

8

Big Data: Architecture and Its Enablement

Bruce Johnson

CONTENTS

Objectives	156
Abstract.....	156
Introduction	156
Data Standards.....	158
Data Governance	158
Data Architecture	159
Enterprise Architecture.....	160
Health Care Specific Analytics Topics.....	162
Targeted Analytic Solutions.....	162
Database Designs.....	163
EDW	166
ODS	167
Data Marts	168
Data Mining	168
Super Marts	168
Analytics	168
Ties to Governance	171
Best Practices	172
Conclusion	174
References.....	175

OBJECTIVES

After reading this chapter the reader should be able to:

- Articulate the need for and advantages of an enterprise-wide data architecture encompassing the data life cycle
- Describe the main database designs for big data
- Describe the relationship between data architecture and data governance

ABSTRACT

The concept of big data is just that: a concept for the value an organization can realize from in-depth analysis of all data. The concept of big data is therefore not a database or data architecture but is more the solutions that leverage any and all data, wherever they come from. In health care, the concepts of big data are enabled only in organizations that focus on data—capture, management, and usage. Health care data is extremely broad, deep, and complex, yet the needs for data access are even greater and ever evolving. To meet such needs, effective data architecture must be intertwined with a formal data governance program. This combination unlocks analytics and begins to leverage big data. It is emerging as a critical best practice to all health care analytics efforts.

INTRODUCTION

Big data takes on new meaning when applied to health care. In most industries, the collection, usage, and analysis of data are the foundation of leveraging facts to understand, grow, and improve business performance. Finding ways to thus incorporate unstructured data, images, and geographical uniqueness leads to the concept of *big data*. Health care data is almost unmatched by other industries due to its challenges in collection, breadth, width, and mass amounts of unstructured data. Most industries leverage basic reporting and analytics for operational effectiveness and

focus on advanced analytics and data mining to identify future direction, process improvement, and R&D.

In contrast, health care has almost limitless basic reporting and analytics needs that are dependent upon data capture challenges. Data capture is complicated by the various areas of practice, unique data specific to those areas, and individualized characteristics that each patient possesses. While standards for data collection across practice areas abound, there are few if any standards that readily can be applied across health care organizations to serve as a solid foundation for data architecture and design.

In the past decade, significant requirements for quality reporting have led to mass efforts within health care organizations to create, adopt, and internalize tracking, monitoring, and process improvement over all areas of care. This focus has given those organizations that are able to build solid data architectures much better capability to focus on improving patient care, patient outcomes, and patient satisfaction as well as operational effectiveness. Yet many organizations are still in the early stages of mass capture of data. Growth and changes by electronic health record (EHR) software vendors have made this easier in recent years but also have added massive expense to establish this foundation of data capture. Limited funds are left for the production of reporting and analytics, let alone the advanced analytics that drive research and quality improvement (Millard, 2014).

Advanced analytics in health care require the blending of the mass amounts of structured data captured by EHRs with images, textual notes, genomics/proteomics, video/audio, and geographical data. With this wide variety of data types comes an almost limitless opportunity to leverage every advanced analytics technique and mass computer processing to correlate, identify, study, and analyze. This multiplicity becomes evident when considering the mass numbers of resources devoted to medical research and quality improvement around the world. These processes are paramount to disease research and management, drug development and monitoring, individualized medicine, medical devices, and bending the cost curve that has plagued our health care systems.

At the heart of the need to serve big data to the health care world lays data architecture. Effective data architecture must lay out the life cycle of data, from definition to capture, storage, management, integration, distribution, and analysis. The need for coherence across this life cycle requires explicit linkage among data standards, governance, architecture, and analytics. We now focus on understanding each of these concepts in relation to the others.

DATA STANDARDS

As health care practices and specialties exist across organizations, data sharing becomes critical to effective data analysis and research. To share data requires consistencies in data terminologies, ontologies, and data keys—that is, agreement as to what “words” we use, what those words mean, and how they are uniquely defined for proper definition. Think of communication among people who speak different languages: if there are 10 people in a room all speaking a different language and they know only their own language, it is nearly impossible for them to communicate effectively. While sometimes the words may sound the same, different meanings can cause misinterpretation. And when dealing with the health and well-being of our patients, it could spell disaster. While there has been significant migration to data and reporting standards, there is still little in the way of common data definition across vendor- and internally developed systems to ease sharing. Within any organization that has built an effective data architecture, a solid data model and associated standards/definitions exists to provide the foundation for data standards.

