A CTI WHITEPAPER

March 2020

ARCHITECTURE FOR A GOVERNED DATA LAKE

That transforms your business

Kurt Rosenfeld

kurtr@ctidata.com

CTI

Your Data Is Our Passion

FIND TRUST UNDERSTAND COMBINE EXPERIMENT SHARE

OPERATIONALIZE PROTECT PROVISION DESTROY

On the fly

TABLE OF CONTENTS

GOVERNED DATA LAKE AS-A-SERVICE

PREAMBLE

This paper introduces the reader to a framework that has one goal: to make data useful and accessible to "business engineers." We identify a business engineer as a person that improves how any work activity is performed. This could be reducing exceptions in the submission of insurance claims, or improving flight patterns to reduce fuel consumption, or in life-changing ways such as identifying people with probable risk of diabetes so that interventions may be applied before it's too late.

These business engineers innovate by analyzing and manipulating data, which nowadays holds the "keys" to every process and interaction that occurs in modern societies. The problem then is how to make this volume and variety of data much more relevant and reachable, something typically solved by "technology engineers", with its inherent bottlenecks and impedance (what I need and what you gave me are not the same). This is no more: we can arm business engineers with freedom to explore the evolving and changing nature of data to solve business problems faster.

The paper is not intended as a primer on data lake platforms and their business rationale. We assume the reader has such understanding, if not, numerous articles exist on the topic – this paper addresses how to make the lake genuinely useful for business engineers.

We introduce a new model for data governance, new capabilities for interacting with the data lake, and a new playbook for IT, business engineers, and business leaders to make it happen. This collective framework – a Governed Data Lake – has emerged from our work over the years leveraging data to improve the business performance for our customers and the evolving technological approaches that work well, and those that don't.

Please know that while we use a diabetes management scenario to illustrate the framework, there is nothing unique to this scenario, it simply serves to convey concepts.

[Sidebar: we like the label "business engineer" because of its contrast with "technology engineer". But our workplaces generally refer to such people as "business analyst" or "data analyst" and we use these interchangeably along with "business professional", "information worker" and simply "user".]

RETHINKING THE DATA LAKE

The impact of data transforming how business and society functions has progressed enormously over the last few decades, yet recent advances have fallen far short of their potential because only a very few – the data experts in our organizations – can "deal with data."

Organizations know that enabling their business professionals to be "information workers" will herald new waves of innovation. The buzzwords "citizen analyst" and "citizen integrator" and "data democracy" all point to this desire. The objective is to arm business users with a means to experiment and construct new data insights and data processes on the fly. It's no accident this goal aligns with the corporate mantra of "digital transformation," which seeks to harvest data and reimagine business processes and customer outcomes.

The Data Lake arrived on the scene because our data technologists recognized the classic data assembly workhorse – the data warehouse – lacked the necessary agility. Yet despite significant efforts, data lakes have failed spectacularly at meeting business expectations.

There are two reasons why. First, **the baseline technology (Hadoop and its cohorts) is too complicated.** Despite heroics to surmount that challenge, the fatal blow is the second reason: **no governing architecture.**

While some organizations tout success, what they've achieved is enabling a small group of experts. They may be doing important work, they may have breakthroughs, but just like a lab researcher, no-one else can participate. As stated earlier, it is the collective contribution of knowledge workers that will drive far broader business innovation, much more than the lone researcher.

The Cloud vendors have addressed the technical challenges by abstracting and templating the data lake platform eliminating the IT engineering burden. Technology, therefore, is not the focus for our work here; it is the governance impediment we address.

The data lake objectives – data variety, immediacy, capacity, agility – with little thought about the data itself, left a mess. If you think of data like tools, there are lots of different things you can do with them. But if you throw all your tools into one big toolbox, how do you find the right tool for the right job? Do you even know if the tool exists? What if you don't know what the tool does? What if someone else is using it? What about the tools that don't work anymore?

Inherent in the overhead of the data warehouse was a deliberate evaluation and structuring of data – cumbersome and slow going, but "consumption ready" as a result. We need to marry the agility of the data lake with an agility in finding, understanding, and trusting what is in the lake, so it too is "consumption ready."

We call this combination a **Governed Data Lake** and this paper describes how such a data ecosystem is manifest by intertwining the technologies of Data Lakes, Data Catalog, Data Governance, and Data Security as an architecture and set of processes, and the critical role that IT assumes to make this possible.

The last decade has established these technology innovations; as we usher in the new decade, their use will fast become mainstream and organizations that apply and scale the methods we introduce here will have a distinct advantage.

THE DIFFERENT DRIVERS FOR BETTER DATA

Data governance addresses the objective of data trust. Much of this paper centers around digital transformation initiatives as the business driver for needing such trust, but there are other drivers that we should keep in mind listed briefly here.

Digital Transformation	The application of data-driven optimization to business processes requires the data has understood structure, meaning and accuracy
Business compliance	Certain business activities must be conducted in accordance with certain (often regulatory) stipulations, with compliance evidenced by robust data trails (or lack thereof) via traceability of a compliance directive to the relevant business activities to the relevant data.
Data privacy	Regulatory data "rights" that prevent improper "digital" behavior, typically regarding the citizenry, require organizations safeguard their data about you, convey what they record about you, and provide you a means to change such.
Business risk	Accurate data is essential for assessing the (in)adequacy of investments that mitigate potentially severe disruptions; and the (im)proper use of investments for their potential reward.

RETHINKING DATA GOVERNANCE

Allow me to share a story using the world of healthcare. This past summer, we ran a survey collecting input from over 200 people in health care who are leading improvement teams or processes. We spoke to physicians in leadership roles, heads of nursing, heads of population health, heads of clinical quality, financial leaders, and IT leaders. The purpose: to zero in on the data challenges most impacting their work.

The #1 issue from the survey – Data Governance. What's striking is that people like the "head of nursing" are talking about the need for data governance. This wasn't so five years ago. The importance of data is top of mind in most every facet of the workplace. Digital Transformation – shorthand for "data is at the center of modernizing business processes" – is so rooted that everyday people talk about data's importance; sounds small, but it's seismic.

