Introduction

Data management for analytics is an end-to-end process, that is, it covers every step from the original capture of data at its operational source to the final preparation and enrichment of that data for analytics. Data can be viewed as a product that moves through a factory – a data management factory. Raw inputs from operational data sources systems are governed, acquired, and integrated into a data management foundation. This data integration module provides an overview of issues to be considered when bringing raw data together to be used for analytics. Please refer to the other modules of the Data Management – A Foundation for Analytics series\(^1\) for discussions on data governance, data enrichment and enhancement, and data storage.

The intended audience of this module is managers and data practitioners who are facing new or expanding data integration demands. Healthcare data management practitioners face the same data integration challenges as any other industry, but in most cases face a more heterogeneous data ecosystem with various levels of standardization. With disparate systems and a variety of data entry points, this module provides a check list of key bases that must be covered in the data integration strategy. Highlighted in this module are (a) Data Control Management and Aggregation, (b) Data Cleansing and Quality, (c) Data Transformation and Reformatting, (d) Cataloguing Data Stores and Data Classification, (e) Master Patient Index, (f) Obtaining External Data, and (g) Data in Motion.

Data Integration in Healthcare

The development of a successful Clinical and Business Intelligence (C&BI) program relies on Data Integration, which is the capture, cleansing and storage of data from clinical and financial sources. Increasingly, to support collaborative and accountable care models, data must be pulled from a wider

array of sources that includes multiple platforms, various clinical sites of care, and various clinical and business domains.

This poses particular benefits and challenges to the data integration process. Data integration helps create a more complete picture of the patients’ history, care and outcomes and facilitates communication across multiple sites of care and multiple providers within the same organization. It also supports clinical and business analytics around the value and impact of services provided to the patient population. Data integration challenges arise due to the “dirty” nature of the data to be integrated (necessitating scrubbing and QA processes), as well as the need to ensure patient identities are accurate when matching is done across source systems (master patient index).

Data integration essentially begins at the “front end” with the user entering data into a source system. Data entry standards, policies, validation, continuous auditing, and feedback all contribute to improving data quality, reliability, and applicability across the enterprise. Improving data competency among the front line staff is essential to achieving data that is as “clean” as possible, making later steps of the process easier and more reliable. This is an area that can be reviewed at the data governance level.

The process of moving the data from the source to the storage area is known as Extract, Transform, and Load (ETL). After data is entered at the source, it is systematically extracted, consistent with specifications to make it cohesive for analytical purposes. The post-processing relies on transformation rules which include cleansing / scrubbing logic, as well as logic to make the source data conform to organizational data standards. Data loading includes establishing a loading cycle with timely and routine processing. Continual monitoring and tuning of the system is essential. As data is processed, quality assurance steps ensure data is received and is consistent with the standards.

**Issues for Consideration in Data Integration**

a. **Data Control Management and Aggregation.** The first step is to determine which data needs to be integrated, potential methods of integration and parameters for the data. Data Control Management and Aggregation includes the methodologies for sourcing data, defining the necessary levels of granularity in the data, and methods for rolling up data to structure it for integration or analysis.

   • **Data Interchange Methods.** Determine whether standard interchange channels are available that will support the necessary data. In many cases, HL7 standards and / or existing Interface Engines will support the transmission of the required fields. If not, other more customized data extract means may be needed.
• **Data Specifications.** Develop data specifications that meet the business needs for reporting and analytics. For example, you may need to pull data from multiple physician practice systems and combine with hospital data to do financial models on particular patient populations. The data specifications will list the fields that each source needs to provide (with a data definition), as well as how it is to be submitted (field names / definitions, file format, naming conventions, etc). It will be important to track the specifications if they change over time because sources may be on different versions, which will complicate data integration for reporting.

