HealthCatalyst
ignite outcomes improvement™

Comparing the Three Major Approaches to Healthcare Data Warehousing:

A Deep Dive Review

by Steve Barlow

Co-founder &
SVP Client Operations

Health Catalyst®

Health systems are being asked to deliver better care and boost productivity while simultaneously reducing costs and waste due to the U.S. government's Affordable Care Act (ACA). Government mandates, subsidies, and health insurance marketplaces are all part of the mix of incentives to increase compliance for the ACA and ultimately improve the delivery of healthcare and reduce Medicare spending.

The task to improve healthcare presents a significant challenge to providers, health systems, and payers. But according to the Institute for Healthcare Improvement (IHI), if health systems focus on achieving the objectives of a framework called the Triple Aim, they will be able to optimize their performance and meet the government's requirements. As stated by the IHI's website, the Triple Aim objectives are as follows:

- Improve the patient experience of care (including quality and satisfaction)
- Improve the health of populations
- Reduce the per capita cost of healthcare

Copyright © 2017 Health Catalyst

Before health systems can focus on achieving the Triple Aim, however, they need to redesign their system with the following components listed on the IHI's website: (1) focus on individuals and families, (2) redesign primary care services and structures, (3) focus on population health management, (4) implement a cost control platform, and (5) achieve system integration and execution.

> A healthcare-specific enterprise data warehouse provides complete and accurate information from across an entire organization.

Changing health systems to fit this new model is possible. But it is critical for health systems to choose the most appropriate data warehouse for healthcare's specific needs as they redesign their systems. This is because the traditional approach of cobbling together reports that pull in data from the many various source systems is too costly and time-consuming. A better approach is for health systems to increase their reliance on complete and accurate information from across the enterprise-wide data ecosystem of their organization, which requires a healthcare-specific data warehouse. Once this data warehouse is implemented, health systems will be able to store and mine their enormous amounts of data to achieve the Triple Aim.

WHY ACHIEVING THE TRIPLE AIM IS CRITICAL FOR HEALTH SYSTEMS

A 2014 report from the Commonwealth Fund reveals that the U.S. healthcare system is the most expensive in the world and ranks last for access, efficiency, and equity in comparison to 10 other nations. The 10 nations from the report include Australia, Canada, France, Germany, the Netherlands, New Zealand, Norway, Sweden, Switzerland, and the United Kingdom. In addition, a 2013 report from the Centers for Disease Control and Prevention (CDC) states that personal healthcare expenditures in the United States total $2.3 trillion, with expenditures for hospital care accounting for 31.5 percent and physician and clinical services accounting for 20 percent of all national healthcare expenditures. Yet despite the high cost of U.S. healthcare, Americans are not any healthier than citizens of other industrialized nations, nor do they enjoy greater longevity. Consider these facts about the costs of healthcare and mortality:

- The 2013 report *U.S. Health in International Perspective* compared the life expectancy of Americans to the citizens of 17 high-income peer countries from Western Europe, Australia, Japan, and Canada. The findings showed that life expectancy for American males ranks last, and life expectancy for American females ranks next to last.
- Preterm-related causes of death accounted for 35 percent of infant deaths in 2009 as stated on the CDC website.

Copyright © 2017 Health Catalyst

- At least 44,000 and perhaps as many as 98,000 Americans die in hospitals each year as a result of medical errors according to the book *To Err Is Human: Building a Safer Health System*.
- Medicare could save at least $12 billion per year by reducing preventable readmission cases that are readmitted within 30 days according to the 2007 report to Congress *Promoting Greater Efficiency in Medicare*.

These facts highlight merely a few of the problems healthcare is facing. To respond to these issues, there are many data warehouse choices being developed and marketed to health systems. Knowing which model will be the most effective and provide the best return on investment (ROI) can be difficult until the advantages and disadvantages of each option are understood. Then, with this knowledge, health systems will be able to make a well-informed decision about their investment.

3 TYPES OF DATE WAREHOUSE MODELS: ENTERPRISE MODEL, INDEPENDENT DATA MARTS, AND LATE-BINDING™ ARCHITECTURE

Currently there are three main types of data warehouses from which health systems can choose to store and mine their data. The data warehouse models are as follows: the enterprise model, the independent data mart model, and the late-binding™ architecture model. While all three models offer a data warehouse solution, some have unique attributes that make them ideal for healthcare.

1. Enterprise Model

Bill Inmon, called the "Father of Data Warehousing" on his website, developed the enterprise model for data warehouses. This is a complex, top-down design that includes the construction of a big centralized data warehouse from the outset of the planning stages.

By using the enterprise model approach, it is necessary to determine in advance all of the data elements anyone would ever need to use for data analysis, such as safety and patient satisfaction data. Analysts are forced to make lasting decisions about the data model in the beginning without being able to plan for changes in the short- or long-term. And then they need to structure the database accordingly, which can take months or even years to complete.

