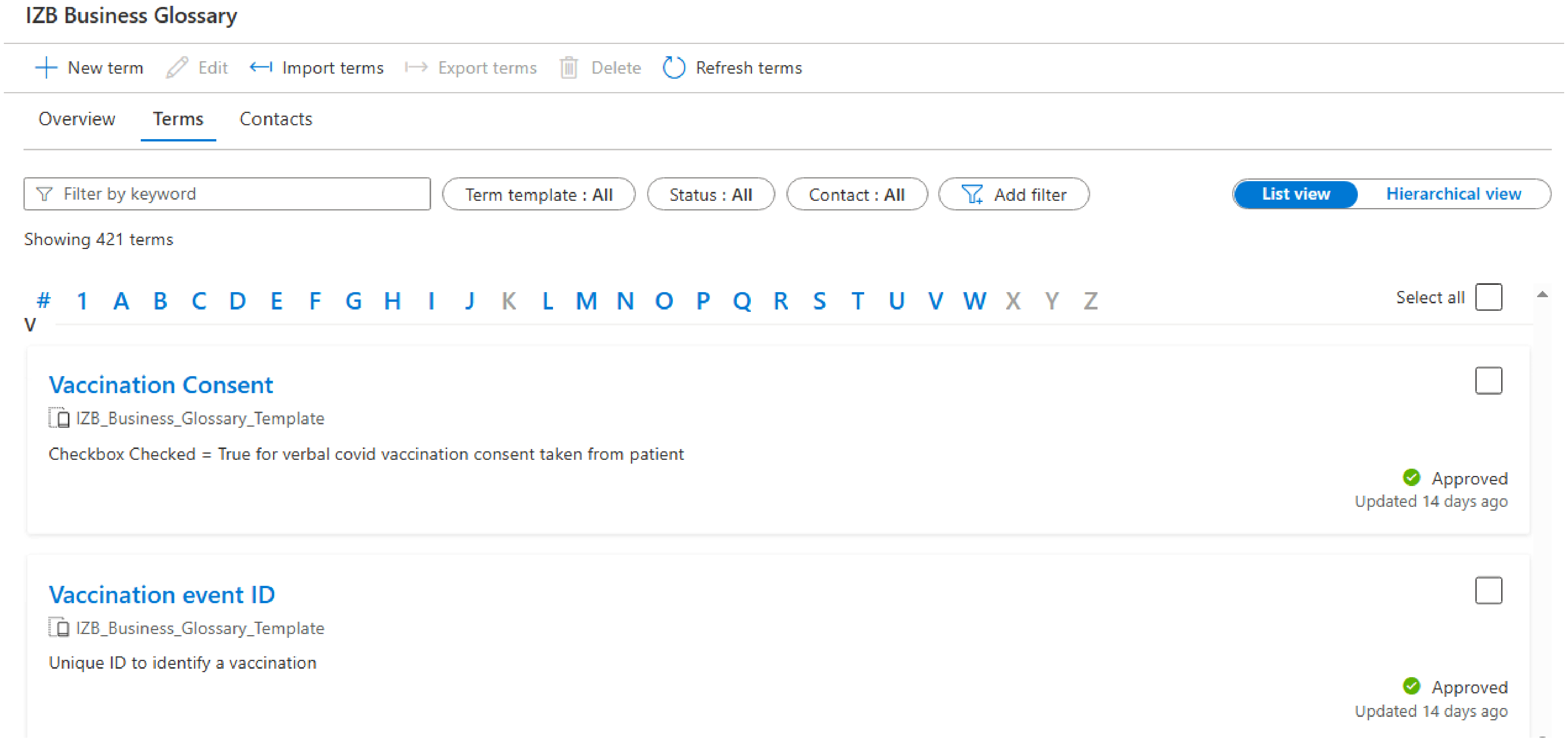
Azure Databricks Table

Updated on April 24, 2025 at 7:30 AM by automated scan (ctlg_cdph_vaccines_prod_01)

