Cerner HealtheDataLab Overview

First Published July 30, 2020

Updated August 30, 2021

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

Contents

Abstract

This paper shows how HealtheDataLab can help streamline the data science workflow in a secure, elastic environment for organizations that use healthcare data. HealtheDataLab is a data science ecosystem that helps users at diverse care venues rapidly develop new insights, and push those insights back into clinical workflows.

Many organizations spend most of their data science resources bogged down in the cleaning, managing, and organizing of data. HealtheDataLab helps minimize this work and simplifies critical tasks, such as creating patient cohorts to answer specific research questions.

Introduction

Many healthcare organizations focus on initiatives to use advanced analytics and intelligence to help analyze data and derive insights, with the goal of improving patient outcomes. These initiatives typically require access to large amounts of data, and to computing and processing power. Most organizations are faced with the need to organize scattered datasets and access around a centralized environment to extract, cleanse, normalize, and validate data.

HealtheDataLab addresses common challenges in the traditional data science workflow that can collectively slow down dataset development, data analysis, and deployment of valuable insights into clinical and operational workflows.

Dataset development breaks down into a number of tasks. Each step can prolong insight delivery.

- **Data ingestion** – Clinical and business systems use a wide variety of data formats, and frequently display various types of anomalous behavior. Accessing and retrieving data can be an enormous challenge, especially without impacting the performance of critical systems.
- **Data identification** – Identifying this data, what it means, and how to best normalize these clinical concepts can consume thousands of hours of labor. For example, as a researcher, you only want to know if a patient has been diagnosed with asthma. Your research efforts may not require additional diagnosis such as an ICD-10 diagnosis code, a SNOMED code, or a CPT code.
- **Data cohort development** — Extracting a group of patients with the characteristics you need, while not contaminating them with undesirable characteristics such as comorbidities or geographical clustering.
- **Data ingestion infrastructure** – Ensuring adequate capacity, both in terms of storage and data processing, for data that can grow unexpectedly in terms of size and complexity.

Once you have a dataset ready to be analyzed, the core of the data science workflow is designed to do the following:

- Analyze the dataset to answer the research question(s)
- Build a predictive model to more accurately forecast spend

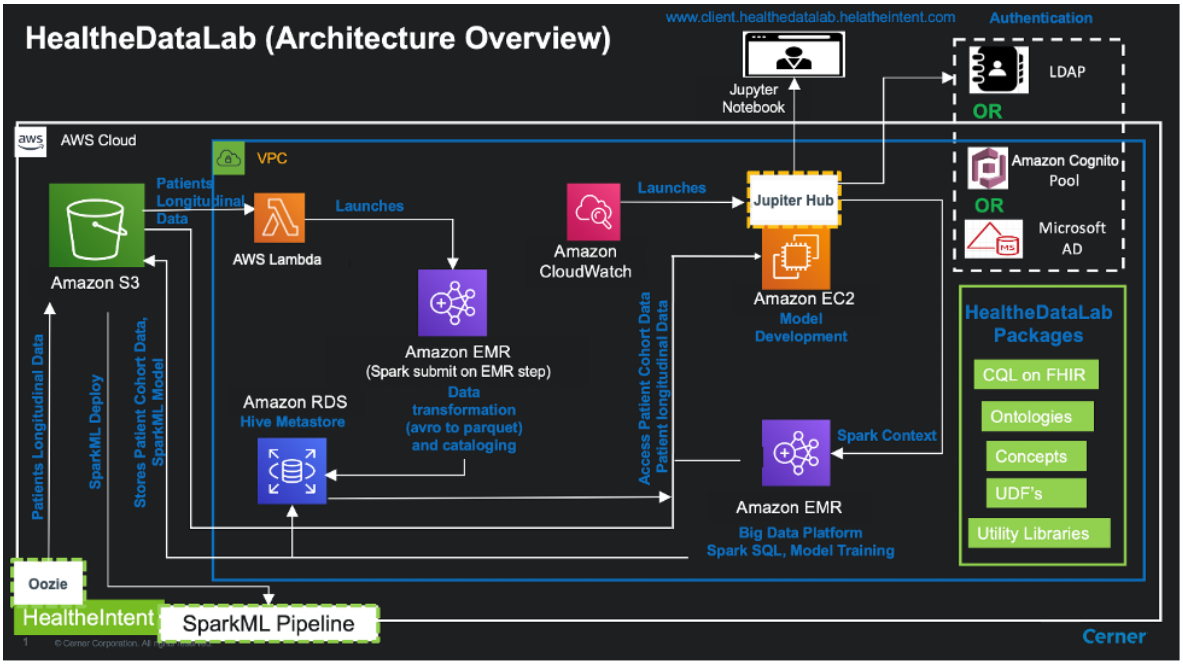
- Quickly validate findings to provide relevant insights back to your organization in a secure environment
- Use cutting-edge artificial intelligence (AI) and machine learning (ML) tools to build out features
- Collaborate with other data scientists, researchers, and colleagues to advance individual research efforts

