Igneous

TECHNICAL CASE STUDY

Accelerating Image Analysis and Cancer Diagnosis with Igneous

Igneous
2401 Fourth Ave., Suite 200
Seattle, WA 98121
(844) IGNEOUS
igneous.io

Table of contents

Introduction

Artificial Intelligence (AI) is being utilized for a variety of tasks today, from self-driving vehicles to optimizing workflows in manufacturing operations to detecting malware on the internet. Deep learning is a form of AI where multi-layer neural networks are utilized to transform input data into progressively more defined and useful output. Deep learning differs from machine learning (ML) in that ML focuses on the development of task-specific algorithms that can be applied to specific problems. In contrast, deep learning focuses on extracting information at multiple levels. Moreover, deep learning networks can change their processing of data over time and without supervision to improve the results that they produce.

Deep learning has recently been applied to computational pathology, a broad set of computational methods whose goal is to extract quantitative information from medical imagery and related clinical data. The application of deep learning to computational pathology can provide significant value in the diagnosis of disease and other medical conditions, as well as in the design of therapy for these conditions. The use of deep learning for computational pathology is particularly well-developed in oncology, where it is used to support diagnostics and treatment formulation for various forms of cancer. PAIGE.AI, a spinout of Memorial Sloan Kettering Cancer Center (MSKCC), is one of the pioneers in the application of deep learning to cancer detection and diagnosis, with the goal of providing a decision support tool to aid medical professionals in the rapid diagnosis and treatment of oncology cases.

A critical part of any computational pathology solution is management of the petabytes of data needed to train the deep learning models, and to back up and restore this data and the artifacts of the AI models when needed. By combining the storage technologies from Igneous and Pure Storage, the PAIGE.AI solution is able to achieve aggressive recovery time and recovery point objectives, while protecting the artifacts generated by the deep learning networks from accidental deletion. The combination of Pure Storage FlashBlades and Igneous data management software provides the high-performance backup/restores capabilities and long-term retention necessary for PAIGE.AI's data protection needs. While the PAIGE.AI solution focuses on the analysis of pathology images, the capabilities provided by Igneous and Pure Storage can be applied to any kind of medical imaging, regardless of the modality that produces it.

Audience

This paper is written as a brief on how the Igneous data management and backup/restore solution can be utilized to support computational pathology workloads, as well as the data sets (medical imaging and diagnostic data) utilized by these models. We also present a use case on how Igneous tools are used by PAIGE.AI to manage its computational pathology data through the AI/ML workflow. Management of these data sets (and the artifacts produced during the machine learning process) is critical to providing the "rollback" capabilities needed as models are tuned and tested against new data sets. The combination of Igneous and Pure Storage FlashBlades not only ensures that these rollbacks can occur, but that they can occur both quickly and without significant management or IT intervention and support. The workflow descriptions in this paper are illustrative of how these data protection solutions can be created and utilized.

About Igneous

We deliver the only UDM as-a-Service solution enabling data-centric organizations with visibility, protection and data mobility at scale, wherever datasets and workflows live. Our customers see, organize and understand all of their unstructured data – anywhere. Our customers protect petabytes of data on a single cloud-native platform – at scale. Our customers automate movement of datasets – for everyone needing them. We combine all UDM functions into a single, API-enabled, cloud native solution.

The right data, in the right place at the right time. Find out more at igneous.io

About Pure Storage FlashBlade

The Pure Storage FlashBlade provides a data platform to support the need for high-performance access to dense unstructured data for on-premise rapid backup and recovery. Pure Storage FlashBlade additionally provides high-performance in a dense-form factor for other unstructured data primary use cases such as machine learning, deep learning and electronic design automation (EDA), among others. For more information, please follow this link (https://www.purestorage.com/products/flashblade.html).

About PAIGE.AI

PAIGE.AI helps pathologists to be more efficient, researchers to be more quantitative, and patients to be more confident in their diagnosis. Artificial Intelligence is at PAIGE's core. Their experts have a decade of experience behind them in building large scale machine learning systems for computational pathology. The PAIGE.AI solution is trained by the world's foremost cancer experts on hundreds of thousands of digital slides. The goal is to guide pathologists, clinicians and researchers via PAIGE's robust clinical decision support system. Clinical experts will gain massive efficiencies and reproducibility of their data.

