PHEMI WHITE PAPER

Zero Trust Data

Solving the Data Dilemma

In this white paper

PHEMI Big Data. Shared. Simply. Securely.

Foreword

Big data encompasses a range of powerful technologies, but comprehensive mechanisms to ensure privacy and data security have been slow in emerging. This may make some organizations think that they have to choose between privacy and innovation — an either/or proposition, for the sake of big data analytics. But that is not the case: privacy and data security are perfectly compatible in a doubling — enabling, win/win manner, through a strategy that includes collaboration, protected data re-use and data sharing.

In this white paper, PHEMI ties together notions relating to privacy as reflected in the 7 Foundational Principles of Privacy by Design, with emerging new models of network security, namely, Forrester's notion of the "Zero Trust Network." The result, "Zero Trust Data," is a privacy and data security framework that is both powerful yet flexible and effective. The Zero Trust Data approach presents a win-win solution for healthcare, business, and government operations.

Zero Trust Data is a compelling approach to the problem of how to strongly protect data while allowing it to be shared in de-identified form. In my view, implementing an approach such as Zero Trust Data will be essential to any organization implementing a big data strategy — where privacy, data security, and governance must form the foundation of their operations.

Dr. Ann Cavoukian
Executive Director
Privacy and Big Data Institute
Ryerson University

Zero Trust Data

Solving the Data Dilemma

Big data holds the potential to transform businesses in virtually every industry. But it also raises new concerns—and substantial new responsibilities—around information privacy, security, and governance. To get the most value from your data, you need to be able to share it. How can you balance the need for access to information with the need to keep it protected?

The Growing Data Dilemma

Digital and cloud technologies have unleashed a new wave of disruption. Small startups become global brands, and longtime incumbents fall by the wayside, seemingly overnight. Look closely, and you'll see a common theme: those with the best insights from their data are winners.

Big data technologies are powering this revolution by radically increasing the volume, variety, and velocity of data collection. With new database technologies, you can store and retrieve virtually anything and everything: documents, web pages, photos and videos, code fragments, virtual machines, and much more. You can unlock data silos, mine data like text and social media, and respond faster to changing business needs, without getting slowed down by traditional database limitations.

But along with these new capabilities comes serious questions: what are your responsibilities with respect to this data? Why should anyone trust you with it? How can you share it while meeting your governance and compliance mandates? Trying to answering these vital questions about data security and privacy, while recognizing the need to share data to get more value from it, is the central "Data Dilemma" facing business leaders today.

> "We do not have to sacrifice privacy or shackle innovation for the sake of big data analytics. Through careful planning and the application of privacy techniques, such as those embodied in Privacy by Design, organizations can use data for its desired effects, while at the same time protecting the personal information contained in the data. It is indeed possible to have both Big Data and Big Privacy."
>
> — Dr. Ann Cavoukian, Information and Privacy Commissioner of Ontario, Canada

You can only be a truly data-driven organization if you can share information—with analytics systems, marketing, partners, and others. But if people are going to trust you with their data, you need to demonstrate that you have a framework in place to share it only for legitimate purposes, and can enforce that control.

Unfortunately, the big data systems commonly used today aren't there yet. Most are built on Hadoop, an open-source software framework that provides an effective file system for distributed storage and processing, but limited data management and access controls. The sophisticated access control mechanisms needed for privacy and governance—controlling precisely who can access data; what forms of data they are entitled to see (for example, de-identified vs. identified); and where, when, and how they are allowed to access it—are left for other systems to address.

So organizations turn to bolted-on network- or application-based mechanisms that can't keep pace with the growing ocean of information that organizations now collect, or the proliferating ways they want to use it.

Fortunately, a model for solving the Data Dilemma already exists, using well-defined principles advocated by leading security organizations like the National Institute of Standards and Technology (NIST) and Privacy by Design (PbD). (See sidebar.) Rather than develop an entirely new approach to big data security and privacy, you can use proven approaches used to defend modern networks, and extend them to the data layer. It's a concept called Zero Trust Data.

