CASE STUDY

How MSK Accelerates Cancer Research with Dremio's Open Data Lakehouse

At a Glance

The Customer

Memorial Sloan Kettering Cancer Center

Challenge

Memorial Sloan Kettering's Cancer Data Science Initiative (CDSI) engineering team faces challenges with getting cancer research data to end users in a timely manner.

Solution

The team chose Dremio's data lakehouse platform to simplify data access and unify cancer research data for end users.

Results

- While solving lower-level data management problems, the team laid the foundations for a data mesh architecture at MSK, as data has become discoverable, available, and more trustworthy
- Showing the potential to build trust between data producers and data consumers
- Data became available within hours to days as opposed to weeks to months
- Virtually eliminated ETL processes, no more data copies

The Business:

Memorial Sloan Kettering Cancer Center (MSK) is the largest private cancer center in the world and has devoted more than 135 years to exceptional patient care, innovative research, and outstanding educational programs. Today, MSK is one of 52 National Cancer Institutes designated as Comprehensive Cancer Centers, with state-of-the-art science flourishing side by side with clinical studies and treatment.

CDSI is a strategic initiative to accelerate cancer research and discovery through advanced analytics. Their primary responsibility is to provide vital research data to data consumers, empowering them to deepen their understanding of cancer and patient outcomes.

The Challenge:

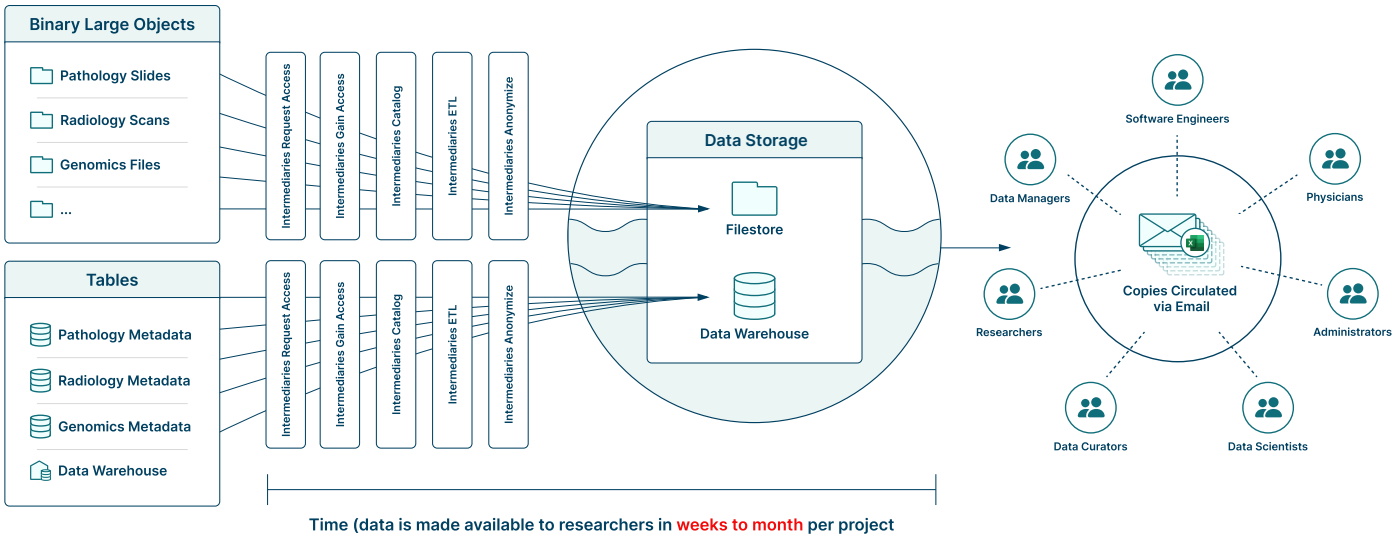
The CDSI engineering team set out a goal to build a scientific data management and compute infrastructure for accelerating cancer research.

The problem was, MSK's existing data infrastructure was not able to deliver research data fast enough to the data consumers - diverse teams made up of technical and non-technical users ranging from engineers to scientists, pathologists, radiologists, physicians, and administrators.

Datasets needed for research at MSK were multi-modal and highly dimensional in nature and dispersed across the enterprise. Large binary objects were copied into file storage, while operational tabular data from relational databases were copied into a proprietary data warehouse.

When data was required, the CDSI engineering team needed to act as intermediaries between data consumers and data producers, or between other intermediary teams who provided access to secondary stores such as the enterprise data warehouse. The engineering team was overwhelmed with individual data requests, and custom ETL pipelines were created in silos for different stakeholders. Depending on the project, it sometimes took weeks to months before the data became available. Once data was made available, it was not uncommon for data consumers to create copies of the data, often circulating via email, making it very hard to track the different versions of data that came into existence for each research project. The existing data infrastructure was inefficient for data access and needed a more mature governance model.

