The healthcare industry, in particular the data and analytics sector, has an obligation to patients to make the best use of data collected on their behalf. That requires not just collecting the data in a secure and private environment, but also making use of that data and driving insights from it. HIPAA is typically considered a security and privacy regulation, but HIPAA also refers to the need for easy access to data to improve healthcare quality. Therefore, balance is a timely topic.

Those who are responsible for developing healthcare data security frameworks in data warehousing should focus on the interplay, or the balance, between data utilization and data security and privacy. Four areas affect this balance and are worth discussing:

1. Monitoring
2. Data de-identification
3. Cloud environments
4. User access
Figure 1 shows this balance as a seesaw, which is often the perceived relationship between data utilization and security; but it's not always a zero-sum game. It's not necessarily true that more data utilization results in less security and privacy, or vice versa. Some processes help with both; some hurt both. Think of a rising tide that lifts all boats or an ebb tide that leaves them high and dry.

Balancing utilization and security is top-of-mind for CIOs, who, in 2016, are investing heavily in data and business analytics (27 percent), security (29 percent), and cloud computing (30 percent).

Quite often, IT and security have a different focus from other groups in the organization in terms of securing data and making it convenient to access. Picture a stronghold, surrounded by a moat, surrounded by a barbed wire fence. The security and IT folks are burying data in a hole inside the stronghold, while clinicians, nurses, data analysts, and executives are outside the fence needing to get in. The security is nearly flawless, but this is the type of imbalance that organizations need to prevent.

One administrative solution is to establish a dotted-line relationship between an individual from the analytics/data warehouse team and leadership from the security/IT team, and vice versa. This cross-reporting structure is an effective measure to help organizations quickly identify projects or situations where this kind of divide or imbalance can crop up.

As added context, security and privacy consists of multiple layers that include physical, preventive, detective, and administrative controls. The Health Information Trust Alliance (HITRUST) uses 14 other control categories based on ISO 27001 standards. Preventive controls address the critical issues of ransomware and email phishing, and there are a lot of materials available about these controls. Our focus here is primarily on detective controls because there’s a shortage of educational material on this topic. Also, some detective controls are particularly relevant to healthcare analytics.
DETECTIVE CONTROLS WITHIN THE SECURITY FRAMEWORK

Balancing Act #1: Monitoring

The first area that affects the balance between security/privacy and data utilization is monitoring, which impacts both sides of the fulcrum in a positive way.

The 2016 Data Breach Investigations Report defines an incident as “a security event that compromises the integrity, confidentiality, or availability of an information asset.” A breach is defined as “an incident that results in the confirmed disclosure (not just potential exposure) of data to an unauthorized party.”

According to this report, the most prevalent incident with a confirmed data breach is misuse of privilege, accounting for 32 percent of all episode types. Healthcare data is particularly private and sensitive, so individuals accessing information for the wrong reasons is more prevalent in healthcare than any other (retail, manufacturing, and finance, among others).

According to the 2016 HIMSS Cybersecurity Survey, only 60 percent of acute care providers audit the logs of each access to patient health and financial records. So there’s considerably more room to move from an implementation perspective.

Logging is simply writing event data, such as someone accessing a record or logging into a machine. But logging is simply “checking the box” in an audit, which is a minimal level of security. What’s needed is monitoring, putting tools, such as search or BI capabilities, on top of those logs. In the analytics space, this is done every day to make sense of raw data, so it should not be a large time investment for an analytics team. But human review of monitoring applications is also needed, much like when building an analytics application. It’s not enough just to build it and send it out the door; it needs follow up.
Once it becomes apparent which metrics are more relevant and are being measured regularly, alerts can be setup using incident resolution tools (e.g., WebHooks, PagerDuty, Azure OMS) that simplify metrics tracking and require less human review time.

Certain types of data are relevant when it comes to monitoring. Different layers of the analytics stacks include everything from who is logging into the network at the lowest level, to who’s logging into VMs, to what’s their performance and whether they have anti-malware at the VM or hardware level.

<table>
<thead>
<tr>
<th>Stack</th>
<th>Examples</th>
<th>Example Metrics</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytical Applications / Reports</td>
<td>Web-based, Qlik, Tableau, BO</td>
<td>Usage, click paths, performance</td>
<td>Security / Privacy</td>
</tr>
<tr>
<td>Analytics Environments</td>
<td>Specialty focused environments for Predictive Analytics, NLP, Image Analysis</td>
<td>Performance, run times, model metrics (rmse, accuracy)</td>
<td>Performance / Efficiencies</td>
</tr>
<tr>
<td>Database / Data Store / ETL / Compute</td>
<td>SQL Server, Oracle, Data Lake</td>
<td>Queries, Access (AD), ETL run times</td>
<td>Development</td>
</tr>
<tr>
<td>VMs / Hardware</td>
<td>OS (Windows / Linux), Virtualization (HyperV, VMWare)</td>
<td>Event logs (installs, invalid logins, failed applications), performance logs</td>
<td></td>
</tr>
<tr>
<td>Network</td>
<td>Switches, Firewalls, Routers</td>
<td>Invalid logins, suspicious login patterns (IP-analysis)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: Monitoring within five layers of the analytics stack.

