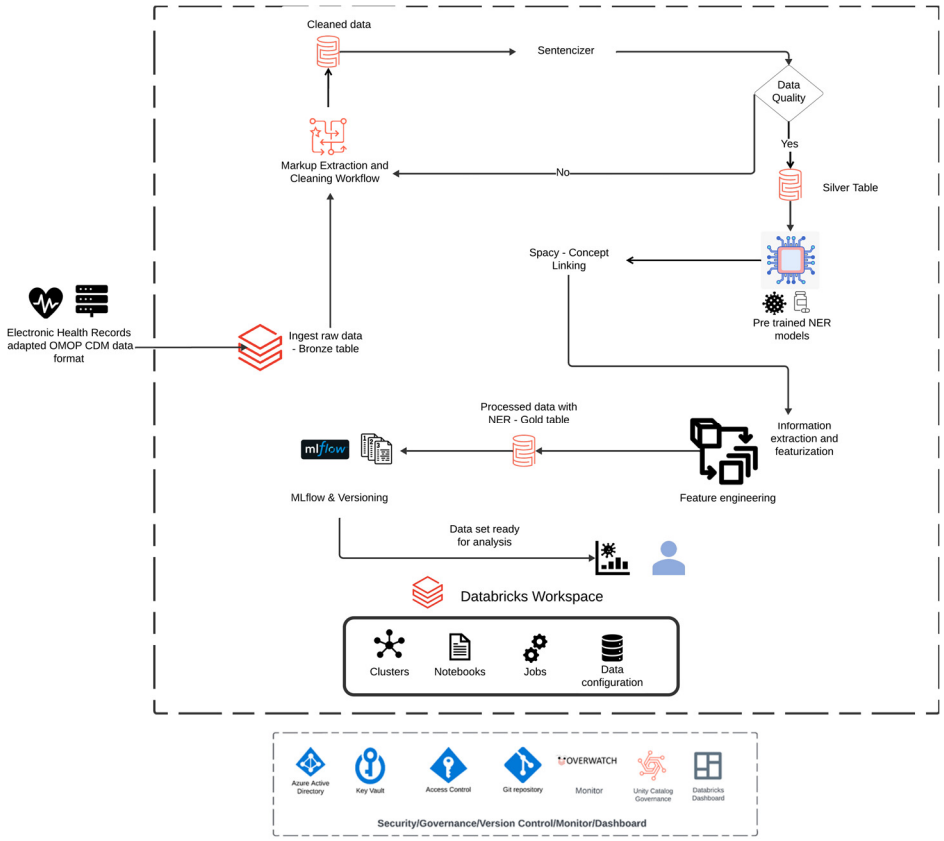
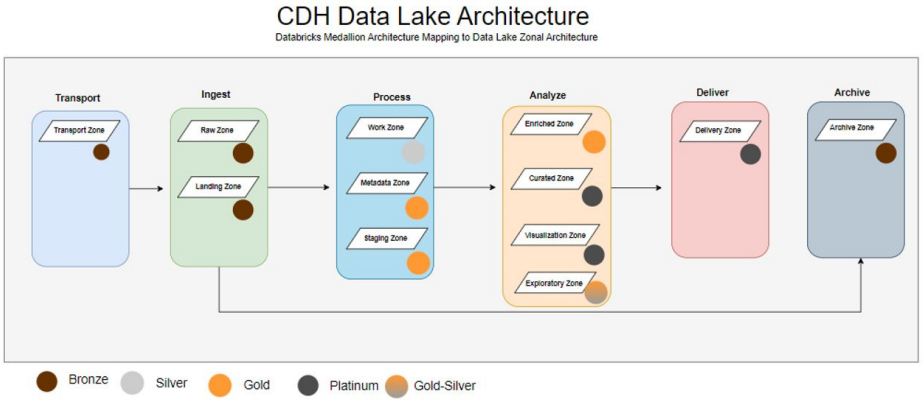
Clinical Narratives Architecture Overview

Application Architecture

The application architecture diagram provides a visual representation of the data lifecycle for the featurization process of the clinical narratives project.





CDC Data Hub Data Architecture - Clinical Narratives uses the "Analyze" Zone

https://lucidgov.app/lucidchart/fae79244-e47d-42d6-8134-d1de6ee5f118/edit?viewport_loc=299%2C829%2C3906%2C3084%2C0_0&invitationId=inv_6695385e-a88c-4505-909c-748b287a94e7