Before Dremio



The Solution:

The team went beyond data engineering and functioned as a team of research software engineers that blended infrastructure development, data products development, and scientific analysis. This combination of responsibilities made them uniquely well-positioned to match the needs of the cancer research community at MSK to the right set of technology solutions for making structured and unstructured data available for research in a timely manner.

The team evaluated several commercial and open-source solutions including Hive, Databricks SQL, Starburst, Denodo, and Dremio for querying structured data and Hadoop (HDFS), Open Stack's Swift, and MinIO for storing unstructured data. They had several requirements:

- On-premises deployment since data and compute infrastructure was housed on-premise
- Interfaces that would satisfy the needs of a diverse set of data consumers given the varied composition of each team (technical and non-technical)
- Eliminate the time spent creating and maintaining siloed ETL pipelines and thereby provide faster access to data
- A means to version data given the iterative and evolutionary nature of research
- Documentation support in the form of data sheets for datasets
- A no-copy architecture to avoid the problem of tracking multiple user copies
- A simple, yet mature governance model

At the start of the journey, the CDSI engineering team desired a data architecture that would eliminate unnecessary ETL, democratize research data, and simplify data governance. They later realized these needs aligned with the core principles of data mesh, which aims to simplify ETL by cutting out intermediaries between data producers and data consumers altogether, enabling data producers to become data product owners, making data widely accessible through a self-service platform, and simplifying governance.

Ultimately, the team chose MinIO as their primary object store and Dremio as the query engine. Although they considered several options, Dremio and MinIO satisfied their one-hour rule for evaluating new technology. Both had an easy barrier to entry and supported their on-prem deployment, with a clear path to the cloud.

One compelling reason MSK chose Dremio was because of Dremio's data reflection capabilities. Many of the data sources for MSK were located across operationally tuned

relational databases and were not built for complex analytical queries hitting the systems. With data reflections, end-users get accelerated analytical queries without impacting the hospital's operational systems. All of this was possible while effectively maintaining a single copy of the source data with Dremio's semantic layer.

Another reason was the modern, intuitive low-code/no-code interface from Dremio's semantic layer to inspect, curate and integrate multi-modal data. Non-technical users like physicians and administrative staff at MSK were used to Excel and Tableau-like views to inspect data, and Dremio provided a user-friendly way to perform basic data inspection operations without having to write a line of code. Engineers found it easier to transform and integrate data with SQL when compared to Pandas, and this offers a low-code solution to data curation tasks. Researchers were able to continue building ML models in Jupyter notebooks over data connected to Dremio. Powered by Apache Arrow, data consumers retrieved and processed data much faster than JDBC/ODBC.

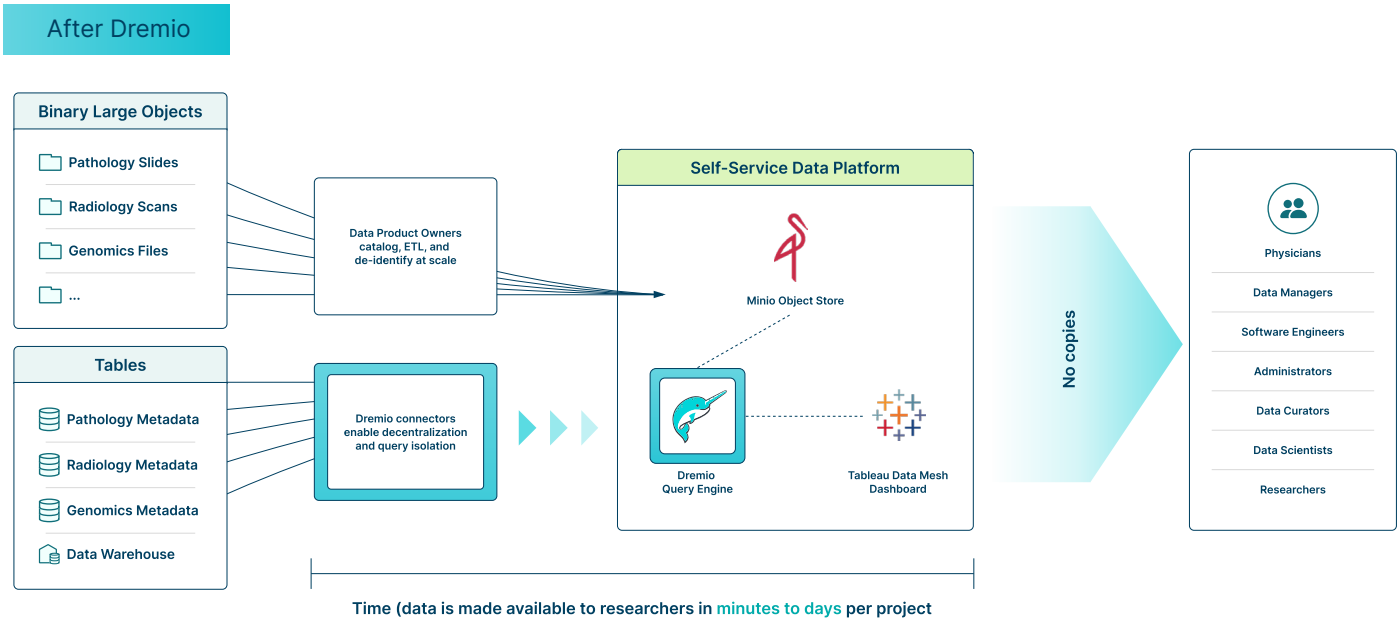
Data discoverability and trustworthiness are critical components of MSK's research. Dremio's adoption of Apache Iceberg and Nessie is allowing engineers to explore versioning of datasets for iterative tasks that exist in evolving research projects. The catalog wiki and tagging features were well-suited for taking shareable notes on dataset evolution. The data lineage capabilities helped researchers understand data even better. The evolving wiki notes help in the development of datasheets for datasets, a powerful yet largely unmet practice in the research community.

