White Paper

Enterprise data strategies for pharmaceutical organizations

Derive actionable insights from your data lake with a virtual lake house

Ray Deiotte, NetApp

May 2021 | WP-7343

Abstract

As pharma companies work to transform their business with technology and data, the need for common access and consistent delivery of information is paramount. Traditional data lakes and data warehouses can fall short of expectations. A virtual lake house (VLH) strategy provides a comprehensive data perspective while enabling incremental transformation without having to replicate, move, format, or even curate existing data.

TABLE OF CONTENTS

LIST OF FIGURES

The challenge of data for pharmaceutical companies

The data landscape for pharmaceutical companies has exploded in recent years. With the expansion of multiomic sequences, clinical trial data, and manufacturing surveillance, and with the addition of real-world data (RWD) and real-world evidence (RWE), the volume, variety, veracity, and velocity of the data have evolved. More importantly, the decisions that are made based on that data have changed dramatically.

Unfortunately, the pace of data management and strategies for data usage have not kept up with demand. Most organizations still operate in the realm of SQL databases with silos of information hidden behind walls of organizational confusion. Other organizations use warehouses to combine data from multiple databases into a multidomain repository that supports elevated decision making for a few key decision makers. The remaining organizations create data lakes with tools like Hadoop at their core.

Warehouses and data lakes provide a centralized data repository. However, both strategies are lacking in at least one of the following areas that are necessary to deliver the impact or return on investment that pharma organizations want:

- **Breadth.** The typical warehouse consists of the contents of two to five individual databases. The data is organized into a schema that links all the sources into a form or a set of forms that can be queried for decisions that are supported by the source and/or warehouse data. Although they are useful, most warehouses span only one or two business units. This limited breadth in coverage hampers the organization's ability to empower enterprise decisions.
- **Depth.** Many pharma organizations struggle with having a multitude of incomplete data sources. This lack of depth can prohibit enterprise insights. Usually, this problem relates to misalignment of temporal periods across multiple datasets, but it can often be a simple volume problem that can skew analytical results through bias of evidence.
- **Clarity.** The issue of clarity is most significant for data lakes and lake-like initiatives. When data is dumped into an infrastructure without any concern for its purpose or organization, the collection of data ends up looking more like a swamp than a lake. In this situation, it's difficult to extract meaningful insights because there is no understanding of data within the collection or how that data may be related, even loosely. Keeping a lake pristine is critical to the success of major data consolidation initiatives. Many factors contribute to a lack of clarity, including governance, curation, and stewardship. Without clarity, data investments are wasted.

A VLH strategy

The optimal data management strategy for enterprise pharma is a VLH. A VLH combines components of data virtualization, data warehousing, and data laking into a single data platform. This strategy spans on-premises and cloud resources. It can use existing infrastructure and data management capabilities without introducing tremendous changes or costs up front.

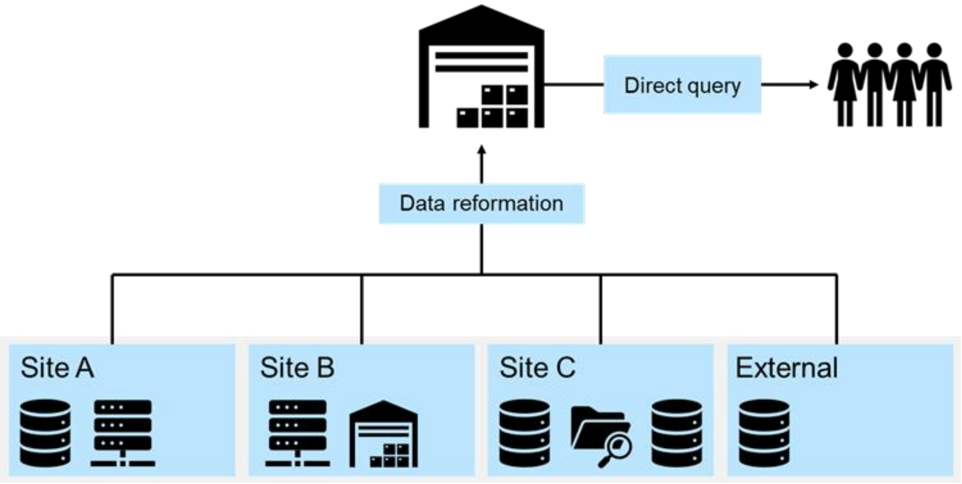
What is a VLH?

Before we get into the details of a VLH, it's helpful to ground ourselves in the traditional concepts of data warehouses and data lakes first.

