Near Real-Time Incremental OMOP-CDM ETL System

Seongwon Lee[1,*], Wanhee Lee[2], Seunghyung Lee[2], Ju Young Kim[2], Rae Woong Park[1,3]

[1]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea;
[2]EvidNet Inc., Pangyo, South Korea;
[3]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, South Korea

Background

Currently, EMR or claim data in a medical institution have been annually or quarterly converted into OMOP-CDM as a batch job, in general. OMOP-CDM is intended for population-level analysis, but if CDM conversion cycle is in sync with real-time medical data, the analytic scope could be expanded to research area that require recent data, such as COVID-19 studies and pragmatic clinical trials. In this study, as a first step for it, we introduce a near real-time incremental OMOP-CDM ETL system, which is already used in South Korea.

Methods

Strategies for Near Real-Time CDM ETL

Prior to develop a near real-time OMOP-CDM ETL system, we established the following ETL strategies:

- Execution cycle: Daily / weekly / monthly
- Scope of data: Incremented data since the last (full / near real-time) ETL
- Backup: Archiving the latest three CDM versions
- CDM synchronization strategy: Full ETL in a year cycle + near real-time ETL
- Concept mapping strategy: Major mapping in a year cycle + minor mapping in a quarter cycle

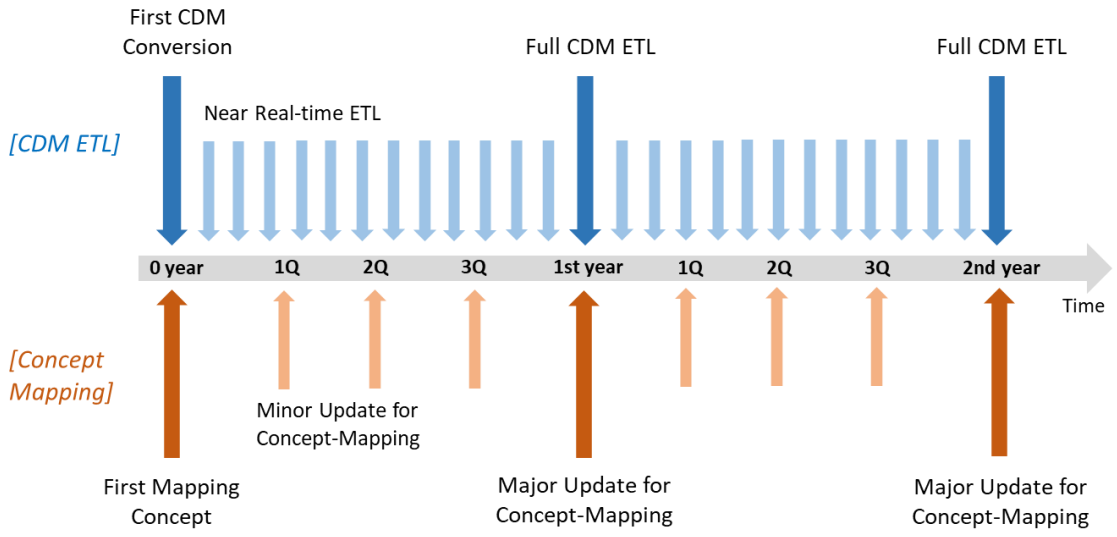

Figure 1. Strategy of ETL and Concept Mapping

ETL System Architecture

We developed the ETL system using open-source softwares of Digdag, Embulk, Apache Spark SQL, Apache Drill and Parquet file format. Figure 2 shows the architecture of our ETL system.

From original database of EMR to the CDM database, four staging intermediate files are produced: 1) an Operational Data Store (ODS) File, in which original data is loaded as is; 2) an Interface File, in which the data corresponding to the CDM is loaded according to the mapping rule between the original database and CDM database. At this stage, no data transformation is performed; 3) an Intermediate File, in which original meta data that hospitals have in different format (e.g., gender_concept_id, visit_type_concept_id) is transformed to OMOP concept; 4) a CDM File, in which original medical OMOP concept (e.g., condition_concept_id, drug_concept_id) is mapped to OMOP concept. All these staging files are generated as Parquet file format.

The whole ETL process is activated and managed by Digdag workflow engine and Embulk supports bulk loading from original data of EMR to ODS Files and from CDM Files to CDM DB. Apache Spark SQL facilitates actual ETL execution as memory-based big data processing engine and Apache Drill enables performing SQL queries on various types of original data.

