World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Unified data integration and record identification framework for diverse data sources in healthcare demographics

Durga Chavali *

Manager, IT Applications, Trinity Information Services, Trinity Health, 20555 Victor Parkway, Livonia, USA.

World Journal of Advanced Research and Reviews, 2024, 22(01), 1733–1738

Publication history: Received on 15 March 2024; revised on 21 April 2024; accepted on 24 April 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.22.1.1250

Abstract

Challenges of bridging data from different sources with diverse data formats are faced by organizations in the modern data management environment. Problems with disparate data sources leading to different formats and inconsistencies mean it can be challenging to get the right matching of data records, especially when information errors such as typos are present. The current lack of a standard pattern for data integration and record identification presents a major problem in ensuring the accurate identification of individual records across disparate sources. The variations in data formats and the abundance of errors, such as typographical mistakes in names, dates of birth, and gender, add to the complexity of this problem. Organizations face the challenge of ensuring the correctness and consistency of data across multiple datasets without a formalized methodology.

Keywords: Unified data integration; Record identification; Patient demographics data; Health care data framework

1. Introduction

This paper aims to create one unified and well-defined approach or methodology for processing data. The goal is to have a set of rules, processes, and procedures that can be used similarly for various datasets and sources. Data integration is the process of combining data from different sources into a single view. Here, we aim to achieve data being integrated and processed consistently from different sources like databases, spreadsheets, APIs, and others. This is called linking the records that represent the same person in these varied data sources. For instance, if one data source uses the "John Smith" to refer to a person, and another source has this person listed as "J." Smith, "the pattern should be able to identify that these two references are about the same person. The main aim is to make sure that the system can identify those records that are related to a unified person and put them together. It includes de-duplication (eliminating duplicates), matching similar records, and maintaining a high level of integrity in the identification process. The problem comes from the fact that the data comes from different sources, with different structures and representations. Some data is structured in tabular form (for example, a spreadsheet), while others are unstructured (like text data from social media). Moreover, data formats, including date formats, naming conventions, and data element types can differ greatly. The purpose statement defines the overarching goal of the project, which is to define a standard and efficient method for managing this data diversity, making sure that individual records about the same person are correctly identified and processed.

One challenge with demographic data arriving in different formats is that it presents a problem of data integration, accuracy, and consistency. There is a large range of highly resolved images available to investigate the expansion sequence and periodicities with a high degree of accuracy. A list of the issues is provided below with details on their definitions.

* Corresponding author: Durga Chavali

Data Inconsistency: Different sources may employ different naming conventions, abbreviations, or alternatives for representing personal information. One source can record a person's full name as “John Smith,” while the other can have “J. Smith" or even "Smith, John." These variations make it difficult to identify and link records that refer to the same person.

Data Fragmentation: Different formats can result in fragmented data. For example, details like numbers and addresses can occupy different fields or have different structures from one source to another. This division makes it difficult to build comprehensive profiles for individuals.

Data Quality Issues: Data quality problems are often linked to a variety of data schemas. Some sources may contain missing or inaccurate information, while others may have old data. The accuracy of the assembled data collection may decrease when data of different standards are put together.

Data Redundancy: With no standardized format, data may be duplicated across various sources. Multiple records of the same person can result in inefficiencies and inaccuracies. In this regard, spotting and removing duplicates is necessary but tough when not speaking of a precise pattern.

Along with the above there are other key elements like Processing Complexity ,Data Matching, Data Privacy Risks, and Scalability Issues

2. Methods and Techniques

2.1. Data Extraction

This crucial stage turns on the gathering of information from various sources, such as databases, spreadsheets, Application Programming Interfaces (APIs), as well as other repositories. Collected data is then systematically abstracted from these various sources in its raw and untreated form. Finally, the raw data are stored in a designated staging area, which is a temporary repository created to support the subsequent processing. The staging area acts as the transition area where the raw data is temporarily kept before it proceeds to undergo more cleaning and transformation. This temporary storage permits a detailed data preparation process and enables subsequent processing steps to be performed on data that is consistent, uniform, and ready for analysis. Moreover, the staging area also plays a crucial role in ensuring the original data is not modified, offering a structured space for preprocessing activities including cleaning, normalization, and validation. The detailed handling of the data at the stage building area also lays the ground for downstream stages in the data integration and processing pipeline that will lead to the conclusion of correct and trusty Standardization.

