Standardization, automation and monitoring system for ETL process to improve CDM data quality
Abstract

- Each institution has a different way to store data, vocabulary and entity since it runs on various EMR system which leads to inefficiency takes up time.

- **EvidNet** had standardized the ETL process with incremented knowledge and know-how from the repeated CDM conversion of 12 institutions, and now we have developed a rule check package named **EvidFormer** to reduce time for CDM establishment as well as to carry out efficient conversion work.

- **EvidFormer** provides scheduling and automation service of ETL process as well as ‘all in one solution’ to enable work management monitoring service for high-quality CDM data with minimal work time.
1. ETL process standardization

- [PROBLEM]
  - From the variation of EMR systems, the vocabulary, entity, and data are stored with type and context variation, which lead to possible variations in vocabulary mapping, information and application rules in the process of converting hospital clinical data into the CDM.

- [SOLUTION]
  - We have standardized the process so that any hospital data can be converted through the same conversion process by grasping data attributes based on the conversion experience of various hospitals.
A layer(1..N) was constructed according to the purpose and characteristics of the conversion path from source data to CDM data.

1. We created an ‘Interface’ table that was most similar to the source data and contained all of the meta information needed for conversion to the CDM.

2. We create a ‘Meta rule’ table by performing data filtering classification and applying transformation rules.

3. After creating the ‘Meta’ table by deleting the data to be excluded from the generated interface table and applying the conversion rules, the conversion path was defined to create the ‘CDM’ table by joining with the term mapping table.
2. Data validation

[PROBLEM]
- Achilles heel checks the errors of whole data, wrong part and correct the ETL conversion code only after completing CDM conversion. The conversion time may take longer than expected as the reconversion is performed at the end of the ETL process.

[SOLUTION]
- We has developed a rule check package for each domain by layer.
  - If the package finds an error data during the conversion process, it corrects the error and then continues to convert it to the final CDM.
  - This eliminates unnecessary processes such as re-conversion and satisfies data quality and time efficiency at the same time.
3. Process automation and visualization

- [PROBLEM]
  ✤ Since one person performs the whole conversion operation, mistakes can occur in the work process, which can affect data quality.
  ✤ Most users also request detailed explanations of the workflow and data verification results to determine the reliability of the data.

- [SOLUTION]
  ✤ Scheduling and automation service of ETL process was developed based on standardized conversion path.
    ✤ This eliminates human errors that can occur during the conversion process.
    ✤ This made the conversion work faster.
  ✤ And we are developing a monitoring service to visualize the conversion and the data verification results for work management.
1. Workflow setup

### Definition

- timezon: UTC
  - `timezone : UTC`
- `include : sub/env.dig`
- `setup`: `echo : start ${session_time}`
- `disp_current_date`: `echo : ${moment(session_time).utc().format('YYYY-MM-DD HH:mm:ss ')}`
- `prepare`
  - `Task1`
    - `sh: ${[w[d]}/tasks/env.sh -h $[db_host] -U $[db_user] -d $[db_name] -s $[db_schema]`
  - `sprint_2`
    - `mt_rule_person_task1`
      - `sh: ${[w[d]}/tasks/psql_exec.sh -f ${[cdm dir]}/sprint_2/mt_rule_person.sql`
    - `mt_person_task2`
      - `sh: ${[w[d]}/tasks/psql_exec.sh -f ${[cdm dir]}/sprint_2/mt_person.sql`

2. Workflow run

### Workflow

- **ID**: 118
- **Name**: cdm_all
- **Project**: edl
- **Revision**: 19e46dc4-2c69-45ac-b352-980387a5ee9

### Scheduling

<table>
<thead>
<tr>
<th>ID</th>
<th>Revision</th>
<th>Project</th>
<th>Workflow</th>
<th>Next Run Time</th>
<th>Next Schedule Time</th>
</tr>
</thead>
</table>

3. Workflow monitoring

### Session

- **ID**: 20
- **Project**: edl
- **Workflow**: cdm_all
- **Revision**: 19e46dc4-2c69-45ac-b352-980387a5ee9
- **Session UUID**: 5a5d4f44-5a034-65b67-4a983b7a36e
- **Session Time**: 2019-09-11T08:37:39+00:00
- **Status**: Pending
- **Last Attempt**: 2019-09-11 17:28:15 (10 minutes ago)

### Timeline

<table>
<thead>
<tr>
<th>Task</th>
<th>Execution</th>
</tr>
</thead>
<tbody>
<tr>
<td>setup</td>
<td></td>
</tr>
<tr>
<td>+disp_current_date</td>
<td></td>
</tr>
<tr>
<td>+prepare</td>
<td></td>
</tr>
<tr>
<td>+task1</td>
<td></td>
</tr>
<tr>
<td>+sprint_2</td>
<td></td>
</tr>
<tr>
<td>+mt_rule_person_task1</td>
<td>5m 29s</td>
</tr>
<tr>
<td>+mt_person_task2</td>
<td></td>
</tr>
<tr>
<td>+person_task3</td>
<td></td>
</tr>
<tr>
<td>+mt_rule_provider_task4</td>
<td></td>
</tr>
<tr>
<td>+mt_provider_task5</td>
<td></td>
</tr>
<tr>
<td>+sprint_4</td>
<td></td>
</tr>
<tr>
<td>+mt_rule_condition</td>
<td>10m 31s</td>
</tr>
<tr>
<td>+mt_rule_device</td>
<td>10m 31s</td>
</tr>
<tr>
<td>+mt_rule_drug</td>
<td>10m 31s</td>
</tr>
<tr>
<td>+mt_rule_measurement</td>
<td>10m 31s</td>
</tr>
<tr>
<td>+mt_rule_note</td>
<td>10m 31s</td>
</tr>
<tr>
<td>+mt_rule_observation</td>
<td>27s</td>
</tr>
<tr>
<td>+mt_rule_order</td>
<td>55s</td>
</tr>
<tr>
<td>+mt_rule_payer_plan</td>
<td></td>
</tr>
<tr>
<td>+mt_rule_procedure</td>
<td></td>
</tr>
<tr>
<td>+mt_rule_specimen</td>
<td></td>
</tr>
<tr>
<td>+mt_rule_visit</td>
<td>5m 04s</td>
</tr>
</tbody>
</table>
Conclusion – All in One Solution

- As shown in the previous section, ETL, data validation, and monitoring can be provided through a single tool (all in one solution), ETL personnel are expected to be easy to manage and highly time efficient with our proposed EvidFormer.

- The end users can get answers and explanations for their fundamental questions whether the CDM data is reliable or how the local source data enters the CDM from our proposed solution.
“Thank you”

Dahye Shin, Seol Paik, Sehee Chang, Junghyun Do, Youmi Lee, Hye jin Kam

EvidNet, Inc., Seongnam-si, Gyeonggi-do, South Korea

*contact : dahye_shin@evidnet.co.kr