IDC PERSPECTIVE

Yale Unlocks the Transformative Impact of Clinical Data with Disaggregated Architecture and AI

Raghunandhan Kuppuswamy

EXECUTIVE SNAPSHOT

FIGURE 1

Executive Snapshot: Disaggregated Architecture and AI from NetApp and NVIDIA Helped Yale with Clinical Data

This IDC Perspective discusses Yale's technological architecture and the problems that were initially encountered and how NetApp developed a technological solution to the problem. The document also includes a brief discussion of the solution's architecture and how AI assisted in resolving specific COVID-19-related issues.

Key Takeaways

- NetApp designed the next-generation architecture (disaggregated architecture) to advance Yale's computational health platform into an agile system more suited to integrate AI and ML in all operations.
- The disaggregated architecture helps use memory and storage efficiently and reduce licensing costs to application stacks that support the costs. Base Command platform is a product offered by NVIDIA. The program supports AI development processes on premises or in the cloud.
- NetApp wanted to start small and scale when things went well, so the company began copying data to the new platform and gave users access. When everything went as planned, the company began scaling the process.

Recommended Actions

- **AI resources:** To realize the full potential of AI, one must have the means to manage it. With the correct data and resources, AI's maximum potential may be realized. If you are unable to maintain an in-house team, search for platforms that provide AI solutions. This will be the ideal starting point.
- **AI audits:** If both technology and resources operate as planned, they would do wonders for the organization. To determine if they are functioning as planned, however, you must do process checks at the key places and conduct audits on a regular basis. This will assist the business in making the most efficient use of its technology and resources, hence enhancing its ability to produce results.
- **Security:** Data security is the most critical factor. Organizations should make data security their top priority and ensure that patient data is maintained and protected with the utmost care.

Source: IDC, 2022

August 2022, IDC #US49496322

SITUATION OVERVIEW

Artificial intelligence (AI) adoption has passed the tipping point, with CY20 driving an acceleration of AI adoption as organizations had to confront everything from real-time liquidity forecasting to processing massive amounts of structured and unstructured data associated with COVID-19 pandemic protocols.

AI in Healthcare

Virtually every industry has adopted AI. The number of applications of AI in the healthcare sector has exploded in recent times. AI remains prevalent in radiology, especially for diseases such as tuberculosis (TB), but it is increasingly being used for other purposes, including as part of preventive health checks, the discovery of new drugs based on a patient's symptoms and issue patterns, immunotherapy for cancer patients, and determining the most effective treatment for a dialysis patient.

Machine learning (ML) is gaining traction in the healthcare sector. As a result of COVID-19, an increasing number of healthcare organizations have begun to implement ML in their IT stacks in order to comprehend patients' symptoms and COVID-19 patterns. Even small healthcare institutions are experimenting with machine learning for insurance and billing to learn from it and make recommendations for the future.

In the healthcare industry, wearables like Fitbit and smartwatches are gaining market share. It can analyze data to provide users and their healthcare professionals with information regarding potential health risks and issues.

Yale School of Medicine and Its Problems

Yale School of Medicine (Yale) was one of the few medical institutions to see AI's potential early and embrace it. It has around eight hospitals in its clinical system. Every hospital has its own labs and pharmacies. The data generated by the hospitals that includes electronic health record (EHR), clinical imaging, genetics and molecular data, high resolution and high frequency physiologic monitoring data from ICUs and emergency departments, labs, and pharmacies are huge and complicated. COVID-19 furthermore enhanced the complications with its variants and symptoms. What started out as the Alpha variant was rapidly mutating. As the cases increased, the data generated went high.

The greatest issue that needed to be resolved was centralizing and integrating data from all sources so that it can be utilized on any platform with ease. Data standardization was a further challenge.

Technology Setup Pre-NetApp Era

Yale had mostly standard bare metal servers that ran Hortonworks Data Platform. Hadoop was the preferred choice for data consumption, processing, and analysis for several years. Hadoop had problems of its own. One of the biggest problems was that licensing costs were going up, and some of the licensed items didn't have much value because the exact same thing was free on the Apache open source community. So, for example, Spark is now used in almost every process, and there isn't much added value in buying a license from a different provider to get something that is freely available. Also, it was hard to add storage for cold storage and even to move files.

At Yale, most of the architecture runs on premises, mostly because of budget constraints. So what matters is the cost of running the business versus the cost of buying assets. The datacenters are all on premises. There were also direct connections to the public Azure cloud, but they were on a private subnet and linked by fiber. But this was limited.

As the servers and other hardware aged and began to fail, it became difficult to keep them operational.