2.2. Standardization process

In data gathered from different sources, it is usual for information to vary in terms of format and quality. Standardization is one of the most important processes which is aimed at reducing the heterogeneity of the data by converting it in a consistent and standard format. This transformative method entails handling of inconsistencies in naming protocols, date formats, and others across the dataset. Standardization is about resolving inconsistencies that can ever arise because of capitalization, abbreviations or alternate spelling. This rigorous cleaning process guarantees full compliance with a uniform set of rules, promoting clarity and consistency in the following analyses. Similarly, normalization of date formats requires resolving differences due to the various ways of representing dates (i.e., order of date parts, e.g., MM/DD/YYYY as opposed to DD/MM/YYYY or inclusion of time stamps). Through this alignment, data become harmonized and interpretable, allowing proper temporal analyses. In addition, standardization affects the rest of the data elements in which deviations in units of measurement, coding schemes, or any other varied representations are harmonized. This rigorous effort guarantees the information not only consistent but also ready to be analyzed and make sense, thus limiting the chances of misinterpretation or failure in upstream processes. The ultimate purpose of standardization is to create a unified dataset that does not contain any inconsistencies, to form an accurate and informative basis for data analytics. It increases the overall data quality, thus making its suitable to different analysis techniques, and helps the researchers and analysts extract interpretable patterns and insights from the standardized data.

2.3. Key Identification

The unique key or identifier is assigned to each entity within a standardized dataset during this crucial stage. This ID is a determining indicator for every person by which the system can distinguish between individuals belonging to the same or similar names or other features. The main goal of the assignment of such unique identifiers to the items is to have a reliable and unambiguous way of identifying the individual items in the entire data set at hand, thus making the

identification process straightforward. This measure ensures that there is no mixing up of the items or misidentification. In many cases, unique keys are selected around commonly known identifiers such as social security numbers, and national identification numbers, or use different attributes like name and date of birth. Social security numbers and national identification numbers are especially valuable as they are usually unique to each individual, thus offering a dependable and standardized approach to individual identification. In such situations where these identifiers are unavailable or not applicable, a combination of attributes such the full name of the person along with the date of birth can be used to create a composite unique key. Assigning these unique identifiers is key for safeguarding data completeness and matching the records with each other. It prevents confusion that individuals may have the same personal identifying information. This step provides the basis for referencing consistently and unequivocally throughout the remaining stages of processing, analysis, and interpretation by establishing a systematic approach for individual identification. Apart from this, it increases the integrity and precision of the dataset as a whole, leading to the production of valuable insights from the standardized and distinct records of individuals.

2.4. Time variation

The awareness that information about individuals undergoes alterations over time that include changes in residence, contact numbers, and occupation status is the reason that the Time Variant step in data management is needed. This phase concentrates on efficiently dealing with and integrating these temporal variations in individual data. The creation of timestamps as part of the data entries constitutes the Time Variant strategy. This is achieved by associating each piece of data with a timestamp that indicates when it was last updated or changed. Stamping the data entries with timestamps allows the system to form a chronological history of the changes thereby adding a temporal attribute to the data. This method guarantees that historical data is still available and trackable for analysis and reporting Consistently, researchers, analysts, or users of any system are capable of following the narrative in the world of data over time, detecting patterns, identifying trends, and comprehending the temporal behavior of the information. Timestamping also leads to a historical view of the data, which allows us to consider the development of individual attributes. In essence, the Time Variant step contributes to the generation of a dynamic and time-aware dataset. It recognizes the dynamic aspect of individual data and gives a means to record and manage these changes in a systematic manner. This temporal dimension augment the dataset usefulness by providing a profound comprehension of specific accounts and a good grasp of evolving knowledge throughout time.

2.5. SID Hub

The SID hub, being the main unit of the data management framework, is a central database designed to preserve district Unique Identifiers (UIDs) given during the key identification step. This central repository serves a major role in providing simple cross-referencing of individuals across different data sources and datasets. The SID Hub provides a meeting point for different data sets connected by the same identifiers. This connectivity is also important for record linkage since it allows a systematic linkage of knowledge related to the same person from several sources. The SID Hub performs the cross-referencing of unique identifiers by keeping a comprehensive and updated record of assigned unique identifiers. This enables analysts, researchers, or systems to move through heterogeneous datasets with confidence in identifying and linking records that pertain to the same individual. Furthermore, the SID Hub acts as a foundation stone for the disambiguation process, which aims at resolving possible ambiguity or redundancy issues caused by similar or even the same information coming from different sources. The control of unique identifiers ensures the accuracy of the data linkage exercises which lead to the development of a more consistent and reliable picture of individuals in the sample data. In sum, the SID Hub performs the critical task of ensuring data consistency and cohesion by serving as a centralized resource for unique identifiers. Its efficient running provides effective referencing, linking, and disambiguation thus leading to quality and correct analyses based on details from different databases.