Data warehouses

In traditional data warehouses, source systems send data to the warehouse in some periodic fashion (the number and variety of these techniques are a long list and will not be covered in this discussion). The source data is reformed in accordance with a preestablished structure or structures in the warehouse. That data is then available for users to directly query to build analytics and to make decisions.

Figure 1) Data warehouse architecture.

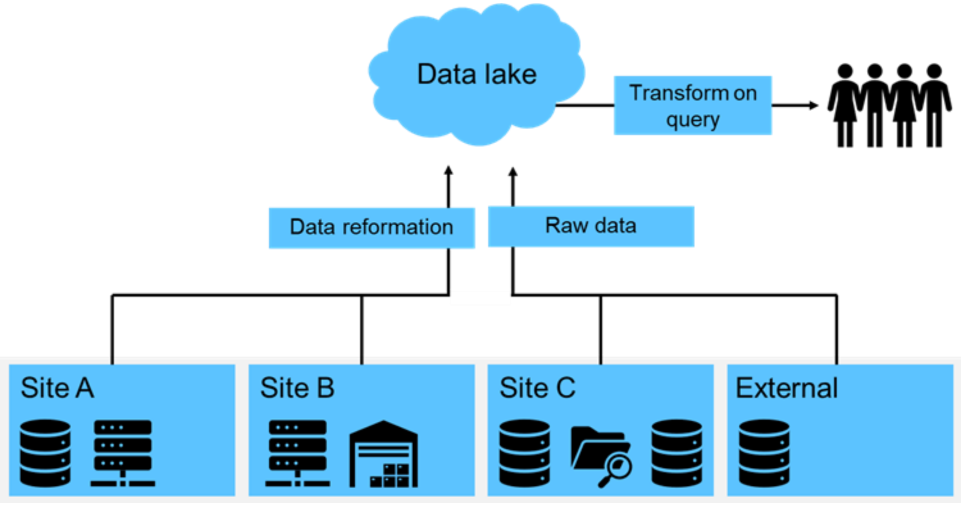


The warehouse has significant drawbacks—the structure of the warehouse must be known a priori, and data must conform to that structure. Adding more data and use cases or otherwise altering the data structure of the warehouse after it has been initialized can wreak havoc on end users.

Data lakes

A data lake is set up like a warehouse, but with two major differences. First, the data lake has no defined structure. Data can be extracted from source systems and imported directly into the lake. This approach has the advantage of allowing datasets to be added to the lake ad hoc, which tremendously expands the utility of the platform and the decisions that can be affected by the data in the lake. The second difference between a data lake and a data warehouse is that users must perform data engineering and/or transformation on a query. Because the lake has no established structure and multiple data sources can be joined to answer questions, more responsibility falls to the data consumer. This responsibility can be seen as a disadvantage because of the need for a more literate, savvy consumer. However, there are many tools that help simplify this process by providing visual data engineering, management, extraction, and analytics that don't require much workforce upskilling.

Figure 2) Data lake architecture.



Data lakes have two significant drawbacks:

- Without good end-user tooling, data lakes can fast become data swamps, in which data is no longer usable or consumable because nothing can be discovered or effectively extracted.
- A traditional lake requires establishment of all the hardware, software, storage, and tooling before the system can be usable. Following this big bang approach of system implementation is prohibitive to many organizations, and the risk from the outlay of capital for a data lake is typically unpalatable.

VLH

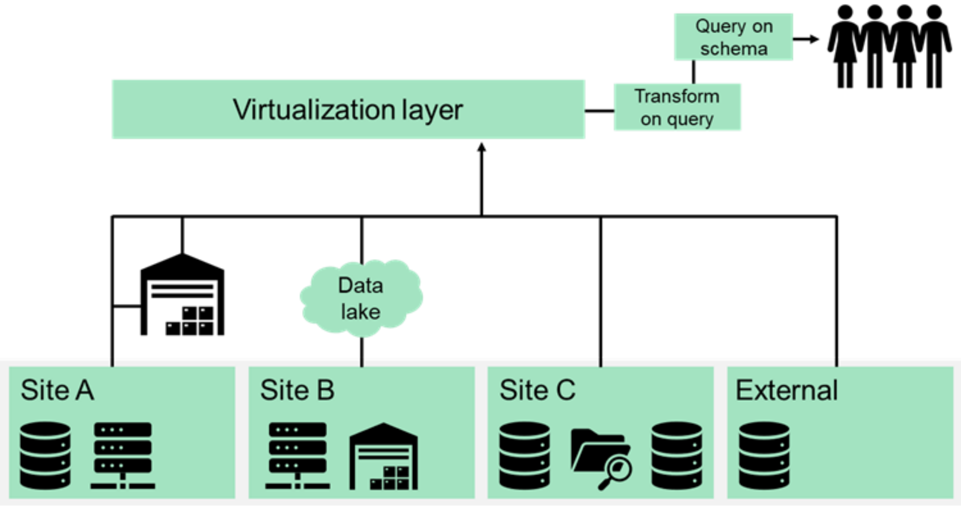
That brings us to the VLH. A VLH uses the best of data lakes and data warehouses and introduces a third concept: virtualization. Data virtualization allows access into native datastores through a connector that's managed by the virtualization layer. Data can then be queried, joined, and otherwise used according to any of the following:

- Structure of the datastore
- Structure of the virtualization layer
- How the consumer engineers that data with other data in the layer

The virtualization layer can front-file structures, databases, warehouses, lakes, object stores, or even APIs and coalesce the data into a discoverable, manageable, exploitable virtual lake. None of the data actually persists in the virtualization layer. This approach can potentially cause some complexities that need to be planned for:

- Accessing multiple stores in multiple locations can cause some latency in data access. For applications that require rapid response (<1 second), virtualization isn't the answer. Landing virtualized data into a data mart that persists or is transient can solve this problem. This kind of sandboxing is great for machine learning and advanced analytics and can be accessed through the virtualization layer. It reduces latency to tolerable levels.
- Data cataloging and governance are absolutely critical and must be addressed with as much diligence and fervor as for the platform itself.

Figure 3) VLH architecture.



Benefits of a VLH

The VLH provides the ultimate in flexibility and scalability while retaining the power of lakes and warehouses—without succumbing to their limitations. This strategy offers multiple benefits to pharma organizations, including:

- Iterative development
- Centralized control without centralized reorganization
- Tunable breadth and depth
- Incremental governance and expertise
- Self-service

Iterative development

By using the virtualization capabilities, you can incrementally engage data sources into the VLH as your business needs dictate. You can organize data into static marts, warehouses, or databases for repeated use, or you can coalesce data dynamically (querying directly from source repositories). Your assets can then be stored in one location until it makes sense to relocate them to a place that provides the most value.

Centralized control without centralized reorganization

When moving to a lake or a warehouse, data must be reorganized from the source. This step improves control and optimization of the data in the new repository. With a VLH, you gain control over the data that you need when you need it. A VLH enables incremental delivery of capabilities into either a mart, a warehouse, or a lake that scales over time with your business needs. A VLH also enables caching or real-time access into data sources without operational disruption or duplication of costs or data. Data sources with existing users can remain in place until the source is end of life or until your organization is ready to clean and engineer the data into a form that's better suited to the use cases that involve it.

Tunable breadth and depth

With a VLH, you can include or exclude as much or as little of the source data systems as you like, without having to refactor or reposition the data. You save time and money by avoiding extensive pre-engineering and hardware and software costs. You can move data into a mart or a warehouse and form the foundation for a permanent data source as necessary.

Incremental governance and mastery

Data governance and analytics require a lot of manual work. Incremental gains are difficult to achieve. Cataloging entire data sources can be complicated and challenging. The traditional one-by-one methodology takes a tremendous amount of time and does not account for the utility of the data.

With a VLH, your active data is immediately subjected to governance. Subject-matter experts create well-defined concepts of intended use, allowing this process to be performed quickly and effectively. Your entire enterprise is incrementally subjected to data governance without slowing down value generation.

Self-service

Traditional self-service models involve business intelligence developers, report writers, data extractors, registry creators, and data engineers for each data request. Depending on the volume of requests, this process can take weeks or months. This time lag forces your data consumers to make uninformed decisions or to postpone decisions, which could severely hurt or limit your organization.

With a VLH, self-service access to data and analytics is accomplished with ease and control. Dataset creation and data engineering are decentralized, so your business data stewards can manage your organization's data requests with common curation, and governance standards. Data experts can develop solutions for your consumers, and more individuals can support this process, bringing down response times for custom-curated datasets.

Although this example is not pure self-service, most data and analytics customers benefit from data catalogs and analytics anthologies. You can use them to find the data and analytics that meet your needs. A VLH also supports rapid data inclusion, so you can quickly discover and share internal and external data sources. With the growth of RWD and RWE, this flexibility and inclusion for self-service are crucial.

A data fabric—the key to success

To build an effective VLH, you need a data fabric that spans edge, core, and cloud. This data fabric must enable seamless data movement, access, consistency, and availability across whatever environment your enterprise operates within.

Most data fabric solutions abstract the fundamental data layer away. A successful data fabric provides a single access point or hook into a global file structure or object stores on the premises and in the cloud. A data fabric should also enable you to move object stores that support databases. This capability is critical for the virtualization layer of your VLH and to accelerate access and governance. It also provides security and privacy controls as part of a larger, comprehensive security strategy.