For purposes of this report, we examine the analytical applications stack: the database/data storage/ETL/compute level, specific analytics environments, and analytics applications that have access provided to them on top of the database.

The Triple Benefit of Monitoring Analytics Products

There are three kinds benefits when putting monitoring in place on top of data that’s relevant at the analytics layer of the stack:

1. Enhanced security and privacy
2. Improved performance and efficiencies
3. Improved product development

Benefit of Monitoring #1: Security and Privacy

When going through HITRUST certification or another type of audit, there are levels of assessments related to measuring and managing. Specifically, HITRUST has five levels when assessing a particular control:

- Policy – does it exist
Process/Procedures – do they exist

Implemented – are policies and procedures actively in use

Measures – observing the control

Managed – improving the control based on learnings from the measures (the highest level of control assessment)

By putting monitoring in place across these different segments, an organization automatically performs the measuring piece. By reviewing the monitoring data and putting alerts in place, the organization performs the management piece for specific controls depending on what’s being logged. This helps to achieve the high levels from an audit perspective.

It also streamlines the recertification process. SOC 2 and HITRUST have certain windows of opportunity for when the recertification or audit is valid, typically a year or two depending on the type being sought. The next year, the process must start over. There are also interim assessments at six months or annually. Recertification and interim assessments are time consuming, requiring retesting of a number of controls and showing data to auditors to prove strong performance across multiple controls. With monitoring in place, this whole process becomes very streamlined because it creates an existing dashboard that shows, for instance, times where certain events have occurred over the past six months.

We have spoken with a number of healthcare systems, specifically those that have had audits of their EDW and analytics environments. The top issues were the ability to audit who had access to exactly what data at any given time, and the ability to audit appropriate use.

Let’s examine some best practices that are relevant to addressing appropriate use and access.

**Appropriate Use:** Regarding appropriate use, one best practice is to pull in log data at the database level to see who is making what query, what SQL query is being run, what database is being tapped at what time, and what table is being queried. Then pull all that into a dashboard so it’s very simple to analyze and manually review where there are situations or individuals querying with a filter on a single person or patient name. Other fields, like SSN and date of birth, are worth noting, but patient names are particularly relevant because that’s a strong indication of a search for a specific family member or celebrity.

This best practice should not be limited only to the database level, where there are potentially dozens or hundreds of users. It should also be done at the analytical application level—whether working in Qlik, Tableau, or Web apps—where there are hundreds or thousands of
users. Everyone’s actions should be logged and monitored within these types of applications.

In the example in Figure 5, the monitoring dashboard shows access by user name, with a filter applied that shows patient name, as well as date and time of access.

Figure 5: Monitoring dashboard shows user access by field name.

**Access:** Generally, there are secure processes in place for initially granting access, but a typical problem that healthcare systems face is reviewing and maintaining access rights. Should someone who has access to data today still have it six months from now? There are three best practices for automating access review:

- Query the Lightweight Directory Access Protocol (e.g., Active Directory) and review who is in which access groups.
- Query database access (SQL Server) or application access (Qlik, Tableau, Web) to see which access groups have access to which databases, tables, or applications.
- Query SQL queries (IDERA) and application usage (Qlik, Tableau, Web)

A data steward can then review the information from these queries every quarter, six months, or year, and determine if only relevant users have access to highly sensitive data. This process is also very appropriate for data in HR categories that don’t typically grant broad user access.
Benefit of Monitoring #2: Performance and Efficiency

ETL processes are logged and monitored to see how often they succeed or fail and how long they take. Oftentimes, there will be a spike over a few days in the number of failures associated with an ETL process, which may mean it’s time for a data architect to review those specific SQL queries and test the connections to determine what’s wrong. There can also be a spike in the amount of time it takes for an ETL job to run, as noted in Figure 6.

![Figure 6: The ETL log shows spikes in run times and anomalies with failed processes.](image)

A one-off spike isn’t too worrisome, because it could mean that someone is testing a new query or pulling in a new data source. But when the average time shows a sustained increase, then it’s time to ask questions. Are ETL jobs being scheduled at the same time? Do they need to be staggered? Are SQL queries non-optimized within the ETL jobs? Do they need to be altered?