DATA GOVERNANCE = DATA CLARITY = BETTER BUSINESS OUTCOMES

The premise is simple – good data habits are extremely impactful to business success, and data governance establishes those habits. If we go back to our survey participants, this is their real gripe; people living these "digital transformations" are saying their work is undermined if the data, it's meaning, and whence it came is uncertain.

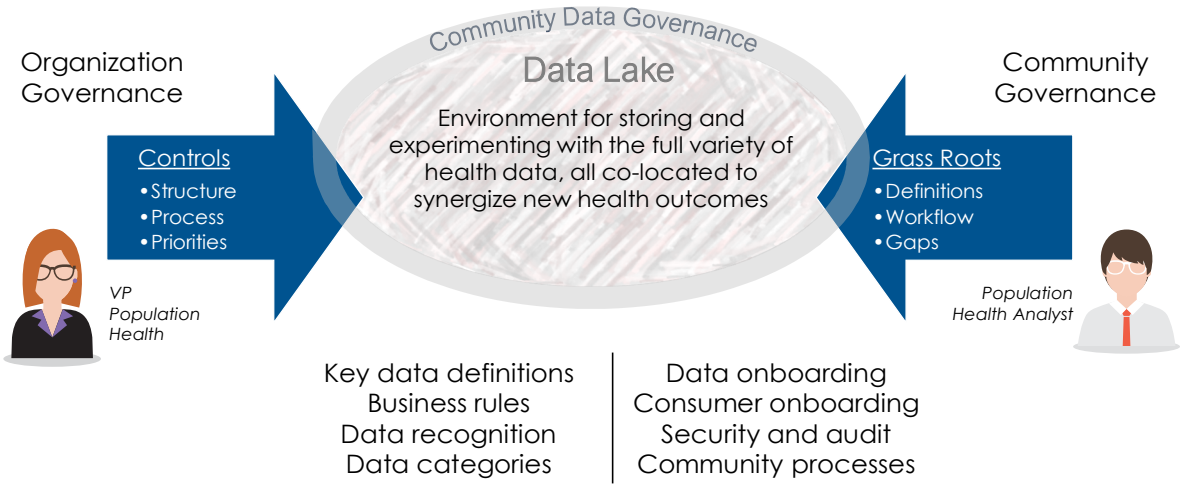
But here's the problem: most governance initiatives are organized as programs. Anything that is organized as a "program" implies a top down organizational structure driving the change, and the governance effort misses the organic vitality of connecting with the day-to-day reality of people getting things done.

Know what I need to know about the data that matters most to me and make it easy to get my hands on it. What's so hard about that? Not much, if I am the expert in the data – where it lives, its weird codes, how it was generated, and such. But if I am the expert, shouldn't I help to govern it? And shouldn't someone else who is also an expert help as well? And wouldn't it be best if I can "govern" as I work with the data? That is when I am most intimate with its details and idiosyncrasies.

This leads us to a critical maxim: **data governance programs struggle because:**

> The business of governance is conducted separate from the business of using data
> – and –
> Governance programs are not organized as a community process

We need to foster a "govern at the point of use" self-sustaining program of good data habits, **a culture of Community Governance**, and it is core to the Governed Data Lake architecture.



The experts in the data – in this case a population health analyst – is with whom the community driven governance occurs. As they do their work, important "tribal knowledge" regarding the data is captured. Meanwhile the department leaders – in this case the VP of Population Health – oversee a process for creating consensus and clarity on the captured tribal knowledge as they focus the organization on its priorities. Both parties propagate valuable governance outputs as they go about their day-to-day work.

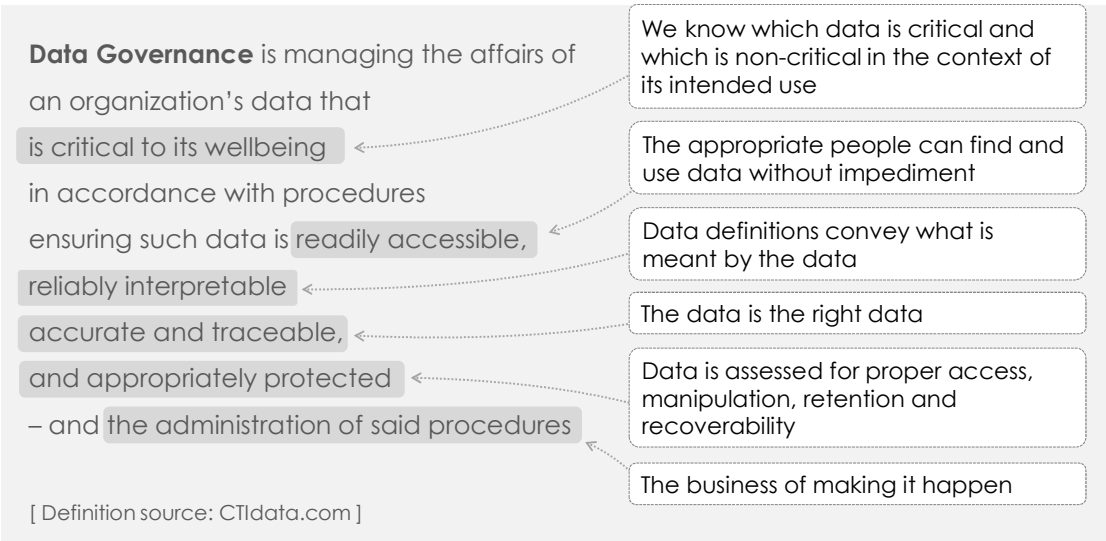
The other wonderful aspect is the **community governance process is inherently self-sustaining**: governance is integrated into how people work from the start, rather than being a separate endeavor.

Data Governance is an investment in organizational infrastructure and accumulates with an amplified impact over time. Except for compliance purposes, governance processes should be:

- User-centric (rather than bureaucratic),
- Adaptive, and
- Selective

STRATIFYING THE GOVERNANCE OF DATA

Ask a group of people to define "data governance" and you'll get a variety of responses. It's ironic that the term itself suffers from ambiguity, so here is how we define it:



One kind of obvious and critically important aspect is not all data needs to be equally governed, rather it is governed commensurate with its use. Therefore, we start by stratifying our data landscape. All data exists somewhere on this continuum:

Data towards the left has evolving and **varied interpretation** and therefore not dependable for important business work, and data towards the right has more **stable interpretation** and therefore more appropriate for important business work.