After the data science workflow is complete, you can put that data to work. Now you can:

- Embed insights into clinical workflows, such as prioritizing a list of patients for care management based on risk assessment
- Monitor performance, rapidly deliver updated models, and improve quality
- Minimize technical resources required to deploy models and algorithms in production environment

Service description

HealtheDataLab provides an environment for model and algorithm development.



HealtheDataLab architectural overview

Information is delivered via the Cerner HealtheIntent platform to the HealtheDataLab data lake hosted on Amazon Simple Storage Service (Amazon S3) From the data lake, patient longitudinal data is extracted using the AWS Data Pipeline service and processed to create metadata describing the patient records. The processing is done using Amazon EMR and then stored as a Hive metastore in a relational database hosted by Amazon Relational Database Service (Amazon RDS).

Data scientists authenticate into a secure environment using Amazon Cognito to interact with and analyze data via a Jupyter notebook. Jupyter is a popular open-source tool for doing data analysis and developing machine learning models. Cerner has packaged the open-source tool as an effortless service that doesn't require any server administration knowledge to use or maintain in the client environment.

To provide a responsive interface and scalable resources, each notebook is attached to an Amazon EMR cluster. This feature enables access to a diverse portfolio of functionality. The Cerner packages and user-defined functions are particularly valuable for supporting data scientist activities.

- Bunsen, an open-source project from Cerner, enables users to load, transform, and analyze Fast Healthcare Interoperability Resources (FHIR) data with Apache Spark. This project has Java and Python API operations that help you convert FHIR resources into Spark datasets for exploration in HealtheDataLab and across systems.
- Clinical Quality Language (CQL) is a programming language designed to express logic in the clinical domain. It can be used within both the clinical decision support and clinical quality measurement domains. HealtheDataLab provides a runtime implementation of the language that runs queries in parallel on a Spark cluster over Hive databases of FHIR resources.
- HealtheDataLab supports working with SNOMED, LOINC, and other clinical ontologies. It can import the content to query with a series of commands to maintain hierarchy, versioning, and relationships within the data.
- HealtheDataLab users plug into the Cerner Discern ontology concepts and contexts with convenient user-defined functions that can be run directly in Spark SQL. Concepts represent a group of codings that originate from the same or different coding systems, while a context is a group of related concepts used to better understand and identify conditions.

The development of models, including training and related activities, occur in the Jupyter notebook and attached Amazon EMR cluster.

Once an algorithm or machine learning model has been developed and verified, it can be hosted using Spark Pipeline on HealtheIntent, or on any other Spark Pipeline instance.

Benefits

Cerner HealtheDataLab supports the development and integration of rule-based, symbolic, and machine-learned algorithms for risk prediction, modeling, patient care guideline identification, and more.

The ability of HealtheDataLab to ingest data from the Cerner HealtheIntent platform offers several benefits. The HealtheIntent platform takes data from more than 1240 data sources and applies data cleansing, standardization, concept normalization, and person matching to deliver a longitudinal patient record and populations for end users. This negates the need to write custom extract, transform, load (ETL) jobs for all of these tasks, including jobs for each new data source, jobs to cleanse and standardize data, jobs to normalize data, and then jobs to perform person matching and build a longitudinal patient record.

The collaboration of Cerner with AWS to deliver HealtheDataLab offers the following advantages.

- Amazon S3 enables storage scaling without performance degradation into the exabyte range.
- Amazon EMR and Amazon RDS allow the rapid extraction of metadata from patient records in a manner that scales with the amount of data provided.
- Amazon RDS provides a robust, highly-available platform for storing metadata long term.
- Jupyter, Amazon EC2, and Amazon EMR provide a scalable, performant, robust mechanism for rapidly analyzing data, developing ML models, and performing other high impact tasks in a scalable, robust manner.
- The HealtheDataLab packages provided for use in the Amazon EMR Spark environment simplify and accelerate many healthcare-related data tasks.

