Clinical Narratives Entity Relationship Diagram

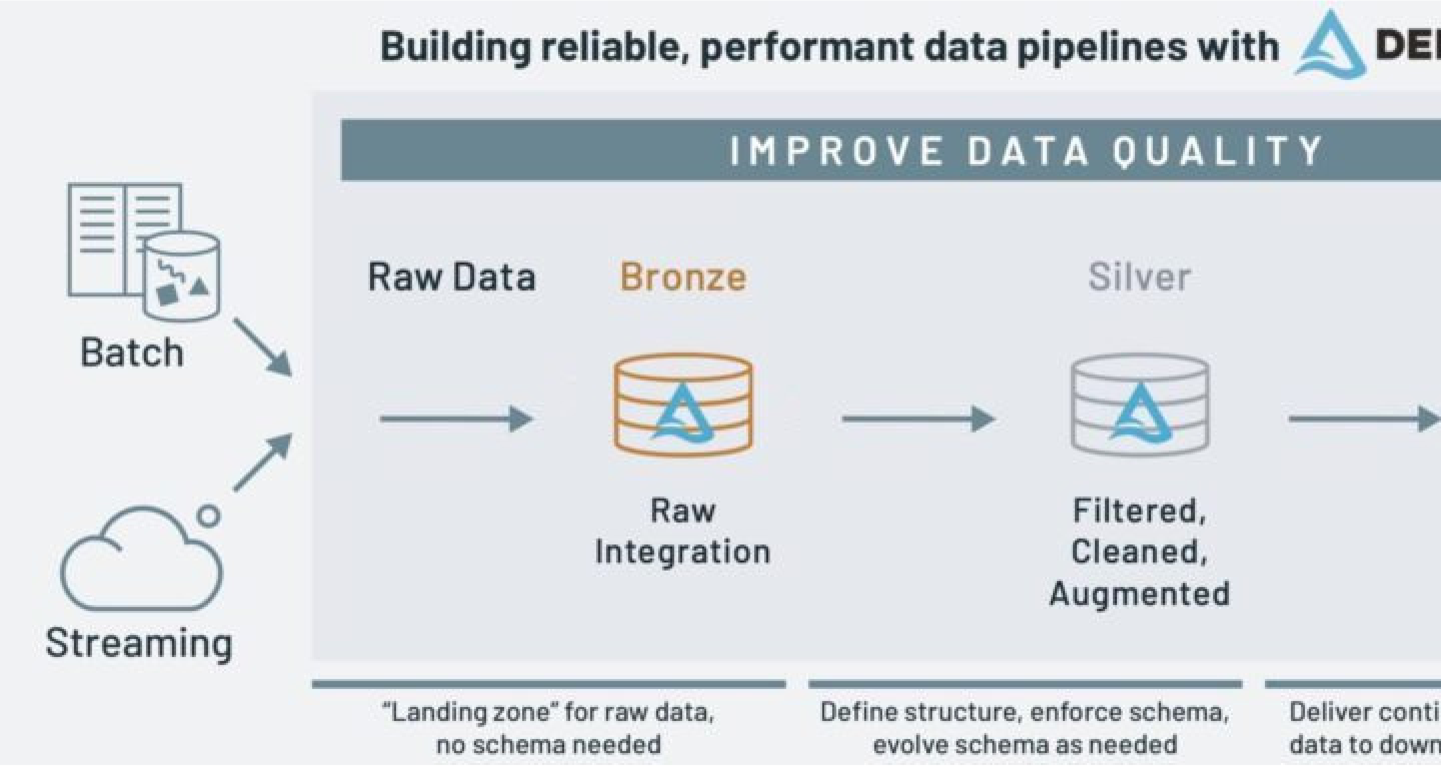
1. Introduction

The purpose of the Entity Relationship (ER) diagram for the Clinical Narratives project is to provide a structured visual representation that aids in several critical aspects of data management. The ER diagram details all tables and illustrates how they are accessed for read and write operations by various notebooks. This level of documentation is invaluable for onboarding new team members or serving as a reference during system maintenance and updates. Furthermore, the diagram demonstrates how a global temporary table is constructed using inputs from JSON files, variables set at the Databricks Cluster level, and information schema.

By mapping out how data moves through the system—from its initial input to processing and finally to output—the ER diagram facilitates visualization of complex data flows. It clearly shows how all notebooks interact with the tables for reading and writing purposes, enhancing understanding of data interactions within the project.

