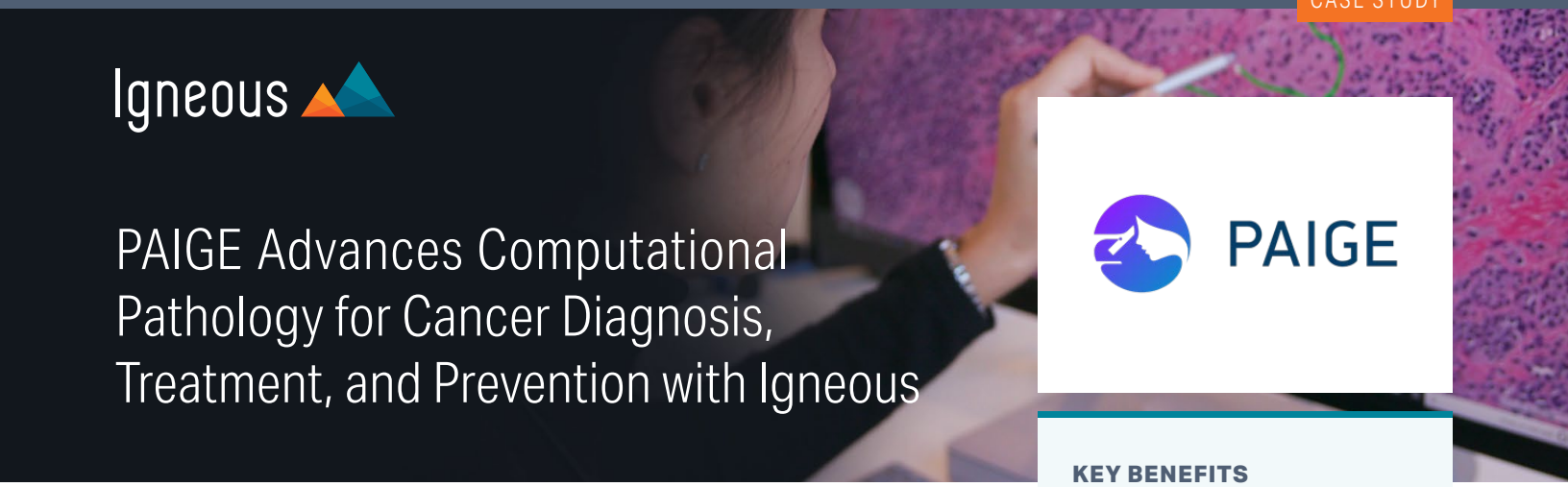
CASE STUDY

Igneous



PAIGE Advances Computational Pathology for Cancer Diagnosis, Treatment, and Prevention with Igneous

PAIGE

PAIGE's mission to revolutionize the diagnosis and treatment of cancer through machine learning requires an extremely large dataset of high resolution slide images. The team needed to not only protect all of this unstructured data, but also programmatically move and process small subsets of the overall dataset on demand for high performance computations. PAIGE chose Igneous to act as their unstructured data management system of record for its intelligent indexing, scalability, and as-a-service model, as well as its data movement capabilities. Igneous is integrated tightly within PAIGE's machine-learning-based data workflow with Pure FlashBlade and NVIDIA for compute and image processing.

Machine Learning Requires Modern Data Management

Research organization PAIGE aims to revolutionize the diagnosis and treatment of cancer through cutting-edge artificial intelligence. Using machine learning algorithms, PAIGE helps pathologists be more efficient, researchers be more quantitative, and families be more confident in diagnosis at a lower cost. PAIGE's team of 20 engineers and researchers is based in New York City at the Cornell Tech campus on Roosevelt Island.

"What PAIGE seeks to do is to harness the power of artificial intelligence and large datasets to help pathologists diagnose diseases more efficiently, more accurately, and more reproducibly. In addition, it's not only about the diagnosis or helping the pathologist; at PAIGE we're really trying to revolutionize the practice of pathology to help us better prognosticate and treat the cancer to directly help the patient," said Patricia Raciti, a pathologist at PAIGE.

At the heart of PAIGE's ML algorithms are enormous, petabyte-scale training datasets of high resolution tissue scan images. Training the algorithm with a wealth of data, and the right kind of data, is key to developing an accurate and efficient ML model for accelerating the diagnosis of cancer; the algorithm is only as good as the data it's fed. While PAIGE's petabytes of tumor scan data make its mission possible, they also present significant challenges.

"One of the major problems that we're tackling is, you have these huge tissues slides assembled by pathologists. We scan them at a very high resolution, so a single slide has billions of pixels. If you have just 40 of these slides, that's already larger than most datasets in the world. The difficult part about dealing with this is, first of all, how do you store so much data?" said Ran Godrich, a research engineer at PAIGE.

KEY BENEFITS

- Enables data owners and AI engineers to quickly organize, find, and move specific subsets of the overall dataset, accelerating research workflows
- Provides scalable data protection for petabytes of unstructured data onsite and offsite
- API-based integration with Pure FlashBlade for high performance compute and NVIDIA for image processing
- Delivers data analysis collaboration between pathologists and engineers to access and utilize datasets directly
- Igneous as-a-Service delivery reduces organizational costs, allowing the organization to focus resources on computational pathology, not systems management

SOLUTION OVERVIEW

- Igneous DataProtect and DataFlow to manage 8PB+ of unstructured data
- Integration with Pure Storage FlashBlade and NVIDIA for image processing
- Dedicated Igneous Customer Success team

Beyond Storage and Protection to Dataset Organization, Movement, and Privacy

However, PAIGE's data management challenges extend beyond the daunting problem of storing petabytes of unstructured data. Storing and protecting the data is only one piece of the puzzle when it comes to ML workflows. For the data to actually be useful in training the algorithm, researchers need to be able to effectively find and move small active subsets of the overall dataset for compute and processing.

