Paper DH05

A Data Lake is not a CDR and a CDR is not a Data Lake... Understanding the Differences and the Value of Each

Chris Decker, Instem, USA

ABSTRACT

According to Wikipedia a CDR is a database that consolidates data from a variety of clinical sources to present a unified view of a single patient while a 'Data Lake' is a centralized repository that allows organizations to store all their structured and unstructured data at any scale in its native format, without the need for preprocessing or modeling. Within our industry we struggle to align on the definition of these two things and how both can be leveraged together to meet the needs of an organization. This presentation will provide a foundation for understanding each and how they can be used within your organization as well as a look towards the future.

INTRODUCTION

Over the last few decades, the industry has used different words to describe the organization of data into a central place. Way back in the 1980s the term 'data warehousing' was introduced to capture data in a centralized and structured database. As the types of variability of data grew the concept of a 'data lake' was introduced and most recently even the phrase 'data fabric' has been a term used in the IT community to define an architecture facilitating end to end integration. Within our small world, we introduced the acronym CDR (we do love our acronyms), a Clinical Data Repository, to store our files in a organized and secure way and work with them in our clinical data lifecycle.

.

We struggle to define these different concepts the same way and understand how each can be leveraged to meet the needs of our organization. This paper will provide a high-level overview of each concept, the challenges and advantages of each, and how they can work together to help your organization achieve their goals. Finally I'll present a challenge for you to think outside the box about what data in the future needs to look like for us to really advance clinical research.

WHAT IS A DATA LAKE?

ChatGPT defines a data lake as a 'centralized repository that allows organizations to storage and manage vast amounts of structured and unstructured data at scale'. The fundamental concept of a data lake is to store all kinds of data in their raw format and to easily be able to grow that data lake to any size. Three decades ago, we didn't have the volume and variability of data we do today, nor did we have the advanced technologies to store, transform, and explore that data that are available to us.

Commented [TD1]: Store..

The key concepts of a data lake include:

- The ability to store large amounts of data whether they are structured (e.g. rows and columns data sets) or unstructured (e.g. images)
- The ability to store this information in its native format so you don't have to transform the data before you put it into the data lake
- Flexible, scalable, and cost -effective storage allowing you to easy add more data to the lake of all kinds
- Access to data science tools designed specifically for working with data lakes including AI, ML, and searching capabilities

While data lakes are powerful for storing the broad range of data collected today, leveraging them can be challenging for several reasons. First, managing the quality of that wide range of data can cause problems and leave you with uncertainty about the quality of the data you are using for downstream decisions. Very much like real world data, you need to understand the quality of the data ingested into your data lake and make sure the quality of that data is not impacted during the ingestion or transformation of the data. Second, the variability of the data types can be a blessing and a curse as you need to manage the metadata that defines the raw data which can introduce significant complexity. You need to spend time up front making sure you have adequately defined the metadata required for each data source and how you are going to manage it. Finally, understanding how you are going to control access to the wide range of data sources making sure the right people can see the data they are allowed to see is critical. Your organization must define a strong governance model for how you will ingest, manage, and make the data visible to consumers.

WHAT IS A CDR (CLINICAL DATA REPOSITORY)?

A clinical data repository (CDR) was a term that was introduced specifically within our clinical trial data lifecycle and is not something commonly known outside of our industry. ChatGPT describes it has a centralized database designed to collect, store, and manage clinical trial data from various sources and provides a structured and secure platform for the data. In our world we needed a term that went beyond just a storage system with folders and files and includes capabilities to manage the data securely and work with it transactionally to do the work of creating derived data and reports. It includes folders and files to hold the structured artifacts with a flexible security model, versioning, high performant processing, and the ability to access those files from various programming and analytical tools.

At its core a CDR is…

- A file-based repository to store the data, code, and outputs delivered for a clinical trial
- Heavily controlled and secure environment making sure each user has the right access to the right data for their clinical trial needs
- Very focused on the clinical trial workflow from generation of derived data sets and the summary reports that produce the analytical results
- A system that integrates with their Statistical Computing Environment (SCE) which is supposed to control the workflow around the clinical data lifecycle

While a CDR is very focused on the clinical trial data lifecycle it also has several challenges. The first challenge is when do the capabilities of CDR stop and an SCE start. This variability in people's understanding of these two solutions and the capabilities each should have has been a challenge for vendors in building solutions. Second, the continued increase in the volume and diversity of the data in clinical trials beyond traditional CRF data can cause problems that are partially solved through an easily scalable cloud environment. However, the part of having diverse data that doesn't fit into a folders and files mentality can limit the data you can use within a CDR. Finally, our existing legacy artifacts (i.e., rows and columns data sets and TFLs) and processes (i.e., transforming the data more times than we need to) leads us to implement CDRs with old ways of thinking.