> In the healthcare analytics environment, the enterprise model is difficult, expensive, and time-consuming to construct.

For certain industries, such as manufacturing, banking, and retail, or when there is a need to design a new transaction processing system, this model may be appropriate. But in the healthcare analytics environment, the enterprise model is difficult, expensive, and time-consuming to construct because data

Copyright © 2017 Health Catalyst

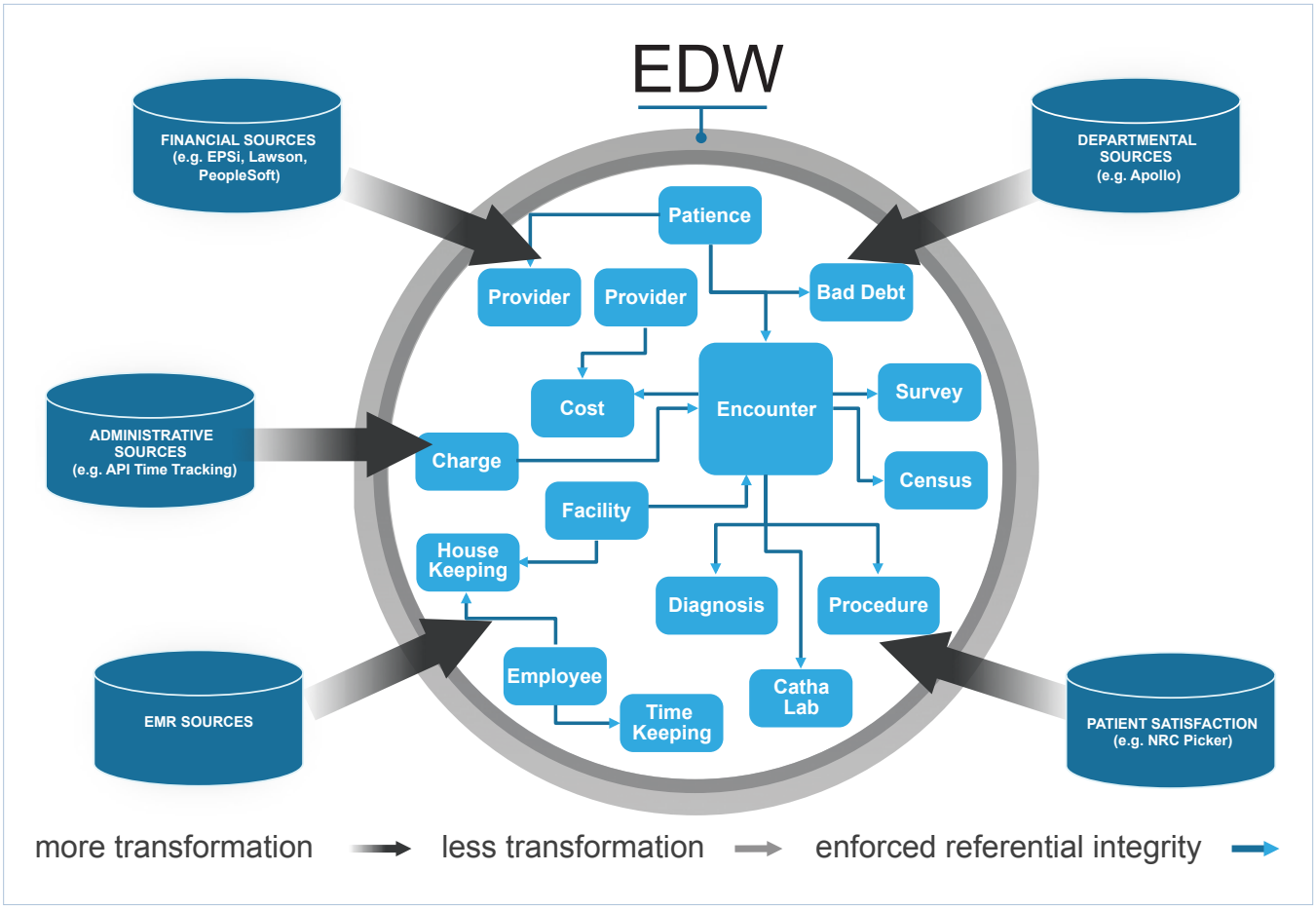

Figure 1: The traditional enterprise model requires significant transformation of the data model from the source systems (dark blue ovals) as it is loaded into the enterprise data warehouse (light blue rectangles).

architects must first build a comprehensive blueprint of various data elements, such as medications, labs, and billing data, for example. To further complicate the process, the blueprint is not necessarily based on data that has already been captured — just the forecasted data needs. As a result, the data model may include superfluous data elements. Additional limitations of the enterprise model approach are listed below.