Privacy by Design

A Privacy by Design (PbD) approach requires you to take into account seven foundational principles throughout your system. But how do you know whether your system implements PbD principles? Here's a checklist:

1. Metadata.
All data should be tagged on ingest with enough descriptive information to allow adequate privacy, sharing, consent, and lifecycle management, plus compliance with any other governance requirements.

2. Role-based access control.
User and application access to functionality and operations is adequately restricted by system roles.

3. Policy-based data access.
Access to and visibility of data is restricted by permissions and authorizations, and controlled by access policies.

4. Automatic policy enforcement.
The system automatically enforces policies and governance; manual intervention is not required. Enforcement is not relegated to applications built on top of the repository. There's a single point of management to ensure policy enforcement.

5. Transparency.
Data stewards and privacy officers can directly view and verify the system implementation of governance policies.

6. Auditability.
The system automatically tracks system activity, and maintains a detailed, tamperproof audit log of data access and system operations.

7. Data immutability.
Data in the repository remains available in its original form, regardless of what digital assets are derived from the original through transformation.

8. Ability to anonymize.
The system should be able to de-identify, encrypt, mask, obfuscate, or redact personal information, and allow the data steward or privacy officer to choose which version of data appears to which users.

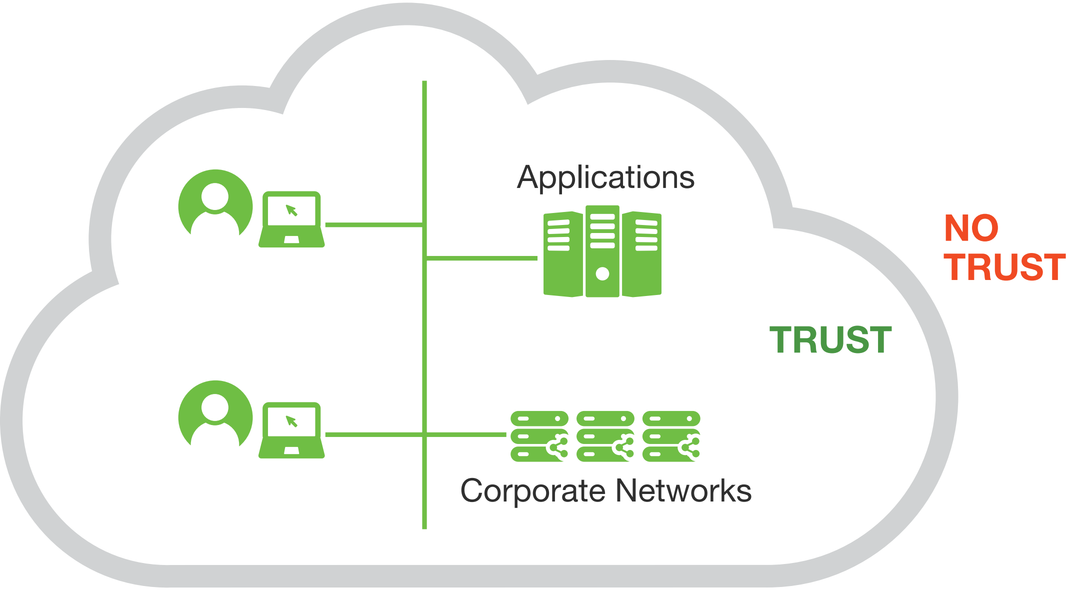
Privacy by Design is recognized as the global privacy standard in a landmark resolution by the International Conference of Data Protection & Privacy Commissioners. Visit privacybydesign.ca.

From History to State of the Art

Zero Trust Data is modelled after the “Zero Trust” strategy now widely adopted in the world of networking, and developed to grapple with changing ideas of trust. From the beginning, networks employed “trust” zones and “no trust” zones, bounded by a perimeter. If you were trusted, you were granted access. If not, access was denied.

In the earliest days, unidirectional firewalls provided a basic layer of protection (Figure 1). You had to have the right credentials to get through the security perimeter, but once you did, you were considered “trusted” and had access to everything.