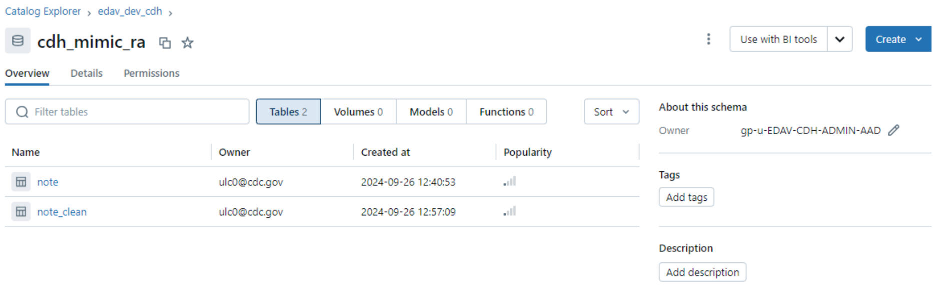
2. Data Storage and Tables

All MIMIC data is stored in EDAV Databricks Dev Workspace and is the Dev "Analyze" Zones of the CDC Data Hub (CDH) Data Lake Architecture using the "Medallion Architecture Zones"


Building reliable, performant data pipelines with DE
IMPROVE DATA QUALITY
Batch
Streaming
Raw Data
Bronze
Silver
Raw Integration
Filtered, Cleaned, Augmented
"Landing zone" for raw data, no schema needed
Define structure, enforce schema, evolve schema as needed
Deliver conti
data to down

The CDH_MIMIC is the "Enriched Zone" and tables are populated by CDH Engineering from source.

CDH_MIMIC_RA is a "Curated Zone" for MIMIC, and these tables are populated by CDH Engineering using parameterized workflows accepted by Analyst. This schema is writable only by CDH Engineering. These tables are "bronze" medallion tables, the "raw" data format. Every project will have a different "raw" format, which must be preprocessed into the modified version of the OMOP CDM format for Clinical Notes, written to the Silver "Curated" Zone by Engineering upon acceptance of the ETL process by Analysts.



cdh_mimic_ra

Name	Owner	Created at	Popularity
note	ulc0@cdc.gov	2024-09-26 12:40:53	
note_clean	ulc0@cdc.gov	2024-09-26 12:57:09	

About this schema

Owner gp-u-EDAV-CDH-ADMIN-AAD

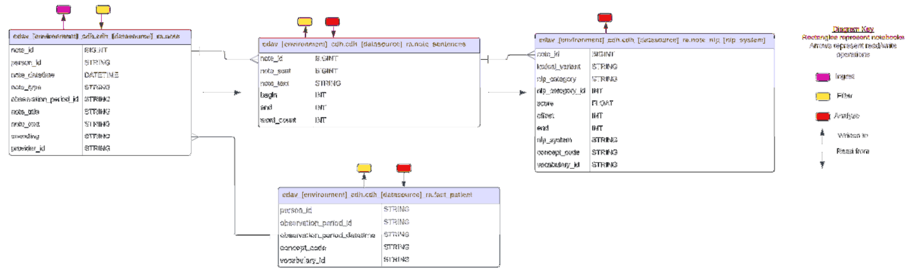
Tags

Description

CDH_MIMIC_EXPLORATORY , the "Exploratory Zone", writeable by analysts and engineering, is the staging/work zone for MIMIC and area for ***note_clean*** quality control with staging workflow, as well as configuration tests of NLP output. Upon acceptance, CDH

Engineering may refactor code to current CDH engineering specs, and will run the any accepted workflow with output to the silver zone, and the staging tables should be purged.

3. Entity Relationship Diagram



edav_[environment]_cdh.cdh_[datasource]_ra.note	
note_id	BIGINT
person_id	STRING
note_datetime	DATETIME
note_type	STRING
observation_period_id	STRING
note_title	STRING
note_text	STRING
encoding	STRING
provider_id	STRING

edav_[environment]_cdh.cdh_[datasource]_ra.note_sentences	
note_id	BIGINT
note_sent	BIGINT
note_text	STRING
begin	INT
end	INT
word_count	INT

edav_[environment]_cdh.cdh_[datasource]_ra.note_nlp_[nlp_system]	
note_id	BIGINT
lexical_variant	STRING
nlp_category	STRING
nlp_category_id	INT
score	FLOAT
offset	INT
end	INT
nlp_system	STRING
concept_code	STRING
vocabulary_id	STRING

edav_[environment]_cdh.cdh_[datasource]_ra.fact_patient	
person_id	STRING
observation_period_id	STRING
observation_period_datetime	STRING
concept_code	STRING
vocabulary_id	STRING

Diagram Key
Rectangles represent notebooks
Arrows represent read/write operations
Ingest
Filter
Analyze
Written to
Read from

4. Table Descriptions

Below are descriptions of the 5 tables (4 required, 1 optional) that comprise the minimum viable product (MVP) for the Clinical Narratives Project.
Some changes or additions to these tables may be warranted depending on the specific dataset used, and the research question that is being investigated.