Results:

By going with Dremio, MSK was able to create a more efficient data platform for accelerating cancer research. Depending on the project, data became available within minutes to days from the ability to connect data consumers on the research side of the hospital directly to data sources on the operational side of the hospital with query isolation and user-friendly self-service interfaces.

The speed and agility with which engineers and researchers could use the data also skyrocketed. Dremio has allowed the CDSI engineering team to virtually eliminate siloed ETL pipelines and user data copies. Data products for research consumption can be quickly assembled in the semantic layer from data that is directly sourced from data producers on the operational side of the hospital.

Dremio plays a key role in MSK's data mesh journey, which centers on governed self-service and trustworthy data. The built-in semantic layer offers data catalog, lineage, wiki, and tagging capabilities to foster collaboration and data sharing at MSK. Both data producers and consumers are able to create data products from domains, similar to



views, and share governed data across teams within the organizations. By directly connecting data producers and consumers, Dremio makes data easy to discover, understand, and trust.

The CDSI engineering team's desire for a data mesh architecture started with 5 users on Dremio's Enterprise Edition. It soon onboarded 6 research teams. In just over one year, the platform served 150 users with diverse backgrounds from MSK's research community.

With the fundamental data mesh principles implemented with Dremio and MinIO, the CDSI engineers feel they now have the solution to take on the next big challenge of creating a true data mesh - the cultural shift across teams in the organization needed to enable data producers to become first-class data product owners.

Conclusion:

Memorial Sloan Kettering's CDSI engineering team faced challenges managing data copies while providing faster access to data to researchers. By building a scientific data management and compute infrastructure around Dremio's data lakehouse, the CDSI engineering team reduced tasks and processes that took weeks and months to perform down to hours or days. The CDSI engineering team was able to achieve its goals by using Dremio to eliminate data copies and serve as the unified access layer for research tabular data. Dremio was able to set the stage for the CDSI engineering team's vision toward a full-blown data mesh architecture at MSK.

ABOUT DREMIO

Dremio is the easy and open data lakehouse, providing self-service analytics with data warehouse functionality and data lake flexibility across all of your data. Use Dremio's lightning-fast SQL query service and any other processing engine on the same data. Dremio increases agility with a revolutionary data-as-code approach that enables Git-like data experimentation, version control, and governance. In addition, Dremio eliminates data silos by enabling queries across data lakes, databases, and data warehouses, and by simplifying ingestion into the lakehouse. Dremio's fully managed service helps organizations get started with analytics in minutes, and automatically optimizes data for every workload. As the original creator of Apache Arrow and committed to Arrow and Iceberg's community-driven standards, Dremio is on a mission to reinvent SQL for data lakes and meet customers where they are on their lakehouse journey.

Hundreds of global enterprises like JPMorgan Chase, Microsoft, Regeneron, and Allianz Global Investors use Dremio to deliver self-service analytics on the data lakehouse. Founded in 2015, Dremio is headquartered in Santa Clara. CNBC recognized Dremio as a Top Startup for the Enterprise and Deloitte named Dremio to its 2022 Technology Fast 500. To learn more, follow the company on GitHub, LinkedIn, Twitter, and Facebook, or visit www.dremio.com.

Dremio and the Narwhal logo are registered trademarks or trademarks of Dremio, Inc. in the United States and other countries. Other brand names mentioned herein are for identification purposes only and may be trademarks of their respective holder(s). © 2023 Dremio, Inc. All rights reserved.