Delayed Time to Value

The enterprise model approach creates additional expense and time for the health system because of the considerable transformation required to force fit the data into the net new data model. This delayed time to value is a significant downside of the enterprise model approach. For example, complex calculations and derivations tend to add increased work and time to an analyst's job. In comparison, other data models allow for inputs to a calculation to be loaded directly into a data warehouse, allowing for greater flexibility and faster delivery of reports.

Copyright © 2017 Health Catalyst

Obscures Data Quality Issues

The enterprise model has rigid acceptance criteria for the data, causing the need to clean and scrub the data each time it is loaded from the primary system into the secondary system. This method obscures data quality issues and delays the improvement of data quality issues at the primary source. For example, suppose a health system wants to measure gestational age for mothers because studies have shown that inducing labor before 39 weeks increases the risk of complications. What the health system will discover as it goes to pull the data from its EMR is that the data has been captured in many different ways. Some entries may show "39.2," "39 weeks and 2 days," "39W and 2D," or "thirty-nine weeks and two days." The variations go on and on, making it impossible to easily and accurately use the data without a significant cleansing effort.

Data Is Bound Early

Top-down data warehouse models require early binding of the data. (Data binding is a technique in which raw data elements are mapped to conceptual definitions.) Early binding means the data is mapped into a predefined data model as it is brought into the warehouse, which limits the ability to make changes to the data in the future. For situations where data rules are relatively static, nonvolatile, and do not frequently change, early binding may be appropriate. Industries that employ early binding models include manufacturing, communications, retail, and financial services.

When health systems try to bind every data element to business rules early, however, they face a time-consuming and expensive approach to data warehousing. It is also difficult to make changes to the data. This is because business rules and vocabulary standards in healthcare are among the most complex in any industry, and they undergo almost constant change, resulting in high volatility.

In fact, there are only a limited number of core data elements that should be bound early because they are fundamental to almost all analytic use cases. Because they are fundamental and not volatile, it is appropriate to bind those

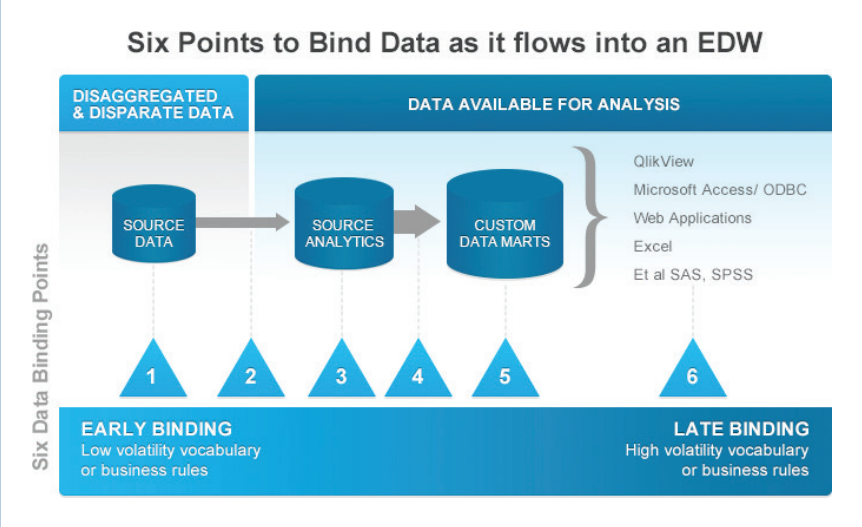

Figure 2: This illustration details the six points in a data warehouse where data can be bound to rules and vocabularies. Rules and vocabularies with low volatility can be bound at points 1 and 2 (early binding); points 4 and 5, are appropriate for those with high volatility.

Copyright © 2017 Health Catalyst

data elements early. All other data elements should be bound late (i.e., when clinicians are trying to solve a problem) because of their volatility. For example, length of stay (LOS) in a hospital may sound straightforward on paper, but surgeons might define LOS as point of incision to discharge from the post-anesthesia care unit (PACU), and cardiologists might define it as emergency department (ED) arrival to discharge. Because the LOS definition will change for different use cases, the objective is to bind it later. The following examples show which types of data are volatile and which types are not volatile.

Volatile data that should be bound late:

- Calculating length of stay (LOS)
- Attributing a primary care provider to a particular patient with a chronic disease
- Calculating revenue (or expense) allocation and projections to a department or physician
- Data definitions of general disease states for patient registries
- Defining patient exclusion criteria for disease and/or population management
- Defining patient admission, discharge, and transfer rules

Nonvolatile data that may be bound early:

- Facility identifier
- Provider identifier
- Patient identifier
- Gender
- Date
- Time of arrival

> When there is boundless scope and complexity with a data warehouse project, the final product may miss the mark or completely fail to deliver the functionality leadership expects.