2.6. Reconciliation step (Maintaining Historical Data Values)

Historical data preservation is the core element in comprehending and tracking changes over time within the dataset. This retention serves a dual purpose: firstly, it enables the creation of the audit trail that ensures transparency and accountability in the data management processes and secondly, it helps to reconcile the unique identifiers (SID values) for downstream reporting purposes. Here, historical data values are systematically stored and linked each one to the person's unique repository identification number (SID Hub). This relationship links the current and past status of the information of an entity. In turn, this offers a complete record of its history. An essential goal in information storage is preserving historical data together with unique identifiers for different purposes

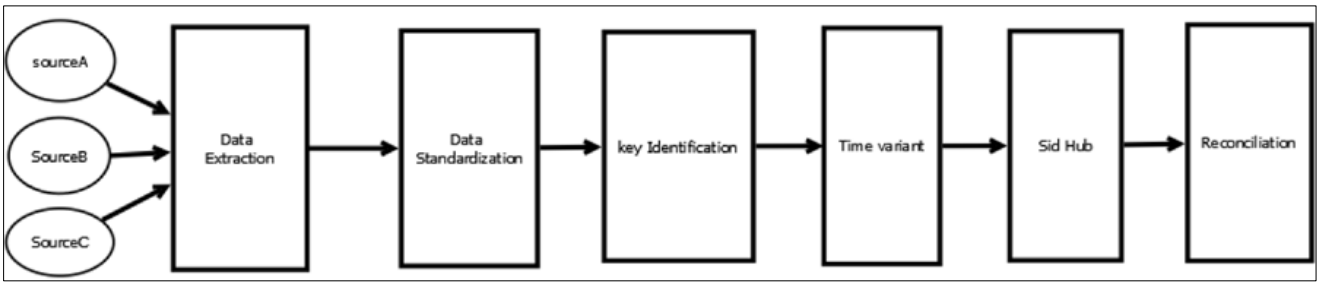

Figure 1 Framework diagram

2.7. Evaluation

In our work, the proposed design and framework received an in-depth analysis using data from seven sources or subsectors within healthcare. This collection of sources covers health-related demographics with datasets varying in size from 50,000 to 120,000 records, providing specific information about patients. The purpose of utilizing diverse sources was to thoroughly evaluate the performance and resilience of the framework under a range of data scenarios and demographic profiles. The combination of multifarious datasets characterized by diverse sizes and features was employed to provide a holistic evaluation of the scalability, flexibility, and efficiency of the proposed design regarding the processing of realistic healthcare data. SQL Server 2017 database management system was utilized to manage the datasets. Based on the RDBMS, the healthcare demographic data were properly structured, stored, obtained, and processed. We made sure that the Data manipulation and querying functionalities needed for the proposed framework through the use of SQL Server 2017 were readily available and efficiently executed. Also, the design's logic and operations were implemented using Transact-SQL (T-SQL) stored procedures. T-SQL, an extension to SQL, provides a set of powerful procedural programming constructs, which enables T-SQL formulation of and execution of complex data processing logic in the SQL Server environment. The implementation of T-SQL stored procedures enabled the orchestration of the data processing steps, allowing for smooth execution of the engineered logic within the database.

By leveraging these technologies and methodologies, our study aimed to validate the effectiveness of the proposed design in real-world healthcare data settings. The utilization of diverse datasets, SQL Server 2017, and T-SQL stored procedures contributed to a comprehensive evaluation, ensuring that the designed framework was not only theoretically sound but also practical and reliable in managing and unifying healthcare demographic data from multiple sources.

Table 1 Example of the person demographics data from various sources

Person_SID	Person_LastName	Person_FirstName	Person_MiddleName	Person_NamePrefix	Person_NameSuffix	Person_DOB	Person_DOD	Person_Gender_STD	Person_Home_Phone	Person_Work_Phone	SSN
119978202 1	FEENEY	ANDREW	NULL	NULL	NULL	8/7/1988	NULL	M	NULL	NULL	1233
119978202 4	COMMODORE	DARRYL	NULL	NULL	NULL	9/18/1973	NULL	M	NULL	NULL	2345
119978203 6	PARVU	JENNIFER	NULL	NULL	NULL	2/26/1974	NULL	F	NULL	NULL	4567
119978204 2	GRAMKE	AMY	NULL	NULL	NULL	10/17/1959	NULL	F	NULL	NULL	NULL
119978204 9	BENEDETTI	VINCENT	NULL	NULL	NULL	9/3/1962	NULL	M	NULL	NULL	1111