Figure 1. Basic Perimeter Defense



Bidirectional firewalls expanded the no trust zone, ensuring that users outside the network had the right credentials before allowing their response to a communication originating in the trusted zone. Even if a communication initiated from within the perimeter, the response from outside had to be explicitly permitted. Over time, organizations continued shrinking trusted network zones, using mechanisms like secure subnets and remote connections (VPNs).

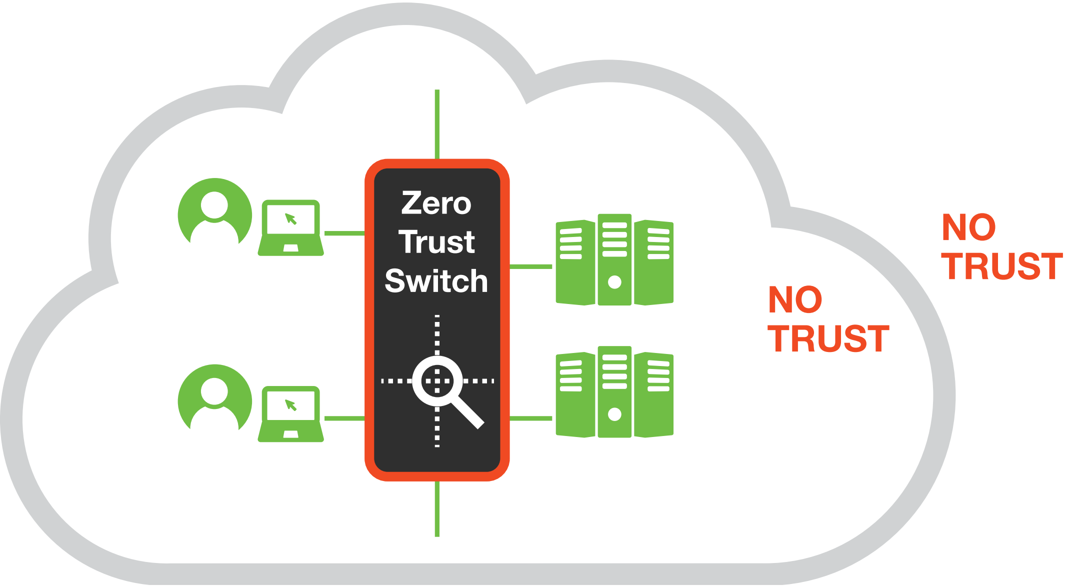
But as security threats became more sophisticated and harder to detect, organizations began to recognize that any trust zone in the network posed too great a risk. A new model was needed to guard against advanced attacks, as well as internal users misusing resources.

The Zero Trust Network

A concept initially introduced by Forrester, a Zero Trust Network is architected so that no one is trusted, anywhere (Figure 2). Zero Trust Networks embody the following principles:

- Never trust, always verify: Every connection is examined for appropriate authorization, even within a network segment. And all traffic is inspected and logged all the time.
- Embed security into the architecture itself: Security is no longer an afterthought. Segmentation extends across network layers and is enforced with secure switches and deep packet inspection.
- Control access on a need-to-know basis: Access to any resource, from any host, must be explicitly authorized.

Figure 2. Zero Trust Network



Modern Zero Trust Networks achieve a much more granular concept of trust by examining the context of each network connection—verifying the “who,” “what,” “where,” “when,” and “how” of every access request. This behavior is often embodied in the Zero Trust Network switch. The contextual attributes are the key to a Zero Trust strategy, because they allow you to go beyond simply walling off segments or resources, and verify that any behavior on the network constitutes legitimate use, based on policy.

An HR employee may be authorized to access HR resources. He has no business accessing Finance resources. A product manager may be authorized to access sensitive product schematics from her office workstation. She shouldn't be able to download them to her mobile device over a public WiFi network.