Keep in mind "interpretation" is the understanding conveyed by the values in the data (its codes, its numbers, its labels). If the interpretation is evolving, we do not have a consistent understanding across people. For some data that may be ok, it all depends on the intended purpose.

We can codify this notion of data interpretation as the meaning, structure, rules and quality (MSRQ) of the data; an aspect of business metadata (data that describes data) defined as:

	Description	Example
Meaning	The business meaning of the data values	The value "AP-6" means a patient is "admitted for observation", ….
Structure	The technical layout of the data	XX-N, where X is a single alpha and N is a single digit
Rules	The business rules that constrain the data	Code AP-6 is used only when the patient is an internal referral, …

Quality	Qualifier regarding the data accuracy	Codes from oncology are inconsistent and should be ignored

On the left of the continuum, we have dynamic and evolving MSRQ and on the right we necessarily have stable MSRQ. Keep in mind it's the uncertainty and variability in the MSRQ that moves us along the continuum, there is no law that says data must progress as such: some data may never be more than an experiment.

We now demarcate the continuum into four regions as follows:

Regions	Description	MSRQ
Certified	Drives critical business activities and changes must be carefully orchestrated to align with enterprise dependencies	Managed
Curated	Specific communities depend on the data for their work and MSRQ changes must be coordinated to avoid disruptions	Controlled
Experimental	Exploratory, may involve team collaboration, volatile, potentially throw-away with no long-term utility	Tribal
Individual	No purpose or visibility beyond that of an individual user	Implied

Degree of Change ↓ — ↑ Increasing data criticality

Increasing data reliability ↑ — Degree of Gov ↑

Many organizations are behind when assessed on this scale. Their Certified data, for example their core record keeping systems, may be ok; but their Curated data, for example their departmental databases and spreadsheets that are used to "get things done", are haphazardly governed.

DATA CATALOG, THE ESSENTIAL TWIN

At this point, we have established the fundamental characteristics of data governance:

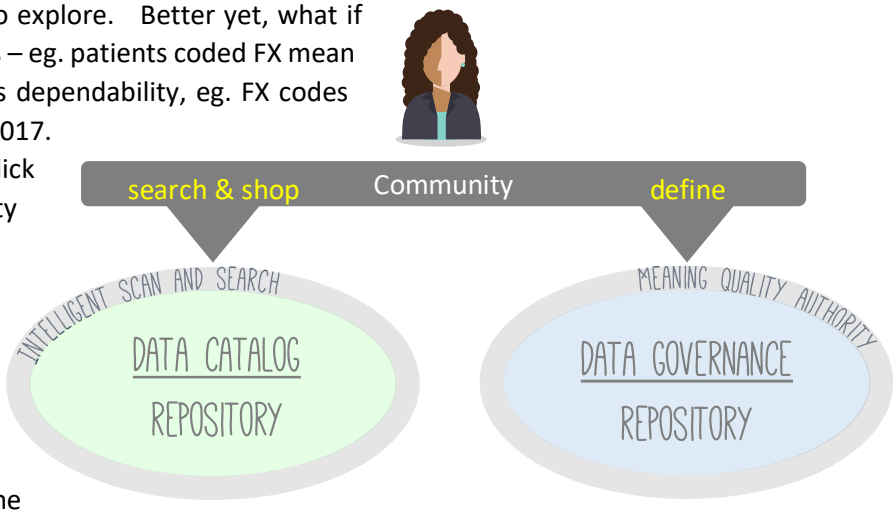
- We have a clear definition of what data governance is
- We recognize governance at the point of use as critical for forming good data habits
- We have a rubric for classifying the variability in metadata and its relationship to data governance

Now that we have a way to understand data, how do we go about making it accessible?

Suppose you work within the chronic disease improvement team of a hospital. You need to identify for all patients receiving medical services over the past 18 months those at risk of becoming diabetic, and further segment this group by different intervention programs assessing the costs of intervention versus the cost of the disease treatment and its related complications (co-morbidity). This patient list then needs to be assigned to care coordinators to act upon, and the patient progression tracked and benchmarked.

How do you even start? In most organizations, you'd need deep familiarity with the institution, the multiple data sources available, be a skilled data expert, and have a good clinical headset. In short, this is a very high bar. [similar examples are in every industry, imagine your own company for a moment.]

Now, what if our clinical thinker could use a search interface and get back a list of relevant data assets to explore. Better yet, what if the list included meaning definitions – eg. patients coded FX mean "financial hardship" – along with its dependability, eg. FX codes are not used consistently before 2017. Furthermore, what if they could click on "financial hardship" to get clarity on that definition too?



The **Data Catalog** repository provides this search capability while the **Data Governance** repository maintains the trusted meaning definitions. How the Data Catalog works is beyond the scope of this paper, but conceptually it scans and indexes the variety of data sources in the organization (including the Governance repository) and infers data types: be it generic types such as currency, date, or address – or business specific types such as financial hardship codes, medical prescription codes, or any code that can be pattern matched. In this way the user can search "high deductible patients", for example.

Along the way, the data is also profiled and quality assessed. Profiling and quality go hand-in-hand. It arms the business user with immediate visibility into the range and frequency of values ("hey, here's a diagnosis code P6-A that occurs infrequently and we have no idea what it means, need to sort that out"). It also arms the business user with a quality score driven off the data governance rules that have been established for the data.

BRINGING IT TO LIFE

We can examine how it all works – the data lake, the catalog, the governance process – by tracing a business workflow overlaid on this Governed Data Lake ecosystem.

