OMOP ETL Made Easier

INTRO:
ETL into OMOP CDM is hard. Using a combined event table makes it easier. ETL code writes to the same location regardless of the domain of the 'Maps to' target concept. Afterward a simple set of views or queries deconstructs the combined event table into Conditions, Procedures, Drug Exposures, etc.

Getting Started
The event combined table is created by identifying all the common and unique attributes of the OMOP event tables. All event tables share several common attributes such as primary key, the person id, visit and visit detail links, start date time, mapped to concept, source value, etc. Then there are attributes unique to the specific tables such as number of refills or days supply from Drug Exposure.

Two additional attributes, default_domain_id and domain_id, are added and will be used to determine the target table when extracting from the event combined table into the specific OMOP event tables. The default_domain_id is determined by the source data. This represents the OMOP table the ETL would write to if it did not have to apply the OHDSI mapping rules. The domain_id is filled with the domain id of the mapped to target concept.

I found the easiest way to identify the common and unique attributes of the OMOP event tables is to annotate each table and its attributes on a single Excel worksheet. This work was done using OHDSI CDM version 5.3.

The ETL
The transform code that reads from the source data and writes into the event combined table is deterministic. The ETL code reads from a source attribute and writes to the same attribute in event combined regardless of the vocabulary mapping. The ETL code tries to map the source concept, but this is only to populate the domain attribute in the event combined table, no additional logic as a result of the mapped-to domain is required.

To create the views that represent the OMOP CDM tables work backwards using the Excel workbook definition of the event combined table. For each view, the "where clause" determines if a row belongs to the target table by the domain_id, if it is defined, otherwise the default_domain_id

RESULTS
Using an intermediate table that is a composite of all the columns of the event tables simplifies the ETL process. It allows the ETL code to disregard the OHDSI mapping rules and eliminates the need to write redundant procedures for each possible target table.

Advantages
- Code is deterministic
- No need to repeat code based on target domain
- Easier to compare source to Event Combined table.

How Domain and Default Domain Works
The default domain is which table the row would go into if there was no mapping. The domain is where the source code is mapped to.

COALESCE( domain, default_domain_id) is where the data will end up in the CDM

Decompose
CREATE view Procedure  
(   procedure_occurrence_id
 ,procedure_concept_id
 ,procedure_date
 ,procedure_datetime
 ,procedure_source_vocabulary_id
 ,procedure_source_concept_id
 ,procedure_source_value
 ,procedure_datetime
 ,procedure_date
 ,procedure_concept_id
 ,person_id
 ,modifier_concept_id
 ,modifier_source_value
 ) AS  
SELECT
  event_primary_id
 ,person_id
 ,event_source_vocabulary_id
 ,event_source_concept_id
 ,event_concept_id
 ,event_start_date
 ,event_start_datetime
 ,event_source_value
 ,event_occurrence_id
 ,visit_occurrence_id
 ,visit_detail_id
 ,provider_id
 ,provider_i
 ,quantity
 ,modifier_concept_id
 ,modifier_source_value
 FROM event_combined;

Don Torok, Mark Khayter, Misha Bronshvayg