Boundless Scope

Healthcare business processes can be complex, and teams can spend anywhere from six months to multiple years mapping their organization's information systems to a single enterprise-wide data model. To account for this intricacy, the data model can become enormous in scope and complexity and may miss the mark in terms of having the functionality leadership expects. In fact, Gartner estimated in 2005 that as many as 50 percent of data warehouse projects

Copyright © 2017 Health Catalyst

would have only limited acceptance or fail entirely through 2007. When a project fails, a lot of money may be spent on something that is never used or never even gets launched.

2. Independent Data Marts

The independent data mart approach to data warehouse design is a bottoms-up approach to data modeling.

With this data model approach, the organization starts small, building individual data marts as places to store specific information for each hospital department. Then the independent data mart draws further department-specific data from the various primary source systems to provide the data that each department needs.

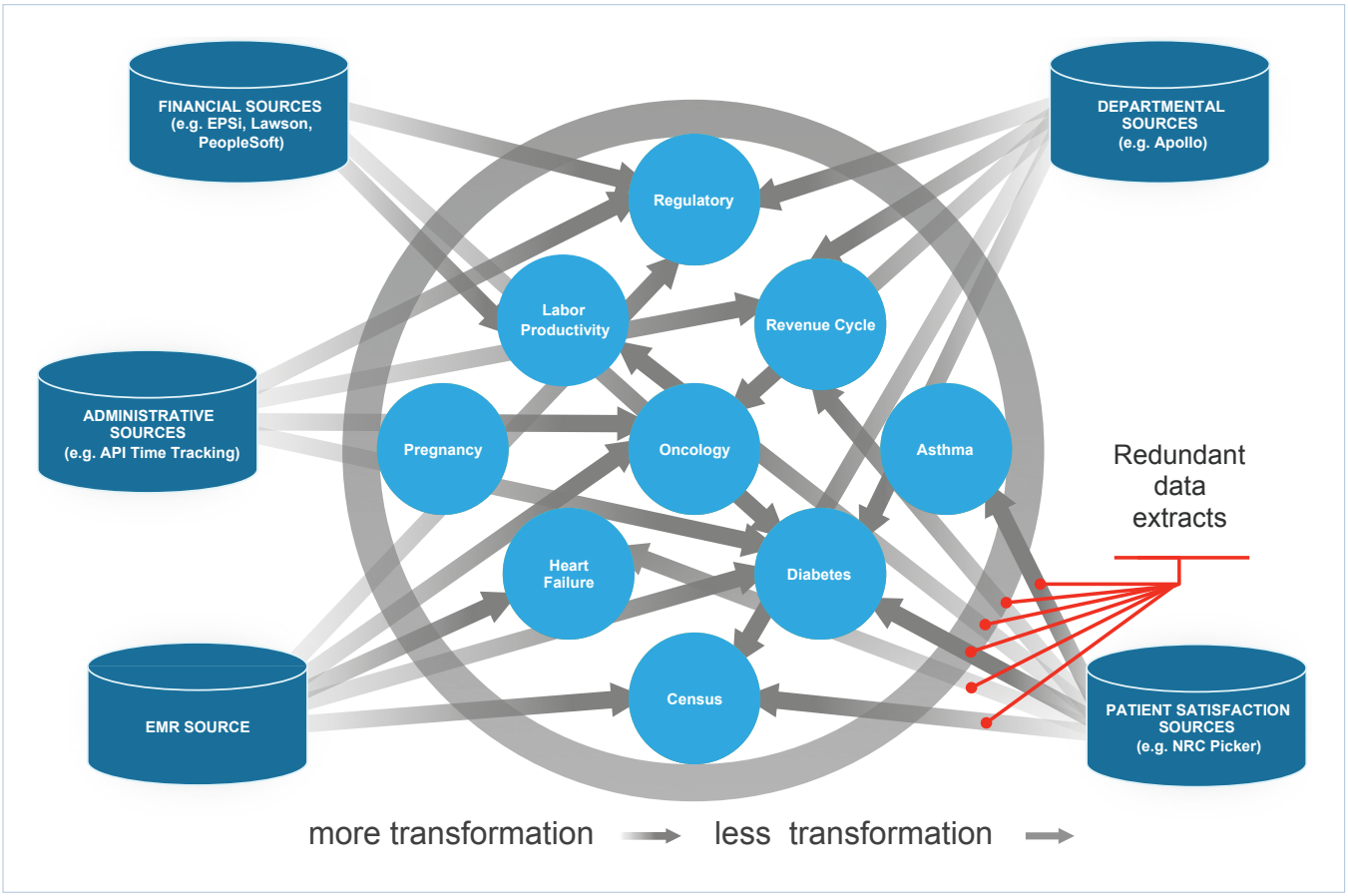

Figure 3: With the independent data mart model, an organization builds an analytic data mart for a particular department — such as heart failure — gathers the data it needs directly from the source systems, and maps it to different areas.

Using the independent data mart model approach, analysts can create and build analyses and reports, which are then combined to form a large data warehouse. For example, if there is a need to analyze revenue cycle or oncology, then a separate data mart is built for that specific department.