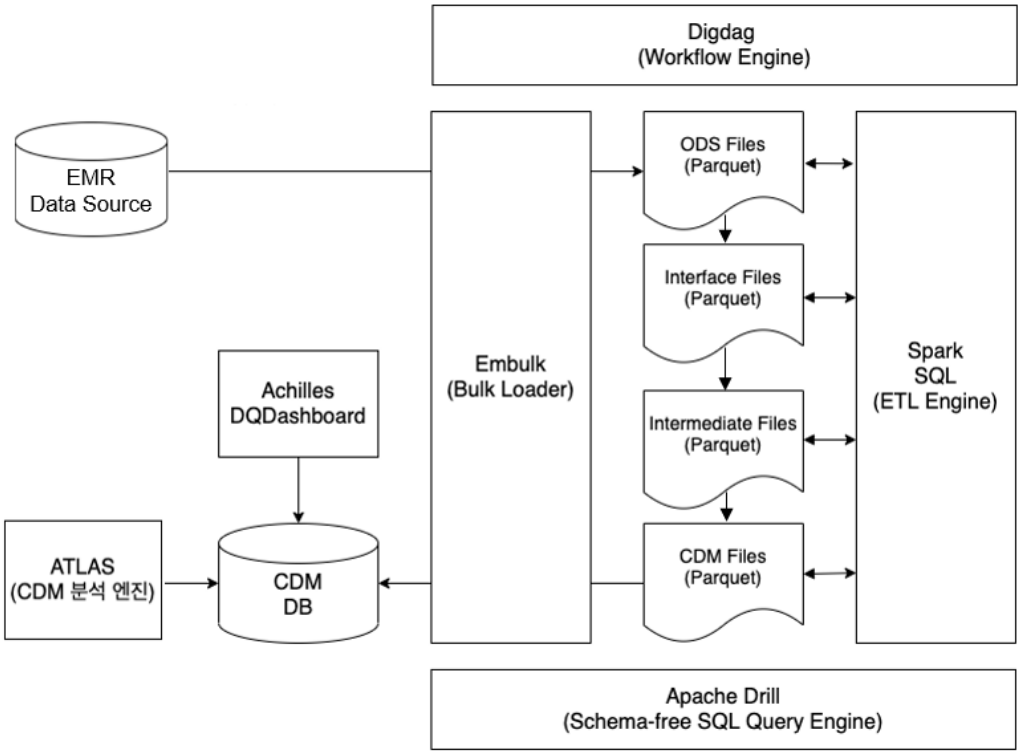

Figure 2. ETL System Architecture

Near Real-Time ETL Process

We improved the ETL system to support not only full CDM ETL, but also near real-time incremental CDM ETL. We incorporated the near real-time ETL workflow shown in Figure 3 into the ETL system. The most prominent change is the extraction of the original data, which is the source of ETL job.