Showing 71 of 71 items

| Column name | Classifications | Sensitivity label | Glossary terms | Data type | Column description |
|---|---|---|---|---|---|
| VAX_EVENT_ID | | | | STRING | Unique ID to identify a vaccination |
| EXT_TYPE | | | | STRING | Indicates the type of external source or system from which the COVID-19 vaccination recor... |
| PPRL_ID | | | | STRING | Personalized Patient Record Locator ID. A unique identifier assigned to each individual's pe.. |
| RECIP_ID | | | | STRING | Unique ID for a Recipient |
| RECIP_FIRST_NAME | All Full Names | | | STRING | First name of recipient |
| RECIP_MIDDLE_NAME | | | | STRING | Middle name of recipient |
| RECIP_LAST_NAME | All Full Names | | | STRING | Last name of recipient |
| RECIP_DOB | Date of Birth | | | DATE | Recipient's date of birth |
| RECIP_SEX | Person's Gender | | | STRING | Sex of recipient. |
| RECIP_ADDRESS_STREET | All Physical Addresses | | | STRING | Recipient's street address |

# Business Glossary– Example



Showing 421 terms

V

**Vaccination Consent** □

IZB_Business_Glossary_Template

Checkbox Checked = True for verbal covid vaccination consent taken from patient

✓ Approved
Updated 14 days ago

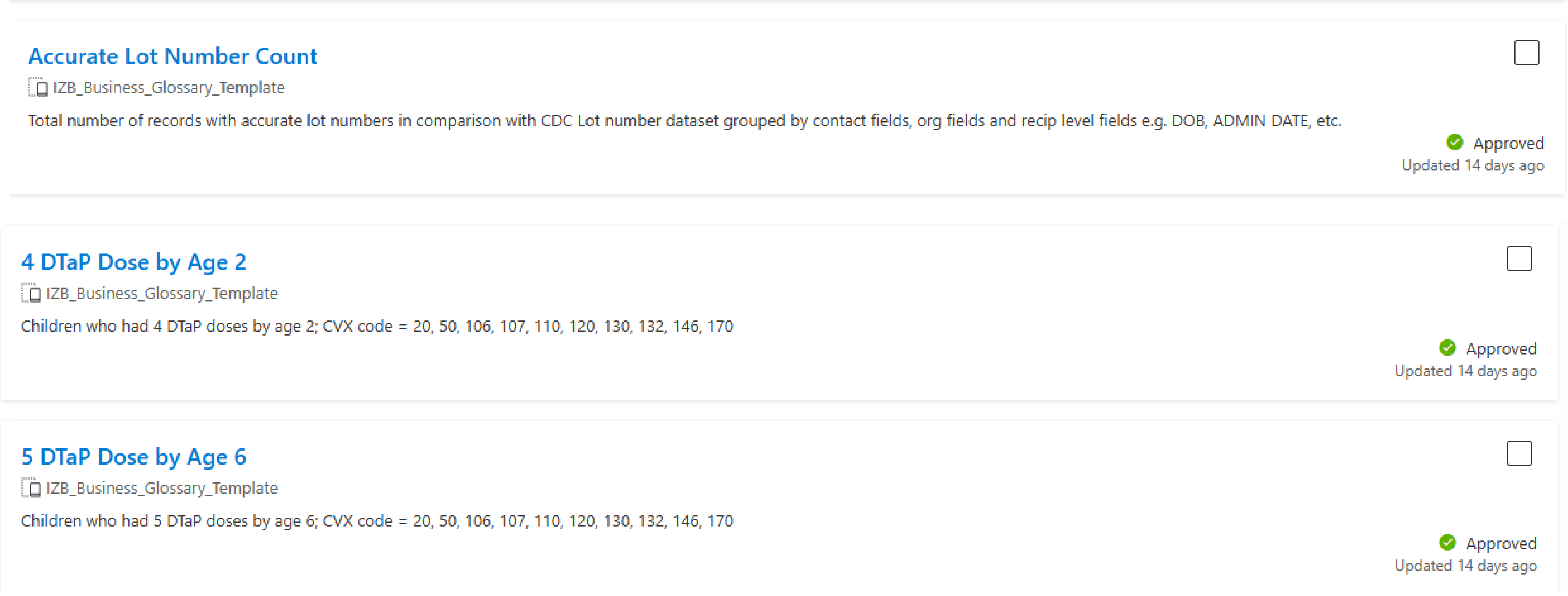
**Vaccination event ID** □

IZB_Business_Glossary_Template

Unique ID to identify a vaccination

✓ Approved
Updated 14 days ago

# Business Glossary– Example



### Accurate Lot Number Count

□

IZB_Business_Glossary_Template

Total number of records with accurate lot numbers in comparison with CDC Lot number dataset grouped by contact fields, org fields and recip level fields e.g. DOB, ADMIN DATE, etc.

✓ Approved
Updated 14 days ago

### 4 DTaP Dose by Age 2

□

IZB_Business_Glossary_Template

Children who had 4 DTaP doses by age 2; CVX code = 20, 50, 106, 107, 110, 120, 130, 132, 146, 170

✓ Approved
Updated 14 days ago

### 5 DTaP Dose by Age 6

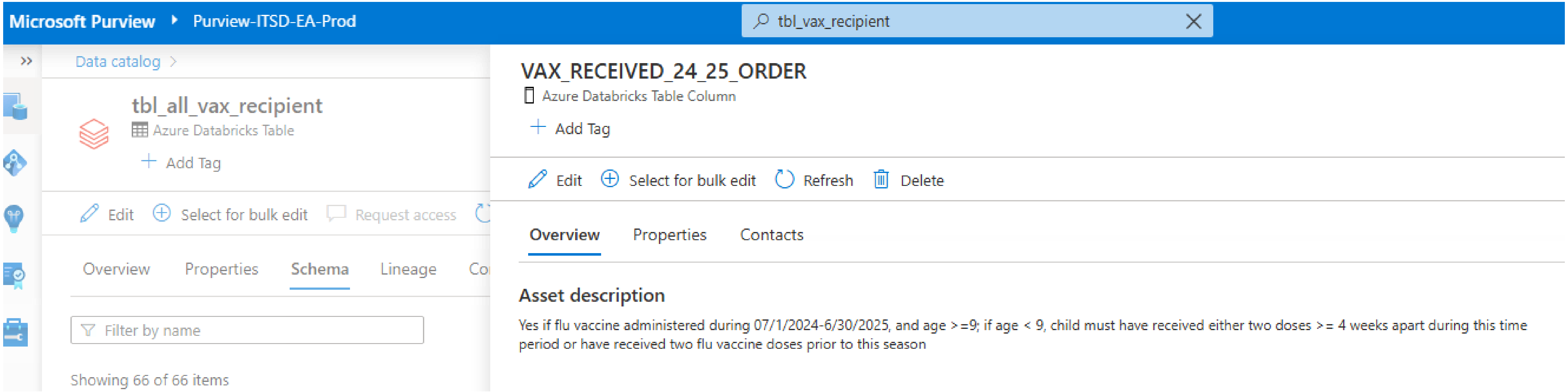
□

IZB_Business_Glossary_Template

Children who had 5 DTaP doses by age 6; CVX code = 20, 50, 106, 107, 110, 120, 130, 132, 146, 170