4.1 Table Name: edav_[environment]_cdh.cdh_[datasource]_ra.note (Catalog Explorer)

Enriched Zone Table transformed to OMOP CDM NOTE data model

Scope: All Notes for datasource

Granularity: One Record per note_id

Linkage: person_id, provider_id link back to other datasource records

Column Name	Data Type	Description
note_id	BIGINT	Unique Indentifier for Each Note

person_id	STRING	Patient ID for each note
note_datetime	DATETIME	Note Date/DateTime
note_type	STRING	Source Specific Note "Type" (optional)
observation_period_id	STRING	OMOP CDM Observation Period, may be inpatient or outpatient visit. May be linked to Fact Tables (Optional)
note_title	STRING	TItle of Note (Optional)
note_text	STRING	Text Version of Note
encoding	STRING	Source provided encoding of note (Optional)
provider_id	STRING	Provider ID for each note (Optional)

4.2 Table Name: edav_[environment]_cdh.cdh_[datasource]_ra.note_sentences (Catalog Explorer)

CDH Defined intermediate table for NOTE data requiring text clean. This table should populate clean_text. This table should be used with an inner join to NOTE table for NLP processing.

Scope: All NOTE records

Granularity: NOTE level, one note_id per table

Linkage: NOTE, downstream tables by note_id

Column Name	Data Type	Description
note_id	BIGINT	Unique Note Id
note_sent	BIGINT	Ordered Sentence Number
note_text	STRING	Cleaned Sentence Text
begin	INT	starting position of sentence
end	INT	ending position of sentence
word_count	INT	word_count of sentence

4.3 Table Name: edav_[environment]_cdh.cdh_[datasource]_ra.note_nlp_[nlp_system]

THIS IS THE DELIVERABLE TABLE FOR THE PROJECT OMOP CDM NOTE_NLP data model for NLP results across methods. Individual Tables/Analyst approved results will be added to this table by CDH_Engineering. This table schema may be extended upon review for NLP systems. One of nlp_category/nlp_category_id must be populated **additions to CDM definition

Scope: All note_id

Granuarity: lexical_variant (Named Entity)

Linkage: note_id back to NOTE table

Column Name	Data Type	Description
note_id	BIGINT	Unique Note Id
lexical_variant	STRING	Entity
nlp_category	STRING	NER/NLP Output**
nlp_category_id	INT	NLP ID for nlp_category**
score	FLOAT	NLP Method Score (Optional)
offset	INT	Entity offset within TEXT
end	INT	endpoint of Entity (Optional)
nlp_system	STRING	Name of NLP System used e.g. SciSpacyLG_UMLS
concept_code	STRING	assigned concept code (Optional)
vocabulary_id	STRING	concept code vocabulary (Optional unless concept_code is not NULL)**

4.4 Table Name: edav_[environment]_cdh.cdh_[datasource]_ra.fact_patient

This table is used to record the visit diagnosis for specific person_id, it assists in the calculations of OKR for this project.

Scope: All person_id diagnoses where person_id is patient id

Granularity: visit level, one entry per date and diagnosis or procedure code

Linkage: person_id

Column Name	Data Type	Description
person_id	STRING	Person ID
observation_period_id	STRING	Observation period ID
observation_period_datetime	STRING	observation_period_datetime
concept_code	STRING	Diagnosis or procedure code
vocabulary_id	STRING	Type of coding system

Optional tables

4.5 Table Name: edav_[environment]_cdh.cdh_[datasource]_exploratory.note_clean (Catalog Explorer)

CDH Defined intermediate table for NOTE data requiring text clean. This table should populate clean_text. This table may be used with an inner join to NOTE table for NLP processing. This table is for QA purposes only.

Scope: All NOTE records

Granularity: NOTE level, one note_id per table

Linkage: NOTE, downstream tables by note_id

Column Name	Data Type	Description
note_id	BIGINT	Unique Note Id
clean_text	STRING	Cleaned Text
encoding	STRING	encoding (optional)
word_count	INT	word_count