The Cerner Discern ontology concepts and contexts enable the selection of patients based specific morbidities or other health criteria without having to translate those morbidities into ICD-10 codes or specific lab result groupings.

- Deployment through the Spark Pipeline to the HealtheIntent platform provides the ability to integrate insights into electronic health record (EHR) agnostic clinical and operational workflows. Cerner customers have integrated more than 65 different EHR solutions with the platform.
- HealtheDataLab provides the ability to upload datasets and insights through a simple utility command from the Jupyter Notebook into a visualization tool using Tableau Software.

This addresses a chronic problem with data science, informatics, and business intelligence teams within healthcare organizations: how to rapidly deliver and update actionable insights for use by the larger organization.

Use case scenarios

The Advocate Cerner Collaborative was established to accelerate innovation in population health management. With a focus on the efficient allocation of clinical resources, predictive analytics was shown to be a key strategy in identifying high-risk patients earlier.

For example, clinicians know that patients with heart failure have a higher risk of being admitted to the hospital, but it would help clinicians to know if there is a subset of patients who are at high risk for short-term admission, whose care could be better managed at home.

Advocate and Cerner developed a risk score using HealtheDataLab and integrated it into the Advocate care management system to prioritize patients for care management outreach. A year later, the team created additional risk scores to offer this type of program to people living with other conditions, such as chronic obstructive pulmonary disease and asthma. The collaborative was able to remove the technological barrier and focus on operating efficiencies.

Children's Hospital of Orange County (CHOC Children's) also used HealtheDataLab to build risk scores. CHOC Children's wanted to better support families with children at the highest risk of readmission. With an organized and unified dataset, and advanced computation inside the solution, CHOC Children's was able to quickly build and improve its readmission model in weeks, rather than months or years like its legacy system. The organization tackled building 11 additional models in just a year with only a small team of data scientists.

> "Previously, it would take a couple of months to develop a model, extract the data and iteratively run through the algorithm. Now it only takes us a couple of days. Since streamlining our workflow, our data science team has gained tremendous efficiency. We no longer require assistance from multiple groups for all of our activities, and we're not constrained by hardware limits."
>
> - Louis Ehwerhemuepha, Ph.D., Data scientist, CHOC Children's

Another example of results with HealtheDataLab is the AWS and Cerner collaboration to create a model for predicting the onset of chronic conditions, such as congestive heart failure[1]. The model was demonstrated to be capable of predicting the onset of congestive heart failure months in the future. With this model, providers can implement strategies designed to reduce risk factors, such as controlling high blood pressure, high cholesterol, and diabetes. This model can also assist researchers in evaluating interventions that have the potential to delay or avert the development of other conditions with high mortality, morbidity rates, and significant costs.

The data preparation and analysis were both carried out using HealtheDataLab.

- Data from HealtheIntent was syndicated into Amazon S3 as Apache Parquet files in a FHIR-inspired data model.
- The database information was stored in the Hive metastore.
- The EMR File System (EMRFS) was used to access the data from Amazon EMR instances with Spark, where data analysis and processing was performed.
- Jupyter notebooks were used as the interface and PySpark was used to analyze the data.
- Once the feature sets were isolated with information from demographics, conditions, lab results (vitals), and procedure tables, they were saved as Pandas data frames and NumPy arrays.

The power of this model is not limited to congestive heart failure. It is generally applicable to predicting the onset of other chronic conditions.

Related services

- HealtheIntent
- Amazon S3

- Amazon EMR
- Amazon RDS
- Amazon EC2
- Amazon Cognito
- Jupyter
- Apache Spark

Conclusion

HealtheDataLab accelerates the development of models and algorithms for organizations that use healthcare data. Time-to-value is an important component of any data activity, whether you are developing improved clinical interventions or financially planning how to deliver healthcare in the future. If you have an interest in using HealtheDataLab, contact Cerner for help with getting started with HealtheDataLab.

Contributors

Contributors to this document include:

- Brian Niemeyer, Sr. Partner Solutions Architect

Document revisions

Date	Description
August 30, 2021	General updates
July 30, 2020	First publication

Notes

[1] Effectiveness of LSTMS in Predicting Congestive Heart Failure Onset