BI developers and managers need to decide where their teams should spend their time. This can be done somewhat anecdotally, such as when physicians, nurses, and administrators request new analytics projects. A more evidence-based method is to see what data source is being utilized the most. What databases and tables are most queried? Who are the query users? The answers to these questions can determine which databases should be the focus for improving data quality and which tables should receive additional indexes for improving performance.

Benefit of Monitoring #3: Product Development

Eric Ries, in his principles of The Lean Startup, describes a process that can be applied to product development. Building a product doesn’t end once code is ready to ship. It’s a full cycle that starts with an idea or hypothesis. The idea gets coded, built, and shared with users. Most importantly, it is specifically measured and the idea or hypothesis is reassessed based on the learnings from those measurements.
The goal is to minimize the amount of time it takes to get through this learning cycle. In our security framework, monitoring is the measurement piece in this cycle. Some of the relevant metrics are simple things, like session counts, login frequency, number of distinct users, and cohort analysis to see who logged in during a given period and when they returned.

Other advanced monitoring features include simple, frictionless user-experience surveys, and A/B testing on applications to determine if new functionality increases utilization.

A great example of this type of monitoring comes from Uber, which consistently runs surveys and A/B testing. In fact, Uber riders are required to complete rating surveys after every trip before they are able to book the next one.

Similarly, when building a Web application, it’s very easy to require users to rate their previous experience when using the app.

A few years ago, a team at Health Catalyst was building a new advanced analytics application. The chart in Figure 10 shows the number of user sessions during testing, with a small spike as we rolled out the app to a small training group. The team iterated the product and rolled it out to a much larger test group. The average usage was expected to remain high, but the number of sessions quickly returned to almost zero, where it had been two weeks earlier. The team dove in to understand the hypotheses it had and why they were wrong.
What we learned was surprising. The amount of time users spent in the app (Figure 9) was heavily weighted toward the Performance sheet (screen), however the development team anticipated a much greater weighting toward the Provider sheet. Thus, the dev team spent the majority of its time enhancing the latter. With a monitoring dashboard that showed user session details, we were able to sit down with the heavy users, find out what influenced their click paths, and then further iterate the product according to that feedback.

![Dashboard](image)

**Figure 9:** Testers spent their time on the Performance sheet, but developers spent their time perfecting the Provider sheet.

**Balancing Act #2: Data De-Identification**

Data de-identification is actually a negative balance (recall the “ebb tide”) on both data utilization and security/privacy. There are two ways to de-identify a dataset to satisfy HIPAA requirements.

- **Safe Harbor Method:** HIPAA defines 18 elements that must be removed or transformed from data before it is considered de-identified. This means removing anything more detailed than the year (month/day/hour/minute). When it comes to healthcare analytics and clinical quality improvement, date elements are particularly important for understanding the sequence of episodes of care and the timing between them.

  Similarly, any geographic details more granular than the state where care was provided must be removed (the exception is zip code data as
long as the zip code contains more than 20,000 individuals). This makes it difficult to develop network optimizations or referral patterns.

Furthermore, HIPAA requires that the data user not be able to use a dataset by itself or in combination with any other dataset to re-identify an individual. This means removing any rare ICD-10 codes and procedures that could be used to isolate an individual.

**Expert Determined Method:** This means partnering with a statistical expert to ensure a very small likelihood that any individual record within the data could be used to re-identify a patient. HIPAA doesn’t allow many data details, so there are no one-size-fits-all processes for transformation. Therefore, this method can be expensive and time consuming. In healthcare, we often work with wide datasets, such as lab results, medications, and specific timing of encounters. With these wide datasets, oftentimes each individual row is unique because of the very large dimensionality. The expert determined method uses a process called k-anonymity, where it’s ensured that “k” number of records appear exactly alike, thus making it impossible to distinguish and re-identify an individual. But difficulties in de-identifying arise with wide datasets because of this “curse of dimensionality,” which leads to a tradeoff between anonymity and utility.

One non-healthcare example of a failed attempt to de-identify data comes from the New York City Taxi & Limousine Commission when it was required to release taxi log data under the Freedom of Information Act. These records showed pickups and drop-offs by latitude and longitude, date, time, frequency, driver name, fare amount, and many other fields. The Commission suppressed many of the fields and hashed the medallion numbers, but neglected to salt those hashes. Within hours of the data release, curious “analysts,” using rainbow tables, reverse identified the medallion numbers and revealed all of the data fields that had been de-identified. The significant fallout from this was that individual riders could be identified by pickup location and home address destination.

This is an illustration of the difficulty in predicting how individuals will utilize a dataset that has been de-identified and why the expert determined method can be a tough balance.

