Name:	ZS Associates
Affiliation:	ZS Associates
Email:	harshal.soni@zs.com; ajinkya.patale@zs.com
Presentation type (select one):	Poster

Automating CDM Conversion Using Machine Learning

ZS Associates, Evanston

Abstract

OMOP CDM data model helps standardize healthcare data (EMR, claims, registries, etc.) and makes it easier to analyze outcomes at a large scale. Any healthcare data source needs to be converted into CDM format before it can be analyzed. Converting any data source (EMR, EHR, claims, registries, etc.) is a complex process and involves multiple steps, a lot of which are manual. Our aim is to automate the end-to-end CDM conversion process using machine learning. This will reduce the OMOP conversion of data sources and will help researchers and teams to start their respective analyses quicker.

Introduction

Converting a data source to OMOP format is a tedious task. It involves multiple complicated processes that require involvement of both medical and technical experts. Our team is working toward creating a working model to demonstrate how OMOP conversion can be fully automated using machine learning algorithm. Before that, a brief overview of the existing OMOP conversion process typically followed:

Existing Process for OMOP Conversion

Following are the key steps to perform OMOP CDM conversion for a new data source, including *steps that require manual intervention*:

- Ingest source data set into an environment (Hadoop/RDBMS)
- *Analyze source data (requires manual intervention)*
- *DQM of source data (requires manual intervention)*
- *Identify business/ETL rules for conversion (requires manual intervention)*
- *Create ETL scripts for converting the OMOP data set (requires manual intervention)*
- *Map source codes with OMOP vocabulary (requires manual intervention)*
- Execute ETL scripts
- *DQM and profiling of OMOP data (requires manual intervention)*
- Export converted OMOP data set for analytics

Need for Automation in Conversion

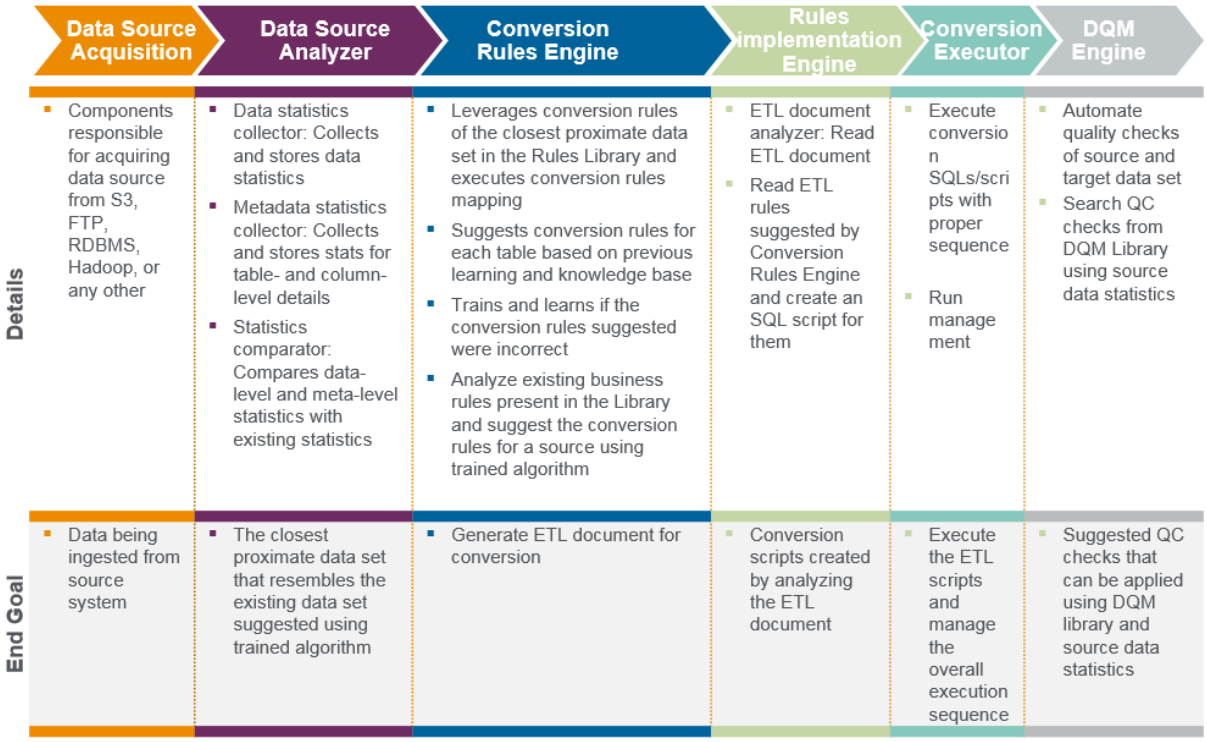
The previous section listed the overall steps involved in CDM conversion. As more than 50% of the conversion is a manual process that requires specialized skills and proper understanding of OMOP and source data, by automating the remaining 50% of the conversion, we can:

- Reduce the cost of conversion by limiting number of resources and efforts
- Reduce the time taken to convert a data set
- Reduce the margin of errors in conversion process
- Streamline the entire conversion, which will give better control to the user