• **Source to Target Mapping.** Once the data specifications are determined, the business and technical team should collaborate to map data from the source systems to the specifications. This may require referencing the data dictionaries for the source systems and / or data profiling to confirm which fields contain the data that meets the requirements. Oftentimes, there may be several fields within one system with similar names, so research is needed to ensure the right field is being used for the business need. Policies will be needed to address scenarios where a source system does not comply fully with the specifications. For example, what if one of the physician practice systems does not store one of the required fields or can only provide it monthly when you wanted it weekly? Should the data be ingested and used where available or should that provider’s data be excluded? This is another data governance issue.

• **Data Architecture / Model.** Determine where the integrated data will be stored and in what structure. If data warehousing is being developed, the benefits and drawbacks of various approaches should be reviewed (e.g., Inmon, Kimball.). Tools are needed for data modeling and profiling and a central data dictionary is essential. Determine if data marts are needed to facilitate reporting, as well as the data roll-ups and dimensions that are needed for analytics.

• **Data Security.** HIPAA data security is usually addressed well in most front end applications, but once data leaves the source system and is integrated with other data for analytics, new risks may arise and protections will be needed to meet patient confidentiality and various regulations. Data integration projects should be careful to do a full assessment of data security risks and how they may be addressed and managed. For example, the State may have tighter restrictions than HIPAA on who can access HIV or mental health data. The electronic medical record (EMR) may protect these data adequately, but once data is pulled out for analytics, it may be open to more users if protections are not put in place.

• **Plan for Data Loading.** Once the interface is in place, loading cycles and how data will be refreshed must be determined. Data may be loaded as Cumulative Files (send everything over each time). Change Data Capture (CDC) methods may be used to identify and load only the data that has changed. Consider how new data is integrated with existing data over time – e.g., it may be truncated and reloaded each time or only changes may be loaded (insert and update). You may also need to implement methods for tracking Slowly Changing Dimensions (SCDs) if your clinical and business users are interested in viewing a data element that changes
over time. This could apply to everything from the patient name to body weight to lab results, depending on the business needs and what the source system stores. For example, a question as simple as “what is the patient’s insurance?” will depend on whether the business need is to know the insurance on the date of service or the patient’s most recent insurance and whether the source system stored it on the date of interest or overwrote it.

b. Data Cleansing & Quality Control. As previously mentioned, healthcare data has a tendency to be “dirty.” This becomes more apparent as the organization attempts to integrate data from various sources and use it for C&BI. Common problems include missing fields (e.g., not collecting phone numbers, date of birth), wrong data (the social security field given may be the patient’s, the guarantor’s, or in the case when the patient is an infant, the mother’s), out-of-date data (insurance is entered one time and not updated each time the patient visits), and non-structured data (i.e., text fields). Attempting to integrate data and use it for reporting brings these issues to the forefront and data cleansing and quality control are used to ensure consistency and standards in each source’s data. It also highlights the need for collaboration and feedback to the source system users to minimize data errors and omissions.

• Establish the data quality standards that are needed by the business users. This includes determining their tolerance for data errors and missing data. Additionally, identify data that you have vs. data that you need for the business purpose. Make sure business users know the limitations of the data and collaborate with the technical team to develop policies and procedures to address the data limitations. For example, a report may be useful to the business customers, even though some data may be missing. In this case, the decision may be to release the reports with a footnote describing the data issues.

• Establish QA standards and processes for data from the source. By identifying the business users’ data quality needs, the technical team can set up automated monitoring of the data as it comes in from the sources. Examples include: Is the source sending the proper / expected date range? Are the critical fields populated and coming in the expected format? Does the new data meet the expected file sizes, record counts, tally, date ranges, etc? This will be an iterative process: as new issues arise in the data (which they tend to do), the QA process is continually refined to identify and address them. Additionally, you will need methods for handling perceived data quality problems. For example, should you reject the data that fails on one or more QA standards, or load it with a flag to indicate there are problems? Again, this depends on the needs of the clinical and business teams.