✓ Approved
Updated 14 days ago

# Metadata Definition– Example



VAX_RECEIVED_24_25_ORDER

Azure Databricks Table Column

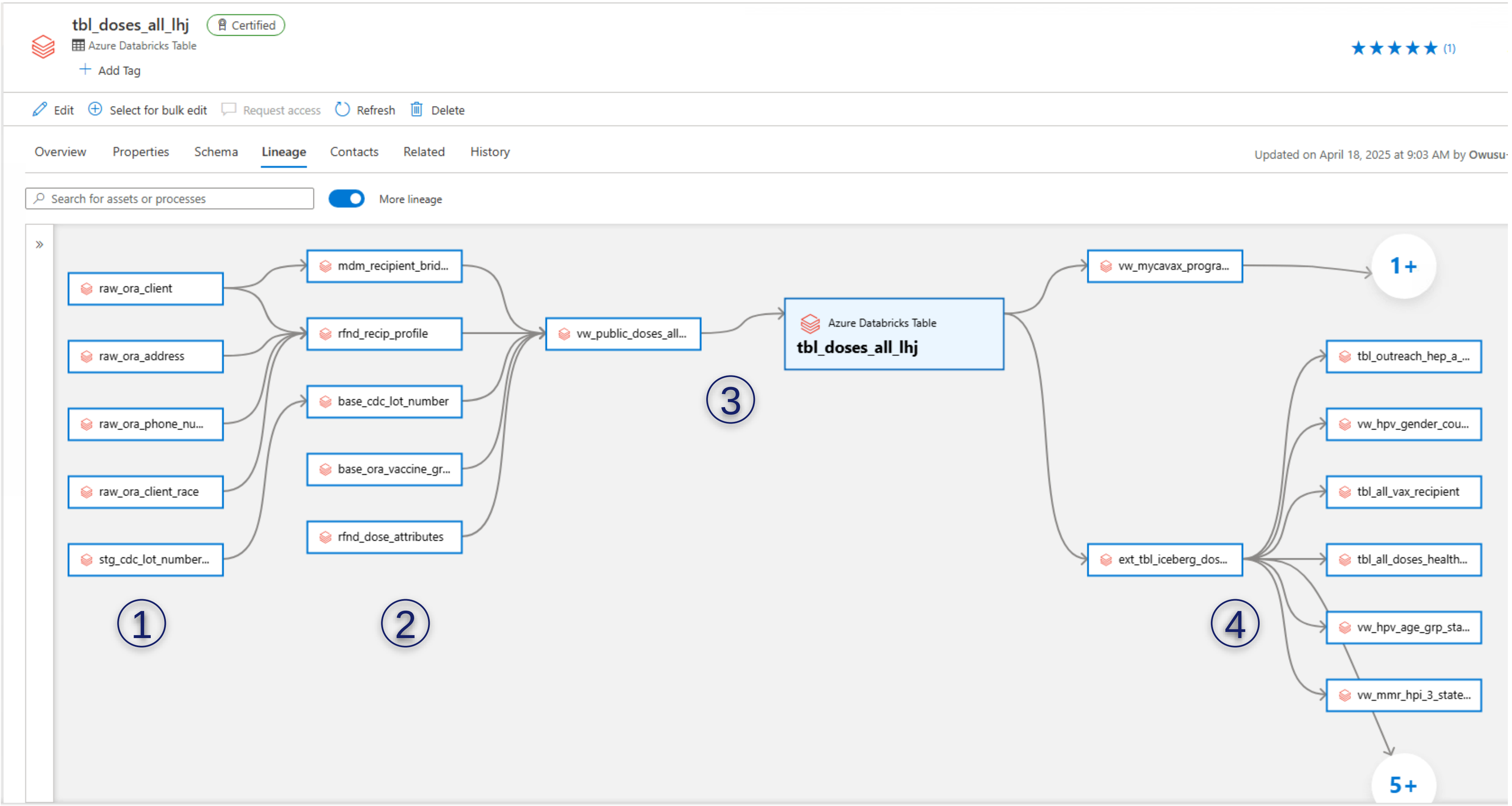
Asset description

Yes if flu vaccine administered during 07/1/2024-6/30/2025, and age >=9; if age < 9, child must have received either two doses >= 4 weeks apart during this time period or have received two flu vaccine doses prior to this season