How to Automate OMOP CDM Conversion Using Machine Learning

With our experience of converting multiple range of EMR/EHR data sets into the OMOP format, we see great potential in automating the conversion process using machine learning. Although seemingly unproductive in the initial stages, it will be beneficial in the longer run. We envision that in the near future, machine learning-enabled CDM convertor tools can increase automation from 40% to 80%. However, this will require considerable amount of time to train the tool with multiple data sets and different requirements. Automation in CDM conversion can be typically classified into four categories:

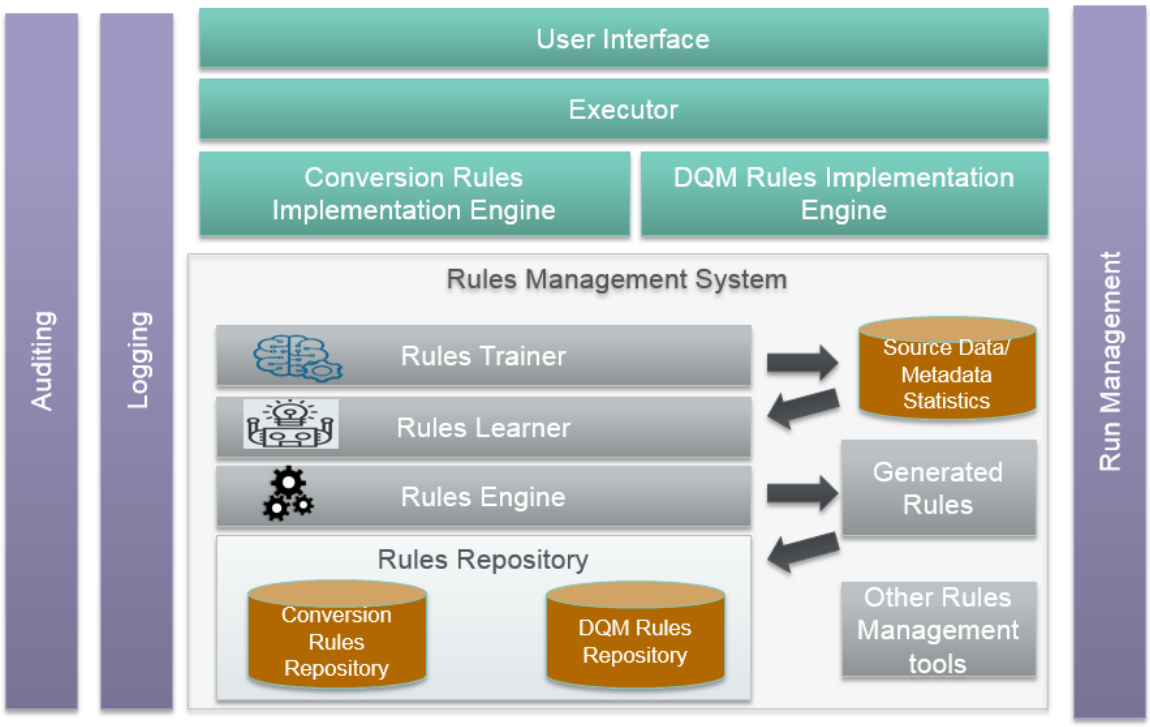
- *Source data analyzer* — perform a complete data profiling on the source data and create a summary of data profiling and metadata information of the tables. This summary is then compared with the summaries of pre-converted data-sources and a match percentage is calculated using machine learning algorithm. This match percentage tells the resemblance of the ingested data source with any previously converted data source. Initially, developers will perform the same steps and notify if the results do not match; this will help the algorithm to train automatically
- *Conversion rules creation* defines the set of modules based on machine learning algorithms that can help create ETL rules of any new data source based on the conversions it had executed and by analyzing the conversion of all other data sources. The initial phase of this automation starts with a Rules Suggestion Engine that analyzes the Source schema, data profiling results of some of the key columns to suggest the optimal rule to be used for conversion. These conversion rules will be in a readable format that can be easily changed to SQL (discussed in the following point)
- *Business rules implementation* — Once the business rules are finalized, one of the modules will be simply creating SQLs out them; there will also be an executor module that will take these SQL scripts and execute them
- *DQM* — For DQM automation, the developer will pre-define the DQM rules for every table and column of the OMOP CDM data set and the DQM automation engine will execute quality checks and publish a report to the user
- *Complete orchestration of the conversion process*: This step will act as master orchestrator responsible for triggering multiple phases and scripts for implementing various steps in the conversion process



High-Level Architecture

There are three crucial components in the architecture of the solution that can help automate the conversion process:

- Rules Management System
- Rules Implementation Engine
- Execution Engine



Conclusion

We have substantial experience in implementing OMOP CDM conversions for many data sets (EMR, EHR, registries, etc.) and we are leveraging our experience to design a solution that can completely automate the CDM conversion process. We believe that this design is just a step forward in the intended direction, and although this may not be the final architecture design, we will be modifying it further once we get more use cases on CDM conversion.