Commented [TD2]: Should possess (?) has been a challenge

DATA LAKE VS CDR - DIFFERENCES

The table below outlines several of the key differences between the capabilities of a data lake vs a CDR. Some of these have been described above but let's expand on a few key points.

CAPABILITY	DATA LAKE	CDR
Data Types	Unstructured/Structured and anywhere in between	Structured & File Based
Data Schema	Raw, unfiltered, and easily updated	Processed, Derived, and difficult to evolve
Users	Data scientists, data engineers	Programmers, Statisticians
Security	Distributed, data point driven, easily complex	File based, groups/users
Workloads	Analytical and Exploratory	Operational and Transactional

Commented [TD3]: Clinical ?

When we reference the data schema, a schema defines the structure and rules of how the data is stored and accessed. In a CDR you are limited by folders, files, and standard security models attached to those folders and files so you are limited on what you can do beyond metadata on the artifacts in the CDR. Within a data lake you have a flexible and scalable schema based on the type of data you are ingesting into the Lake and therefore, have a lot more options for how you store, search, and retrieve information. With regards to the type of work usually done within each tool, usually a data lake is seen as place for exploratory analytics with cutting edge tools where a CDR is often referred to as the boring machinery of generating derived data and TLFs.

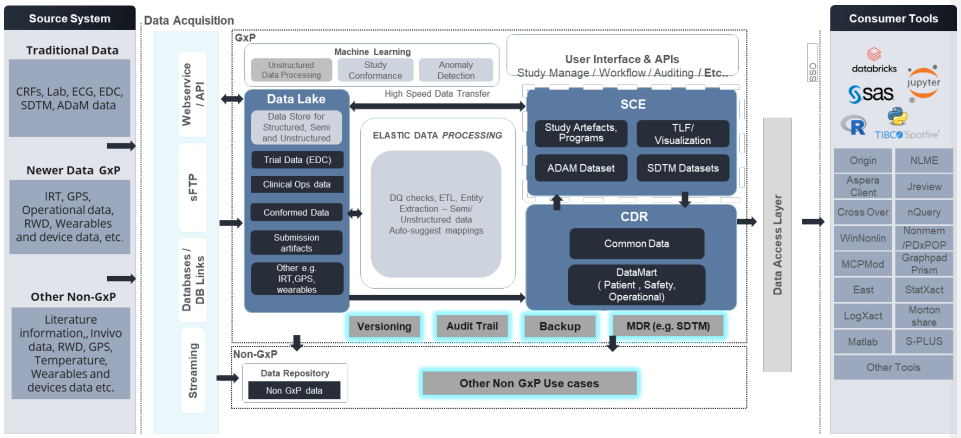
As we think about the differences between the two, frequent questions arise

- Where do you standardize our data?
- Where do you transform the data?
- Does all data go through or the Data Lake or does some go around?
- When and what data produced in the CDR goes back in Data Lake?

These are questions we see clients grapple with as they try to define the best way to leverage both a Data Lake and CDR within their environment. While there isn't one black and white answer to each question there are best practices you can follow.

LEVERAGING BOTH TO ACCELERATE YOUR WORK

As we think about answering the questions above and describing how you might leverage both a Data Lake and CDR we will use a reference architecture to describe the flow of information.



First, within the figure above we must first understand the need for storing data expands beyond our small world of collected patient data and includes digital health data, wearables, operational data, and non-GxP data such as literature and real-world data. The ability to pull this data in a central place for use throughout the organization is key and is the fundamental need for a data lake. Once this information is ingested into the data lake the next step tis to govern and secure that information so you know what you have and can easily share with others. You must understand who the consumers of the data lake are and their needs, so you know how to organize the information. Once you identify those needs, the next step is to make sure you have seamless interfaces to extract the information for the specific need of each stakeholder community. In this case we have a need to pull patient data (i.e. both CRF and non-CRF) into the CDR so you can use that data to build your derived data and generate your key analysis results. In addition, you might need to pull in your operational data such as your protocol information, site management, or other key information to help drive the workflows within your CDR. The CDR should be the single source of the clinical trial results.

Commented [TD4]: is

After you produce your clinical trial data and results the next question is whether the data lake becomes the final source of truth after the study is completed. This really depends on who needs access to the data and what they want to do with it. Let's now come back to the questions above.