Yale wanted to move out of that environment and into something a little bit more modern, thus it chose to develop its in-house compute platform.

NetApp Was the First Choice

Yale wanted to develop its existing computational resources into something that is more efficient, scalable, and forward looking in terms of cloud ready, which will in turn help drive the analytical insights within medicine.

Also, NetApp had an excellent sales and technical support team (see Figure 2). It had the technology (ease of moving data between systems, storage arrays with an NVMe option, and high-speed disc interconnect), support, and scalability that Yale wanted. Yale was looking for a team of experts who could help it through this technology upgrade and in the future if it needed technical or nontechnical help, so Yale chose NetApp.

FIGURE 2

NetApp's Approach

Source: NetApp Inc., 2022

NetApp looked at the problems at Yale and studied how Yale's technology was set up. To solve the problem, the company came up with some goals:

- **Ingest of data at scale:** Transitioning data from a decentralized to a centralized platform (unified data lake), how can one consume data at scale, apply data governance to it, and utilize it to convert, and analyze data to maximize its utility?
- **Building analytical muscle:** Ensure that we have the appropriate tools and skill set. Since the most modern tools and skill sets from five years ago are no longer applicable.
- **Frictionless accessibility:** Ability to access data where and when it is needed, and not always on the same platform (i.e., data lubrication).
- **Governance and security:** As it contains patients' sensitive information, data security should be of the highest order.

Disaggregated Architecture Design

Yale collaborated with NetApp to design the next-gen architecture, and to advance its computational health platform transform into agile next-generation system that is better suited to embedding AI and ML in all its workflows so that it is much more modern and next-generation technology.

NetApp chose disaggregated architecture for the existing computational health platform and named it as 2.0. Disaggregate as the name implies, breaks the traditional computing architecture into smaller standalone systems. Memory and storage will be separate for each system, and all these systems will be networked. With this design we can utilize memory and storage in a far more efficient manner. Also, under a disaggregated architecture, you can be quite prescriptive regarding how you pay for the license, and you are not required to pay for a licensed product if it is open source. The workloads are reduced because we have a unified data lake, and it's easy to adopt next-generation technology, whether it's a computational technology or faster processors or GPU technologies. If you want to collaborate in bursts on the cloud, it's much easier when you have data governance built around your offering, and you can prescriptively use whichever cloud you want.

The goal is to eventually leverage DGX Foundry, a world-class infrastructure solution fully managed by NVIDIA and available with monthly subscriptions. It includes NVIDIA Base Command Platform software and uses NetApp storage (Keystone Flex Subscription), so data scientists who need a premium AI development experience get immediate access without their IT organization having to build a complex AI infrastructure themselves. The program supports hybrid, or disaggregated, architectures and enables users to use the most recent DGX systems and NetApp storage solutions without incurring a capital expense. Base Command Platform software fully supports NVIDIA Clara, a healthcare application framework for AI-powered imaging, genomics, and the development and deployment of smart sensors for healthcare environments.

At Yale, the vision with DGX Foundry and NetApp is to make their platform even more powerful without taking up more datacenter space. The data can be put where it needs to be. With Base Command Platform, users can access and manage their entire environment, including APIs, machine learning operations; manage users, teams, and data; and get access to TensorBoard, TensorFlow, and Jupyter environments. In other words, they get access to the same tools that they already use in-house. In addition, they get the power of NetApp's AFF storage systems, and all functionality that comes with the Keystone Flex Subscriptions, fully managed and supported by NVIDIA and NetApp. Because it is fully

managed, Yale can focus on its research instead of figuring out how to connect everything and make it work and can move data in and out of this service at no additional cost.

Once the architecture was in place, NetApp wanted to start small and scale when things went well, so the company began copying the data to the new platform and allowing the users (physicians/researchers) to access it, and when everything went as planned, it began scaling the process. Workloads were set up to run on both CPU and GPU, and whether a task uses CPU or GPU usually depends on the workload.

Most projects in NLP work well on GPU. Also, parts of a predictive model are trained on the GPU and then moved to the CPU. Most inference is done by CPUs, but some image-based inference is done by GPUs.