• Data cleansing. It may be possible to “cleanse” data for integration by applying scripts to look for certain known issues and replace/repair. This is reviewed below in the Data Transformation and Reformatting section. However, caution is recommended because errors in the source system will persist. For data integrity between the source systems and the C&BI side, it is best for data to be corrected in the source and then allowed to flow through to C&BI. For example, your data may include a birthdate for a patient that is in the future due to a data
entry error (and unfortunately, the source system is not preventing this). The QA process could catch and flag to exclude the bogus date of birth from reports, but that error will continue to persist in the source system. It will continue to be sent over again in the data feed. It will also impact services for the patient. In this example, the patient may be excluded from reports and processes that use the date of birth (such as a mailing list for patients over a certain age in need of cancer screening or patients eligible for a research study). Feedback to the front end is required to continually improve data quality.

- **Changes and upgrades in the source systems.** As your organization attempts to integrate a wider array of sources (e.g., multiple hospitals, private practice physicians, payer files), there will be increasing variability across the sources. The data feed from Hospital A may go offline while they move to a new server. The data feed from Practice B may break when they upgrade their software, sending the wrong data in a field. This necessitates methods to handle the changes and gaps in data and, ideally, automated notifications from the sources when there are changes or offline situations. Otherwise, unexpected gaps and data errors will flow through to C&BI.

c. **Data Transformation and Reformatting.** Once the necessary data elements are identified, data transformation and reformatting may be used to integrate data from the source system to storage and C&BI. This includes methods to ensure data definitions are consistent across data sources and data points within sources.
   - Determine if data fits into the destination table structures and whether it can be corrected or transformed to comply with the standard structure. For example, Practice A’s system may store 1 for male and 2 for female, but your data warehouse stores M for male and F for female. Practice A’s data can be transformed to meet the standard. Or reformatting of source data can be used to restructure the data to meet your standards, such as re-ordering, transposing or pivoting columns to from the source data.
   - Assess data transformation tools, such as ETL tools. Consider their functionality with your sources and destination. Rules may be table driven or integrated in the ETL tool. Set standards for using the ETL, including performance and memory considerations.

d. **Cataloguing data stores and data classification.** At the heart of data cataloguing is metadata, or data that provides information about other data, thus making the latter data relevant and usable for C&BI. Metadata has a key role for data integration. Embedded information about the content, context and structure of records is essential to know when the activity took place (data source, location, date, and time), who it involved (physician ID, patient ID) and who entered the data (user ID). This also includes establishing unique identifiers to index and describe resources (data, documents) so that they can be located and consulted, facilitating data retrieval, maintenance, and audit.
e. Master Patient Index (MPI). One of the keys to integrating data in the healthcare domain is the Master Patient Index. This is a unique ID for each patient in the integrated data and is derived by matching various fields in the patient identity file from each of the multiple source systems.

- For example, Patient X may have identifier 1234 in Practice A’s system and 99999 in Hospital B’s system. You may suspect these are the same person because they have the same date of birth, gender, last name, first initial and address. By applying logic that matches various fields of the patient’s identity, a new identifier (the MPI) can be created that matches ID1234 in Practice A with ID 99999 in Hospital B. Any reports that use the MPI will pull in source data that matches any patient identifiers matched to the MPI. This is a very important means for pulling together physician and hospital costs for a patient procedure, for example. When all providers are on the same system, this is usually not an issue. When providers are on different systems, the MPI is essential for advanced, patient-centric analytics.

- As the organization attempts to integrate a wider array of source systems, there are surprisingly few fields that are reliable across all of them for determining matches for the MPI: names change, social security numbers are often not collected or are not reliable. Third party tools are available that utilize algorithms to determine weighting of fields and confidence levels of the matching. For example, if the confidence level between a new patient and an existing MPI is 85% or higher, you may decide to match the patient to the existing MPI. If less than 85%, you will not match to the existing MPI, but create a new MPI. Manual review and testing must be done to set a confidence level that works within the boundaries set by the organization. This includes a review of false positives (patients matched to an MPI who shouldn’t have been) and false negatives (patients not matched to an MPI, who should have been). Both types of errors will have implications for the patients and usability of the data, so ongoing manual review is needed to audit the accuracy of the MPI.