The New State of the Art: Extending Zero Trust to Data

Zero Trust is a highly effective model for securing networks. But for Chief Data Officers (CDOs), it doesn't solve the Data Dilemma, because it doesn't extend Zero Trust to the data itself. While traditional data warehouses and many Hadoop-based systems provide some protection for data, data has not enjoyed the equivalent of the Zero Trust Network switch, where every request for data access would be tested against against specific characteristics of both the user and the data.

Zero Trust Data extends the Zero Trust model (never trust, always verify) from the network into the data itself. In a Zero Trust Data model:

- Access is denied by default: Privacy and governance policy is encoded directly onto data as it's collected. And every piece of data can be encrypted and inaccessible by default.
- Data requests without proper credentials yield no information: The data system knows whom it can trust to view any data asset, when, and in what context. Without the right access credentials and attributes, you see nothing.
- Data security is enforced independent of the network: Data protection no longer relies on networks or applications to enforce privacy and governance. Those functions are now controlled by the organization's data stewards, and operationalized in the data layer itself.

Implementing Zero Trust Data

A Zero Trust Data implementation is based on the same principles as Zero Trust Networking, but extended to data. And it requires a unique set of capabilities. The first is the ability to decouple users from data:

- Data is described using **metadata**.
- Users (including both people and applications) are described and controlled based on **attributes**.

This may sound like a simple distinction, but it's at the core of your ability to share and gain value from the data you collect while keeping it protected. For example, Active Directory is a great way to dynamically update user attributes and ensure that every system in your network has the same view of user authorizations. But if data is associated with specific users (or more likely specific roles), change management becomes much more onerous, and the applications using that data become more brittle.

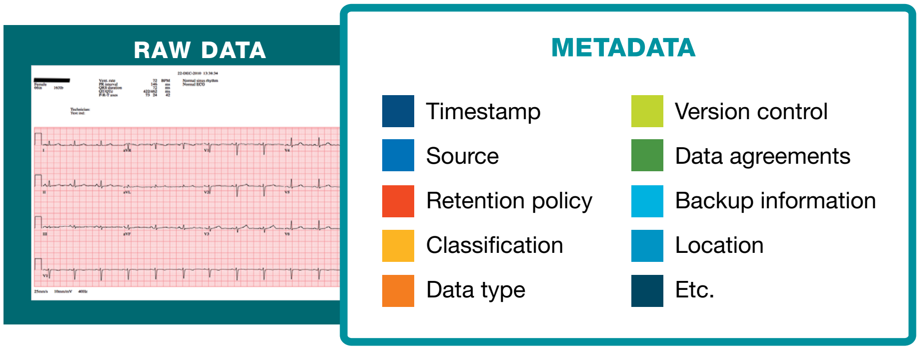
People and roles change all the time. Applications need to be written and updated as fast as new needs arise. By keeping the mechanisms governing privacy and access distinct from the data itself, you can accommodate constant change without slowing down application development. And your data always remains stable.

Metadata

A Zero Trust Data model wraps detailed metadata around every piece of data as it's collected. This goes far beyond the metadata applied by traditional databases, which typically involves a more limited and static description, such as the data type, and possibly provenance.

In a Zero Trust Data implementation, metadata captures the full policy and governance framework in which the data now lives. It encompasses descriptive, structural, and administrative aspects, including detailed indexing, rights management, retention periods, and privacy and confidentiality agreements (Figure 3).

Figure 3. Example of Metadata Around a Digital Asset



Effectively, metadata should define everything you need to know to control usage and governance of big data: what it is, where it's from, and what's allowed to be done with it. And it's through that metadata that you can begin to realize the full value of your digital assets.

Metadata should be infinitely flexible, allowing you to model any number of situations or contexts in how it can be used. And it should be changeable as policies evolve: you should be able to alter it without having to redesign your data model. It's extremely difficult to achieve this when you're relying on a separate database for static metadata. A Zero Trust Data strategy works best when the metadata is wrapped around the data itself.