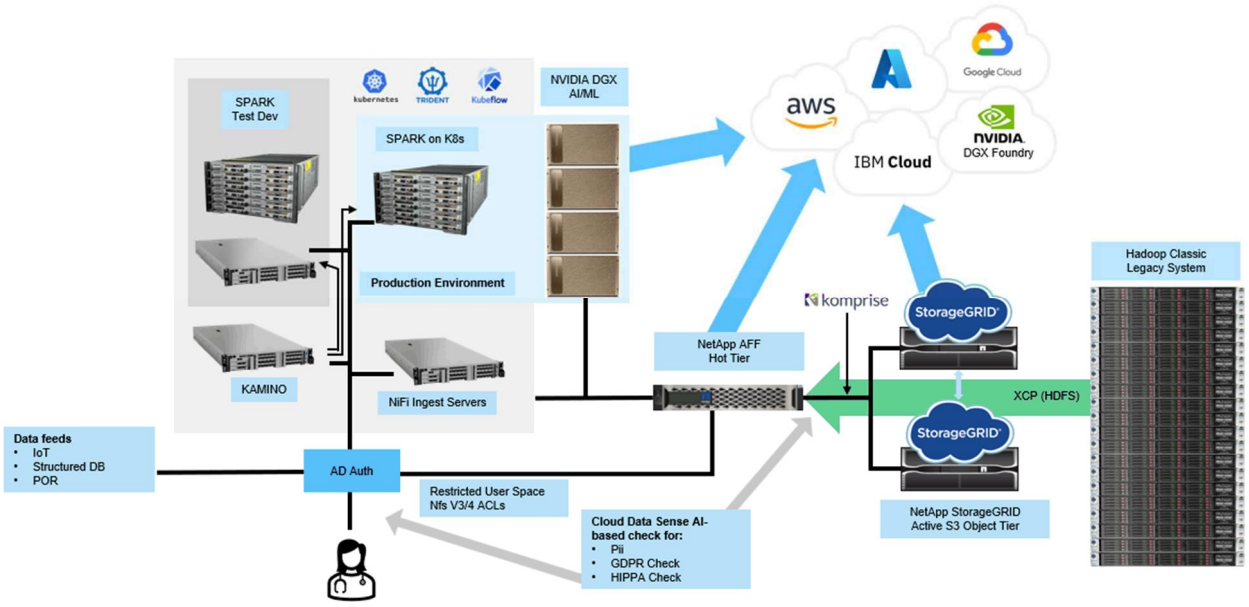
NetApp provides protection against ransomware. It also offers Snapshot (a NetApp ONTAP software feature that makes incremental data-in-place, point-in-time copies of a LUN or a volume with minimal performance impact to revert to known good points in time and FabricPool (NetApp ONTAP software feature) automates data tiering from flash to object storage in the cloud or on premises.

And finally, Kubernetes is used by a lot of platforms and is becoming the standard for how they are set up.

Figure 3 shows NetApp's solution from a technological perspective.

FIGURE 3

NetApp's Solution from a Technological Perspective



Source: NetApp Inc., 2022

Here is a list of the infrastructure hardware and how it is used in this solution:

- **Kubernetes and Spark:** This next-gen disaggregated architecture platform 2.0 is built on Kubernetes and Spark. Those are the core tenants of this. Production environment, test dev, and research environment are housed in the same place.
- **Active directory/ACLs:** Role-based access control and protections of who can do what within the data are handled here.
- **NetApp AFF Hot Tier:** NetApp All Flash FAS (AFF) is a robust scale-out platform built for virtualized environments, combining low-latency performance with best-in-class data management, built-in efficiencies, integrated data protection, multiprotocol support, and nondisruptive operations. You don't have to store data in separate silos. Data can be moved from AFF into different tiers with high speed (about 120–150TB of very high speed).
- **NetApp StorageGRID Active S3:** Data can be migrated to StorageGRIDS's S3 layer and can be accessed without refreshing.
- **NVIDIA DGX AI/ML:** Yale was an early adopter of NVIDIA's DGX technology; however, not all operations require DGXs. This environment is diverse and can manage all data using NiFi and other technologies.
- **Hadoop classic legacy system:** Older Hadoop systems with lots of RAM and CPU still function. With virtualization and the ability to serve data from network appliances, one can shift it to the newer technology and platforms without any issues. Legacy hardware can be utilized until its end of life. Yale can integrate all of this with Kubernetes and assign workloads to specific devices and determine the optimal device for each job.

How Did Yale Benefit from This Solution?

At Yale, technologies have changed from databases and data warehouses to data mart to the data lake (unified data lake). Yale wanted to make sure that as it grew, it could stay flexible and quick to change, so that when the next big thing comes along, like whole exome and whole genome sequencing, which it didn't do 5 to 10 years ago, or COVID-19 testing, which it didn't do 3 years ago, it has the infrastructure to change as quickly as it can.