Data Governance

Data governance for the data itself establishes data standards and continues to grow them. For how it is used, data governance establishes the rules of the data-use game. Data governance becomes the function that owns the quality of data across the organization. The participating policy makers ensure that standards are in place, that data quality is monitored, and that new/emerging data and data sources are always tied into the rest of the data picture for the business. As many industries have very refined data that is easier to capture and always consistent, data governance is sometimes viewed as an information technology (IT) function. In health care, this would equate to technical people having to define the concepts of medical definitions, views, interpretations, and even disease associations. While data governance relies heavily on technical resources to provide the tools and monitoring to ensure data integrity, it is also a business function operated by those who know the business. Data governance requires balancing data security and privacy concerns against the need for knowledge from the data. As an IT graduate myself, I would not want to be responsible for establishing that mission-based balance.

Data Architecture

The concept of a data architecture in health care is interpreted many different ways across the industry. Up until the early 2000s, very few formal data architectures were in use. Health care IT systems were implemented for practice areas or for specific needs as they were identified, and thus a mass of systems with no integration or standards evolved. For example, within a large hospital system, clinics would have their own EHR that could not exchange information with billing systems, pharmaceutical systems, or other clinical systems within the facility. As health care needs for access to data grew, custom solutions were created for each need furthering the conundrum. As such, there was limited experience with multiple architectures that would identify what works best. Each organization thought up what they figured might work, with little to draw from successful examples, or simply purchased a solution from a vendor that the vendor used in a simpler industry but hadn't tried in health care. The more those solutions became a part of the culture of an organization the harder it became to change to a planned approach.

With the emergence of Meaningful Use requirements and broad adoption of EHRs has come a required focus on creating an overall data architecture that serves all the needs of a health care organization, at the same time that the Affordable Care Act (ACA) is exerting pressure for the organizations to change their business. The requirements and needs for analysis have also changed significantly with the emergence of the Triple Aim—looking for ways to drive down costs and improve population health and the patient experience. Thus, the need for integrated data that gives various views across the spectrum of data has become critical. For one organization to satisfy so many data needs requires a formal data architecture.

Many complex analytics in health care have been kept separate from more basic health care data by resources that want to control their data. However, that balkanization limits access to other data and to building consistent, standardized data. The best example here is how to incorporate research data. Researchers are very protective of their data, and rightfully so. However, their data could be insulated within a larger architectural structure at the same time that they are updated with much more accurate and timely clinical data, with greater ease and efficiency than researchers' current practices. Disease registries for research comprise a common example where basic clinical data is not kept up to date,

because of the current practice of manual curation, even though the data exist within the EHR that could provide automated updates. Linking the research database structurally to the EHR would make research systems much more diverse and powerful.

There are three main data architectures, from broader to narrower reach, that can accommodate the needs we have laid out:

- Enterprise architecture: Covers integration of end-to-end data from EHRs and operational data collection systems into enterprise data warehouses (EDWs), whose data are made accessible through topical data marts
- Health care specific analytics topics: Designed to accommodate broad needs across large topics of analytics such as quality, finance, supply chain, operations, and research
- Targeted analytic solutions: Cover specific areas of analytic needs for operations, a specific practice area, quality reporting/analytics, and specific research needs (like disease registries, cohort identification, population analysis, studies, or omics analysis)

We will examine each of these data architectures to further describe the purpose, approach, and design to identify which ones are best for what needs.

Enterprise Architecture

The most significant challenges organizations face in providing appropriate access to all the right areas/individuals in a timely fashion to support organization wide analytics is the data itself. While data in other industries are much more defined and structured, as such they enable technology to tie systems together to leverage data from disparate systems. However, in health care different standards, definitions, and even capture make actually pulling different data fields from various systems virtually impossible. Unimaginable funds have been spent with vendors to try to build the ultimate technology that eliminates the need for data integration/standardization, yet all of the successful solutions have this as a foundation of their architecture. Thus, the enterprise architecture focuses first and foremost on data integration to enable analytics. Highly complex and even private data-like research data can be fully integrated as long as governance puts in proper controls on usage and viewing.

The main focus areas of an enterprise analytics solution are data standardization, data integration, data preparation, and data delivery. Figure 8.1 depicts the connectivity of a formal enterprise analytics environment.