- *Where do you standardize our data?*
  Simple answer: Both. Depending on what you are trying to achieve you can standardize the data in either place. In most cases if our goal is the standard data required for submission (i.e. SDTM and ADaM) then the best place for standardization is most likely in the CDR. If we are talking about standardization for broader purposes across the organization then it might make sense to do some of this work in the Data Lake.

Commented [TD5]: In both perhaps? Otherwise it is not that clear..

- *Where do you transform the data?*
  Same answer as above. It depends on the target for the transformation.

- *Does all data go through or the Data Lake or does some go around?*
  Ideally, to reduce the complexity of your data flow you should always have the data come through the data lake. Even if you aren't doing anything directly with that data today in the data lake you might in the future so having one data flow will be simpler in the long run.

- *When and what data produced in the CDR goes back in Data Lake?*

While this is a decision based on the use case and the needs of consumers outside of the CDR work, some or all the study data and analysis results should go back to the data lake at a point in time that makes sense for your consumer needs. Making this information available in the data lake isolates key work within the CDR and provides that control over who can access this information outside your key CDR users

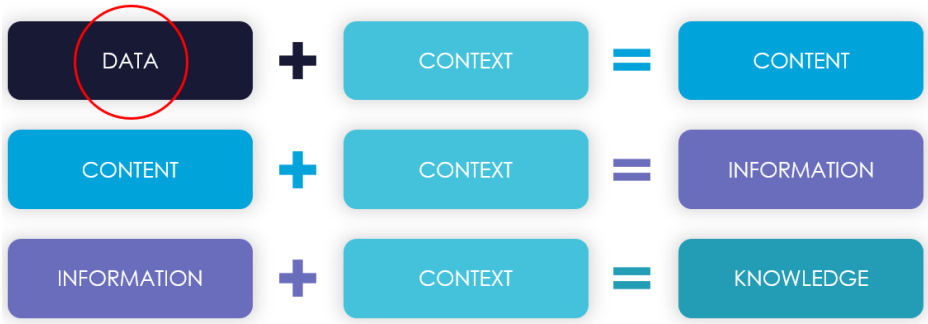
WHAT DOES THE FUTURE HOLD?

If we take a step back from the previous discussion and think about what the future might hold, we need to find a path to expand our thinking about the way we have thought about data in the past. Florence Barkats, a RWE subject matter expert had the following quote which I use often in trying to get people to think about things in a different way.

"In the real world, we want to capture the complexity and diversity of the standard of care without forcing it into an artificial structure. The current technology framework used within clinical trials to collect information often makes researchers constrain data to a structure that is not real and has no context…… and forces them into a rigid non intuitive thought process"

If we explore this quote, what she is saying is that we are trying to fit a round object into a square hole. The clinical information about a patient is complex and multi-dimensional and requires a model that can support these relationships. In today's world we force our data into two dimensions across different domains and we lose the context and relationships of the data limiting our ability to get meaningful information out of the data.

The challenge is that we have been doing this for so long that our systems, processes, and workflows are all centered around this representation of the data – from collection through data cleaning and multiple transformation cycles to the final step of analytics. This overall ecosystem is so well established that trying to change it is almost impossible. However, if we don't keep trying, we are going to continue to be limited in the advancements we can make in accelerating clinical research.



The diagram above points to the need for us to change the data paradigm. Today, we primarily only have data and we had knowledge through a lot of manual human intervention. We need to start adding more context to the data so that we have content – data that tells you what it is about. If we then take that content and had context about the relationships between the data, we have information – data that can tell you how it is linked together. Finally, if we add context to this information, we now have the knowledge to link the data from beginning to end and then can drive automation, gain value of the data that we have never before, and really accelerate the work we are doing.

CONCLUSION

In summary, in todays' world, a Data Lake and CDR are essential and can be leveraged together to address the broad range of use cases within your company. The Data Lake can be the central hub for all your information, structured and unstructured and can be governed to make sure the right people have the right access across the organization. The CDR can continue to focus on delivering the well-controlled and structured folders and files paradigm to execute the clinical trial analyses that need to be completed and included in submissions to regulatory authorities. Understanding the relationship between the two and how they must interact for your specific needs is critical to ensure value out of both.

Commented [TD6]: Double space

Commented [TD7]: Double space

As we look to the future, we must think about how data is captured, linked, and surfaced in other spaces outside of our world and how we need to change and adapt to how our data paradigm needs to look in the future. Today, we

are limited by what we can do because we are stuck is a rigid thought process about the data. Let's look for opportunities to break out of that mindset and think about what could be.

Commented [TD8]: Double space

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Chris Decker
Instem
North Carolina, United States
Work Phone: 919-995-0957
Email: Chris.Decker@instem.com
Web: www.instem.com

Brand and product names are trademarks of their respective companies.