"The problem is, you store so much that you have petabytes or hundreds of thousands of these slides, but you only want to use several hundred of them at a time or several thousands of them. Being able to select only a handful from all of these images is very difficult, especially because a lot of the data points that we receive are loosely connected," said Godrich.

To complicate PAIGE's challenge, the field of health and medicine adds further security requirements that must be met. The team at PAIGE had to not only factor in their internal needs, but also the constraints of the medical field when searching for the right data management solution.

"The medical domain poses a lot of additional challenges for startups in that space. Part of that is privacy or security issues of your data, especially if you work at a very large scale where you get data from a large set of hospitals from all over the world to train your machines. It's important that the underlying structure is not just plain storage, but that it also allows you to manage and organize your data while keeping guard over it," said Thomas Fuchs, Chief Scientific Officer at PAIGE.

PAIGE's forward-looking research required a forward-looking approach to data management that could protect petabytes of data, as well as make it easy for researchers to de-identify, classify, find, and move subsets of that data.

Scalable Data Protection, Easy Data Discovery, and Seamless Data Movement

In the end, the decision came down to the capabilities that would help the team accelerate their research and deliver better results. The team needed flexibility to try new, innovative software; the ability to organize, discover, and move data; and reliable, fully managed data protection that wouldn't drain the resources of the small team.

In addition to Igneous' scalable data protection layer and data movement engine, Igneous appealed to PAIGE because it was cloud- and storage vendor-agnostic. The ability to use any cloud and integrate with any existing or potential primary storage systems grants PAIGE the flexibility to experiment with fledging technologies and move quickly as a startup—all while Igneous' as-a-Service delivery means that the team can focus on the research, rather than systems management.

> *"What Igneous allows us to do is choose whatever software you want, whatever cloud provider you want, and whatever machine learning models you want to build.*
>
> **Ran Godrich**
> Research Engineer at PAIGE

"One of the main reasons that we chose to go with Igneous is because they are cloud agnostic. As a young company, we always want to use the latest offer packages that are coming out every week, every month, in the world of artificial intelligence and machine learning. We don't want to limit ourselves to one," said Godrich. "And what Igneous allows us to do is choose whatever software you want, whatever cloud provider you want, and whatever machine learning models you want to build. And that's really what we're looking for."

PAIGE is protecting and managing 8PB of medical data on Igneous, which acts as the central repository for all of their data and enables seamless data movement to cloud. When small subsets of the data need to be used, Igneous pushes the data to Pure Storage FlashBlade, a high performance storage platform which acts as a "hot edge" for the data to be processed by NVIDIA's image processing software. Igneous also archives the computational results, enabling the "hot edge" to be cleared for subsequent workloads.

Igneous' easy data discovery process powers this workflow. Instead of manually sifting through petabytes of data to find the right set, the research team saves time by using Igneous to query the specific dataset they want and move it to where it needs to go.

"Let's say that we're currently working on some prostate cancer data, where we only want to pull down the prostate cancer slides. What Igneous allows us to do is query only the database that we want, such as prostate cancer slides from 2017. Igneous will package the data for us and it will be an all-in-one data pipeline from Igneous to Pure FlashBlade. This flow that Igneous manages for us is very effective because we only want to work with this small dataset at a time. And that's something that is agnostic to any cloud provider that we want to use," said Godrich.

Fully Managed, as-a-Service Delivery Lets Team Focus on Research, Not IT

The small team of 20 making up PAIGE today does not include any dedicated IT personnel. With any traditional data protection and management solution, this would pose a problem—but because Igneous is delivered as-a-Service and fully managed, PAIGE did not need to hire any IT employees to manage Igneous. PAIGE's research engineers, scientists, and pathologists are able to directly use Igneous to see, protect, organize, and move the data they need without going through an IT team.

> *The openness to solving problems at hand together with the customer puts Igneous apart from most of the classical storage companies.*
>
> **Thomas Fuchs**
> Chief Scientific Officer at PAIGE

"Before I started here, I didn't know anything about storage. I never worked with data in this size and I never actually worked on building these huge architectures. But the people from Igneous were really helpful and patient. They really care about what we're doing here. For me [in my role as a research engineer], they really made the process super easy. I was really nervous coming in, but it was actually very simple and to this day I still only need to run certain processes," said Godrich.

The PAIGE team looks forward to iterating and improving their algorithms, with the support that Igneous' data management platform and integrated workflow with Pure Storage and NVIDIA provide. One of the advantages of working with Igneous that the team foresees is being able to work together to solve problems, due to Igneous' agility as a startup and Igneous' investment in the AI/ML space.

> *That deep involvement of Igneous engineers and experts in actually developing our data infrastructure—not only storage but also the management systems—was absolutely key to building a high performance compute infrastructure.*
>
> **Thomas Fuchs**
> Chief Scientific Officer at PAIGE

"Igneous is unique because, like PAIGE, it's a young and fresh company that's very agile. And because we're in a space which constantly changes, agility is enormously important. The openness to solving problems at hand together with the customer puts Igneous apart from most of the classical storage companies," said Dr. Fuchs.

"One of the key advantages Igneous has over competitors, and one of the reasons we chose Igneous, is that we really felt Igneous is interested in the domain itself. They're interested in helping us build PAIGE into a company that will actually help doctors and patients. That deep involvement of Igneous engineers and experts in actually developing our data infrastructure—not only storage but also the management systems—was absolutely key to building a high performance compute infrastructure."

Contact Igneous

1-844-IGNEOUS / 206-504-3685 / info@igneous.io

2401 Fourth Ave, Suite 200, Seattle, WA 98121, USA / 1-844-IGNEOUS / www.igneous.io

Igneous