Data standardization requires a formalization of data definitions, valid values, terminologies, and ontologies, resulting in standards for how data are captured or integrated. If data are standardized at the source, it is much easier to integrate and leverage. If not, standardization is incorporated into the data integration process. With the mass of different data from difference sources (like practice-specific systems and research study data), it is unreasonable to expect all source data to be fully standardized; thus, data integration is critical.

Data integration leverages business rules and standards (ideally from a formal data governance or standards area) to build the rules and systems that perform the integration. This information is stored in some database designed for storing or delivering data, most typically an EDW.

Data preparation then takes the integrated data and defines the data tables/structures that are designed for ease of getting data out (i.e., data marts, mining marts, and super marts). These structures are designed and built in very similar fashion to the category of healthcare specific analytics topics. While the goal of data integration

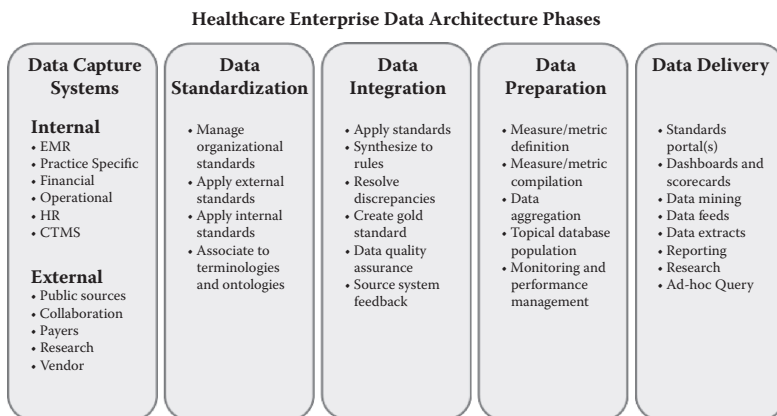


FIGURE 8.1

Components of the data life cycle within an enterprise.

is to get the data in, the goal of preparation is to make it easy for application and tools to get it out.

Data delivery is thus aimed at getting data to the hands of users in an easy-to-use, timely fashion, supporting their needs. Leveraging business intelligence (BI) tools, analytics tools, and portals for information delivery is much faster and more effective with a sound enterprise architecture than with standard, siloed sources. Organizations having bad data often tend to find good tools ineffective and draw a negative response to those tools. On the other hand, organizations with sound architectures of integrated data can even make moderate to simple tools look excellent. Many BI vendors are now focusing on that target market for their tools and being very open with customers about the separation of tools and data.

Health Care Specific Analytics Topics

One common solution approach less comprehensive than the enterprise-wide strategy is to target solutions around specific practices or areas of analytics needs. Having a handful of separate data warehouses that can serve these areas enables businesses to prioritize and work on these efforts independently. When not a part of an enterprise approach, each of these solutions builds its own data extracts to integrate and populate data for the purposes desired. For small- to medium-sized organizations, this can be an effective approach and help them keep all of the associated costs under control. For large organizations or those that have many practice areas, it can quickly get out of control to the point where there are hundreds of resources pulling from the same systems but using data and metrics quite differently while incurring repeated costs for the hardware and services to manage them all.

Targeted Analytic Solutions

Targeted solutions are often vendor products for specific needs. While many vendors will say their data warehouse is enterprise in fashion, it should be quickly discernable that it indeed covers only the purpose at hand. These solutions can be quite valuable for that one need but tend to proliferate, and suddenly there are dozens or even hundreds of them. Taking a single solution approach to all analytics would be akin to not implementing an EHR but instead buying or building a data capture

system for every health care need. When leveraged by an enterprise data architecture, these solutions can be very effective, but as a standard approach for an organization this approach should be taken only by small, very targeted organizations focusing on one practice area. These solutions can be very cost-effective for what they do, but when one considers the mass amount of analytics needs and the constant emergence of new needs the ability to keep up quickly inundates those that choose solely this approach.