Copyright © 2017 Health Catalyst

Data is then loaded into this new independent data mart from the primary source systems to support this new subject area.

There is a benefit to the independent data model approach: it takes less time for the organization to build the data warehouse, and analysts can start to analyze data quickly — a big difference from the two- to five-year lifecycle of the enterprise model. However, it grows very quickly, as do the data streams, until several redundant streams exist. This creates a challenge for those trying to maintain the model. If one underlying source system changes, they have to change each extraction routine that uses that particular source. The many isolated data marts also means there is not an atomic-level data warehouse from which to build additional data marts in the future. There also several other major drawbacks to this model, which are explained below.

Lacks Granular Data

Often, independent data marts do not contain data at the lowest level of granularity or patient-level detail. (In data warehousing, granularity refers to the level of detail stored in a database and how that level relates to other data. For example, one database list might store patient names; another list might store individual patient encounters, a finer level of detail.) Instead, data transformed in a data mart is usually summarized up one or more levels from the lowest level of granularity. This means the data mart may present information about a certain metric falling below the benchmark, but it does not contain the granular data that enables the analyst to dig down and determine why that metric is low. Without that more detailed information, it is difficult to make the data actionable and to determine how to bring that metric up to the benchmark.

Lacks Efficiency

The independent data model causes source systems to be repetitiously and unnecessarily bombarded by data extracts, which slows down the system. Redundant feeds from each source system need to be built to feed each independent data mart. This creates a challenge for those trying to maintain the model. If one underlying source system changes, then each extraction routine that uses that particular source needs to be changed. It also results in a significant drain on the system. Imagine building a new feed from the electronic health record (EHR) into every data mart built: heart failure, pregnancy, asthma, diabetes, oncology … and the list could go on and on. The single-purpose “point” solution may have addressed the immediate need, but as the demand for analytics grows, the mass of redundant feeds from the same primary systems creates a solution that is difficult to maintain.

Requires Early Binding of Data

Like the enterprise model, the independent data source model requires early binding of the data. As data is brought into each independent data mart, the

Copyright © 2017 Health Catalyst

data is mapped into the predefined data model, a process called conformance and normalization. The terms imply exactly what is required: data that was modeled and captured in disparate source transaction systems must conform to a new data model in the data warehouse.

While at first this approach might appear reasonable, in practice, it leads to major problems when applied to the healthcare industry. This is because the data environment is much more complicated than a sales receipt in the retail industry, and the analytic use cases are constantly changing. For example, the process of mapping and conforming data to these early binding models in a healthcare delivery data warehouse typically takes 18 to 24 months or longer. When new data sources are added to the data warehouse — as occurs in mergers, acquisitions, and ACO partnerships — this lengthy time- to-value is repeated again and again.

> The healthcare data environment is much more complicated than a sales receipt in the retail industry, and the analytic use cases are constantly changing.

Likewise, as the complexity of analytic use cases inevitably matures in an organization, the early binding data model must be modified and the source system data must be conformed and mapped again. These early binding data models cannot keep pace with the changes in the analytic environment and the data warehouse subsequently fails to deliver.

3. Late-Binding™ Architecture

A new approach to data modeling to address healthcare's unique data needs is Health Catalyst's Late-Binding™ architecture (Figure 4). Binding data later means delaying the application of business rules, such as data cleansing, normalization, and aggregation for as long as possible. This provides health systems with time to review and revise data, form hypotheses, and determine optimal analytic uses. In addition, there is no longer a need to make lasting decisions about the data model upfront, which is useful, since it is difficult to see what new information will come down the road in two, three, or five years. By binding late, analysts only need to bind the data when there is an actual clinical or business problem to solve. The Late-Binding™ architecture is especially ideal for what-if scenario analysis and is best suited for ever-changing healthcare data and evolving use cases. Similar to just-in-time binding, Late-Binding™ works like this:

> The Late-Binding™ architecture is ideal for what-if scenario analysis and is best suited for ever-changing healthcare data and evolving use cases.

1. First, data in its atomic form is brought from the source systems to the source marts of the data warehouse. There is very little data

Copyright © 2017 Health Catalyst

transformation that occurs at this point, but some transformation, such as the variation in naming standards by the source systems, will be resolved to a single standard. For example, the term “patient” from one source system may be titled “PT_NAME” and from another source system may be called “PT_NM.” In the data warehouse, however, these variations would conform to one standard: “patient name.” It is not appropriate, though, to bind the data to any volatile business rules or vocabularies at this point.