1199782052	MORROW	BENJAMIN	NULL	NULL	NULL	7/11/2009	NULL	M	NULL	NULL	9999
1199782064	GOINS	MELANIE	NULL	NULL	NULL	9/29/1964	NULL	F	NULL	NULL	293743323
1199782075	Morris	Lester	NULL	NULL	NULL	4/15/1953	NULL	M	NULL	NULL	NULL
1199782090	COOPER	GALE	NULL	NULL	NULL	4/8/1958	NULL	F	NULL	NULL	6666
1199782126	MONTGOMERY	DANIEL	NULL	NULL	NULL	8/20/1954	NULL	M	NULL	NULL	4432

3. Results

Table 2 Results for different sources with and without processing the data in the proposed framework.

Source ID	Source Name	old persons	single persons	delta	% Difference
1	Blue Cross Blue Shield of Michigan	661,380	634,872	-26,508	-4.01
3	Columbus United Healthcare	89,914	89,682	-232	-0.26
4	Blue Care Network West Michigan	133,469	132,246	-1,223	-0.92
5	Blue Cross Blue Shield of Michigan Colleague	102,374	100,725	-1,649	-1.61
6	Horizon-Lourdes	43,972	41,640	-2,332	-5.3
7	Florida Blue Commercial	20,177	20,165	-12	-0.06

The findings show that there is an excellent association between the information collected by the healthcare demographic sources and the outcomes generated by the proposed method. The evaluation of "old" and "single" people in various sources, including calculated deltas and percentage differences, has shown a good similarity between the information derived from the sources and the processed results from the system.

3.1. Observations

Consistency in Identified Records: The results demonstrate a high level of consistency of the numbers of seniors and individuals recorded by the health care sources and the system. The calculated deltas, representing the differences, are fairly small, meaning that the records are well-matched. The percentage differences that give insights into the proportional variation within each source between old and single people are usually small. The differences, including those in the Horizon-Lourdes dataset, are still within the acceptable range, which implies a relatively accurate correspondence between the source data and the generated results. The small delta values and the low percentage differences imply that the proposed architecture provides an efficient capture and identification of records while keeping accuracy in the representation of demographical fields. The system shows reliability in handling different types of data because it matches closely with the records obtained from the health data sources. The alignment of records found from the sources and the results produced by the system signifies that the proposed architecture can capture the details of the demographic information on healthcare. Any observed differences could be further investigated to verify the accuracy of data and to possibly identify areas for system improvement. The alignment between the identified records from the sources and the system results serves as a validating of the proposed architecture's effectiveness in unifying and processing healthcare demographic data. The system's outputs closely mirror the records identified by the diverse sources, enhancing confidence in its reliability.

4. Conclusion

Finally, the developed system architecture for information consolidation and record merging has enabled enormous improvements in improving the consistency, correctness, and ease of use of individual records from different sources. The systematic approach followed in the SID Hub, with the inclusion of historical data and temporal considerations, has shown to be well suited for overcoming the challenges derived from the different data formats, errors, and temporal changes. The usage of unique identifiers that are assigned during the key identification step and centrally managed in the SID Hub has enabled rapid cross-comparison and resolved ambiguity. Apart from lowering the complexity of linkage workflows, this centralized repository also guarantees an integrated and precise profile of individuals, offering an all-encompassing view across varied datasets. Additionally, the integration of historical data has offered companies the ability to conduct intelligent historical analyses, meet compliance measures, and keep a clear audit trail. The dataset's temporal dimension not only allows for historical recapitulation but also brings up an issue of accountability and error detection. The general results demonstrate a successful integration of data from various sources to be transformed into a unified and standard format. This development has not only brought efficiency to the process of individual record identification but also provided the groundwork for better decision-making, compliance, and historical analysis. With the evolving dynamics of data management in organizations amidst many formats, errors, and temporal changes, the established model remains a dependable framework for overcoming such challenges. The consolidated records that are achieved through this approach not only assist current data analyses but also provide a beneficial foundation layer for future data-driven programs and insights. Such all-inclusive and integrated techniques for data integration and record reconciliation effectively enhance the overall effectiveness, trustworthiness and strategic value of the business data management processes.

References

[1] https://www.ncbi.nlm.nih.gov/books/NBK253313/

[2] How to Build a Modern Data Platform Utilizing Data Vault | phData

[3] The Importance of Unified Patient Data Access | Hyland