Here's a step-by-step explanation of the process

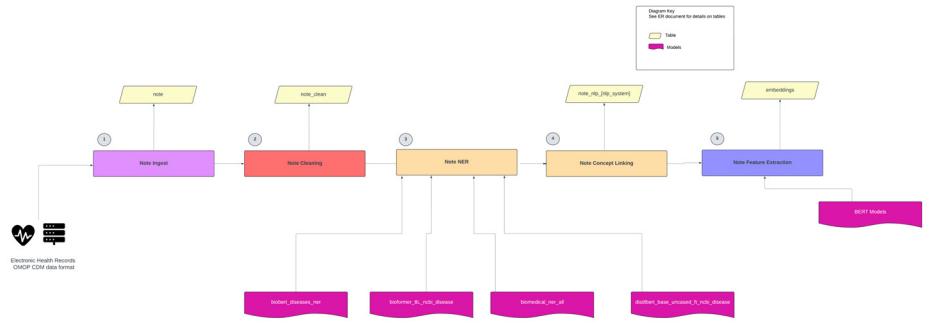
1. Adapted **OMOP CDM data format:** Raw data is considered bronzed medallion*, many EHR do not necessarily come in a standard format. Therefore, identifying relevant tables from the original data set and transforming them into the desired schema to assist standardization approaches is a preparation step.
2. **Markup extraction and cleaning:** To use machine learning and artificial intelligence techniques data MUST be at the silver layer*, this means that should be cleaned, standardized, and transformed into the desired format (see entity relationship diagram section) at the sentence level in such a way that is ready for NER. Because data can come from distinct EHR across regions and facilities, a robust cleaning process should be in place. Initial tests showed that unstructured data from EHR could be in xml, rtf, htm format and include non-printables that need to be processed to avoid inference errors. A sentencizer is added to facilitate the feature extraction in subsequent steps.
3. **Data quality:** The process includes a quality check to promote the data set to the silver layer. Therefore, it includes a unit testing process that supports the analyst and developer interaction to include additional cases for cleaning when needed. Also, a set of test cases is selected to include them as a quality control (QC) procedure to identify additional patterns, and thus embedding a continuous improvement loop in the pipeline.
4. **Cleaned data - silver table:** After QC and approval, the data is stored in the data lake as a silver level medallion table in a CDH defined format following the CDH_[datasource]_RA.NOTE_SENTENCES schema
5. **Process data with NER:** Using the silver medallion table, the NERs are used. Initially, there is a set of four (4) NERs from HuggingFace Transformers hub that were pretrained in the context of biostatistics, medicine and clinical notes. If the user chooses to either

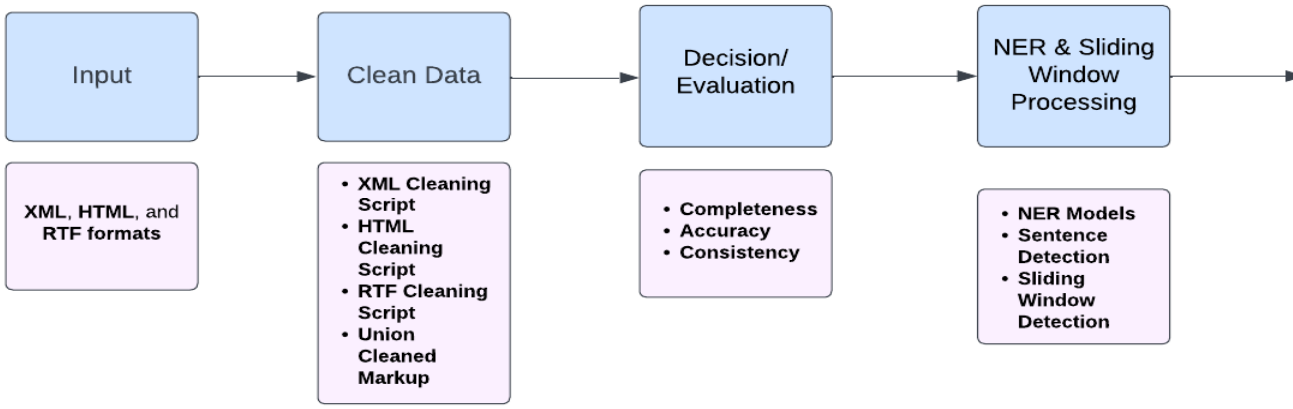
select distinct NERs based on specific needs or train a model on their own with the support of an SME they have the flexibility to do so.

6. **Feature engineering:** Using the NER results, the entity linker from *Spacy* can be used to further disambiguate the identified entities. This would include the Spacy entity linker that supports disambiguation between sentences or words and links them to a Unified Medical Language System (UMLS) to find the best match between specific entities (disease, signs, organizations etc.) to specific concept unique identifier (CUI) or Type Unique Identifier (TUI) codes. The Spacy linker returns these codes and the matched word.
7. **Processed data - gold table:** Store the resulting table containing entities from the NER and the Spacy entity linker table, and with SME support reduce the number of features.
    1. ML Flow and versioning: This subprocess runs for every NER and for the entity linker to keep track of the NER inference performance.
    2. Dimensionality reduction: Dimensionality reduction techniques are outside of the workflow but are available to the user and should be made per cohort basis

Application Flow Diagram

[Flowchart: Lucidchart (lucidgov.app)]





The purpose of this diagram is to show on a higher level of resolution of this application and multiple workflows that can support the overall and specific pieces of processing based on user needs.

Step 1 - Note ingest: Obtain access to the data in the derived OMOP CDM NOTE format which includes the NOTE table. The data would be stored in the data lake by CDH.

Step 2 - Note cleaning: The note cleaning step is a parametrized workflow in data bricks fined tuned for the cleaning operations that identifies the target pattern (xml, rtf, htm) to clean. It also cleans for non-printables and check if a note is already in readable format (UTF). A sentencizer is run to create the NOTE_SENTENCES table, and after passing QC process the table is promoted to the silver medallion level.

Step 3 - Note NER: In this step the set of NERs is run un parallel. They are parametrized and use the table with clean notes at the sentence level as input for identifying entities, this step uses optimized compute (GPU). In the figure there are four (4) Huggingface NERs. The output is a NOTE_NLP table for each NER used with all the entities identified.

Step 4 - Note Concept Linking from Spacy: the purpose of the entity linker is to disambiguate and better standardize the identified entity. This is because clinicians tend to use multiple terms for the same concept. Therefore, the linker is able to return TUI and CUI codes that would provide certainty and consistency across the NER. This table is store at the gold level medallion and could be considered by the SME for analysis

Step 5 - An additional step that would use MLFlow is to perform a feature selection process by means of dimensionality reduction. This is because there is a chance of having a large variety of features that may not be informative. Users should use this approach if they already have a target cohort in mind for which they need to select features for.