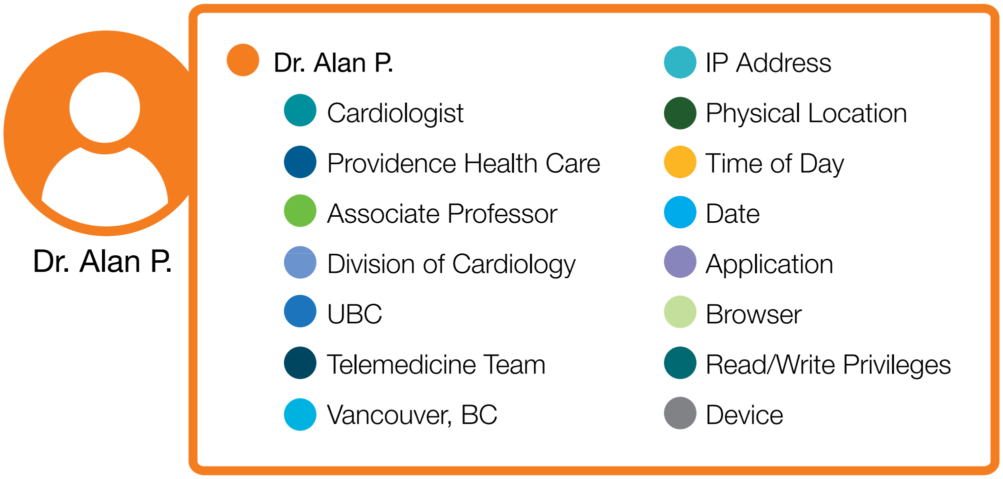
When you use metadata in this way, your data can remain stable and immutable, without placing limitations on how you use it. The metadata can adapt to changing requirements, knowledge, purposes, and contexts. Your data is always linked to that metadata, everywhere.

Attributes

Attributes address the other side of the equation, describing who is attempting to use the data (whether human or application), and in what context. A user has this role, in this department, at this location, using this device. When you can grant or deny access to a system, or some portion of it (such as specific data) based on those contextual attributes, you have access control aligned with the principles set out by the National Institute of Standards and Technology for Attribute Based Access Control, or ABAC (**NIST, 2014**).

A system based on contextual attributes can provide flexible and context-aware access controls in dynamic information systems. ABAC itself is now the access control standard mandated for U.S. government systems. And ABAC-based systems are ideally suited for enforcing security and privacy of data assets, even as you accommodate changing users, roles, devices, and applications. Because within an attributed-based model of access control, sets of attributes can be easily changed or expanded.