CDPH

accenture

# Data Lineage – Example



tbl_doses_all_lhj

Certified

Azure Databricks Table

Updated on April 18, 2025 at 9:03 AM by Owusu-

1 Source Data

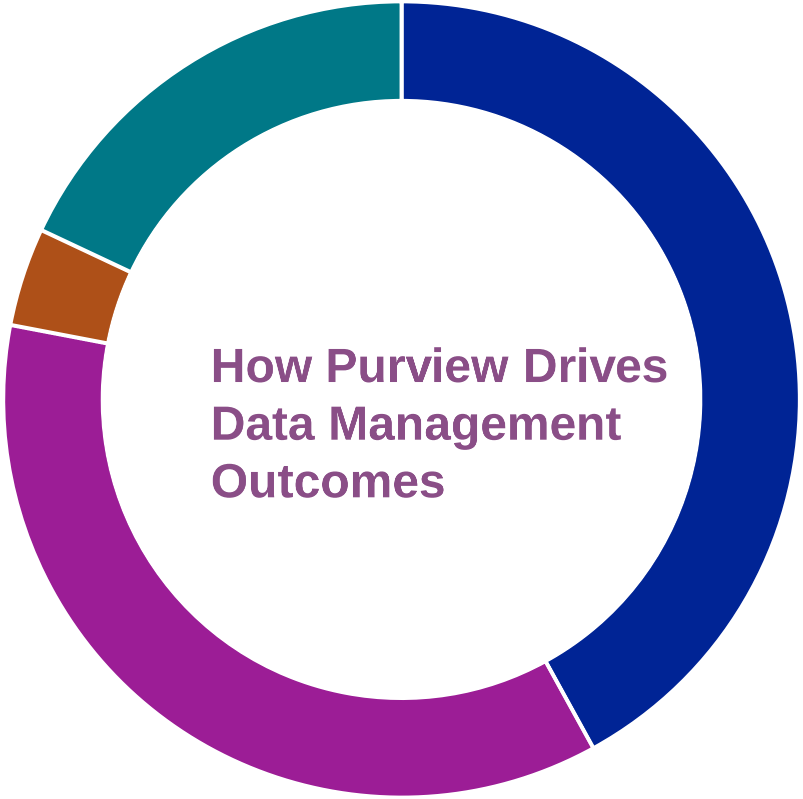
2 Bronze Layer

3 Silver/Cleansed Layer

4 Gold/Modeled Layer

# Data Management– Impact

Accelerating data management efforts through faster insights, better data quality and improved collaboration



How Purview Drives Data Management Outcomes

| | |
|---|---|
| Faster Dataset Search | 42% |
| Fewer Reporting Errors | 36% |
| Reduction in Report Duplication | 18% |
| Faster Vaccine Eligibility Identification | 4% |

# **Data Management** – Value Delivered

The use of Microsoft Purview has many benefits for the California Dept of Public Health, the most notable:

**Improved Trust in Data:** Stakeholders can validate how data was processed

**Faster Root Cause Analysis:** Trace errors to exact data source

**Accurate Reporting:** Consistent definitions and audit trails across reports

**Data Governance & Compliance:** Ensures sensitive data is controlled and traceable

# Master Person Index

All versions, same individual

# **MPI** – What is it?

A Master Person Index (MPI) is used to identify matching individuals based on current and historical personal, demographic, and vaccine information

## Immunization MPI

CDPH has developed an **AI enabled MPI solution** to match immunization records. This solution:

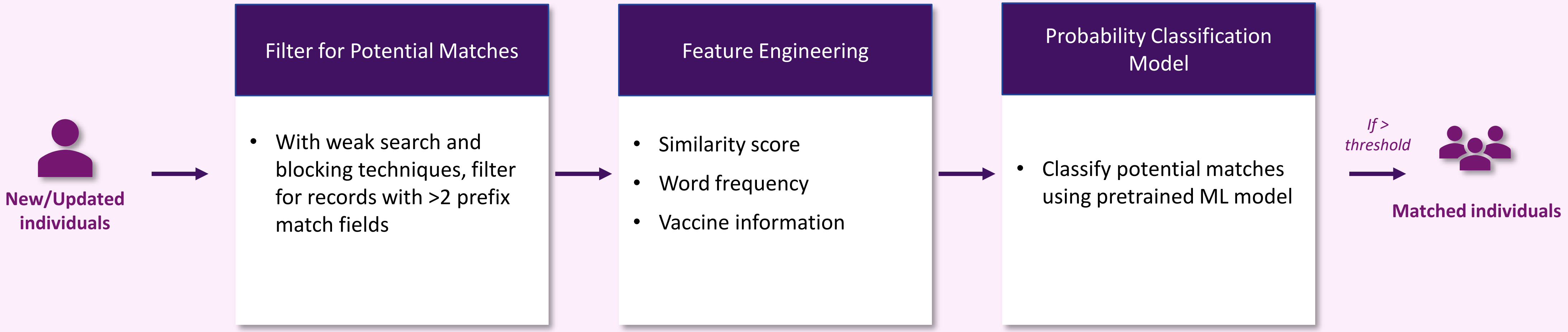
- Employes robust **blocking technique** to efficiently identify potential matching individuals on a set of loose criteria
- Utilizes **customized features** to combat data quality issues such as typo and abbreviations
- Leverages a pre-trained a machine learning model to output **probability scores**

## Key Features

- Enables storing and tracking individuals' entire immunization data **across the lifetime**
- Leverages **customized features** to tackle data quality issues
- Outputs **probability match score** to evaluate the likelihood of the two individuals are matched

# MPI – How does it work?

There are three main steps in determining a match between individuals



New/Updated individuals

### Filter for Potential Matches

- With weak search and blocking techniques, filter for records with >2 prefix match fields

### Feature Engineering

- Similarity score
- Word frequency
- Vaccine information

### Probability Classification Model

- Classify potential matches using pretrained ML model

*If > threshold*

Matched individuals

Frequency: Hourly (production environment)

# MPI –Data Sources & Features

## Personal and Demographic Features

- First name
- Middle name
- Last name
- Birth date
- Phone number
- Email address
- Street address and zip code
- Sex code
- Responsible person names
- Mother’s first and maiden last name
- Multiple birth count