PAIGE.AI – Machine Learning for Image-Based Cancer Diagnostics

PAIGE.AI's name exactly states their mission: to combine **P**athology images and **A**rtificial **I**ntelligence to provide a **G**uidance **E**ngine that accelerates the clinical diagnosis of cancer cases. PAIGE.AI utilizes a database of hundreds of thousands of anonymous digital images and case notes with a novel artificial intelligence/deep learning model and a best-in-class scale-out architecture to provide a clinical decision support system for oncology pathologists, clinicians, and researchers. At the heart of the PAIGE.AI solution is the Artificial Intelligence Ready Infrastructure (AIRI), the neural network that is utilized to run the PAIGE.AI deep learning models during training and for clinical diagnosis. AIRI comes in two sizes ("Mini-AIRI" and the full-size AIRI model), and combines Pure Storage FlashBlades, an Arista-based 40GbE/100GbE network, and multiple NVIDIA DGX-1 platforms (each DGX-1 has eight Nvidia Tesla V100 GPUs). The full-size AIRI system can hold up to 179TB of data (usable space before data reduction). The AIRI system is shown in Figure 1.

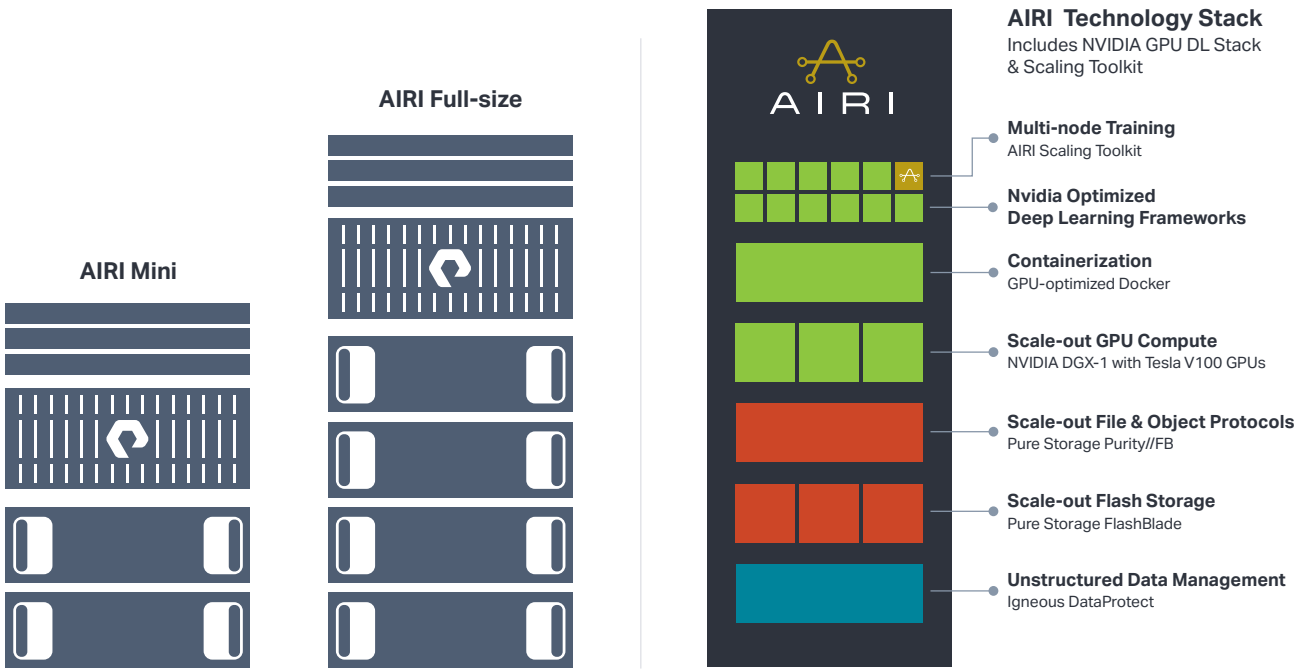


Figure 1: AIRI Systems and Technology Stack

The AIRI Computational Pathology Deep Learning Workflow for Oncology

Producing usable deep learning models is an iterative process where different learning and analytic approaches are applied to the learning dataset, and then validated with one or more test datasets. Each "learning run" results in a significant number of important artifacts that, along with the learning dataset and test dataset(s), need to be archived for later reference and/or use. These datasets are typically in the petabyte range, and require a significant storage infrastructure to manage them. This capability is provided by Igneous software and the Pure Storage FlashBlades. In addition, the Igneous solution also provide dataflow automation to the entire workflow.