**THE DATA CONTINUUM**

Security and utilization experts work in three categories of healthcare data:

- **Full Protected Health Information (PHI):** best for healthcare analytics, ad hoc querying, identifying root cause, and decision support. We urge
systems to store this data in a secure environment following control best practices as established by HITRUST and other sites.

- **Redacted Data (still PHI):** For those who are still nervous about making certain data available to analysts, this provides a decent balance. It’s still considered PHI, but direct patient identifiers (SSN, patient name, patient address, etc.) are removed, leaving fields like MRN available for analytical purposes. This helps support the HIPAA minimum use standard when working with healthcare data.

- **HIPAA De-Identified Datasets:** good for very aggregated data, narrow datasets, and product development.

Figure 10 shows the privacy and security risk of data in each of these three categories, along with the inverse proportion of analytical value of that data. This is a useful visualization of the tradeoffs between these two security framework issues.

Figure 10: Privacy and security risk of data in the three categories of healthcare data.

**Balancing Act #3: Cloud Environments**

Most of the analytic stack will eventually move to the cloud. The reason behind this is that security, as well as diverse and distributed analytics environments, can be created very quickly and cost-effectively in the cloud. In healthcare, this may take a while, but the first pressure that we are experiencing is for specific analytic use cases, such as performing **predictive analytics** on a larger dataset, **Natural Language Processing** (NLP), and image recognition. These often require distributed computing environments,
which are difficult to manage on-premise. It’s easier to have someone else managing them, spinning them up and down, as needed.

The cloud can help with both data utilization and security/privacy. The major cloud vendors (Amazon, Google, Microsoft Azure) perform ISO and SOC II audits and they will sign HIPAA Business Associate Agreements (BAAs), which allows healthcare organizations to fully leverage their audits. However, discretion is still the better part of valor. Figure 11 shows the responsibilities between the healthcare organization and the cloud provider over three variations of hosted services (the customer is responsible for everything in an on-premise environment): Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS).

Regardless of the structure, there is a lot of blue in this diagram, aka, cloud customer responsibility. This means still following security best practices when moving to the cloud, like putting in firewalls, installing anti-malware, installing intrusion prevention, and monitoring.

Because cloud environments are built on the same fabric, it’s very easy to pull in logs and monitor everything, from the VM level to the application usage, on top of those logs. It is much easier to monitor these from a central place. Alerting is also easier to setup within a cloud environment because alerts already exist when logs are pulled into one place. For instance, when a predictive model takes more than a certain amount of time to run, an alert can notify the analyst who can determine if performance is worsening and if the model won’t be ready for downstream ETL jobs.

Another piece to this is that specific policies can be embedded into a cloud environment, for example, setting up network protections on all in-points or

![Figure 11: Customer and cloud provider shared responsibilities for security controls.](image-url)
anti-malware solutions on all VMs. The cloud has tools that will then scan the environment to see what security best practices are in place and make recommendations for changes, if needed.

**Balancing Act #4: User Access**

The final point for creating balance between data utilization and security/privacy is user access. We don’t have a good answer for addressing this issue, but it’s something the industry should be considering for improvement. There are two areas to emphasize:

- **Streamline the permission-granting process:** One audit issue that comes up in conversations with healthcare systems across the U.S. is not that the process isn’t secure when granting users permission to data, but that it takes too long. So users are bypassing best practices. Taking too long to grant access is bad from both a data utilization and security perspective.
  
  - One way to alleviate this problem is to make certain default applications available to everyone within certain security groups, especially if those applications don’t have patient level data.
  
  - When thinking about granting access to a dataset, like HR or clinical data, involve data stewards, someone who best knows that dataset. They will often have a good sense of who should be granted access and how the data will be used.

- **Role-based security:** Simplify this as much as possible. Complicating it results in mistakes because it becomes too challenging to match individuals to security groups. It’s generally better practice to simplify and give fewer people access to data than it is to overcomplicate and make the IT mistake of putting people in the wrong buckets.

**JOINING THE PIECES OF THE SECURITY FRAMEWORK**

Data is useless unless it gets in the hands of analysts, operators, and clinicians. But healthcare organizations need to strike the balance between security/privacy and data exposure. When monitoring, logging must be integrated into a search and BI tool for manual review. While this will take longer, it will lead to multiple benefits in security/privacy, performance efficiencies, and better product development. Data de-identification is typically not a good balance of utilization and security for most healthcare improvement analytics use cases, because of how it weakens the value of data. Cloud environments, if set up properly and with caution, will lead to a better balance. The final cog in the security framework system is the need to streamline user access and permission setting to realize faster time to value.
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.