

Enabling the Breakthroughs of Tomorrow: The Data Lakehouse for Clinical Trials

Venugopal Mallarapu, eClinical Solutions, Mansfield MA, USA

ABSTRACT

Clinical trials continue to grow in complexity amid decentralization, digital trials, and ever-increasing surge of data from multiple external sources. Meanwhile, leaders in life sciences are under immense pressure to reduce cycle times, gain oversight, and mitigate issues and delays in the clinical data lifecycle. Modern research requires a scalable clinical data lakehouse architecture. In this session, attendees will learn about the key components of a clinical data lakehouse, why it's the ideal compared to the data lake or data warehouse, and what is different about the data lakehouse in a regulated environment. Find out how the data lakehouse blueprint supports the use of advanced techniques like AI/ML to accelerate data science and secondary use of data, beyond the submission pipeline. Explore how a clinical data lakehouse not only delivers the critical scalability for faster time to value and decreased cycle times, but also enables the scientific innovation to power tomorrow's breakthroughs.

INTRODUCTION

Clinical Research has been continually transforming through adoption of new pathways, tools, technologies, and techniques. Events like COVID has accelerated adoption of Decentralized Clinical Trials substantially as per the research report published by GlobalData on "Clinical Trials: The Movement Towards Decentralized Clinical Trials", from 250 trials in 2012 to 1425 trials in 2022 (by Sep 2022. Another report by Institute of Medicine (US) RoundTable on R&D found that an average clinical trial generates up to three million data points and counting.). Such developments are increasing the complexity of data collection, cleanup, aggregation, and analysis.

It has become a universally accepted notion that data is the currency of the life sciences industry, and how we aggregate and analyze clinical data plays a critical role in getting treatments to patients faster. In turn, data technology architectures are evolving to support the deluge of data and complexity of today's clinical trials while aiming to reduce cycle times.

INDUSTRY CHALLENGES & TRENDS

Irrespective of our religious beliefs and whether we believe in creation or evolution every human being trust finding remedies for problems ranging from everyday health concerns to the most serious diseases on earth. As a human race we have come a long way in making progress to find treatments for a lot of maladies and our ability to make rapid progress when challenged with grave situations with the most recent COVID pandemic is a testament to how far medicine

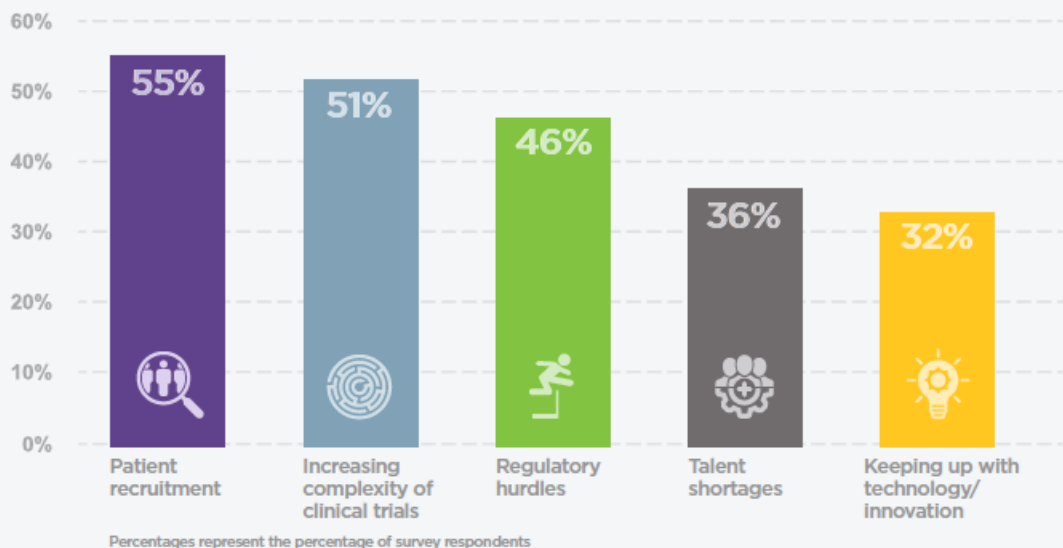
has come. Despite this progress Life Sciences remains a complex industry and continue to be challenged with new and evolving issues that goes with the time.

Advancing clinical research is challenging. Historically we have struggled to recruit patients for our studies. As mentioned above, the complexity of the trials continues to increase. In a recent survey by PPD, of more than 150 drug developers worldwide, these top 5 challenges below were highlighted.