Data is the core of everything. Yale was able to collect data from various processes such as clinical, operational, patient-centered clinical trials, scientific research, and COVID-19 testing using the platform and perform data discovery, data integration, and data enrichment in real time. Also, the platform effortlessly streamed unstructured data like clinical notes and imaging alongside structured data like vital signs and prescriptions. Several NLP processes on the platform, extract concepts from clinical notes in real time and insert them into the data model. These real-time data enrichment systems aided in the early detection of diseases.

Within COVID-19 area, Yale was able to combine the data that was coming from EHR and other clinical systems with the research team's data (which focus on immunobiology). The data was then standardized and with this data Yale built an application called KAMINO. This combined data helped doctors to research more on the COVID-19 patients, their symptoms and early treatment patterns. This research data helped Yale to decide on its approach toward COVID-19. With the help of its computational platform, Yale was quite successful in handling the COVID-19 pandemic.

Yale is using this platform for a few different things, like building new language models of clinical notes to find embeddings and computer phenotypes, like long COVID-19, and doing image-based analysis,

like looking at EKGs and trying to use convolutional neural networks to predict specific outcomes or looking at echocardiograms and comparing that with genomic data.

Also, from an AI point of view, Yale has researchers and other contributors who work on machine learning libraries. Some people use the libraries like tools. Yale also has data robots and other tools for those who want to do less direct code manipulation and more AI exploration instead of development. Yale made sure that all these tools and capabilities were built into the platform. Also, there are application layers, code repositories, and continuous integration/continuous delivery (CI/CD) pipelines to manage and integrate with all the other systems that will support these clinical AI initiatives.

One of the most important takeaways from this is that the computational health platform is an integral part of the Yale School of Medicine, and it's a sign to the world that, given the way that medicine is evolving today, there is no longer a world in which technology does not play a role in guiding physicians' decisions. They are all decisions based on clinical data and AI that do not rely on a single person to figure out, but instead provide augmented assistance in several ways. These platforms of the future generation, aided by AI, can be used to better identify the appropriate medicines for COVID-19, what works and what doesn't, as well as the likelihood of bad results based on the patient's morbidities, comorbidities, and ethnicity.

And then finally, integrating with artificial intelligence. It's becoming more and more important, and one of the most important things about it is that you don't get locked in again if you can use a platform as a service. So, the whole point of the idea of disaggregated architecture was to be able to switch to the best technology at any time. And really, with NVIDIA DGX Foundry with NetApp, it fits right into that overall theme since you're not locked into an on-premises solution..

The work that is done at Yale will have an impact on many different fields, not just medicine but also the design of autonomous vehicles and modern factories.

ADVICE FOR THE TECHNOLOGY BUYER

The faster an organization adopts AI, the more it can use it to improve its processes and get more benefits from it.

IDC recommends that healthcare businesses do the following to maximize the AI's potential:

- **Understand the data:** To get the most out of the data, you must first comprehend the data. The better you comprehend your data, the more chances it provides. Therefore, the first step is to comprehend the data: what data it is, where it came from, and what we can do with it. Once we have this information, we may consider the appropriate technology to handle the issue.
- **Know the capabilities of your existing IT stack:** First, we must determine our IT infrastructure and where I may begin implementing AI inside the firm. To gain a better understanding of this, we can evaluate the use cases of firms that are already active in this space, which will serve as an excellent starting point.
- **Note that AI requires resources:** To realize the full potential of the technology, one must have the means to manage it. With the correct data and resources, AI's maximum potential may be realized. If you are unable to maintain an in-house team, you can search for platforms that provide AI solutions. This will be the ideal starting point.
- **Perform AI audits:** If both technology and resources operate as planned, they would do wonders for organization. To determine if they are functioning as planned, however, you must

do process checks at the key places and conduct audits on a regular basis. This will assist the business in making the most efficient use of its technology and resources, hence enhancing its ability to produce results.

- **Emphasize security:** Data security is the most critical factor. Organizations should make data security their top priority and ensure that patient data is maintained and protected with the utmost care.

LEARN MORE

Related Research

- *Manage AI/ML Business Risks and Thrive with Responsible AI* (IDC #US48235521, September 2021)

Synopsis

This IDC Perspective discusses the technological infrastructure and the problems that were initially encountered, and how NetApp developed a technological solution to the problem. The document also includes a brief discussion of the solution's hardware technologies and how AI assisted in resolving specific COVID-19-related issues.

"Yale and NetApp have successfully identified the problem's root cause and developed technology solutions. Yale handled data-related issues extraordinarily well, particularly in unforeseeable situations such as the COVID-19 pandemic," said Raghunandhan Kuppuswamy, research manager, AI and Automation Software research at IDC.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com