Figure 4. Example of User Attributes



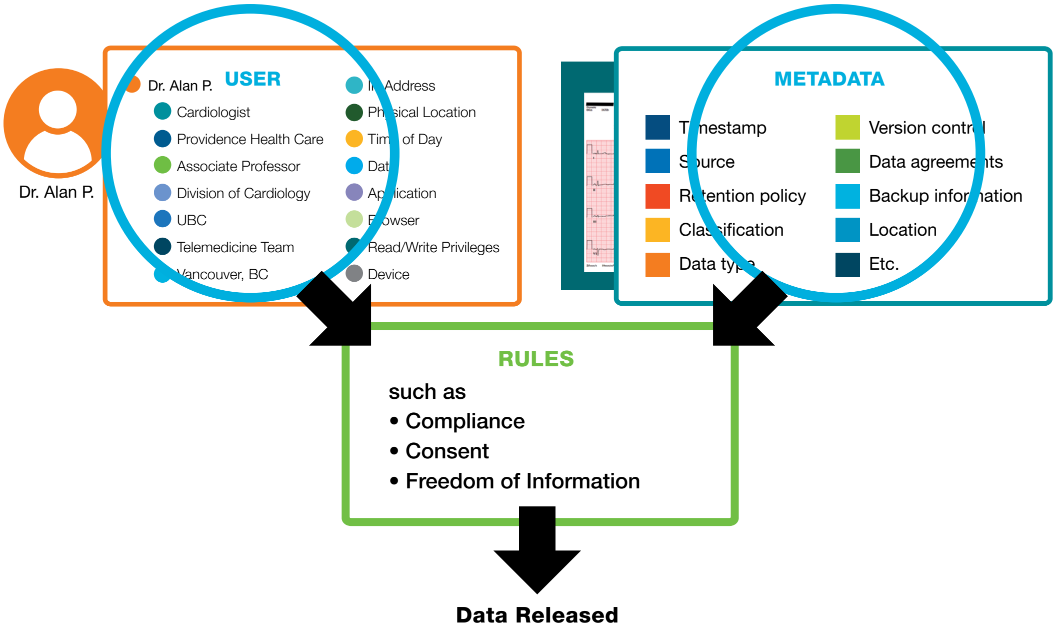
Applying Policy

The final step is to compare user attributes to metadata using access rules to provide contextual access to data based on policy. Here, a rules engine reads the metadata wrapped around a particular data asset, interrogates the user attributes, and executes the policy governing who can see that data and in what context (Figure 5).

In a Zero Trust Data model, the rules engine lives at the data level, not in the application. And it's not controlled by application developers, but by the organization's data stewards, who understand all the necessary parameters around data provenance, sharing, consent, and retention. Once again, this relieves applications of having to contend with this often complex and changing authorization logic, so they are cheaper, faster to develop, and less likely to break.

When you can control data access with policies that consider both user attributes and metadata characteristics, you've achieved Zero Trust Data.

Figure 5. Policy-Based Data Access



Other Zero Trust Data Requirements

Sophisticated metadata and user attribute intelligence are core enablers for Zero Trust Data implementations, but they are not enough on their own to fully capitalize on the big data you're collecting. An effective data system must address other considerations to provide the speed, scale, and flexibility needed to continually realize value from your data.

Seeing the Same Data through Different Lenses

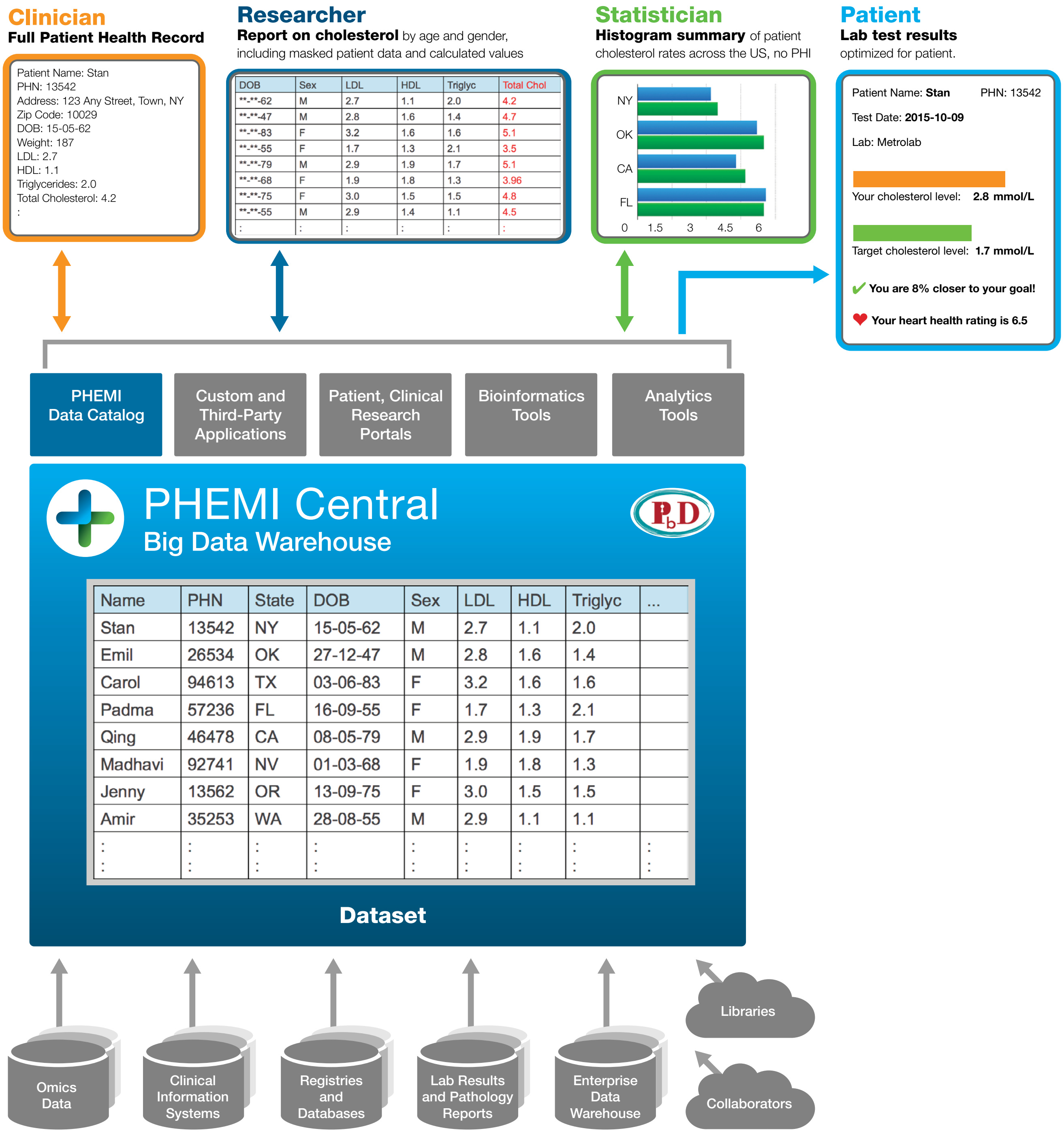
To effectively use and share data, the data system should also provide "data virtualization" through in situ processing. You should be able to take the same data asset and repurpose it on the fly to address different needs and different levels of access.