PPD surveyed more than 150 drug developers worldwide to gather insights on the state of the drug development industry.



Drug developers reported that the COVID-19 pandemic and evolving industry landscape have created various disruptions, increasing challenges in executing clinical trials.



Source: Top 5 Challenges and Opportunities Drug Developers Face in Executing Clinical Trials by The Editors of the Insights Hub, PPD clinical research business of Thermo Fisher Scientific

DATA COMPLEXITY AND DATA-CENTRICITY

According to a study by Tufts Center for the Study of Drug Development on “Protocol Design and Performance Benchmarks by Phase and by Oncology and Rare Disease Subgroups” published in 2022, the number of data points collected in a Phase III study has reached almost 3.5 million. The table below from the paper highlights the ever-increasing complexity of clinical trials. The table below, an excerpt from the whitepaper highlights the increasing complexity across various scientific design characteristics of clinical studies.

Scientific design characteristics by phase

Means (Coefficients of Variation)	Phase I	Phase II	Phase III
Total Endpoints	15.6 (0.62)	20.7 (0.62)	18.6 (0.54)
Total Eligibility Criteria	31.7 (0.30)	30.9 (0.36)	30.4 (0.37)
Number of Distinct Procedures	29.0 (0.41)	33.5 (0.32)	34.5 (0.40)
Total Procedures Performed	168.8 (0.86)	259.7 (1.16)	266.0 (0.90)
Non-Core Procedures as Percent of Total	19.0% (1.11)	26.1% (0.71)	23.1% (0.81)
Total Protocol Pages	86.7 (0.30)	107.6 (0.37)	115.9 (0.34)
Total Case Report Form Pages	186.4 (0.75)	163.5 (0.80)	177.5 (0.87)
Total Datapoints Collected	330,420 (1.99)	2,091,577 (1.79)	3,453,133 (1.56)
Total Data Sources	3.7 (0.41)	4.2 (0.40)	4.0 (0.39)

While the amount of data collected is increasing, as highlighted in the table above, the source of the data is also shifting. As per a report published by Gartner, 80% of study data was entered by the site in the past while 20% came from other sources. However, in the modern studies this had flipped and 80% of the data is coming from other sources while only 20% is being entered by the sites. This can be attributed to the advent of Decentralized Clinical Trials (DCTs), Real World Data (RWD) and other digital ways of collecting data. DCTs specifically has increased the amount of data collected passively from patients through sensors and devices.

BUSINESS OUTCOMES AND DATA AFFAIR

The above-mentioned challenges and trends have resulted in rapid digitalization of clinical trials and increased the data complexity. This has resulted in complex digital ecosystems. At the same time the outcomes that organizations desire, within life sciences, focus on cycle time reduction, overall business efficiencies and long-term value creation that is innovation led. Also, the advent of Artificial Intelligence, Machine Learning, Generative AI etc. are creating compelling needs to have robust Data Infrastructure and Analytics capabilities to drive the business outcomes in the age of Digital Clinical Trials. Therefore, there is a chasm that needs to be crossed to achieve these business outcomes.

At the same time, the data affair is getting more complex. If you ask the question of “what’s shaping our data ecosystem?”, the answer will be many things. A closer look will reveal the following as major reasons:

1. Breakthrough innovations and rapid technology advancements
2. High volume, variability of EHR/EMR and RWD from multiple sources

3. Pervasive use of AI/ML or the intent and investment to automate and generate insights and
4. Need to provide analytics to patients and sites, beyond the needs of sponsors.

To manage this there are some critical needs that should be fulfilled, as it relates to data:

1. Data Complexity
2. Data Volume
3. Data Quality
4. Data Privacy and Security
5. Interoperability and Standardization
6. Data Governance and Consent
7. Analytical Tools and Expertise

Addressing these challenges requires collaboration between stakeholders, investment in technology infrastructure, adherence to data privacy regulations, and the development of advanced analytics capabilities.

TECHNOLOGY CHALLENGES IN DATA AGGREGATION

When it comes to clinical data aggregation and analytics in the life sciences industry, there are several technology-focused challenges to consider. These include:

1. **Data Integration:** Technology plays a crucial role in integrating diverse clinical data from various sources, such as EHRs, wearable devices, and clinical trials. Implementing interoperable systems and data integration frameworks enable seamless data flow and consolidation.
2. **Data Storage and Management:** The exponential growth of clinical data necessitates advanced storage and management solutions. Cloud-based storage platforms, distributed file systems, and data lakes provide scalable storage and efficient management of large volumes of clinical data.
3. **Data Processing and Analysis:** Technology is essential for processing and analyzing complex clinical data. High-performance computing, distributed processing frameworks (e.g., Hadoop, Spark), and specialized analytics tools facilitate efficient data processing, feature extraction, and statistical analysis.