## Customized Features

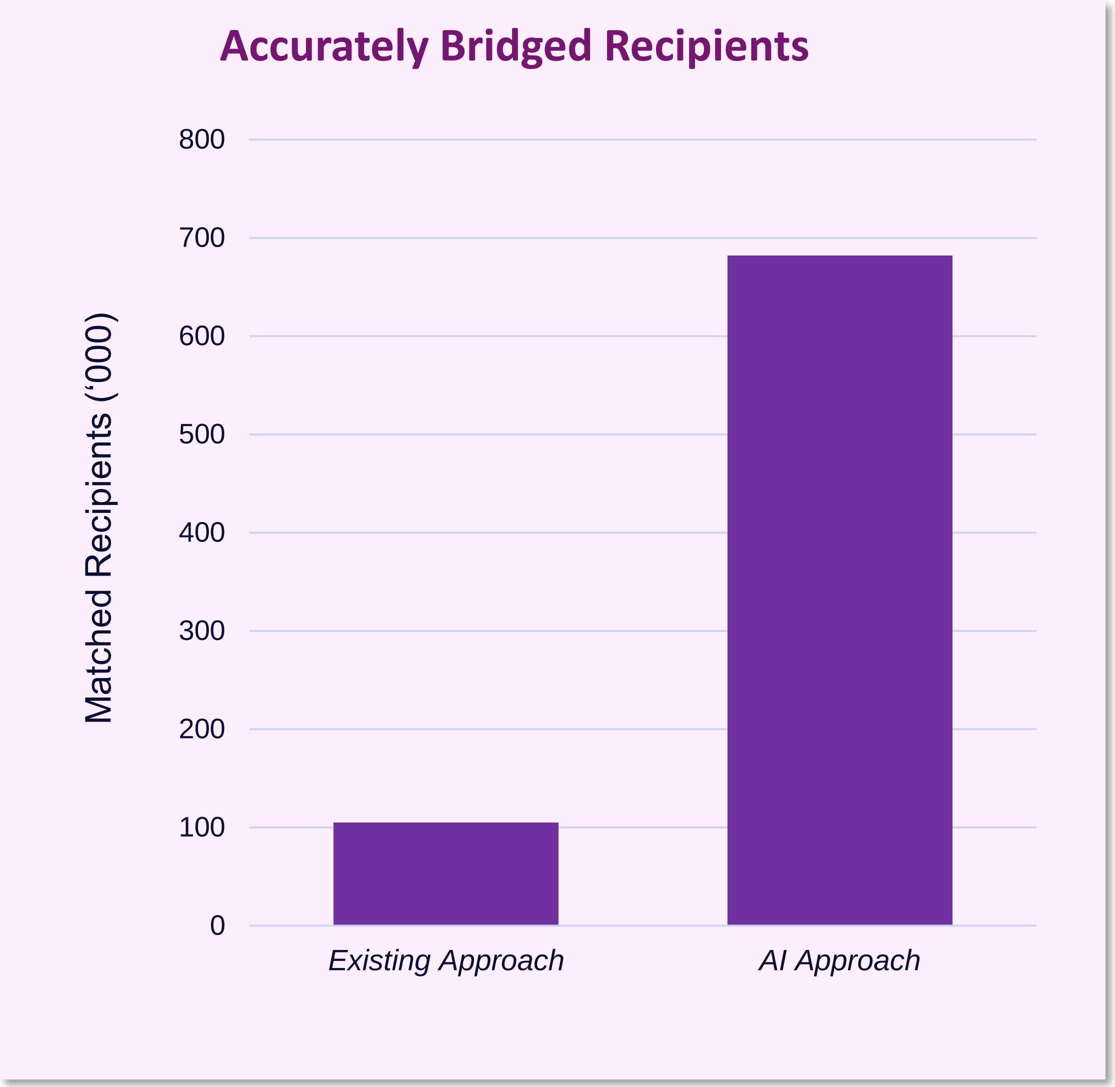
- Frequency of individuals’ first name, phone number and street address
- Weighted Soundex (phonetic algorithm) for individuals’ first and last name
- Calculated distance between addresses
- Baby Name Placeholder

## Vaccination Features

- Vaccine lot number and admin date
- MRN chart number
- Provider ID

# MPI – Impact

- **Automated** the record merging process
- Rapid identification of records with **data quality issues**
- Generated **probability scores**, that can be used to flag uncertain pairs for manual review
- Supported identification and tracking of **family relationships**



# **Search API** – MPI in Real-Time

Search API is a scalable, real-time system that efficiently searches and matches individuals PII in a large database leveraging the MPI solution