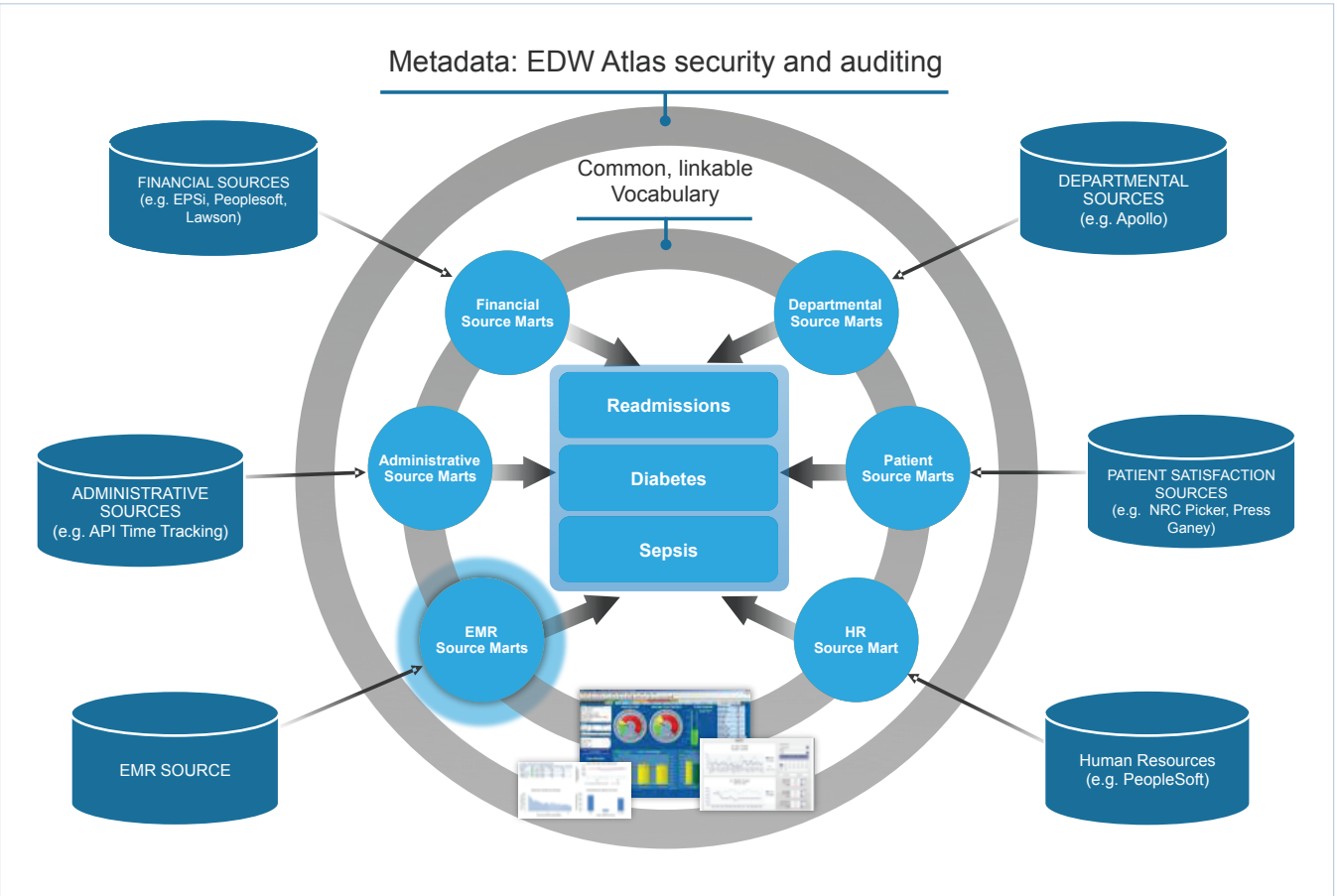

Figure 4: The Late-Binding™ architecture accelerates time to value by requiring less transformation as data moves from source systems to the EDW and comprehensive integration is done selectively at the subject area mart level (center rectangles).

2. Next, by using an incremental approach to binding data, it is possible to determine the ideal binding points for data rules and vocabulary. Because these data elements do not change often, it is acceptable to bind them early. More volatile rules and vocabularies should be bound as late as possible. By focusing first on the core data elements and then binding to additional rules and vocabulary when a clear analytic use case requires it, data engineers can deliver rapid time-to-value initially as well as later, when new analytic use cases arise.

Copyright © 2017 Health Catalyst

3. Lastly, data from the source mart is extracted to create a subject area data mart. A subject area data mart consists of patient populations (diabetics or asthmatics, for example) or a population of events, such as operating room workflows, admission workflows, or discharge workflows. Within the subject area data mart, some transformation and integration of the data is appropriate, but actual binding only occurs when a specific business driver or use case calls for it. For example, in the subject area mart, there should only be one version of gestational age. After determining the gestational age, binding the data to one definition or rule to support the stakeholders of that subject area data mart is acceptable.

There are many advantages to choosing the Late-Binding™ architecture from which health systems will benefit. They include the following:

Data modeling flexibility: Late-Binding™ Data Warehouse architecture leverages the natural data models of the source systems by reflecting much of the same data modeling in the data warehouse.