4. **Artificial Intelligence and Machine Learning:** AI and machine learning techniques are increasingly utilized to extract insights from clinical data. These technologies enable predictive modeling, pattern recognition, and anomaly detection, aiding in diagnosis, treatment planning, and drug discovery.
5. **Natural Language Processing (NLP):** Clinical data often includes unstructured text, such as physician's notes and medical literature. NLP techniques enable the extraction of relevant information from textual data, supporting decision-making and knowledge discovery.
6. **Data Visualization:** Technology-driven data visualization tools enhance the interpretation and communication of clinical data. Interactive dashboards, visual analytics platforms, and augmented reality (AR) interfaces allow researchers and clinicians to explore and present complex data in an accessible manner.
7. **Privacy and Security Solutions:** Advanced technology solutions, such as encryption, access controls, and anonymization techniques, help protect patient privacy and ensure data security. Blockchain technology also shows promise in enabling secure and auditable sharing of clinical data.
8. **Real-Time Analytics:** Rapid analysis of clinical data in real-time is critical for timely decision-making and interventions. Streaming analytics frameworks and real-time data processing enable continuous monitoring, early detection of adverse events, and personalized interventions.

Addressing these technology challenges requires continuous innovation, investment in infrastructure and software solutions, collaboration between technology providers and life sciences organizations, and the cultivation of data science and analytics expertise within the industry.

MODERN DATA ARCHITECTURES

Data Collection, Aggregation and Analysis is a critical need for clinical research. Doing so with quality and compliance to assure unbiased analysis and outcomes from the clinical trials will ensure use of the treatments will result in safety and prove to be efficacious for patients. Data collection through Randomized Clinical Trials has evolved substantially over the past 30 years going from paper-based trials to electronic and now digital trials. This has improved the ability to collect data fast and ensure better quality but also brought in its own complexities. With the advent of Decentralized Clinical Trials (DCTs), the Volume, Velocity, Variety and Veracity of data has increased substantially.

Challenges in handling Clinical Trial Data

Therefore, handling patient and clinical operations data in clinical trials can present several challenges like:

1. Data Security and Privacy
2. Data Collection and Management
3. Data Standardization and Interoperability
4. Data Integration and Analysis
5. Timeliness and Real-time Monitoring
6. Data Quality Control
7. Regulatory Compliance
8. Data Sharing and Collaboration

Addressing these challenges often requires a combination of robust technological infrastructure, well-defined processes, skilled personnel, and adherence to industry best practices and standards.

Use of Data Warehouses in Clinical Data Aggregation & Analytics

Data warehouses play a crucial role in clinical data aggregation and analytics. They serve as centralized repositories that gather and store large volumes of structured and unstructured data from various sources such as electronic health records (EHRs), laboratory systems, imaging systems, and other clinical databases.

In the context of clinical data aggregation, data warehouses integrate data from disparate sources, transforming and standardizing it into a consistent format. This process enables clinicians and researchers to access comprehensive patient data in one location, facilitating a holistic view of a patient's health history, diagnoses, treatments, and outcomes.

For clinical data analytics, data warehouses provide a foundation for conducting advanced analyses and deriving valuable insights. Researchers can leverage data warehousing techniques, including data mining, statistical analysis, and machine learning, to uncover patterns, identify correlations, and generate predictive models. These analytics can be applied to various domains, such as clinical research, population health management, disease surveillance, and quality improvement initiatives.

By employing data warehouses in clinical data aggregation and analytics, healthcare organizations can improve decision-making, enhance patient care, accelerate research efforts, and identify opportunities for process optimization and cost reduction.