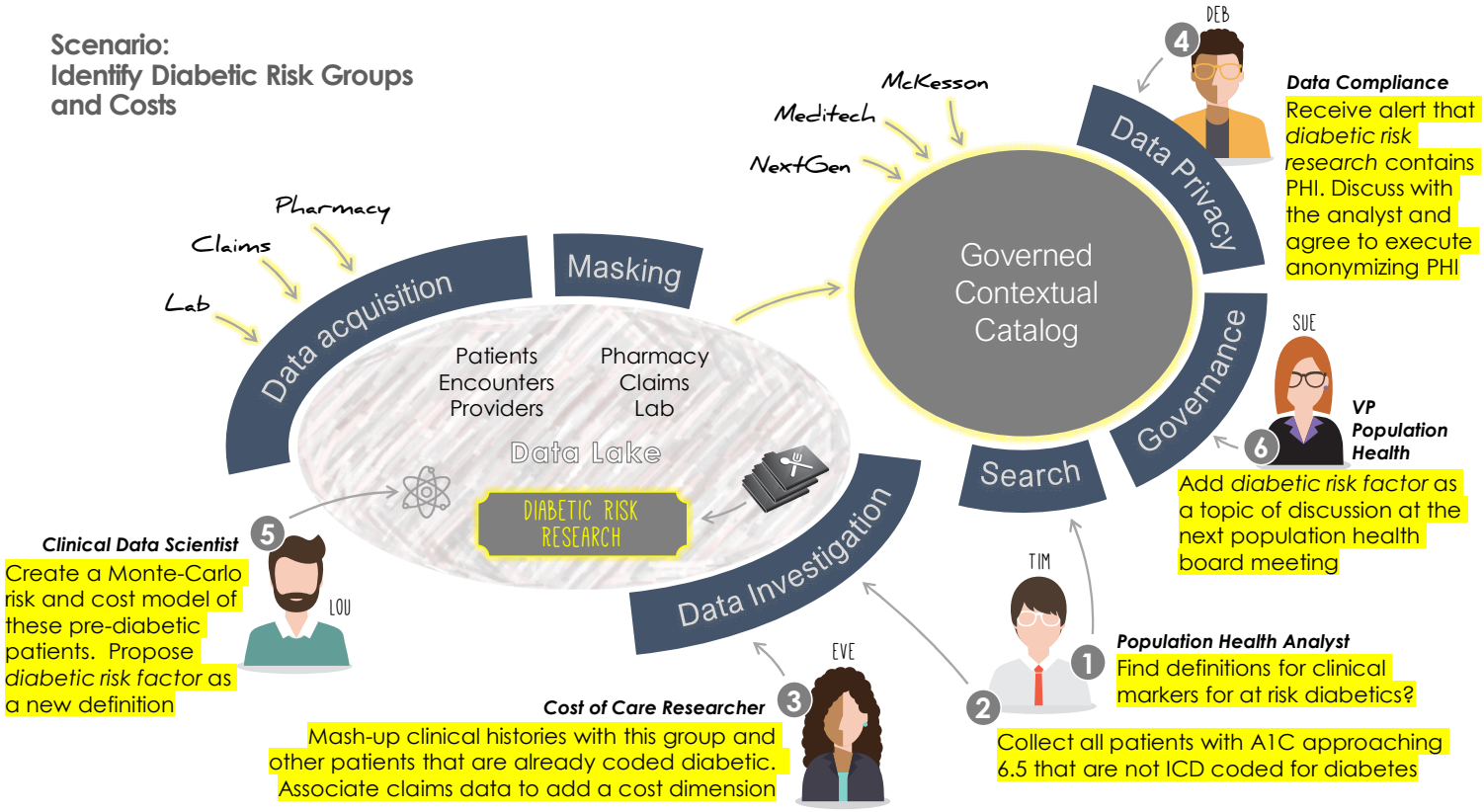
How data arrives in the lake is important but it's helpful if we first understand the governance and organizing structures of the lake which the arriving data occupies. Therefore, for now, we assume the lake has already been populated with a variety of data as-is from its various sources including databases, files and analytic reports – and leave for later how the data arrived.

[Sidebar: in general technologists have inappropriately prioritized their planning on how to get data in with its technically interesting challenges, rather than on how it should be governed which is where the value emerges]

In the workflow below, five different community roles collaborate in creating, enriching, securing and governing a new patient dataset in an evolving process that dovetails with the way people naturally work.

For the walkthrough, we'll use our earlier scenario: identify patients at risk of becoming diabetic and assess the costs of intervention versus the cost of disease treatment and its related complications (co-

morbidity). This patient list then needs to be assigned to care coordinators to act upon and the patient progression tracked and benchmarked.



1. The request lands on the desk of our Population Health Analyst, Tim. The first thing he does is search for definitions of diabetes risk. This hits the glossary stored in the Catalog. He finds a data set of diabetic patients created by another analyst but it's not useful, he needs to find patients that are not yet diabetic but show a risk of becoming so. He also finds a definition of diabetic clinical markers that could be helpful and investigates the thresholds of such markers.

2. Tim now looks for patients with a history of progression in these markers, filtering only those receiving medical services in the last 18 months (i.e. considered an active patient for insurance guidelines), and not coded for already diabetic. This becomes his starting list of, say, 1261 patients. Out of curiosity, he downloads zip-code census data from the web and blend this into the data to overlay socio-economic aspects. Finally, he gives this new data set a label and description for future search and retrieval.

3. At this point our population health analyst asks a colleague in the value-based programs office to help attribute costs to this list. The analyst, Eve, preps a list of patients that were coded diabetic 12-36 months ago, along with the services they have required. She blends in her own standardized services and pricing sheet (from Medicare) to create a normalized costing schedule rather than use the actual billing the patients have been charged.

4. Meanwhile, the data compliance manager, Deb, is notified of new data containing PHI. She contacts the originator, Tim, to discuss the data and if the PHI can be masked. Tim agrees to remove key fields but needs to keep the MRN (medical reference number) since the data may be used by care coordinators in the future. Deb adds a security restriction to the dataset and makes a note in the dataset conversation.

5. The population health analyst, Tim, now engages a clinical data scientist, Lou, in his department to create a risk model with varying risk scenarios that forecasts how many of the 1261 at-risk patients will become diabetic. The model incorporates some of the socio-economic factors that Tim had added earlier. The result is a risk-quotient applicable to specific demographic bands. The quotient has been back tested as a 95% accurate predictor and asserts that of the 1261 patients, 59 will become diabetic in the next 12 months and a further 188 thereafter.

   Consistent with the Community Governance culture, Lou submits a new glossary term "diabetic risk factor" with a description of the algorithm, it's purpose, and where it can be retrieved.

6. The population health VP, Sue, is notified of the glossary submission and discusses it with Lou. She deems it so important that she decides to bring the definition to the next board meeting.

