Building a Healthcare Data Science Platform in the Cloud

Grant DeLong, BA, Elliot G Mitchell, PhD
Geisinger, Danville, PA

Introduction

As organizations increasingly rely on data-driven insights to inform decision-making, the need for robust data science platforms has become paramount. Traditional on-premises infrastructure may not meet the scalability and agility requirements of modern data science workflows. Cloud computing offers a compelling solution, providing elastic resources and a wide array of services tailored for data-intensive workloads. However, the flexibility of cloud tools offered by multiple competing vendors creates a very large design space and many decisions. In this submission, we present a case study of the Geisinger AI Lab's initiative to build a new data science platform in the cloud. Through this case study, we aim to provide strategies and best practices for other organizations embarking on similar endeavors to harness the power of cloud computing for data science in the healthcare domain.

Methods

Our platform's design was centered around three key goals: portability, scalability, and security. **Portability** allows our team to remain flexible in the face of an evolving cloud provider market. This flexibility is crucial as business decisions may necessitate switching from one cloud provider to another, such as from AWS to Azure, without the need to rewrite our codebase. **Scalability** is essential due to the large volumes of healthcare data we handle and the need to scale our projects. Building performant models requires scalable compute infrastructure that can dynamically adjust to our variable resource demands. **Security** is a top concern because our source data contains Protected Health Information (PHI) and must comply with HIPAA regulations.

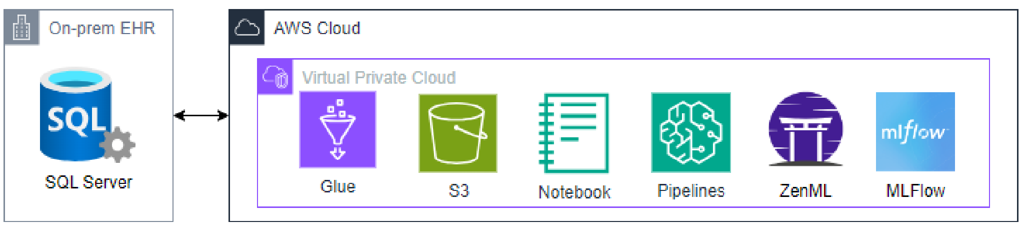

Figure 11. An overview of the Geisinger AI Lab's cloud architecture, combining cloud resources like AWS SageMaker and Glue to access on prem data, and using open-source tools like ZenML and MLFlow to ensure portability, scalability, and security.

Results

To achieve our design goals, we implemented several key strategies. For **portability**, we prioritized the use of open-source tools in our technology stack, ensuring adaptability and avoiding vendor lock-in. We chose ZenML[1] as an open-source tool to define and deploy our machine learning pipelines. This allows us to configure various cloud stacks (e.g., AWS Sagemaker, Azure ML) for pipeline execution, enabling us to remain cloud agnostic. Additionally, we use MLFlow[2] for experiment tracking and model registries, leveraging its wide support among cloud vendors to maintain flexibility and interoperability. For **scalability**, we selected serverless orchestrator stack components in our ZenML stack, such as Sagemaker Pipelines, to configure the compute necessary for our workloads, including memory-optimized or GPU-enabled instances. ZenML containerizes our pipeline code into Docker images, which are then used by the orchestrator to create the execution environment scaled to the specified compute type. This ephemeral compute infrastructure spins down after pipeline execution to control costs, ensuring efficient resource utilization. For **security**, we undertook a comprehensive approach to ensure best practices and compliance. Our collaboration with internal security expert teams was iterative and collaborative, following the principle of least privilege. We defined roles with specific conditions on IP addresses and S3 buckets, enforced encryption, and utilized customer-managed keys to enhance data protection. Furthermore, we enabled traceability and auditing through ZenML, which logs each pipeline run and each output from pipeline steps as artifacts in our cloud storage. These measures collectively ensure that our platform not only protects sensitive data but also maintains robust compliance with HIPAA regulations.

Conclusion

Our healthcare data science platform in the cloud is designed with portability, scalability, and security as its core principles. We share this case study with the aim of aiding other organizations in designing their cloud environments for data science workloads.

References

1. ZenML https://www.zenml.io/
2. MLFlow https://mlflow.org/

Abstract
This case study details the Geisinger AI Lab's cloud-based data science platform, designed for portability, scalability, and security. By using open-source tools and serverless architectures, the platform ensures flexibility, efficient resource use, and HIPAA compliance. We offer strategies and best practices for leveraging cloud computing in healthcare data science.