clearscale

aws partner network

Premier Consulting Partner

The American College of Radiology Builds Secure and Scalable Data Lake

ACR American College of Radiology

Executive Summary

For nearly 100 years, the American College of Radiology (ACR) has worked to advance the science of radiology and support healthcare professionals all over the country. Today, the organization represents over 40,000 medical providers, including radiation oncologists, interventional radiologists, medical physicists, and nuclear medicine physicians.

Previously, the ACR and ClearScale worked together on a data infrastructure project. ClearScale delivered a cloud landing zone prototype with automated governance controls, secure architecture, and modular design. Given the success of that round of work, the ACR asked ClearScale to help with the next step which involves building a sophisticated data management ecosystem that makes the most of the Amazon Web Services (AWS) platform.

As an AWS Premier Partner with a Data & Analytics Competency, ClearScale was the ideal partner to complete the work.

> "ACR's focus was to bring speed and agility to end-to-end data pipelines for faster and continuous data delivery for analytics. We were looking for a partner that could work with our team to build a data lake that would allow us to process and add new datasets easily. ClearScale helped in a variety of areas including creation of a serverless data platform to ingest data from various data sources, automated data cataloging, and creation of a scalable datastore for business analytics and reporting."
>
> **Shree Periakaruppan**, Director of Data Engineering and Analytics, ACR

The Challenge

There were several data architecture challenges in particular that the ACR wanted ClearScale to help address.

First and foremost, the organization did not have a single unified data warehouse system that encompassed all enterprise data assets and maximized the ability to quickly analyze data and pursue advanced knowledge discovery across all enterprise systems. As a result, assembling data for integrated analyses could be time-consuming and labor-intensive.

There was a lack of a comprehensive, dynamic metadata catalog that would allow for a quick and easy window into the current state of all enterprise data assets at the most atomic data element level. Different teams needed to maintain their own system-specific catalog requiring that complex inquiries that cross systems involves different individuals and distinct metadata resources. Not only does that make identifying data assets time-consuming, but the process of tracing complex data transformations could be labor-intensive, particularly when involving multiple systems and complex data lineages.

Other goals included the desire to provide more fine-grained security controls such as column-level permissions that would allow one database to serve multiple constituencies, rather than controlling access through collection-level permissions that may require multiple copies of data to accommodate complex security requirements.

With these objectives in mind, the ACR requested the following:

- Enhanced cloud landing zone capable of supporting subsequent solutions
- Integration with ACR's CRM to ingest data and allow for more advanced reporting solutions.

The ClearScale Solution

ClearScale approached the ACR's deliverables in several stages. The cloud consultancy leaned heavily on AWS's suite of managed services and data management solutions to fulfill ACR's needs.

Increasing Landing Zone Control

ClearScale first focused on giving the ACR more granular control over its cloud landing zone. The team added preventative and detective rules using AWS Service Control Policies (SPCs) and Config Rules. With these rules, the ACR can now easily track such events as Network Configuration Updates, identify unsafe ports in use, and set up alarms that trigger when violations occur. ACR's IT team no longer has to spend substantial time proactively monitoring its AWS landing zone.

After making these changes, ClearScale moved on to the bigger task at hand: delivering a secure data lake.

Building the Data Lake

The ACR wanted a secure data lake built on Amazon S3 that stakeholders could use to analyze data from various sources using standard SQL. The two most important data sources at this stage were Salesforce and an on-premises MSSQL database cluster.

To enable data updates from Salesforce, ClearScale used Amazon AppFlow, a fully managed integration service that makes it easy for users to transfer information between popular SaaS applications. On the database integration side, ClearScale implemented AWS Database Migration Service (DMS).

When used together, these solutions enabled ClearScale to create connections to remote data sources, configure automatic cron-triggers to AppFlow instances, and implement ongoing replication to keep source and target databases in sync.