Paige.AI and MSK are currently scanning about 30,000 pathology slides each month into their database, and the AIRI configuration is sized to support the scanning of up to 100,000 archive cancer slides per month. The role of the Igneous data management software is to manage the access to this data (including the movement of the data through the workflow), and the protection and long-term retention of this data. Igneous enables Paige.AI to protect the work product of the machine learning process through easy, automated policies, allowing for possible reuse or data restoration in case of accidental deletion. The combination of Pure Storage and Igneous high-performance storage for the AI/ML learning process, as well as an easy to-use method to protect and store the highly valuable results of the machine learning process. The workflow can be broken down into the following steps (see Figure 2 for a diagram of the workflow):

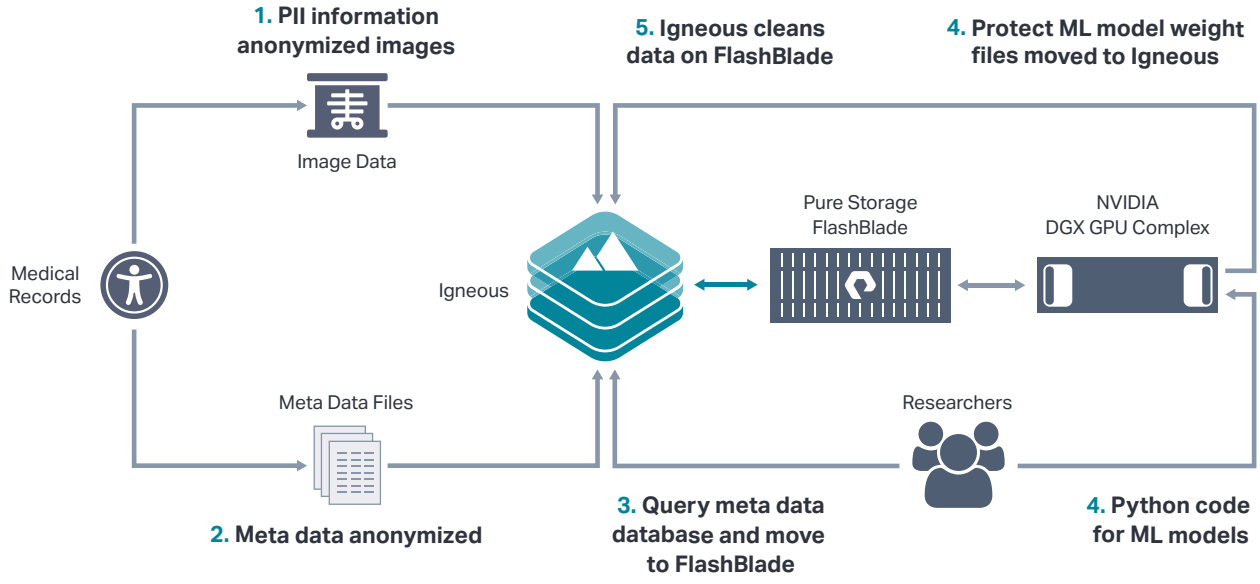
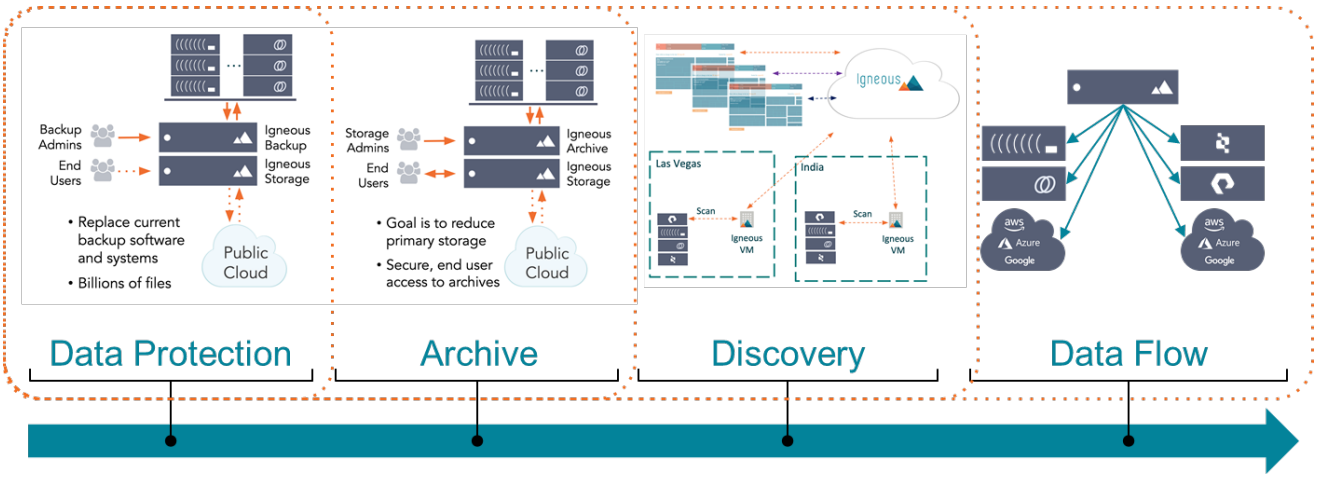


