Journal of Biomedical Informatics 84 (2018) 179–183

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A scalable method for supporting multiple patient cohort discovery projects using i2b2

Evan T. Sholle[a], Marcos A. Davila[a], Joseph Kabariti[a], Julian Z. Schwartz[a], Vinay I. Varughese[a], Curtis L. Cole[a,b,c], Thomas R. Campion Jr.[a,c,d,e,*]

[a] Information Technologies & Services Department, Weill Cornell Medicine, New York, NY, USA
[b] Department of Medicine, Weill Cornell Medicine, New York, NY, USA
[c] Department of Healthcare Policy & Research, Weill Cornell Medicine, New York, NY, USA
[d] Department of Pediatrics, Weill Cornell Medicine, New York, NY, USA
[e] Clinical and Translational Science Center, Weill Cornell Medicine, New York, NY, USA

ARTICLE INFO

Keywords:
Secondary use
Electronic health record
I2b2
Scalability
Cohort discovery
Clinical data warehouse

ABSTRACT

Although i2b2, a popular platform for patient cohort discovery using electronic health record (EHR) data, can support multiple projects specific to individual disease areas or research interests, the standard approach for doing so duplicates data across projects, requiring additional disk space and processing time, which limits scalability. To address this deficiency, we developed a novel approach that stored data in a single i2b2 fact table and used structured query language (SQL) views to access data for specific projects. Compared to the standard approach, the view-based approach reduced required disk space by 59% and extract-transfer-load (ETL) time by 46%, without substantially impacting query performance. The view-based approach has enabled scalability of multiple i2b2 projects and generalized to another data model at our institution. Other institutions may benefit from this approach, code of which is available on GitHub (https://github.com/wcmc-research-informatics/super-i2b2).

1. Introduction

The secondary use of patients' electronic health record (EHR) data for cohort discovery is increasingly common in academic medical centers [1]. One popular approach is i2b2, which stores data from EHR and other systems in a star schema and allows investigators to run queries in a user-friendly web-based tool [2]. After creating one i2b2 instance containing de-identified data for 2.7 million patients at Weill Cornell Medicine, we deployed i2b2 to meet the needs of specific groups of investigators requiring identified data for patient subsets of interest to them. For each group of investigators, we planned to create separate i2b2 "projects" as described by multiple institutions [3–5] in an i2b2 online forum [6]. The project-based approach involved creating separate standard i2b2 databases for each project – "i2b2demodata" for patient data, "i2b2metadata" for ontologies to query i2b2demodata, and "i2b2workdata" for user data such as saved queries – as well as the shared databases "i2b2imdata," which contains project-specific data (e.g., cohort definitions), and "i2b2hive," which contains project metadata [2].

While implementation of the project-based approach showed initial promise, scalability problems quickly emerged. One project for 256,473 patients focused on digestive care research required 10 h of extract, transform, load (ETL) time and nearly a terabyte of disk space. Another project for 27,659 patients for the leukemia program required three hours of ETL time and a quarter of a terabyte of disk space. Although the projects contained fewer patients than the institution-wide i2b2 project, the patients in the new projects had more data likely due to the acuity of the patients, as "sick patients have more data" [7]. In contrast, our institution-wide project containing records for 2.7 million patients required 22 h and four terabytes of storage. With more projects for other investigator groups planned, we needed a scalable solution for disk storage and ETL time in order to deploy updates to the investigator community on a regular basis.

In each project, the database with the most records and disk space usage was i2b2demodata, which contained the OBSERVATION_FACT table, an entity-attribute-value store for all clinical data. Although each project contained different cohorts of patients, patients commonly existed across multiple cohorts; thus, the same patients and data existed across separate OBSERVATION_FACT tables in separate project-specific i2b2demodata databases. To prevent duplication of data and enable

* Corresponding author at: 575 Lexington Avenue, Third Floor, New York, NY 10022, USA.
E-mail address: thc2015@med.cornell.edu (T.R. Campion).

https://doi.org/10.1016/j.jbi.2018.07.010
Received 22 December 2017; Received in revised form 16 May 2018; Accepted 11 July 2018
Available online 19 July 2018
1532-0464/

custom ontologies in i2b2, Patel and colleagues have described sharing of common database tables using SQL views [8]. Although their approach supports sharing an OBSERVATION_FACT table with multiple ontologies, it does not address how to restrict specific projects to particular patient cohorts, a critical requirement of our overall institutional strategy for mass customization of data resources to support groups of investigators. To our knowledge, the literature does not describe an approach for i2b2 or other platforms for the secondary use of EHR data that reduces record duplication, enables restriction of patients, and supports customized and shared ontologies for multiple groups of investigators. The objective of this study was to develop a scalable multi-project i2b2 approach and compare it to the status quo with respect to ETL time and disk space usage as well as query performance time.

2. Materials and methods

2.1. Setting

As described elsewhere [9], Weill Cornell Medicine (WCM) is a clinical research facility and medical college on Manhattan's Upper East Side. Over 900 WCM physicians treat upwards of 2.7 million patients at over 20 distinct sites, including NewYork-Presbyterian Hospital (NYPH), a long-time WCM affiliate and teaching hospital. WCM makes use of multiple EHR systems, including EpicCare Ambulatory in the outpatient setting and Allscripts Sunrise Care Manager at NYPH. WCM's Information Services and Technologies Department includes the Research Informatics team, charged with the extraction and transformation of EHR data to support clinical and translational research. The Research Informatics team makes use of an established technical infrastructure relying on Microsoft SQL Server 2014 to acquire and integrate EHR data, which we employed for the current study.