DATABASE DESIGNS

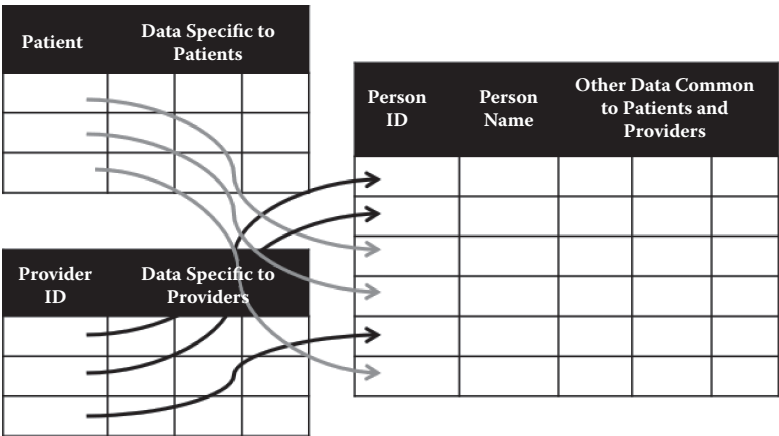
Several aspects to database design need to be understood and compared before determining the exact analytics architecture for an organization. While each can be very effective with the right architecture, they can be equally ineffective in the wrong architecture. Each of these databases needs to be modeled, designed, and implemented somewhat differently based on their inherent uniqueness. The incorporation of complex data types, images/video, and unstructured data also needs to be carefully considered. While all of these data can be fully integrated, they may overwhelm the structure. Leveraging connections and linkages from the various databases to the complex data sources, like imaging picture archiving and communications (PAC) databases, is often best as long as it doesn't impact the source system negatively—looking up one record at a time shouldn't cause that. For targeted mining of said data, it is effective to pull subsets to data-mining solutions for in-depth analysis. For many analytics and certainly population health and research, external data are required as well. Big data is evident in combining clinical and operational data with externally available data on populations, biomarkers, and even aggregate statistics.

Before explaining the individual designs, we will address the *data modeling techniques* that are most effective. Most data warehousing professionals think of data modeling as either dimensional (typically used for data marts or EDWs that leverage common dimensions) or normalized in nature. Normalization is actually defined more as the level of relationship detail that is established in entity-to-entity relationships (historically called entity relationship modeling) and thus is very similar to what occurs in dimensional models as well. Highly normalized models

are driven down to a very detailed level of relationships—such as a patient would have one-to-many addresses or even one-to-many email addresses. While this may be true, it is often counterproductive to establish so many email addresses for a patient when you really want to know where to get a hold of them (and thus focus only on their primary address).

Denormalized models tend to have very few relationships, and as such all patient data would fall into a single patient entity (Figure 8.2). The concept of abstraction is generally more associated with denormalization (although it can be seen in fully normalized solutions as well, which often presents enormous challenges and is only really suited to operational systems, not data warehouses). It takes an entity to a generic level—like patients are people and by categorizing them as such all patients would be in the Person entity. While this generic representation makes modeling easier, the characteristics of the various people you need to include in the system are quite different, and thus the meanings of the individual fields in the generic table tend to easily get misconstrued, making IT programming and end-user data analysis much more difficult.

One other key modeling consideration is the concept of views of data. One of the most common mistakes in modeling health care data is to try to model the concept of an *encounter*. It gets so large that every data item that relates to care is bulked into an encounter. It is also complicated by the concept that an inpatient encounter is different from an outpatient encounter.



(a) Denormalized and Abstracted

FIGURE 8.2
Differences between normalized, denormalized, and abstraction in database design.
(Continued)

Patient Address	All patient addresses		
Patient Email	All patient email contact info		
Patient Phone	All patient phone contact info		

Patient	Data Specific to Patients		

Provider Address	All provider addresses		
Provider Email	All provider email contact info		
Provider Phone	All provider phone contact info		

Provider	Data Specific to Providers		

(b) Normalized and Non-abstracted

FIGURE 8.2 (Continued)

Differences between normalized, denormalized, and abstraction in database design.

When modeling, this duality of the encounter concept often leads to disagreements from the subject matter experts who provide the context required to model and ultimately only one viewpoint wins. I have seen organizations spend years trying to model encounter only to give up. However, if we separate the view from the data we can enable all views to be considered for the appropriate need. This separation is a critical component of a successful enterprise data model working effectively with effective data delivery for servicing all needs with the least effort but most accuracy and consistency.

The main database designs for a health care data architecture are EDW, operational data store (ODS), data marts, data mining, and super marts. Depending on the overall architecture, each of these can be used by itself or together (for the most effectiveness). A technical enterprise data architecture outline is included in Figure 8.3 to reference the main data movement required to support enterprise health care analytics.