Figure 2: PAIGE.AI Deep Learning Workflow

1. **Image Workflow:** This part of the workflow moves scanned images from primary storage, anonymizing the images (removes any patient identifying information), and installs the images in Igneous storage until the data is ready to move to AIRI for analysis by the machine learning program. Images within the MSKCC workflow are contained in an object called an "image pack". Each image pack contains the high-resolution medical image, a thumbnail, and a barcode that acts as the image file name. When a batch of scans have been completed by the Pathology Department, they inform the team managing the image transfer that the scans are ready, after which the transfer team starts the image workflow. As a part of this workflow, the images are anonymized by deleting any personal identifiers, and generating a randomized number for the anonymized slides. Once this occurs, the images are moved to the Igneous storage where they are stored in a data bucket, and the database is updated with the filename.
2. **Diagnosis and Patient Data Workflow:** This part of the workflow occurs after the Image Workflow is complete; it anonymizes the patient records (in CSV format) and transfers the data to Igneous storage. This workflow starts after completion of the Image Workflow. A set of three CSV files are generated by MSKCC (known as CSV1, CSV2, and CSV3); these files contain metadata regarding the imagery and patient case/outcome data. These files are associated with the image files through the original image file name; once matched, the CSV files are also anonymized by stripping out any patient identifying information (PII).

3. **Data Alignment and Transfer to AIRI:** Once the CSV and image datasets are both in the Igneous storage, they are "aligned" (tying specific CSV data to the available image files) by parsing the CSV files and store the metadata results in the database. Once the metadata has been loaded into the database, the movement of relevant data from the Igneous Storage system to the Pure FlashBlade is initiated by querying the database and calling an Igneous API with the list of files to move. This approach still allows Igneous to speed up the data transfer process for Paige.ai from the Igneous Storage system to AIRI.
4. **Protection of ML Artifacts and Code:** This part of the process takes the deep learning source code and artifacts of the AIRI training session and stores them on Igneous, where they can be accessed by researchers, clinicians, and pathologists as required to judge the results and tune the models. During the machine learning process on AIRI at Paige.ai, many artifacts are generated. These artifact sets, and the source code for the deep learning run that generated them, are of high value and need to be protected for potential reuse in the future, against both accidental and/or intentional deletion. The Igneous data protection and restore functions enable the automation of this process through the easy creation and modification of policies through the Igneous user interface (UI). Igneous allows the protected data to be stored and accessed in time series and can handle an unlimited number of version copies.
5. **Cleanup:** The final step in the process is to "clean up" the data on the Pure Storage FlashBlades within AIRI at Paige.ai. This is important to ensure that only desired files are processed by each machine learning run, and that data from a previous run is not inadvertently included in a new deep learning training run. The cleanup process is accomplished through a set of scripts that are run by the Igneous data protection engine on AIRI after the data protection phase has been completed.

Summary: Accelerating Computational Pathology by Improving Data Transfer and Management for Artificial Intelligence and Machine Learning



Artificial intelligence in general, and machine learning/deep learning in particular, are quickly becoming critical technologies in a variety of settings. Machine learning/deep learning has helped to increase the accuracy and speed of cancer diagnosis and effective treatment. This has significant value in that it improves likely patient outcomes. At the same time, computational pathology and deep learning decrease the cost of treatment by identifying disease and potential treatment options far earlier in the disease cycle than is possible by utilizing conventional clinical methods.

The combination of storage technologies provided by Igneous and Pure Storage protect the value provided by the PAIGE.AI solution by protecting the data utilized in the deep learning models, as well as the deep learning models produced by PAIGE.AI's neural networks. The flexibility provided by the Igneous APIs allows Paige.ai to modify their architecture as needed for any given deployment scenario. The Igneous APIs also enable the automation of the PAIGE.AI workflow. The use of Igneous storage and data protection/ restore capabilities provides significant acceleration to this process by streamlining the movement of data to Paige.ai's Artificial Intelligence Ready Infrastructure (AIRA). This ensures that critical code and artifacts are protected for future reference and/or reuse. While this case study focused on PAIGE.AI and the use of deep learning for oncology diagnosis and treatment, these technologies can be applied to a variety of medical imagery use cases.

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