> By choosing a late-binding architecture, health systems will have as much flexibility as they need to tackle a wide variety of use cases.

Data flexibility: Because the data is not bound from the outset into a comprehensive enterprise model, the health system can use that data as needed to create analytics applications with the platform. For example, an analytics application for quality improvement might contain clinical data, patient satisfaction data, and costing data. An application for operational purposes might contain staffing levels and clinical data. An analytics application for research might combine clinical outcomes data with a research registry. The more data the health system feeds into the warehouse, the more options it has for using the data. With this approach, the health system can load the most useful sources of data first. Then they can incorporate more data sources in the future.

Changes saved: With a Late-Binding™ architecture, a record of all of the changes made to the vocabulary and rule bindings of the data models are kept in the data warehouse. By storing this history, it is possible for analysts to conduct retrospective analysis, forecasting analysis, and predictive analytics.

Iterative approach: Late-Binding™ architecture allows analysts to break detailed, high-intensity technical work into manageable chunks. Successful iterations early on in the project allow analysts to build momentum and celebrate and realize their successes sooner before committing additional resources and/or embarking on the next project.

Granular security: The Late-Binding™ architecture's security infrastructure keeps data secure while enabling appropriate access for different types of users. For example, the system could be set up to grant a researcher access

Copyright © 2017 Health Catalyst

only to data marts that have been de-identified. Additionally, researchers approved for access to patient data could be granted access to only the specific data about patients in their study.

The advantages of the Late-Binding™ architecture are significant, and they help overcome the limitations of the enterprise model and the independent data model. The following three use cases demonstrate the effectiveness of the Late-Binding™ architecture for healthcare-specific scenarios that are challenging for the other two data warehouses to support.

1. Lab Administrators Realize Improved Efficiencies

Lab administrators are primarily concerned about the efficiency of their lab. They want the process of returning results to the ordering physician to be as effective and efficient as possible. While they are not interested in data from systems like human resources or financial services, they are interested in having a good data set from their lab services.

Figure 4 shows an example of the data a lab administrator would be interested in. For example, a laboratory information system (upper right source system) can accommodate the many data queries a lab administrator would be interested in conducting to determine the operating efficiency of the lab. They do not need to go anywhere else within the EDW to perform their data analysis.

2. Enterprise Analysts Access Disparate Data Sets Quickly and Easily

An enterprise analyst needs to understand typical volume measures, cost of care measures, and trends in supply usage. To find this data, they need to gather data from many different primary sources, such as the supply system, the billing system, and the EMR system. With a Late-Binding™ architecture, making detail queries from traditionally disparate systems now becomes easy because all of the data elements are linked.

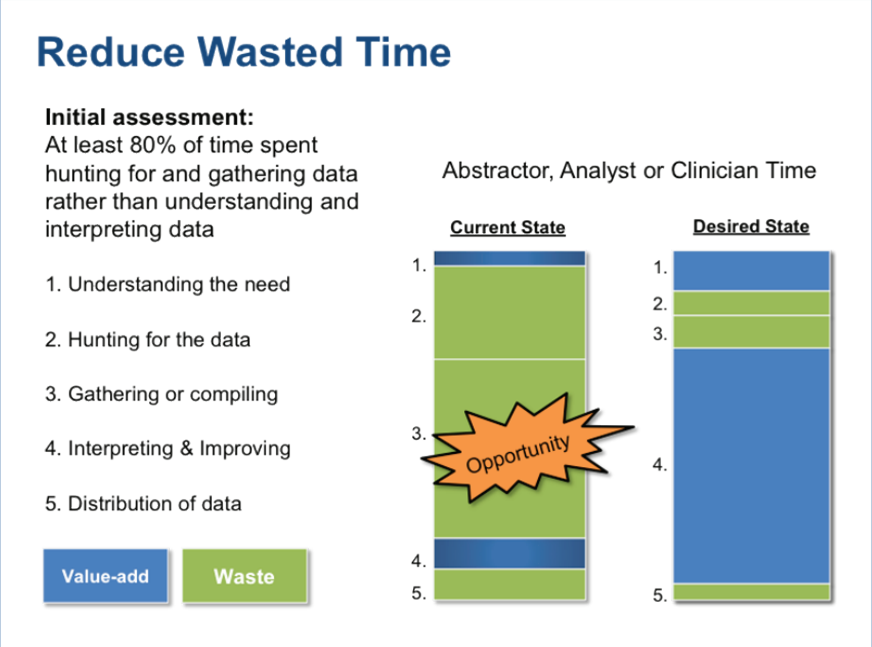

Figure 5: When data analysts work with fragmented source systems in a siloed environment, they spend the majority of their time hunting and gathering data rather than interpreting it.