That's a pretty fantastic workflow. But for this ecosystem to work, critical organizing processes are essential to its startup and ongoing sustainability. Processes that address the questions in the box below – without which the ecosystem will fracture and become unusable.

- How do we onboard data sources?
- How is data classified?
- How do we avoid duplication?
- How is it "approved"?
- How does it become operationalized?
- What if it changes? Who is impacted?
- How do we avoid data clutter? Who is the housekeeper?
- How do we make it self-organizing?

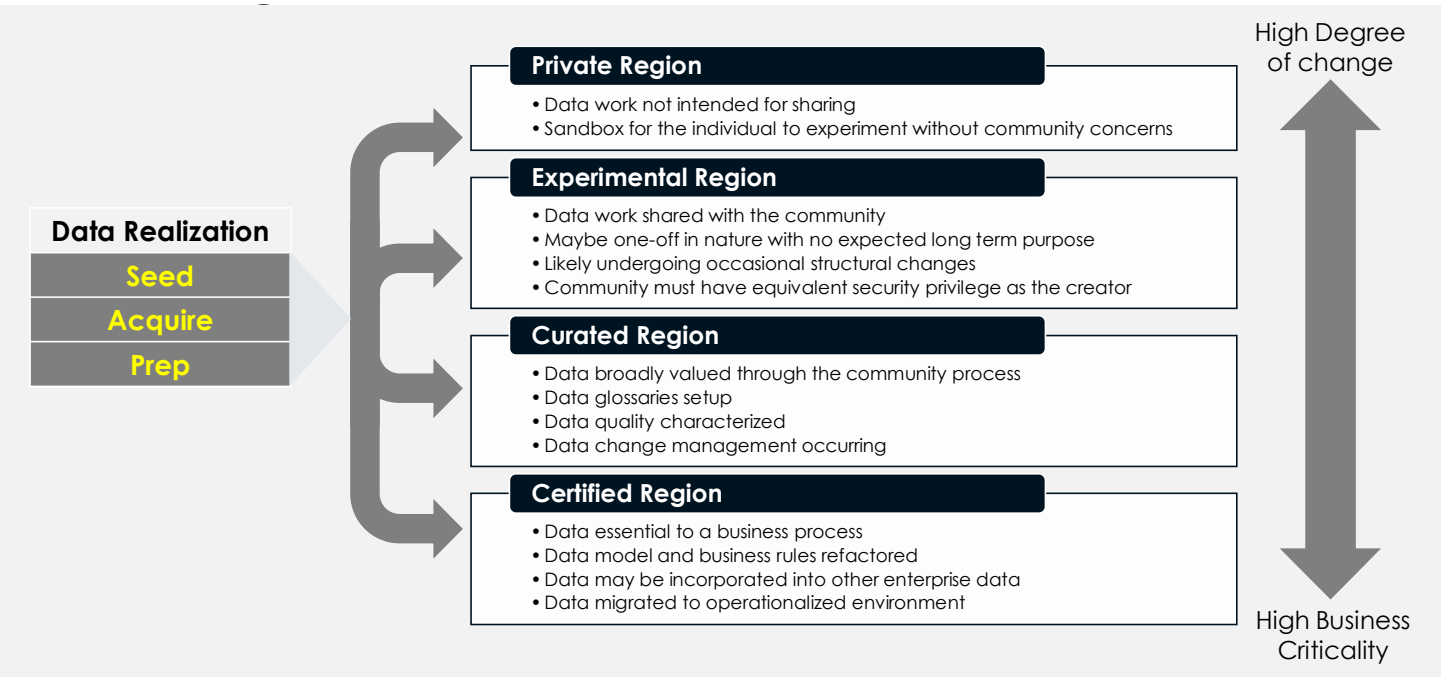
These challenges are where Operational Governance is applied. Where Data Governance is a community function that delivers trust, Operational Governance is an IT function that delivers cohesion and integrity.

This is a slight shift in the traditional IT data role. Typically, IT either provisions data platforms, or provides ready-made data solutions (such as data warehouses and reports).

In this ecosystem, IT plays a provision-and-guardian role – IT provisions and engineers the platform components, then overlays them with robust processes and structures for the benefit of the community. This brings predictability and confidence in the data evolving within the ecosystem and assures its vitality. We close this paper by mapping out these processes and how they cooperate. Before we do so, we must introduce the concepts of Data Regions and Data Realization.

Managing Data's Fit For Purpose using Regions

Returning to the earlier notion of **Data Regions** designed around meta-data maturation, regions are used as containers to recognize that as data is "worked" its impact to the organization evolves.



Note: not all data is required to "progress" – it is perfectly fine for some data to remain experimental and be discarded when no longer needed (perhaps informing new ideas or eliminating misconceptions) while other data evolves to have high utility.

Referring to the diagram above, we'll use the same diabetic risk scenario to explain the setup.

- The private region is a place for the individual to experiment isolated from the community, it's their private sandbox. The Population Health Analyst working with lab data, and patient data, and census data – might start in their private region.
- At some point the data may be ok for exposure and collaboration with the community (e.g. with the cost-of-care researcher and the data scientist) it moves to the Experimental region.
- As the work evolves, and in preparation for review with clinical leadership, it needs to be vetted by other clinical experts. The data transitions from Experimental to Curated. This requires that the data has rigor as evidenced by its accompanying glossary, quality, and known gaps. It also requires that the data is assured via change control processes.
- The data may be further enhanced (under change control) and then approved for use by care coordinators for patient follow-up.
- As the data becomes ingrained in the care coordinators routine, it needs to be operationalized: every week the at-risk list is automatically updated with new patients being added, and the less risky coming off. It is also improved to eliminate those already scheduled for follow-ups. The data moves to the Certified region. Here the data logic will be inspected and likely refactored, and the data generation and distribution process automated.