Disadvantages of using Data Warehouses

While data warehouse architectures offer several benefits for clinical data aggregation and analytics, there are some potential disadvantages to consider:

1. **Complexity:** Implementing and maintaining a data warehouse architecture can be complex and time-consuming. It requires designing and building the warehouse, integrating data from various sources, ensuring data quality and consistency, and managing ongoing updates and changes.
2. **Cost:** Building and operating a data warehouse can be expensive. It involves investments in hardware, software, data integration tools, and skilled personnel. The costs can escalate further as the volume and complexity of clinical data increase.
3. **Data latency:** Data warehouses typically involve periodic data extraction, transformation, and loading (ETL) processes, which introduce some latency. This means there may be a delay between the availability of new data and its inclusion in the data warehouse. In clinical settings, where real-time or near-real-time insights are crucial, this latency can limit the timeliness of analytics and decision-making.
4. **Scalability:** Scaling a data warehouse to handle increasing data volumes can be challenging. As the amount of clinical data grows, the performance of the warehouse may degrade, requiring additional hardware investments or architectural changes.
5. **Data silos:** While a data warehouse aims to centralize data, it can inadvertently reinforce data silos. Different departments or healthcare providers may maintain their own data repositories, making it difficult to integrate and analyze data across the entire healthcare ecosystem.
6. **Data quality:** Ensuring data quality and consistency is crucial for accurate analytics. However, data from multiple sources may vary in terms of completeness, accuracy, and standardization. Data cleansing and transformation processes are necessary but can be complex and time-consuming.
7. **Changing requirements:** Clinical data requirements evolve over time and adapting a data warehouse to accommodate new data types, formats, or analytic needs may be challenging. Flexibility and agility in modifying the data warehouse architecture to meet changing requirements are essential but may require significant effort.

It's important to weigh these disadvantages against the potential benefits of data warehousing in clinical data aggregation and analytics to determine the most appropriate solution for a specific healthcare organization.

Use of Data Lakes in Clinical Data Aggregation and Analytics

A Data Lake is a centralized repository that stores vast amounts of structured and unstructured data in its raw format. It is designed to accommodate data in its original form, without the need for pre-defined schemas or transformations. The main goal of a Data Lake is to provide a scalable and cost-effective solution for storing and processing large volumes of data.

The architecture of a Data Lake typically involves three main components: data ingestion, data storage, and data processing.

1. **Data Ingestion:** This component involves the process of collecting data from various sources and bringing it into the Data Lake. Data can be ingested from internal systems, external sources, or even real-time streaming platforms. The data ingestion process may involve data extraction, transformation, and loading (ETL) techniques.
2. **Data Storage:** The data in a Data Lake is stored in its native format, such as files, objects, or blobs, organized in a hierarchical structure. It allows for flexible storage and accommodates different types of data, including structured data (e.g., databases, CSV files) and unstructured data (e.g., text documents, images, videos).
3. **Data Processing:** The Data Lake architecture enables data processing capabilities, allowing organizations to perform various operations on the stored data. This can include data exploration, transformation, analytics, and machine learning. Technologies such as Apache Hadoop, Apache Spark, and distributed computing frameworks are commonly used for processing large-scale data within a Data Lake.

In the context of clinical data aggregation and analytics, Data Lakes play a significant role in bringing together diverse healthcare data from various sources. Clinical data, including electronic health records (EHRs), medical imaging, patient-generated data, and research data, can be ingested into the Data Lake. By storing this data in its raw form, it becomes readily available for analysis and enables researchers and analysts to explore and derive insights from a comprehensive and holistic view of patient data.

Data Lakes facilitate advanced analytics techniques, such as data mining, predictive modeling, and machine learning, on the aggregated clinical data. Researchers and healthcare professionals can leverage the Data Lake architecture to gain valuable insights, identify patterns, improve patient outcomes, and support evidence-based decision-making in clinical research and healthcare delivery.

Disadvantages with Data Lake Architectures

While data lake architectures offer several benefits for Clinical Data Aggregation & Analytics, they also have some potential downsides. Here are a few considerations:

1. **Data Quality and Consistency:** Data lakes allow for the storage of various types and formats of data without strict schema enforcement. This flexibility can lead to challenges in ensuring data quality, consistency, and standardization. Without proper governance and data management practices, data lakes can become cluttered with irrelevant, inconsistent, or erroneous data.
2. **Data Governance and Security:** Data lakes typically involve a centralized repository where data from multiple sources are stored. Ensuring proper data governance, access control, and security can be complex. Without appropriate security measures, data lakes can become vulnerable to unauthorized access, data breaches, or privacy concerns, especially when dealing with sensitive clinical data.
3. **Data Discovery and Accessibility:** As data lakes accumulate large volumes of data, finding and accessing the relevant information can be challenging. Without robust metadata management and data cataloging, users may struggle to discover and understand the available data assets, leading to inefficiencies in data retrieval and analysis.
4. **Processing and Analysis Complexity:** Analyzing data in a data lake often requires specialized skills and tools. Data scientists or analysts may need to navigate complex data structures, perform data transformations, and address issues like data cleansing, normalization, and integration. The lack of predefined schemas and data models can increase the complexity of analysis tasks.
5. **Cost and Scalability:** Data lakes can store vast amounts of data, leading to potentially high storage costs. Additionally, as the data volume and complexity grow, the computational requirements for processing and analyzing the data can also increase. Ensuring cost-effective scaling and optimization of resources can be a challenge.