We now examine the effective designs of each database.

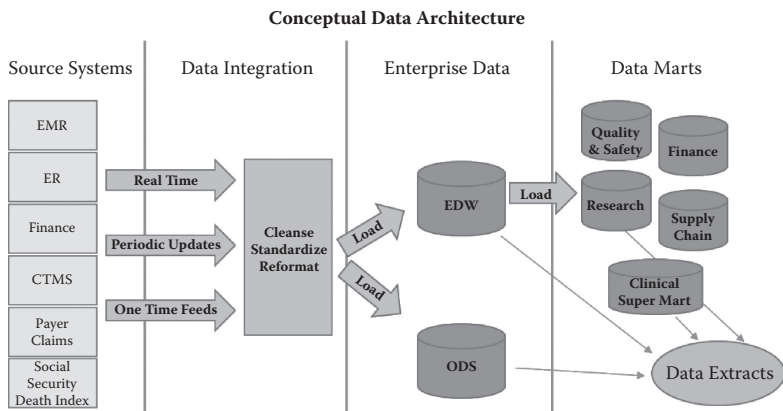


FIGURE 8.3
Main database designs for health care data architecture.

EDW

An EDW is most often thought of as the source of truth. While simply put, the word *enterprise* implies that all data from across the enterprise is contained and in a fashion that enables all enterprise analytics. However, too often databases are called EDWs but contain data for only one or a handful of purposes, discrediting the term and misleading users or executives who are dissatisfied that they cannot get what they need. An EDW is also not a simply a place to drop data—that would be a data dump—requiring IT resources to pull data for each and every need. EDWs are most often designed to be populated from source systems (including external sources) on a regular basis, with specific medical transaction data flowing through daily or weekly. Through the process of data integration with defined data standards, the source of truth aspect emerges. Considering the mass and variety of available health care data that are used for the plethora of purposes, this is where the real-world concept of big data takes its form. Data models for EDWs usually take on a mid-level of normalization, referred to as third normal form (the same for dimensional EDW models). This makes them easier to populate and deliver data for usage. While the most common effective EDW designs are not dimensional, there are some very good dimensional models as EDWs. This requires an organization to have strong data governance and controls over the definition and depth of all dimensions. While many have stated that EDWs evolve over time, data models and databases themselves in fact are not designed to evolve. They must grow in content, but any structural changes

(e.g., adding tables or columns) are difficult to put into effect. For example, if you want to redefine how you identify a patient, you will have to design the tables, which would require reloading all data from the original sources (assuming you have all history) and rebuilding all of your analytics. In all likelihood, the numbers you had may change, leading to significant confusion among all users. Thus, having the right design and approach to your EDW and entire data architecture is a key aspect of success in both the long term and the short term.

One key aspect of an EDW is to get data out only through programmatic access. By connecting data marts to the EDW, data can be reloaded into data marts frequently to keep them up to date. Most analytic usage of data warehouses is focused on tabulating historical data, looking for correlations or trends over time. If users were allowed to access EDWs while they are operating (i.e., during patient care), it could result in impacting the analytics users negatively during load processes or from data changing from request to request, resulting in a query ran one minute having one result and later having a different one. This would likely result in a lack of trust in the quality and content of the data.

ODS

An ODS is very similar in context to an EDW, but slightly scaled down. It is designed to serve immediate, sometimes even real-time needs for operational access to data. In the normal operations of business, key information is required that often resides on different platforms to make quick, accurate decisions or even to monitor what is currently happening across the organization. It still has the needs to standardize and integrate data but is focused only on serving operational purposes. And oftentimes data are populated more frequently, perhaps even leveraging real-time feeds, like patient registration, admit-discharge-transmit (ADTs), or orders. Thus, it is usually much smaller than a full-fledged EDW.

The data modeling that is used for an ODS is usually third normal form or is somewhat denormalized to fit the operational analytics requirements. The biggest challenge comes when organizations try to utilize ODS structures to conduct all analytics. Combining some real-time data feeds with complex analytics will result in users looking at analytics one minute and finding one result and the next, getting a different result. Enabling direct access to an ODS is acceptable as data access for operational purposes in an ODS is targeted at watching real-time data as it changes.