Keep in mind that data is composed of two parts – its metadata and its values -- with the user seamlessly interfacing with both. As such, the fact that the data is "native" to a region is part of its metadata. So, the user knows…

"I am using data that is a list of high-risk patients, and this list is experimental"

[and they know what "experimental" means since that too is defined]

The second aspect is interdependence: the list of high-risk patients may incorporate derived data from someone else's work examining, say, patient lifestyles. The metadata tracks such interdependence, so the user also knows…

> "This list of high-risk patients incorporates lifestyle scoring by the demographic sciences team, and that data is Certified"

This leverage of data work across the community is one of the most transformative impacts of the Governed Data Lake and it is powered by the metadata. At this point, we can assert a further significant shift in IT's relationship with data: away from creating business data models and towards creating business metadata models. The change is liberating and eliminates age-old IT data challenges such as:

- The work to learn the business, develop models (and build analytics) is slow and error prone
- The endless cycle of prioritizing the backlog of requests, leaving people frustrated with the delay
- The black-box conundrum of, for example, how did we calculate re-admissions?

[Indeed, these headwinds are why business analysts resort to their own data solutions.]

In this new role, **IT becomes a data custodian** who enables the community by providing:

- A facility (the governed data lake) to collect, blend and synthesize business data
- A repository that captures the meaning and accreditation of data
- Structures that organize and control the repository
- Structures that organize and control the data lifecycle
- Integrated technologies that make this possible

Current	Future
IT provisions data platforms	IT provisions data ecosystems
IT develops data models reflecting specific business scenarios, and reports/dashboards	IT develops metadata models to capture data meaning, family tree, and lifecycle
Business analysts use the analytic applications to make decisions	Business analysts experiment and synthesize data, describe its meaning, and apply it to their work

This role of **data custodian** is an effective signal of the change: a custodian is responsible for the safekeeping and integrity of an asset as it journeys through life, but not for the journey itself. IT shifts from prescribing how data is used, towards making data more accessible. IT shifts away from how to interpret data, towards making data more interpretable.

Managing Data's Realization

Regions are one aspect of operational governance. Equally important is how data is realized in the lake.

We call this "data realization" as opposed to "data loading" because data comes into being not just through a loading process, but also through user manipulation and experimentation of existing data (thereby creating new derived data), and through algorithms and automated processes generating new data. Regardless of these different forms of data realization, all data must be organized in the lake in accordance with the same principles.

There are three different methods at play. The first is **Seeding** and it is performed by the IT Data Management team (for certain data that is unsuitable for user-initiated acquisition) when:

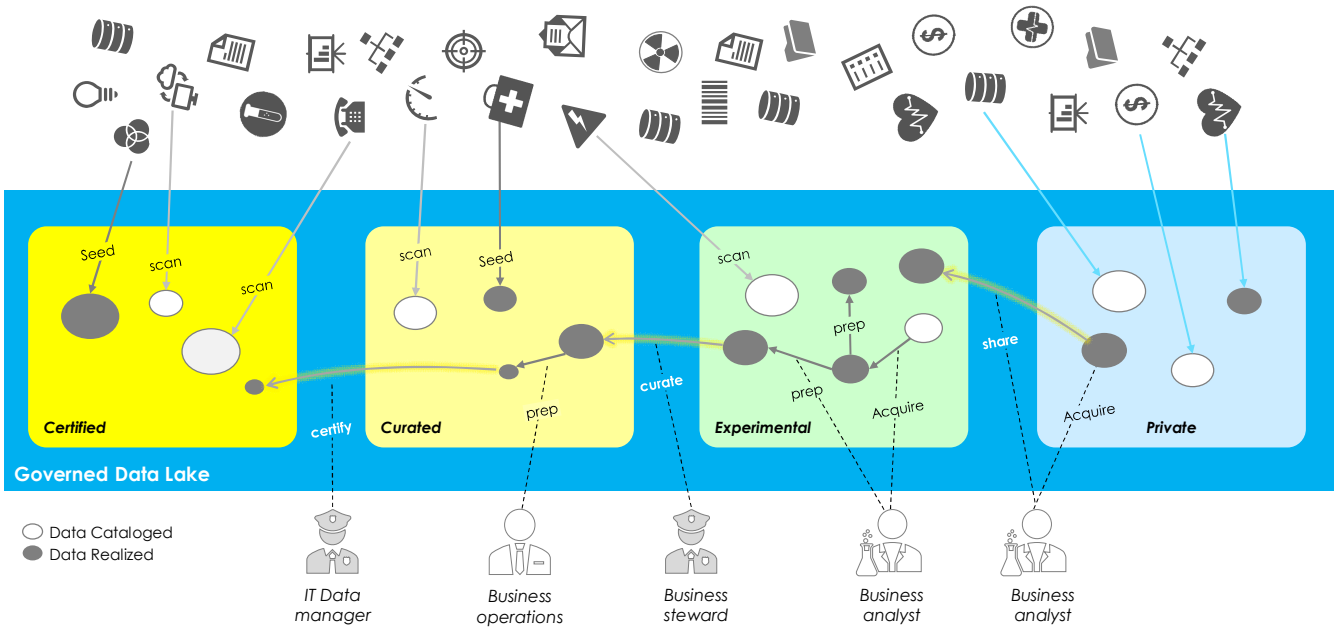
- Complex business rules need to be applied to make the data usable
- The volume of data is such that a subset is a necessary pre-requisite
- Security aspects are varied and require preparatory work
- Certain data has issues requiring specialized fix-up
- The data has high community utility that warrants upfront loading and classification

The second is **Acquiring** and is performed by the business user to bring new data into the lake, which is incredibly liberating and eliminates the pervasive problem of the community creating their own private data platforms (and is essential to community adoption of the data lake in the first place).

The last is **Prepping** and occurs as the user or automated processes manipulate and blend data to create new datasets that are core to their work.

Seed	Acquire	Prep
Eg. the main patient records	Eg. procedure codes	Eg. create the at-risk diabetic list
• Performed by Data Mgt. • Complex or sensitive data • Schema targets selected • Seeding targets assigned	• Performed by the user • Activates ingestion of a data source (no seeding)	• Performed by the user or process • Manipulation of existing data • Created on "go" (no seeding)
• Security classifications assigned • Data classifications assigned	• Security = user privilege • Data classification = user defined	• Security = user privilege • Data classification = user defined

Let's briefly visualize how this process looks below…



We can see how data is elevated from region to region as we discussed earlier, we also see how data is manifest as "just" meta-data (the hollow ovals) which means it is discoverable, versus data that physically exists in the lake (the solid ovals) which means it can be explored and manipulated.