## Search API

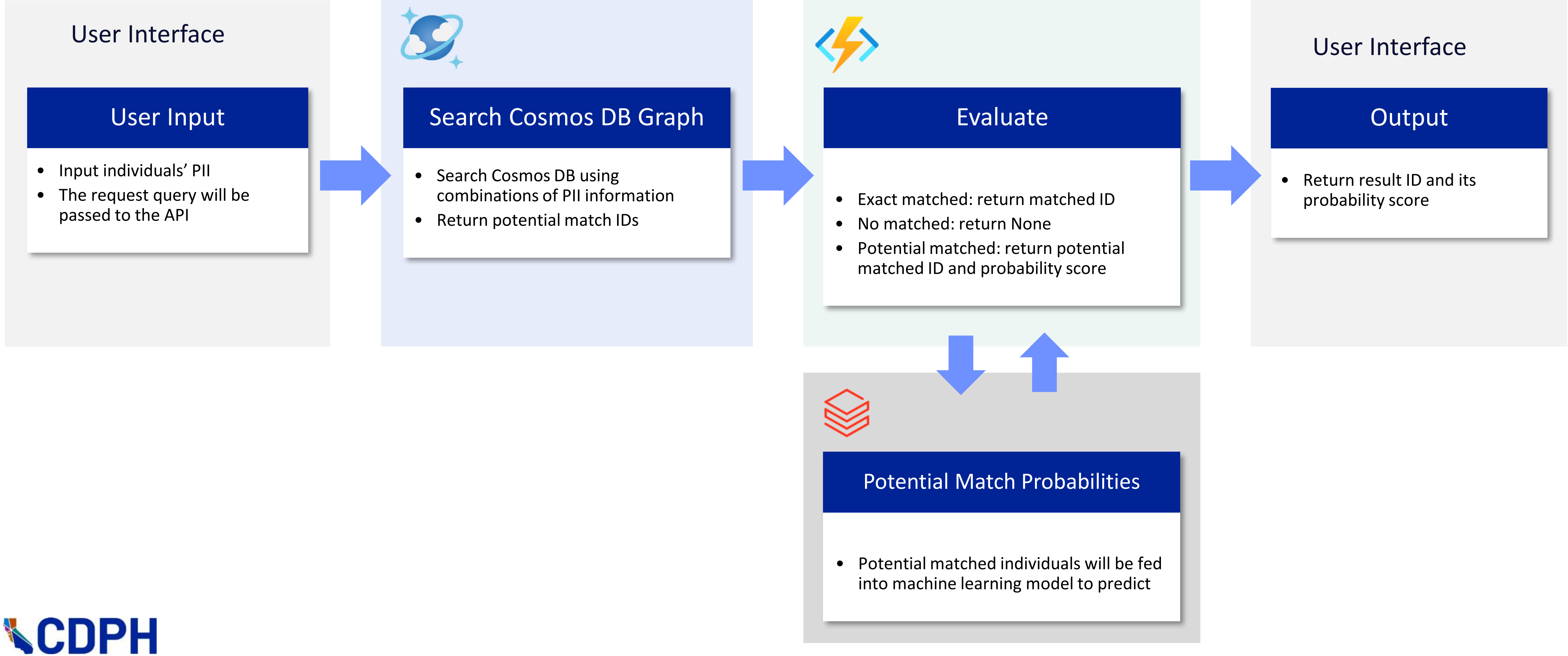
CDPH has developed a **Search API solution** to leverage the MPI matching capabilities in real-time. This solution:

- Has an **interface** that allows end users to input individuals data and search the database
- Uses a **graph database** to store all individuals PII, optimizing the search process for fast, accurate retrieval of individuals' information
- Leverages **machine learning model** in the MPI solution to classify potentially matched individuals

## Key Features

- User-friendly interface that **simplifies access for end users**
- Fast search capability, enabling **near real-time response** for exact matches and no matches
- Return **probability score** to evaluate the likelihood of the two individuals are matched

# Search API – Solution Steps



User Interface

**User Input**

- Input individuals’ PII
- The request query will be passed to the API

**Search Cosmos DB Graph**

- Search Cosmos DB using combinations of PII information
- Return potential match IDs

**Evaluate**

- Exact matched: return matched ID
- No matched: return None
- Potential matched: return potential matched ID and probability score

**Potential Match Probabilities**

- Potential matched individuals will be fed into machine learning model to predict

User Interface

**Output**

- Return result ID and its probability score

# Immunization data, resolved

By integrating modern, accessible, and cost-effective tools early in the data pipeline, CDPH can significantly enhance data reliability and security. These methodologies not only ensure accurate immunization reporting but also foster greater confidence in the data shared with stakeholders, paving the way for more effective public health decisions

CDPH

California Department of Public Health