Data Marts

Data mart structures are specifically designed to ease use in accessing and getting data out a data repository. Leveraging star schemas for simple data access for dashboards and scorecards, they become very effective. In other industries the concept of departmental data marts is considered a standard, but if this approach is taken in health care it could result in thousands of data marts. If not sourced from an EDW, this would mean they all have different data and results, which is terribly inconsistent for the organization. Thus, star schemas should be designed by topic, like a quality-centric data mart, which can serve needs for quality reporting and quality improvement and enable sharing for external reporting and even send data feeds. This topic-specific strategy also limits the numbers of data marts, making their production and response much more effective for serving various areas across the enterprise.

Data Mining

Snowflake designs can also serve groups of large analytics needs but are more focused on in-depth analysis needs. Having a handful of snowflakes by analysis area—like financial and research—enables in-depth analysis and incorporation of really complex data types. The primary goal of these databases is to enable the many analysis tools used across the organization to have a source to go against or extract from.

Super Marts

A super mart is a very large snowflake or star schema that is specifically designed to satisfy many needs from across the enterprise. Seldom should any organization have more than a handful of super marts, as they are quite complex. This limiting of proliferation keeps the creation and maintenance costs and time frames to a minimum while providing mass amounts of data for analysis.

Analytics

Analytics is all about getting accurate and comprehensive data into the hands of those who need them, when they need them, with security and performance that enables them to focus on the domain challenges (and not on the IT and informatics involved). The wide range of analytics available

is described elsewhere in this volume. With the almost unlimited needs for data access and ways to measure and monitor health care data for various purposes, having a defined approach within a technical framework and via an approved toolset becomes a critical efficiency and cost issue. Too often health care data is thought of and delivered only via reports. Reports are focused only on operational review and are somewhat useless for most analysis and monitoring—what is mostly needed across the analytics of health care organizations. Figure 8.4 depicts a few examples of the various areas of need and the methods they might use to get access to analytics. A formal diagram of analytics across health care would be much more complex—so much so that the lines would likely blend into one another from the mass of needs. Which data warehouse tools required depends upon the chosen architecture approach and design. In an EDW design, only some data extracts and feeds would actually access the data warehouse.

Receiving the most value from analytics in a timely fashion starts and ends with access to the right data at the right time. While analysis tools can be extremely effective, they strictly are dependent upon complete and accurate data. There are many analysis and BI tools that all are very effective at delivering data. Yet most BI vendors in particular have had a hard time serving the needs of health care, and health care organizations generally are displeased with them. The real root of this issue lies in organizations hurrying to dump data into the tool and relying on the tool to

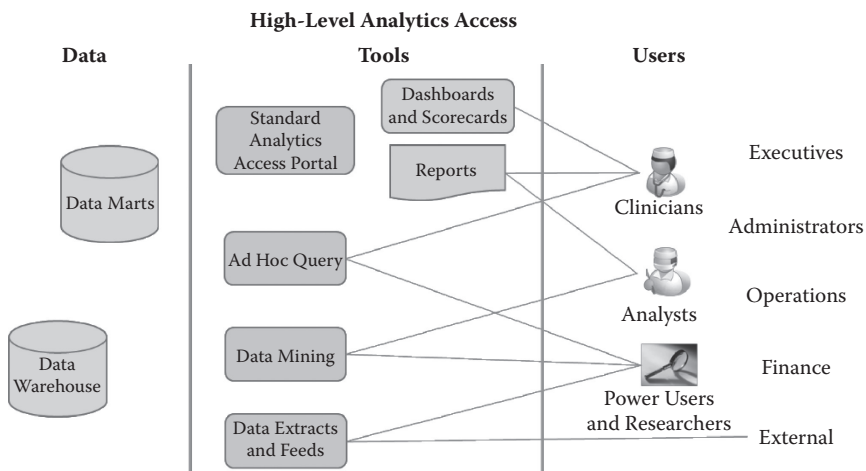


FIGURE 8.4

Relationship among data, tools, and users for analytics.

work its perceived magic while ignoring inconsistencies and inaccuracies inherent in the data. Making each one of the categories of analytics a standard is a critical part of enterprise data architecture.

In general, several categories of analysis tools need to be used across health care:

Portals: This type of access enables the easy distribution of analytics in a consistent fashion. If an organization has a good portal to deliver analytics of all kinds, all users go to one spot to see their data and each user type or usage function can still see a unique view and leverage their tools. Too often each need or area develops its own portal, and like its own data warehouse, it proliferates out of control quickly.