Copyright © 2017 Health Catalyst

3. Care Team Analysts Focus on Optimal Care

Consider the analyst who has been assigned to support a diabetic care team to manage a diabetic population. With the Late-Binding™ model, the analyst can perform all of their analytics in the subject area mart (center rectangles in Figure 4). The subject area mart integrates all of the necessary data from the source marts and isolates the data to the diabetic population only. This approach means fast analysis times because the analyst does not need to pour through every patient record and can easily focus on the queries specific to the diabetic population.

Despite the unique advantages the Late-Binding™ architecture offers the healthcare industry, there are a few considerations to take into account. Companies that develop healthcare-specific solutions are not nearly as large as the top technology companies that develop non-healthcare specific solutions. There is some risk these companies may not be around in 10 to 15 years. To assess their future viability, it is important to look at the strength of the business model, the company's momentum, and the success of their early customers. In addition, by choosing a partner in this category, there will be one more relationship to manage.

CONCLUSION

Early-binding data warehouse models, such as the enterprise data warehouse model and independent data marts, are not necessarily or inherently bad. In fact, they work very well for specific industries, such as retail, banking, and finance because the business rules and vocabularies they are working with are stable and predictable. But the healthcare industry is unique: business rules and vocabulary standards are complex and change constantly.

Health systems need a different type of data warehouse model that addresses its particular challenges. In specific, health systems need a late-binding model that is flexible, adaptable, scalable, and that offers a short time to value. With the Late-Binding™ architecture, health systems will have as much flexibility as they need to use their data to tackle a wide variety of use cases as the need arises. Otherwise, it is a dangerous waste of resources and time to bind to rules and vocabularies that are far beyond the current analytic use cases of the organization.

In addition to being flexible and adaptable, the Late-Binding™ Data Warehouse platform is designed to handle the massive quantities of data in large healthcare organizations. Because of this unique design, health systems will have the tools they need to effectively use their data to make well-informed improvements and decisions. With the era of shared

Copyright © 2017 Health Catalyst

accountability and the need to achieve the Triple Aim, choosing an agile, late-binding architecture makes sense as it provides the ideal architecture for a healthcare data warehouse.

REFERENCES

1. Centers for Disease Control and Prevention. “Health, United States, 2013.” 2013. http://www.cdc.gov/nchs/data/hus/hus13.pdf
2. Centers for Disease Control and Prevention. “Preterm Birth.” http://www.cdc.gov/reproductivehealth/maternalinfanthealth/PretermBirth.htm
3. Corporate Information Factory. “About Bill.” http://inmoncif.com/about/
4. Gartner. “Gartner Says More Than 50 Percent of Data Warehouse Projects Will Have Limited Acceptance or Will Be Failures Through 2007.” 2005. http://www.gartner.com/newsroom/id/492112
5. Institute for Healthcare Improvement Initiatives. “The IHI Triple Aim.” http://www.ihi.org/Engage/Initiatives/TripleAim/Pages/default.aspx
6. Report to the Congress. “Promoting Greater Efficiency in Medicare.” 2007. http://www.medpac.gov/documents/reports/Jun07_EntireReport.pdf?sfvrsn=0
7. The Commonwealth Fund. “Mirror, Mirror on the Wall: How the Performance of the U.S. Health Care System Compares Internationally.” 2014. http://www.commonwealthfund.org/~/media/files/publications/fund-report/2014/jun/1755_davis_mirror_mirror_2014.pdf
8. The National Academies Press. “To Err Is Human: Building a Safer Health System.” 2000. http://www.nap.edu/openbook.php?record_id=9728&page=26
9. Wolf, Steven H., Aron, Laudan.“U.S. Health in International Perspective.” 2013. http://nap.edu/catalog.php?record_id=13497

Steve Barlow, Co-founder & SVP, Client Operations

Mr. Barlow is a co-founder of Health Catalyst and former CEO of the company. He oversees all development activities for Health Catalyst®’s suite of products and services. Mr. Barlow is a founding member and former chair of the Healthcare Data Warehousing Association. He began his career in healthcare more than 18 years ago with Intermountain Healthcare, and acted as a member of the team that led Intermountain’s nationally recognized improvements in quality of care delivery and reductions in cost.

Copyright © 2017 Health Catalyst

ABOUT HEALTH CATALYST

Health Catalyst is a mission-driven data warehousing, analytics, and outcomes improvement company that helps healthcare organizations of all sizes perform the clinical, financial, and operational reporting and analysis needed for population health and accountable care. Our proven enterprise data warehouse (EDW) and analytics platform helps improve quality, add efficiency and lower costs in support of more than 50 million patients for organizations ranging from the largest US health system to forward-thinking physician practices.

For more information, visit www.healthcatalyst.com, and follow us on Twitter, LinkedIn, and Facebook.

3165 East Millrock Drive, Suite 400
Salt Lake City, Utah 84121
ph. 800-309-6800

Copyright © 2017 Health Catalyst