As the user shops the metadata and selects various datasets, the ecosystem will automatically import the physical data from its source into the lake. While this is mostly seamless to the user, it requires time for the import to complete. The metadata generation occurs on a regular basis, the "scan" action in the diagram, ensuring the catalog reflects the current data landscape.

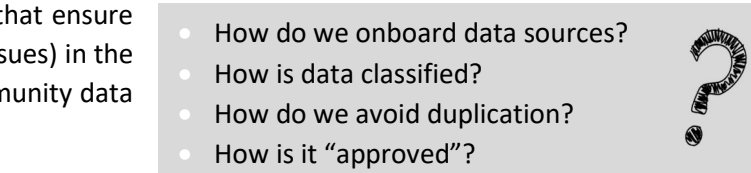
[Sidebar: The technicalities of how this all works are beyond the scope of this paper, our focus is on the framework that makes it sustainable – the need to Seed certain data, the use of Regions to classify data, and the authoritative roles that control Curating and Certifying data.]

GOVERNED DATA LAKE PROCESSES

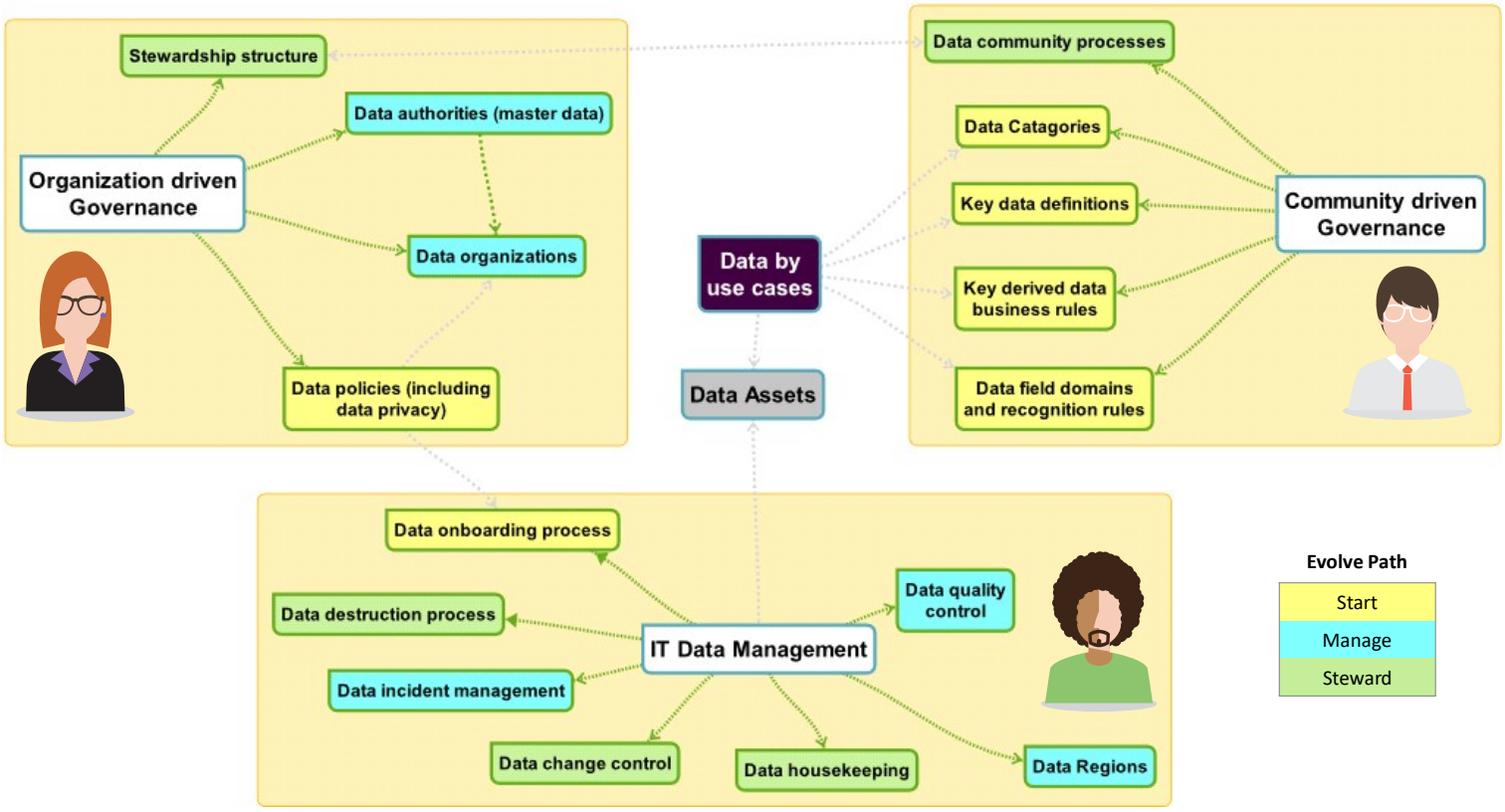
We now introduce the operational processes that ensure the vitality of the lake (addressing the hit list issues) in the grey box) and how they sit alongside the community data governance processes discussed earlier.

As depicted in the diagram below, there are 3 constituents that cooperate – the Data Governance aspects are business owned (the top two boxes), the **Operational Governance** is IT owned (the bottom box). Keep in mind these processes are structured around the Grass Roots Governance model which asserts…

- How do we onboard data sources?
- How is data classified?
- How do we avoid duplication?
- How is it "approved"?
- How does it become operationalized?
- What if it changes? Who is impacted?
- How do we avoid data clutter? Who is the housekeeper?
- How do we make it self-organizing?



Data governance is organized as a community effort, the business of governance occurs



In this structure:

- The community works with data and evolves the metadata
- The organization leaders control and approve the application of data and its metadata
- IT establishes and controls the ecosystem, and serves as the data custodian

The process responsibilities of each group are depicted in the diagram. These processes will take time to season and become part of the everyday. Consequently, we phase-in a "start-manage-steward" maturing sequence that mirrors the need for more controls as the data lake expands. The exact boundary points for such process phase-in will be customized to fit the culture of the organization and the pace of change that is practical. We show in the color coding a desirable scenario, but it is not critical that it be followed as depicted.

Along with phasing-in these processes, we also phase-in the data landscape by use-case. This allows the effort to be segmented and provides a graduated rollout to the community, improving adoption as it intersects with the kinds of work people want to accomplish.