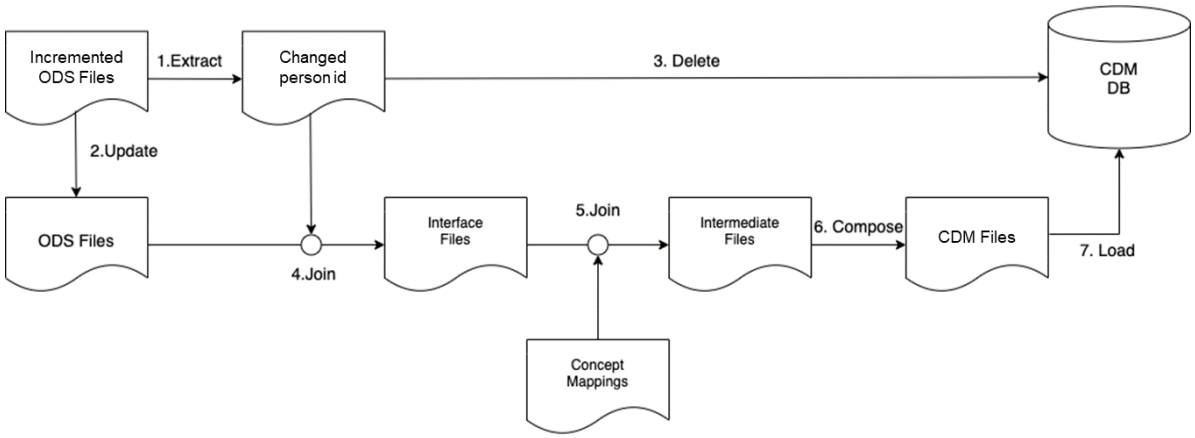

Figure 3. Workflow for Near Real-Time ETL

Near real-time ETL operates on incremental data, and we extract it using one of two ways. The first is to use a Change Data Capture (CDC) software. This is applicable only to hospital that have a CDC software, with which we can extract changed data from it since last ETL job. The second one is to execute a SQL of 'SELECT ~ BETWEEN' on the 'DATE' type column that can inform which data has been changed (e.g., update_date, last_modified).

After incremental data is extracted, 1) we derive a list of changed person_id from it, and 2) update and 3) delete the corresponding data to the list in ODS File and CDM DB respectively, which is the preparation step for actual ETL job. After that, 4-7) the ETL process is performed on the source data of the changed person_id list according to the procedure of the ETL system from Interface Files to final CDM DB.

Results

The near real-time OMOP-CDM ETL system was developed and evaluated in the CDM infrastructure environment of Ajou university hospital, which has an Oracle CDC software. Currently, this ETL system was successfully applied in three tertiary hospitals and is under application in seven hospitals in South Korea.

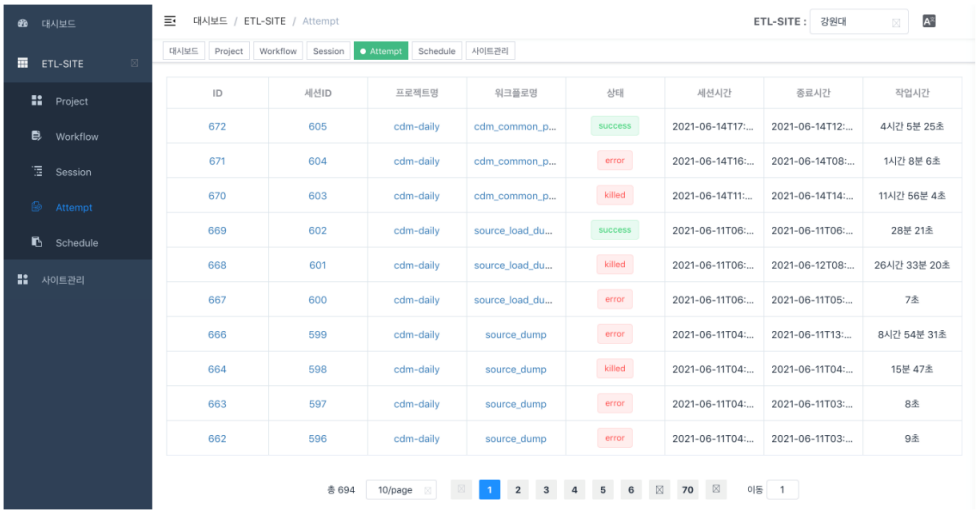

Figure 4. Dashboard of Near real-time incremental OMOP-CDM ETL system

This study has great implication to increase the usability of OMOP-CDM to the research that requires recent data such as COVID-19 studies and pragmatic clinical trials. And if this study proceeds to real-time CDM conversion, OMOP-CDM can be utilized for predictive research on the prognosis and severity of patient that should be performed in a short time.

However, there are several limitations in our near real-time ETL system. First, if hospital do not have CDC software, the source data deleted after the latest near real-time ETL is not selected and cannot be the target for ETL job, so it remains in the CDM DB. The synchronization of CDM will be achieved through full CDM conversion, but CDC software is essential for real-time ETL. Second, if too much data is accumulated in a day, it is inevitably necessary to perform near real-time ETL on a weekly or monthly basis. It is necessary to improve the system so that performing ETL in a short interval does not become a problem.

Conclusion

We developed the near real-time incremental OMOP-CDM ETL system and have applied it successfully to three hospitals. We expect that the near real-time OMOP-CDM ETL system would broaden the scope of OMOP-CDM use to the real-time prediction analysis. In the future, we will verify the system by conducting research using the near real-time converted CDM and continuously improve it.

Acknowledgement

This research was funded by the Bio Industrial Strategic Technology Development Program (20003883, 20005021) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea) and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health &Welfare, Republic of Korea (grant number: HR16C0001).