2.2. System description

Fig. 1 compares the standard table-based approach to the novel view-based approach, which we describe below. First, we extracted source system data, transformed it into the i2b2 star schema format, and loaded it into a single base i2b2 project. Second, we set up individual i2b2 projects, which we called "sub-projects," and assigned specific cohorts of patients to each sub-project. Third, we created SQL views to support each sub-project. Finally, we created indexes to optimize general performance and use of SQL views in i2b2.

2.2.1. Extracting, transforming, and loading source system data into a single i2b2 base project

We extracted, transformed, and loaded data from source systems into an i2b2demodata database called basei2b2demodata. As shown in Fig. 1, the basei2b2demodata database, which contained an OBSERVATION_FACT table and corresponding dimensions tables, served as the main source of patient data for cohort-specific projects.

2.2.2. Creating i2b2 projects and generating project-specific ontologies and patient cohorts

For each sub-project, we created a corresponding prefix, which we assigned to a new i2b2demodata database, denoted as sub-i2b2demodata. As shown in Fig. 1, to distinguish between sub-projects, we created a table in i2b2pm called PREFIX_MAPPINGS associating each prefix with a PROJECT_ID as defined in PM_PROJECT_DATA. Of note, we created PREFIX_MAPPINGS to avoid tight coupling of the PROJECT_ID and the name of the physical databases. For example, for a sub-project called "test project" with a prefix of "test" and a PROJECT_ID of "100," we created a database called testi2b2demodata; edited crc-ds.xml, ont-ds.xml, and work-ds.xml within JBoss [10] to include a new data source for "test project;" and added a record to the PREFIX_MAPPING table for the prefix "test" and the PROJECT_ID "100."

As is standard with i2b2 configuration (https://www.i2b2.org/software/index.html), the XML files crc-ds.xml, ont-ds.xml, and work-ds.xml specified the data sources of each i2b2 cell of each sub-project. For each sub-project, we updated the CRC_DB_LOOKUP, ONT_DB_LOOKUP, and WORK_DB_LOOKUP tables in i2b2hive to specify a project path, schema, and data source from the associated XML files related to PREFIX_MAPPING. The field C_PROJECT_PATH identified the path associated with a sub-project as specified in the i2b2

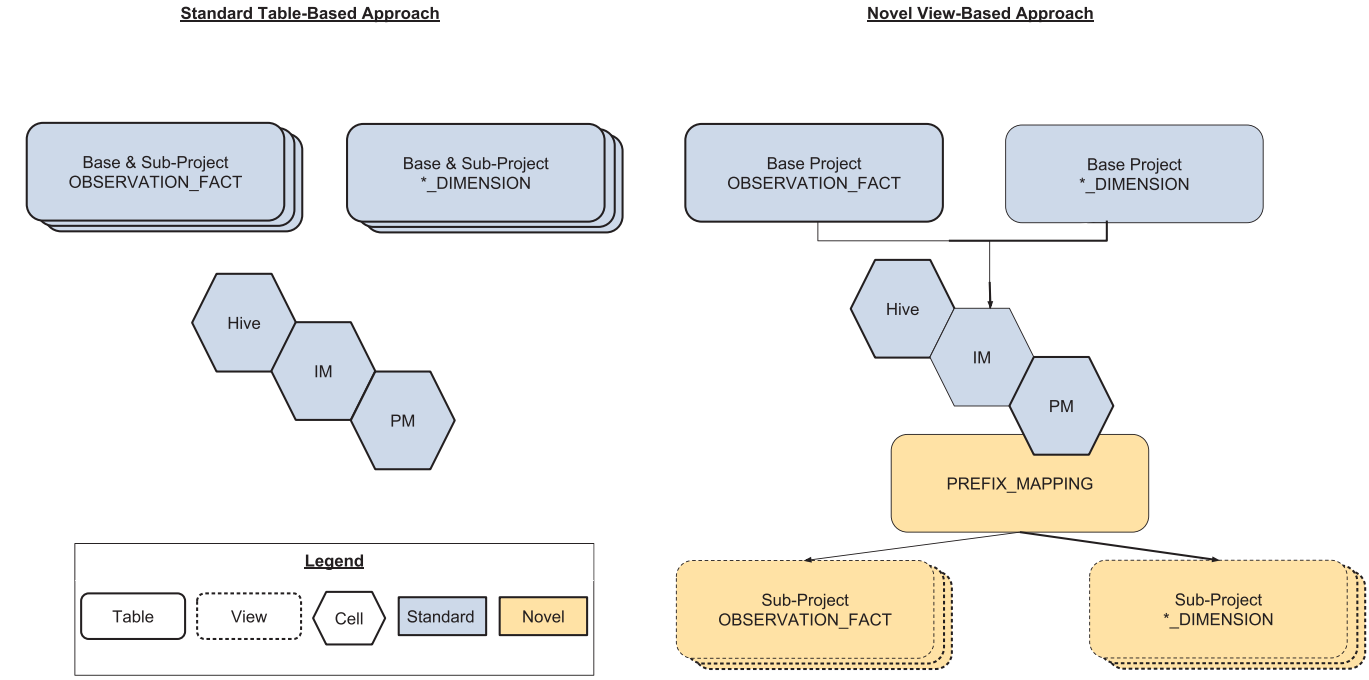

Fig. 1. Graphical illustration of distinction between standard table-based approach and novel view-based approach. * denotes multiple i2b2 objects, such as PATIENT_DIMENSION, and VISIT_DIMENSION.