- Once the related patient records are matched to the MPI, you can go further and apply logic to determine which of the names, addresses, and other data elements from each of the sources are considered the “truth” for the patient. For example, the most recent record could be used or a hierarchy can be used of the most trusted to least. The result is the “golden” record for the patient which can be used for patient-centric reporting.

- MPIs are a key to enabling Health Information Exchanges (HIE) to function, and most HIE products will have an MPI process to match patients across multiple EMR systems. Careful assessment must be done of the accuracy and limitations of the MPI, as clinical decisions may be made based on records that are combined using the MPI.

f. Obtaining External Data. This includes data from outside the provider organization and such items as benchmarks, CMS Medicare files, all payers’ claims data, pharmacy data, and reference data. Some considerations when integrating external data are:

- Determine whether the data must link to your provider data and if so, which key(s) can be used for linkage (MPI, National Provider Identifier, etc.). External data may go into
segregated tables or be added to existing structures depending on the business and reporting needs.

- Determine the data availability from the external source, the cycle for loading and how files will be transmitted. Find out and plan for the data update schedule for key reference data such as CPT, ICD-9, and RVRVS. Establish rules for versioning data from external sources to address changes over time. Work with the business users to set methods to handle situations where the expected data does not become available by the expected target date. For example, if you have a physician productivity report that has to run in January, and the updated RVRVS file is not yet available at that time, you may use the previous year’s RVRVS values. However, this must be noted and communicated to the business users and physicians.

g. **Data in Motion.** Once data extracts are developed from source systems, the data must be moved on a predictable schedule and in a secure manner. This “data in motion” is the engine that enables C&BI. Issues for consideration:
  - Protocols for data transfer (e.g., sftp, ftp, smb, email, https) and encryption standards need to be set with each of the source systems. A data management tool may be utilized to automate data transfer from source.
  - Scheduling of ETL jobs. Frequency of data loads can be determined by the business needs associated with each type of data. The more frequent the loads, it is potentially more costly/time consuming to maintain. Ensure the business users are aware of any time lag between when the data is captured and when it is loaded in the destination store, as this will impact the usability of the data for C&BI.
  - Ensure expected data is received from the sources including all the required files and all the required fields. Monitor data transfer failures and errors and making notifications through different methods such as control files, metadata files, email notifications, system log writing and file system log writing. Communicate back to the source to reach resolution.

**Additional Resources**

Links to more information on data integration:
  - Data Warehouse Information Center  
    [http://www.dwinfocenter.org](http://www.dwinfocenter.org)
  - Healthcare Data Warehouse Association  
  - Kimball Data Warehouse and Business Intelligence  
  - Inmon (data warehouse)  
    [http://www.inmoncif.com](http://www.inmoncif.com)
For More Information
Organizations beginning their work efforts in healthcare data management are also advised to seek out other modules in the Clinical & Business Intelligence: Data Management – A Foundation for Analytics series:

- Data Governance
- Data Enrichment / Enhancement
- Data Storage

The series is available in the Resources and Tools section² of the HIMSS Clinical & Business Intelligence Resource Library.

Task Force Contributors
This paper is a product of the HIMSS Clinical & Business Intelligence Information Management and Data Technology Task Force. Its contributors include:

Task Force Chair
David Dobbs, PMP
maxIT-VCS, a SAIC company

Task Force Contributors
Theresa Burkhart, MS, MHA, FACHE – team lead
Rush Health

David Dobbs, PMP
maxIT-VCS, a SAIC company

Ray Halper, PhD
Rush Health

Jason Oliveira, MBA
Recombinant By Deloitte

John Sharp, MSSA, PMP, FHIMSS
Cleveland Clinic

HIMSS Staff
Shelley Price, MS, FHIMSS
HIMSS

Nancy Devlin
HIMSS