Name:	Yonghui Wu
Affiliation:	Assistant Professor
Email:	Yonghui.Wu@uth.tmc.edu
Presentation type (select one):	Poster

Using an Open-Source Extract-Transform-Load Package to Convert Cerner Health Facts Dataset to the Common Data Model

Yonghui Wu, PhD[1], Jingqi Wang, MS[1], Xiao Dong, MD[1], Guixiao Ding, MS[1], Hua Xu, PhD[1]
[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

Standardizing clinical observational datasets using the Common Data Model (CDM) is a prerequisite for applying study protocols from the OHDSI (Observational Health Data Sciences and Informatics) community. However, existing ETL (Extract-Transform-Load) tools for CDM are usually developed in an ad hoc manner that requires programming background to use in a command line. In this study, we share our experience of using an open-source ETL (Extract-Transform-Load) package, Kettle, to develop graphical data mapping workflows to convert the Cerner dataset to CDM. We introduce the typical components required for ETL workflow design in Kettle. Using the components, we show how to design graphical ETL workflows to map data fields and other complex procedures such as code mapping as well as data aggregation. Compared with the command line based ETL procedure, the ETL pipelines developed in this study are designed as graphical workflows using a user-friendly interface, which is more feasible to adapt for new datasets.

Introduction

Recently, there is an increasing interest in conducting cross-institutional clinical observational studies using multiple datasets. One of the obstacles that hampered the cross-institutional clinical studies is insufficient understanding of the unique and even characteristics from different datasets. To alleviate this problem, several initiatives are proposed to develop strategies and technologies to leverage cross-institutes observational datasets, including Mini-Sentinel, EU-ADR, and Observational Medical Outcomes Partnership (OMOP)[1]. OMOP has developed the OMOP Common Data Model (CDM) to address the data standardization issue. The motivation behind the OMOP CDM is to standardize datasets from diverse observational datasets into a common format using a standard vocabulary. Thus, researchers can develop standard protocols and algorithms to examine different observational datasets across institutes with minimal efforts. To standardize the raw observational datasets to the CDM, researchers often develop Extract-Transform-Load (ETL) tools in an *ad hoc* manner as the source data differed from each other. The existing ETL tools to convert source data to CDM are often command line based scripts, which requires programming background to adapt.[2] In this study, we introduce our experiences of using an open-source ETL package[3] (Kettle) to convert the Cerner Health Facts® Dataset[4] to CDM. The Kettle package provides a friendly graphical user interface to design data mapping workflows as well as design customized functional modules from standard Java jar files.

The objective of this study is to design a user-friendly ETL pipeline to convert the Cerner Health Facts® Dataset to CDM. The Cerner Health Facts® Dataset composed of hospital procedures, diagnostic information, demographics, medical history, admission, discharge, drug prescriptions, and laboratory tests over time. Over 480 facilities contributed de-identified information on about 47 million unique patients since 2000. In this study, we will share our experiences of using an open-source ETL package, Kettle, to develop data mapping flows to convert Cerner dataset to CDM.

Method

Workflow design

A typical ETL workflow in Kettle composes of three components: 1) an input definition component, which defines the input data from the source database or local text file; 2) a data processing component, which standardizes the input data and applies other complex procedures; and 3) an output definition component, which dumps the standardized data to the CDM database according to the data field mappings between the source database and target database. We divided the data mapping workflows into two categories according to the complexity of data processing component. The first category composed of workflows that could map the input table to the output table without complex procedures. (Such as dumping the PROVIDER table in CDM). Figure 1 shows a workflow to map the physician information to the PROVIDER table in CDM, where the source data was defined as the PHYSICIAN table from Cerner Dataset located on an Oracle server and the target output table was defined as the PROVIDER table in CDM database. The second category composed of workflows that require complex code mappings or other procedures (e.g., populate the DRUG_ERA and CONDITION_ERA tables). We encapsulated the complex data structures and procedures into a Java class and packaged it into a jar file to use in Kettle. For the code mapping ETL pipelines, we loaded the code-mapping file into memory in the customized Java class. For the complex procedures, such as populate the DRUG_ERA table, we implemented the algorithms in the Java class as well. Figure 2 shows an example to use the customized jar file in an ETL workflow.

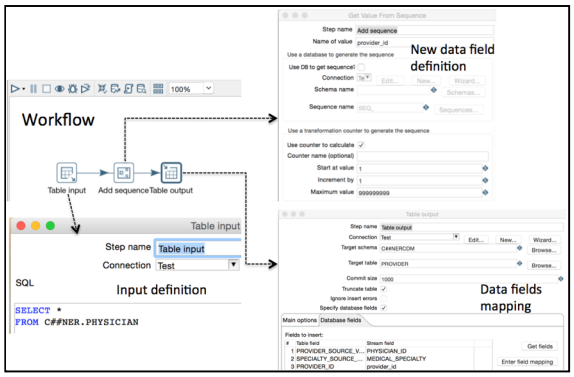

Figure 1. An example workflow of data fields mapping.

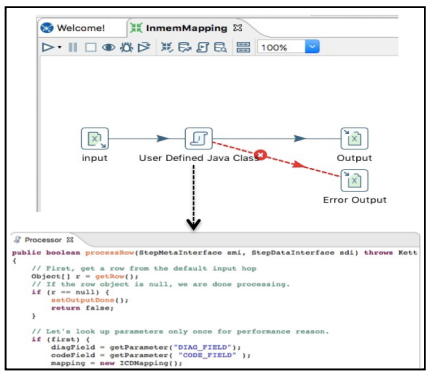

Figure 2. An example of using customized jar file.

The developed ETL pipeline can be executed from the Kettle platform as a workflow or from the server as a background process. The data processing speed can be further improved by running in parallel.

Conclusion

In this study, we share our experience of developing a user-friendly ETL pipeline to convert the Cerner dataset to CDM using an open-source ETL package - Kettle. The ETL pipelines are designed as graphical workflows, which are more feasible to adapt for new datasets.

References

1. Observational Medical Outcomes Partnership (OMOP). Observational Medical Outcomes Partnership Website. http://omop.org. Accessed June 20, 2016.
2. Soysal E, Wang J, Jiang M, Xu H. Mapping Local Laboratory Names into OMOP (Observational Medical Outcomes Partnership) Vocabulary Using Logical Observation Identifiers Names and Codes (LOINC®) Terminology. OHDSI Symposium 2015, October 20th, 2015, Washington DC.
3. Kettle Website. http://community.pentaho.com/projects/data-integration. Accessed June 20, 2016.
4. Cerner Health Facts® Database. http://www.bridgetodata.org/node/1789. Accessed June 20, 2016.