The Start phase kicks the community process into gear. Business subject experts begin capturing the interpretation of their data as they work with it: they continue to do what they do day-to-day but stop making implicit assumptions but rather express and capture those assumptions.

As the utility of the lake widens, and these initial processes settle down, we beigin the Manage phase which introduces data controls. Controls that allow for better organizational boundaries (to meet compliance needs), and controls that stratify the needs for data interpretation and governance (the Regions).

At this point, we have community engagement and structures for controlling the lake, now we overlay oversight processes which we call the Steward phase. These processes control the inherently evolving and changing landscape of the data, and control how the community and business owners cooperate to govern the data. We also introduce further IT data management processes that address rigor regarding the handling of business-critical data and rigor regarding the "end-of-life" of data.

Each process is briefly described below with an example.

Process	Description	Example
Community Responsibilities		
Data categories	Specification for labelling of data	PHI/non PHI, Clinical, Provider, Financial
Key data definitions	Meaning of a data field	In-patient counts are all non-observation patients who stay overnight
Key derived data business rules	Meaning of a derived data field and the logic for its calculation	Diabetic risk is a person we expect to become diabetic in the next 18 months, and is true when the last two A1C lab tests >= 6.5 with BMI >= 25
Data field domains and recognition rules	Specification for automatically recognizing data field meaning	Auto-tag fields for ICD, Location, DOB, MRN (user corrections improve results)

Data community processes	Process for data stewards and community collaboration to enhance data meaning, value, quality, authority, etc	We are leveraging the census asset “p-census-05” prepared by population health in May – suggest we elevate its status to “certified”
Organization Leadership Responsibilities		
Data policies	Policies that govern data criticality, sensitivity, retention, destruction	All clinical research using data containing PHI must be verifiably anonymized or destroyed after 3 months
Data authorities	Process for identifying and establishing authoritative data	Certain data may be recommended by the community or data stewards to be recognized as “key data” by submitting a data elevation request
Data organizations	Specification for segmenting data by organizational characteristics	All patient data regarding affiliate hospitals (clinical, claims, etc.) is to be isolated via the new “organization identifier” and is not to be comingled with internal patient data
Stewardship structure	The organization/committees accountable for improving and certifying specific data	The head of provider credentialing, owns standardizing provider specialty codes and descriptions across all practice centers
IT Data Management Responsibilities		
Data onboarding process	Process for reviewing, cataloging and publishing new data according to its organizational characteristics	All data will be onboarded as follows: schema targets identified, data is classified, security restrictions are set, data domains are tagged, pre-load targets are identified, …
Data quality control	Assessing, warning, restricting, and improving data quality	The Key Data Item “referring provider” captured by the Glastonbury practice center is lacking consistent and accurate NPI codes
Data regions	Specifies the maturity context of data	Data may be in experimentation cycles, in a curated and trusted state, in a certified state, or decommissioned state. See details definitions later in this document.
Data incident management	Process for describing data problems for systematic follow-up	The lab results data is using the old provider ID codes
Data destruction process	Process for initiating and destroying data according to the organization’s standards	Data is targeted for destruction according to the “destruction criteria” and requires prior acknowledgement

		by impacted parties according to the destruction policy
Data change control	Process for managing data change ramifications to downstream dependencies	All encounter data is converting to new location coding, 15 dependent "certified" data assets need impact review
Data housekeeping	Process to address data no longer used, data having too many duplications, data that is orphaned (lineage to a dead source)	A monthly report will list data assets that have not been accessed in over 2 years and initiate a workflow with the data creator to confirm data deletion.

It is important to remember that these processes are phased-in specific to each business use case and its associated data. As the constituents learn to cooperate, "governance at the point of use" takes root and becomes the foundation for establishing good data habits that are sustained and then repeated throughout the enterprise.

In Summary

This paper prescribes a framework that ensures the data lake is useable as an enterprise asset by the community. Data Lakes are valuable not just for their agility, but also their enterprise purpose – we want the diversity of data, and the diversity of people, from across the enterprise, to be co-located (in the data lake) because the data innovations needed must transcend organization and department boundaries.

We established the critical components necessary, including methods for understanding and codifying the interpretation of data (MSRQ), methods for managing how data evolves its impact (the Regions), methods for the community to create and bring their own data aligned with enterprise needs (Data Realization), methods of data governance "at the point-of-use" ensuring proper behaviors are adopted, and repositioned how IT serves as a data custodian ensuring the business integrity of the lake.

These mechanisms can be applied to existing data lakes or be part of the launch for a new lake, when overlaid on modern data catalog, data governance and data privacy tooling.

The topic of data security and data privacy warrants its own discussion and the reader should contact CTI to understand the frameworks we have evolved for overlaying regulatory and other privacy and security needs within the Governed Data Lake framework.

A Philosophical Conclusion

Allow me to philosophize in the close of this paper. The amazing progress of human ingenuity is based on re-use – people coming up with a novel solution that borrows previous ideas, methods or systems as a new kind of application. Now re-use of an idea or a method is knowledge sharing -- the receiver acquires a thought model. Whereas re-use of a system is resource sharing – the receiver acquires a means of performing useful work. Think of it this way: one teaches you how to build a boat, which you then undertake for the purpose of fishing; the other delivers a boat, which you then adapt with nets and lines to use for fishing. Clearly, re-use of a system has scale and immediacy – the user doesn't need to learn and perform boat building.

In the world of high-tech, re-use of hardware systems was always so – components embed other electronic components to do new things. In the world of software, re-use exploded once the internet matured – software had always been crafted as a collection of parts to facilitate the engineering process, the internet allowed the parts themselves to be leveraged by other software systems.

The transformation now in our midst is the re-use of data, data processes, and data algorithms – but not by technology engineers, rather by business engineers. Think about the scale implications of business engineers innovating in this way and its huge impact across our societies.

ABOUT CTI

We transform business performance through innovative data models and advanced data platforms, data governance, data catalog, security and analytic technologies – solving the most complex and interesting challenges for our clients in Healthcare, Pharma / Life Sciences, Financial Services, Higher Education, and other industries.

CTI™ *Your Data is Our Passion* | ctidata.com