Analytics tools: While a goal to have one standard tool across the enterprise is noble, it is also very limiting and ineffective as different tools specialize in different analytics. With the mass of health care data analysis needs, it is inconceivable to limit all analytics to one tool. That said, it is also a logistics nightmare to try to let every user or need have its own tool, which results in the inability to build template and solutions that can be reused for similar purposes causing an organization excessive expenses for servers, storage, and IT resources.

Extracts and feeds: Health care organizations are inundated with internal and external needs for sharing data. Various frequent examples include external reporting, data sharing internally, data sharing externally (for research and collaboratives), and targeted extracts for complex analysis and research. Most health care organizations build each one of these interfaces separately and too often have to build each one from scratch, including identifying and mapping data back to original source systems. This is extremely time-consuming and costly. It also greatly limits their ability to respond to emerging analytics and reporting standards and guidelines. Having a foundational source of integrated data enables an organization to build a framework for feeds from a known source that streamlines their efforts.

Vendor solutions: There are many vendor solutions that target specific analytics or research areas. Many of these solutions have great pre-build analytics that can be easily incorporated and receive significant value from. The concept of leveraging work already done on these

analytics can be a huge timesaver to any health care organization. The challenge is almost always the access to data and making sure it is consistent across the enterprise. An organization with a formal data architecture and integrated data is much more capable of achieving successful implementation and use of these tools, often rather quickly.

Analytics also requires the leveraging of internal and external metrics standards. When dealing with external organizations and governmental regulations, numerous metrics are partly or fully defined. These common needs can enable practice areas to establish standards such as the National Surgical Quality Improvement Program (NSQIP) has done to bring consistency to how data are measured within and across health care organizations (<http://site.acsnsqip.org>). As many standards are emerging or are yet to develop, there is also significant need for individual organizations to develop internal standards for how they view data. Key definitions are required and include what is a visit, how to report quality of care, operational effectiveness—like patient flows and volumes. Many health care organizations have multiple sites and locations that they need to look at collectively. Without those standards, they are often reporting on events slightly differently, which results in confusion of what is happening and management challenges for leadership.

TIES TO GOVERNANCE

Architecture without governance or standards is a project without a purpose. Governance or standards without an architecture is an exercise in discussion and debate—without any ability to identify if it will work but a lot of pride in thinking it is well constructed. One of the significant challenges in the world of data for health care is to bridge the expertise of medicine with technology. This requires resources on both sides to leverage their areas of expertise and to rely on the other to provide theirs. A well-aligned governance organization coordinates both areas working collaboratively under common guidance. Information technology is often thought of by users of all industries as programming. Yet aligning data, needs, and methods across those groups is what is truly required to build effective systems that impact business.

The complexities of health care make this even more relevant and pertinent. Owning data quality in health care is a business function. As the business works with IT to identify how to do that, IT resources build the systems that enable governance to own data quality. When dealing with enterprise data, this becomes even more crucial and forces a shift to an enterprise focus. Measures and metrics are yet another area that the business must define and own. When dealing with defining populations and creating metrics that can be used to measure the health of those populations, the business again must own this. IT works collaboratively with them to create views and solutions that provide that information in a timely and easy-to-use fashion to those who need it. Data governance is also the avenue for ensuring the privacy and protection of data. Data breaches are commonplace in health care today. With significant needs for data, there must be the appropriate controls to ensure the protection of patient identifiable information. Creating controls requires the definition and implementation of data privacy rules as well as the monitoring and reporting of who has access to what data for what purpose. Research has a special case in the place in the privacy discussion. They have additional rules and guidelines they have to comply with in the process of their research and as such need to make sure they are governing all research activities.

BEST PRACTICES

A nonprofit health care best practices collaborative called the Healthcare Data Warehousing Association (HDWA; <http://www.HDWA.org>) consists of members from several hundred health care organizations across North America. It is a volunteer organization that serves to enable sharing and thus advancement in analytics. Organizations that can leverage the work of others can respond much more rapidly to meet their needs. With the pressing need for access to data for reporting and analytics every organization faces, HDWA focuses on providing an environment for organizations to share what they are doing and what they have learned and in turn to learn from the work of others. This collaborative has helped many organizations of all sizes define, refine, and deliver solutions within their own organizations more effectively over the last decade.