It's worth noting that many of these downsides can be mitigated through effective data governance, metadata management, and proper implementation of security measures. Organizations should carefully plan and design their data lake architectures, considering these challenges and addressing them through best practices and robust data management strategies.

Advantage of Data Lakehouse Architecture

A Data Lakehouse is a unified data management architecture that combines the capabilities of both Data Warehouses and Data Lakes. It aims to address the limitations and challenges associated with traditional Data Warehouses and Data Lakes.

Unlike Data Warehouses, which rely on structured and processed data, and Data Lakes, which store raw and unprocessed data, a Data Lakehouse allows organizations to store both structured and unstructured data in its raw form. It provides a single repository for diverse data types, such as relational, semi-structured, and unstructured data.

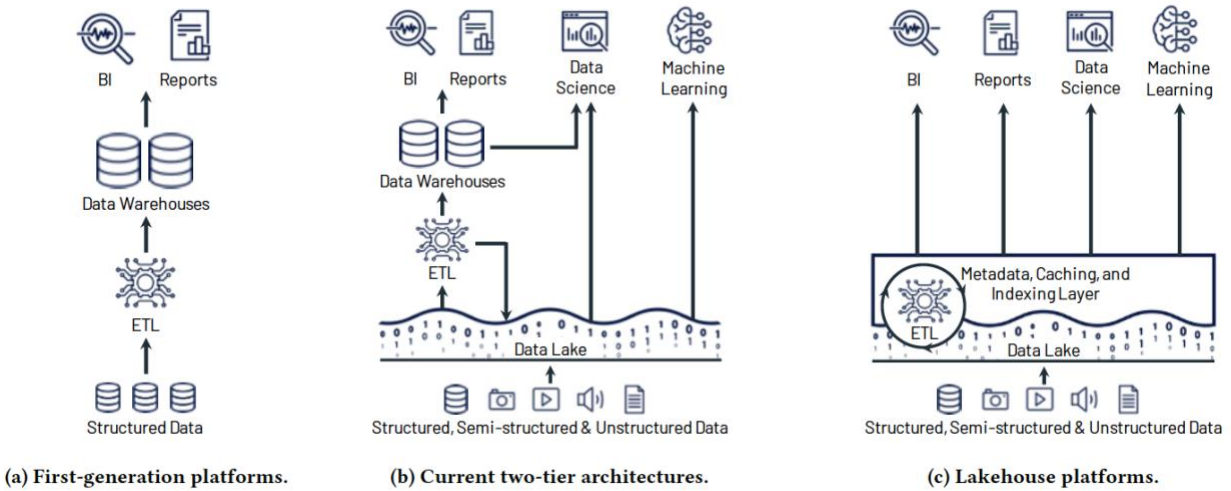
The advantages of a Data Lakehouse include:

1. **Flexibility:** Data Lakehouses offer the flexibility to handle a wide range of data types, including structured, semi-structured, and unstructured data. This flexibility enables organizations to leverage various data sources and formats without upfront data modeling or transformation, reducing data silos and enabling faster data exploration.
2. **Real-time and near real-time analytics:** With a Data Lakehouse, organizations can perform real-time or near real-time analytics on both historical and current data. This capability allows businesses to gain immediate insights, respond quickly to changing conditions, and make data-driven decisions faster.
3. **Cost-effectiveness:** Data Lakehouses leverage scalable and cost-effective cloud storage solutions, enabling organizations to store large volumes of data at a lower cost compared to traditional Data Warehouses. Additionally, the elimination of data duplication and the ability to store raw data without immediate processing help reduce storage costs.
4. **Scalability:** Data Lakehouses leverage cloud-based technologies that provide scalability and elasticity. This allows organizations to easily scale up or down their data storage and processing capabilities based on demand, ensuring efficient handling of large data volumes, and accommodating growth.
5. **Data governance and security:** Data Lakehouses incorporate robust data governance and security features. These features enable organizations to enforce access controls, implement data lineage and auditing, and ensure compliance with regulatory requirements, protecting sensitive data and maintaining data integrity.