Consider an example from the healthcare industry. A patient's physician may be authorized to access a patient's full medical record. An analyst may be authorized to see the patient's age, gender, and cholesterol level, but not individually identifying information. An affiliated statistician may be authorized to view aggregated summaries of patients' cholesterol by gender and age range, but nothing more specific.

Traditionally, to meet the needs of each of these use cases while de-identifying protected data, you would need to create a brand new data set (or data mart). A data analyst would have to manually extract the data, modify it, and then make it available to the data mart—a time-consuming process that results in proliferating copies of the data. Instead, an effective big data system can, in effect, "virtualize" the original data asset based on user attributes and rules. It can present the same data through multiple lenses, displaying only the specific information a user is authorized to see, without requiring manual steps or continuous copying of data assets (Figure 6).

This ability to provide different views of data to different users is essential to obviating privacy breaches. When you're collecting petabytes of information, it's the only practical way to allow for on-demand, self-service access at scale, while hiding protected information on a user-by-user or case-by-case basis. From a business perspective, it means you can continually identify new ways to extract value from your data and execute them much faster, without compromising data security, integrity, or privacy.

Figure 6. Virtualizing Data Assets for Different Users


Clinician
Full Patient Health Record
Researcher
Report on cholesterol by age and gender, including masked patient data and calculated values
Statistician
Histogram summary of patient cholesterol rates across the US, no PHI
NY
OK
CA
FL
0
1.5
3
4.5
6
Patient
Lab test results optimized for patient.
PHEMI Data Catalog
Custom and Third-Party Applications
Patient, Clinical Research Portals
Bioinformatics Tools
Analytics Tools
PHEMI Central
Big Data Warehouse
PbD
Dataset
Omics Data
Clinical Information Systems
Registries and Databases
Lab Results and Pathology Reports
Enterprise Data Warehouse
Libraries
Collaborators

Patient Name: Stan
PHN: 13542
Address: 123 Any Street, Town, NY
Zip Code: 10029
DOB: 15-05-62
Weight: 187
LDL: 2.7
HDL: 1.1
Triglycerides: 2.0
Total Cholesterol: 4.2
:

DOB	Sex	LDL	HDL	Triglyc	Total Chol
--62	M	2.7	1.1	2.0	4.2
--47	M	2.8	1.6	1.4	4.7
--83	F	3.2	1.6	1.6	5.1
--55	F	1.7	1.3	2.1	3.5
--79	M	2.9	1.9	1.7	5.1
--68	F	1.9	1.8	1.3	3.96
--75	F	3.0	1.5	1.5	4.8
--55	M	2.9	1.4	1.1	4.5
:	:	:	:	:	:

Patient Name: **Stan** PHN: 13542

Test Date: **2015-10-09**

Lab: Metrolab

Your cholesterol level: **2.8 mmol/L**

Target cholesterol level: **1.7 mmol/L**

✓ **You are 8% closer to your goal!**

❤ **Your heart health rating is 6.5**

Name	PHN	State	DOB	Sex	LDL	HDL	Triglyc	...
Stan	13542	NY	15-05-62	M	2.7	1.1	2.0	
Emil	26534	OK	27-12-47	M	2.8	1.6	1.4	
Carol	94613	TX	03-06-83	F	3.2	1.6	1.6	
Padma	57236	FL	16-09-55	F	1.7	1.3	2.1	
Qing	46478	CA	08-05-79	M	2.9	1.9	1.7	
Madhavi	92741	NV	01-03-68	F	1.9	1.8	1.3	
Jenny	13562	OR	13-09-75	F	3.0	1.5	1.5	
Amir	35253	WA	28-08-55	M	2.9	1.1	1.1	
: :	: :	: :	: :	: :	: :	: :	: :	: :

Data Immutability

Along these lines, a critical aspect of solving the Data Dilemma is preserving the immutability of data assets from the moment they are collected throughout the entire lifecycle. An effective data system will not just lock down access. It will help you meet modern-day transparency and governance requirements, such as those promoted in the PbD guidelines, by assuring that data assets cannot be changed—even as they are repeatedly virtualized for a variety of purposes.

When your data is immutable, you can:

- Audit, track, and checksum it to verify compliance with privacy and governance requirements
- Keep the data itself stable, no matter what changes around it (metadata, attributes, policy, users, or applications)
- Track provenance of all data (its history, origin, and management across its lifecycle) in a transparent and reliable way

The Zero Trust Data Advantage

Balancing the need to share data with the need to protect it is the single biggest problem facing data-driven organizations. But it's not unsolvable. By adopting a Zero Trust Data model, you can extend granular and context-driven access control all the way into the data layer. So you can tap into the value of your most valuable organizational asset—your information—without compromising the trust of those counting on you to protect it.

With Zero Trust Data, you can:

- **Place control over privacy and data security in the hands of your data stewards**, so the people responsible for data governance are the ones operationalizing it.
- **Embed privacy, governance, and consent policies directly in the data store**, so they are enforced automatically rather than manually.
- **"Virtualize" data assets** so that you can share information with a wider range of stakeholders, while ensuring they see only what they are authorized to see.
- **Make application development faster and less expensive** by offloading privacy and governance responsibilities to the data system.
- **Make applications "thinner," less brittle, and stateless** by processing data at the data layer; applications need only pass on attributes and serve datasets.

- **Improve your overall security** by making all data invisible by default; even a compromised application need not compromise your data.
- **Simplify change management**, since people, devices, applications, and authorizations can all change, but the underlying data does not.
- **Employ more flexible security models** by tying access to rich context: role, location, device, or any other parameter.
- **Enable more sophisticated consent management** by supporting more granular layers of consent (identified, de-identified, histogram), and building a secure intermediary layer between those requesting data and those providing it.

Take the Next Step to Zero Trust Data

Zero Trust Data principles are embodied in PHEMI Central,™ a big data warehouse built from the ground up to address the core data dilemma facing CDOs and data stewards. PHEMI Central empowers you to capitalize on your valuable and growing information assets, within a comprehensive framework for privacy, security, and governance. With PHEMI Central, your organization's data strategists have the tools they need to unlock the full value of your information, and the controls they need to protect it.

To learn more, visit www.zerotrustdata.com