While it is easy for technical resources to focus on the technical aspects of any analytics need, the business purpose and reason come back to the Triple Aim requirements. In quality improvement, for example, if we don't impact one of the three aims we should question why we are not working on things that will since there are so many areas to target. Through my work with the HDWA, I have seen many approaches, solutions, designs, and implementations intended to tackle analytics and reporting challenges in health care. Many of the bigger organizations have had the resources to learn and focus on this challenge.

I have spent time with four in particular (apologies for not including every organization that has achieved analytics successes!) that to me epitomize the success of leveraging analytics to impact their practice and research. Each of those four would tell you that it is not a destination but a journey. They all still know there is much more to be developed and incorporated into their solutions. But all are well equipped through their efforts to respond to additional and emerging needs. Each of these organizations has a somewhat different data architecture, technical tools, and even organizations of their resources. All four are also renowned for their analytics and quality improvement expertise—in alphabetical order) Banner Healthcare, Geisinger, Intermountain Healthcare (2014), and Mayo Clinic (Chute et al., 2010). All four of these organizations have enterprise data architectures with the following features in common:

- Strong engaged leadership who understand the critical role of data to the organization and as such provide oversight, resources, and governance
- An organizational commitment to quality, not just in words but action; evident in the care they provide and also in the solutions they deploy for analytics
- A strong relationship and interaction between the technical teams, data governance, and subject matter experts from across the organization
- A formal enterprise data architecture; while the specific designs of each of these organizations may differ, each of these organizations has formal data governance, a fairly mature EDW, and solutions across analytics spectrums
- Many clinical examples of using factual data for quality improvement that is written, published, and shared via medical media or events

- Two-pronged focus on standard reporting/analytics and advanced/emerging analytics
- Effectively providing for the various reporting/analytics and external data feeds while also leveraging those data to support research and clinical operations
- Willing to share and to collaborate with other organizations on their data architecture and analytics learning, which has helped each of them in turn grow their solutions and expertise

CONCLUSION

Significant amounts of money have been spent by health care organizations, academic centers, and technology vendors to address health care's specific analytics needs, but little ground has been made to provide the foundational environment required to enable organizations to shift and adapt as the analytics requirements around them move—with confidence and accuracy and without radical financial output. Big data is no bigger and more complex in any industry than in health care (e.g., images, clinical notes, documents, research studies, genomics). Big data in health care has many variables and almost unlimited analytics needs that can overwhelm any organization.

While it is perceived that health care lags most all other industries in use of technology and data, this perception derives from the data and information being so much more complex, from data not being captured or simply being unstructured or from participants being inundated with analytics requirements, whose volume fails to promote learning. To further the cause of quality improvement and medical research, it is imperative to enable proper access to big data with sound controls. While it is easy for health care organizations to focus on the most complex data (like omics), it is equally critical to understand the complex views and correlations across all data.

As standardization of data and analytics matures across health care, the technology concepts of automated pulling of data from many sources and assembling with a tool for each need may turn into reality, but as of this writing those methods are successful only on a single-solution basis or in industries where data are much more defined, refined, and consistently applied. Database management systems (DBMS) and associated

hardware are still very limited in their ability to allow access for simple or mass query while also enabling real-time entry without impacting those doing entry. (We would never allow a query on a production system if it meant providers couldn't enter data while in the room with a patient.) As the most successful organizations have made incorporation of a formal data architecture and associated business led governance the foundation of their approach, they still are constantly in learning and improvement mode. Organizations that strive for excellence will focus on factual data to drive hypothesis, quality improvement activities, research, and operational excellence.

REFERENCES

- Codd, E.F. 1971. Further Normalization of the Data Base Relational Model. In Randall J. Rustin (ed.), *Data Base Systems: Courant Computer Science Symposia Series 6*. Upper Saddle River, NJ: Prentice-Hall.
- Chute, C.G., S.A. Beck, T.B. Fisk, and D.N. Mohr. 2010. The Enterprise Data Trust at Mayo Clinic: A Semantically Integrated Warehouse of Biomedical Data. *Journal of the American Medical Information Association* 17(2): 131–135.
- Intermountain Health Care. 2014. Data & Analytic Resources. Available at <http://intermountainhealthcare.org/qualityandresearch/researchmanagement/resources/Pages/dataanalyticresources.aspx>.
- Millard, M. 2014. New To-Do Lists for Post EHR Era. *Healthcare IT News*, January 30. Available at <http://www.healthcareitnews.com/news/new-do-lists-loom-post-ehr-era?page=1>.