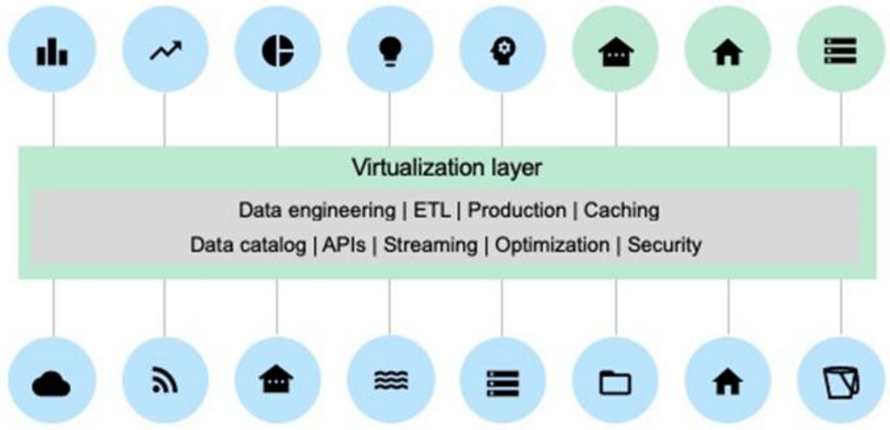
A data fabric that's powered by NetApp® technology revolves around NetApp ONTAP® data management software. Whether data resides on spinning-disk, hybrid, or all-flash storage; on the premises or in the cloud; in files, blocks, or objects, you can manage it with ONTAP. You get seamless movement and access of your data from anywhere in a consistent, well-managed namespace. Native security and privacy capabilities integrate directly with third-party security solutions. And NetApp Cloud Insights helps you gain deeper, richer insight into your data and helps optimize data science workflows in your production environment.

The VLH reference architecture

The VLH reference architecture can be broken out into three distinct layers:

- **Information layer.** Databases, warehouses, lakes, virtualization, governance and cataloging, business intelligence, and other tools make up the information layer. These tools enhance the capabilities in the data layer. They provide a curated view of raw data. A VLH brings them together so that incremental progress can be made toward complete enterprise democratization.
- **Data layer.** The integration of these technologies at the information layer is supported by the unencumbered flow of data at the data layer. Data should be put on optimal technologies to support the information layer.
- **Knowledge layer.** At the knowledge layer, consumption of information becomes knowledge that helps you prescribe and make decisions. It often includes advanced analytics through the use of artificial intelligence (AI) and machine learning (ML), or AI/ML. AI/ML requires well-curated information. Consumption and analytics feed back into the information layer, creating additional information on which to act later.

Figure 4) VLH reference architecture.


Virtualization layer
Data engineering | ETL | Production | Caching
Data catalog | APIs | Streaming | Optimization | Security

Empower your digital transformation

As pharma companies work to transform their businesses with technology and data, the need for common access and consistent delivery of information is paramount. From tactical decision making to strategic planning, comprehensive data perspectives are the key to breaking down barriers and deriving valuable insights.

Traditional strategies of data laking, warehousing, and even marting tend to be either narrow attempts to expose data and information or big bang efforts that take months or even years to implement.

With a VLH, you can transform your system incrementally without having to replicate, move, format, or even curate existing data. Practitioners and decision makers can use capabilities as they come online to propel strategic initiatives and to accelerate development. The incremental nature of a VLH gives you tremendous flexibility in growth, transformation, and technology adoption while delivering consistent value.

NetApp can help you build an effective VLH. Our broad range of tools span edge, core, and cloud, and can be brought together incrementally to meet your business and technology needs. Having one operating system that spans all our capabilities means that you can derive value without wasting time procuring and integrating a new tech stack.

With NetApp and its rich partner community, you get the technology that you need to empower your transformation initiatives and to turn your organization into a data-driven enterprise.

Where to find additional information

To learn more about the information that is described in this document, review the following documents and/or websites:

NetApp resources

- Blog: Cloud Data Lake in 5 Steps
- Blog: A NetApp IT Perspective: Data Science and the Data Lake
- NetApp Community: ONTAP Recipes: Easily Create a Data Lake Using ONTAP Storage

External resources

- Blog (Databricks): What is a Lakehouse?
- Blog (Medium): Data LakeHouse—Paradigm of the Decade
- Blog (Eckerson Group): An Architect’s View of the Data Lakehouse: Perplexity and Perspective

Version history

Version	Date	Document version history
Version 1.0	May 2021	Initial release.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright information

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

Data contained herein pertains to a commercial item (as defined in FAR 2.101) and is proprietary to NetApp, Inc. The U.S. Government has a non-exclusive, non-transferrable, non-sublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b).

Trademark information

WP-7343-0521