By combining the strengths of Data Warehouses and Data Lakes, a Data Lakehouse offers a more comprehensive and flexible data management solution, empowering organizations to derive valuable insights from diverse data sources while addressing the challenges associated with data storage, processing, and analytics.

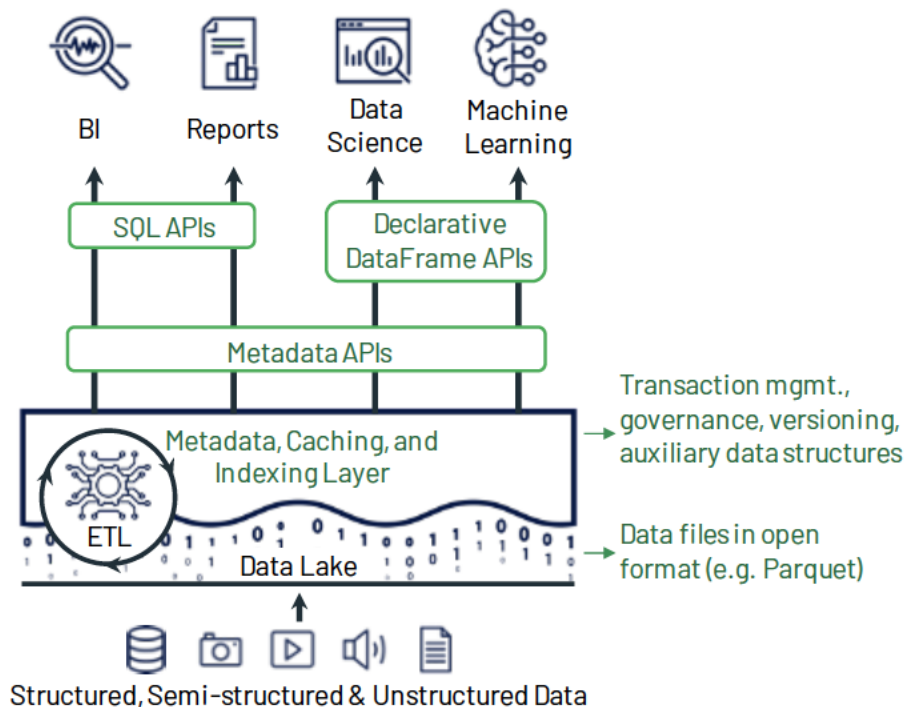
Data Lakehouse Architecture

The following diagram depicts the evolution of data platform architectures to today's two-tier model and the new Lakehouse model.



Sample Lakehouse System Design

The following diagram depicts an example Lakehouse System design, with key components shown in green. The system centers around a metadata layer such as Delta Lake that adds transactions, versioning, and auxiliary data structures over files in an open format and can be queried with diverse APIs and engines.



Data Lakehouse Adoption

Data Lakehouses were gaining popularity across various industries, including the life sciences industry. Data Lakehouses combine the capabilities of data lakes and data warehouses, enabling

organizations to store vast amounts of structured and unstructured data in a single location. This allows for scalable and cost-effective storage, as well as efficient data processing and analysis.

In the life sciences industry, data Lakehouses can be particularly valuable due to the large volumes of data generated from research, clinical trials, genomics, electronic health records, and other sources. By consolidating diverse datasets into a unified data Lakehouse, researchers and scientists can derive insights, discover patterns, and accelerate scientific discoveries.

CONCLUSION

Because a Lakehouse architecture reduces data silos, and increases performance, flexibility, and scalability, it is designed to better support modern clinical trials while ensuring a solid foundation for the future of digitization.

REFERENCES

1. GlobalData's "Clinical Trials: The Movement Towards Decentralized Clinical Trials" research report
2. Top 5 Challenges and Opportunities Drug Developers Face in Executing Clinical Trials: <https://www.ppd.com/blog/challenges-opportunities-in-clinical-trials/>
3. Tufts CSDD White paper on "Protocol Design and Performance Benchmarks by Phase and by Oncology and Rare Disease Subgroups": <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9373886/>
4. [Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics](#) by Michael Armbrust, Ali Ghodsi, Reynold Xin, Matei Zaharia

CONTACT INFORMATION

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

Author Name: Venugopal Mallarapu

Company: Address: eClinical Solutions

Work Phone: +1 919 599 3012

Email: vmallarapu@eclinicalsolutions.com

Website: <https://www.eclinicalsolutions.com/>

Brand and product names are trademarks of their respective companies.