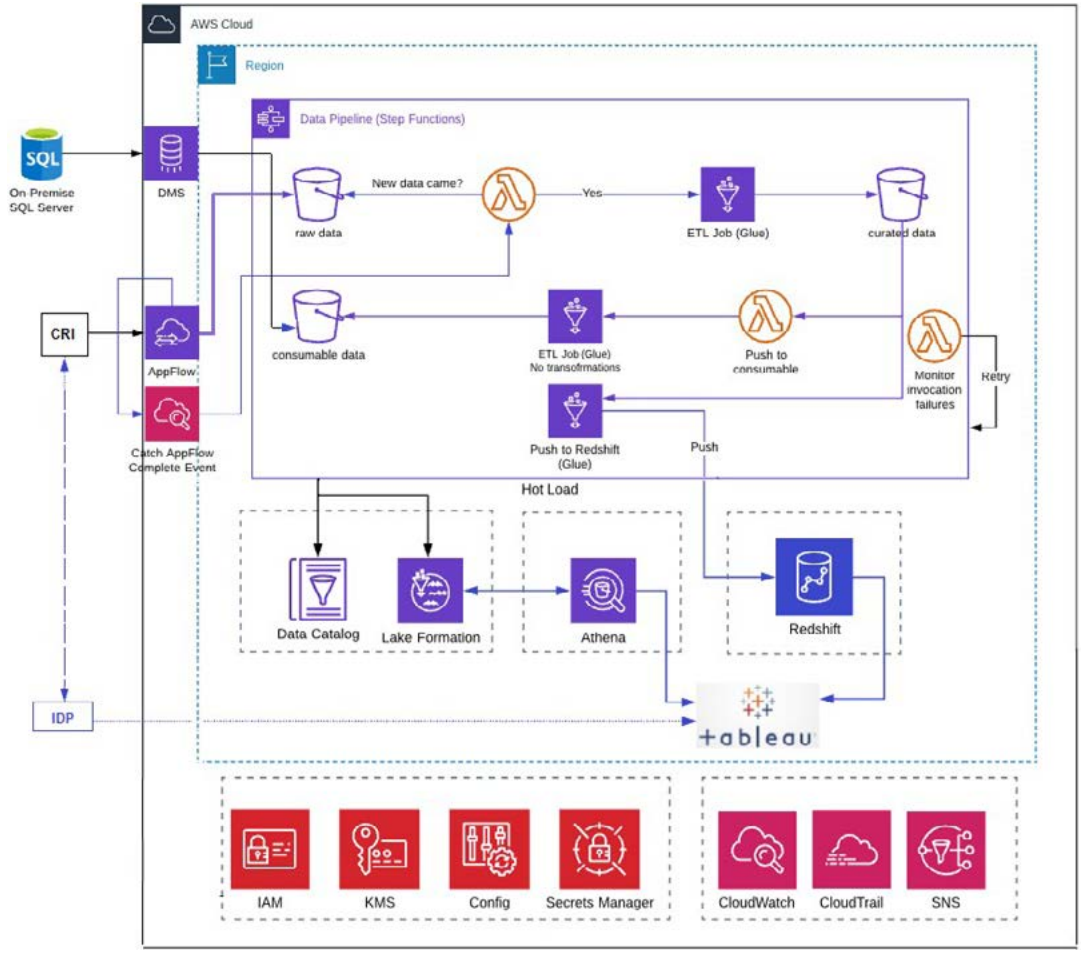
Next, for the data lake, ClearScale created three layers on Amazon S3:

- A raw layer for storing data in parquet format from external sources in an append-only manner
- A curated layer that handled all the data transformations to the raw layer and made the data available for efficient querying in Athena. This data is ultimately moved to AWS Redshift to serve as a source for Tableau dashboards
- A consumable layer for aggregated up-to-date data that serves as a backup in case data is lost from Redshift

To move data from the raw layer up through the consumable layer, ClearScale's data experts used AWS Glue Jobs and AWS Glue Crawler, both of which keep S3 schema up-to-date for Athena queries.

With this revamped infrastructure, the ACR can ingest, store, extend, and publish data from Salesforce and the existing MSSQL database in a secure manner. Furthermore, should the organization want to incorporate additional data sources, it can easily do so.

Architecture Diagram



The Benefits

Thanks to ClearScale's expertise, the ACR now has a secure, scalable, and reliable data lake that integrates seamlessly with various sources. ACR stakeholders have all of the information they need in one place and can easily track how data moves across the ecosystem. Users can dig into the history of any data point and track all transformations. Teams can also prepare reports that feature data pulled from many sources.

The ACR's new data lake increases the overall reliability of the organization's data infrastructure by protecting source systems from overloading and providing a single point of access for users. On top of that, the ACR's IT team can implement advanced security measures to safeguard sensitive data from unauthorized personnel.

Through this second engagement with ClearScale, ACR's cloud landing zone and data architecture evolved from a prototype to a comprehensive solution that will serve the organization well for years to come.