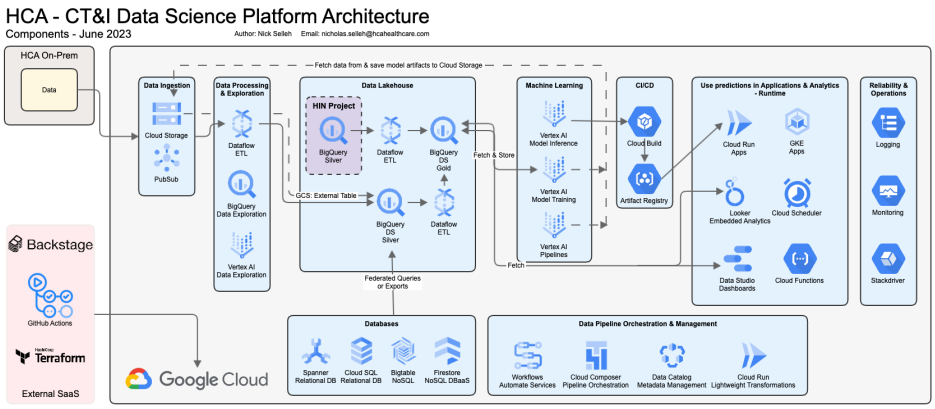
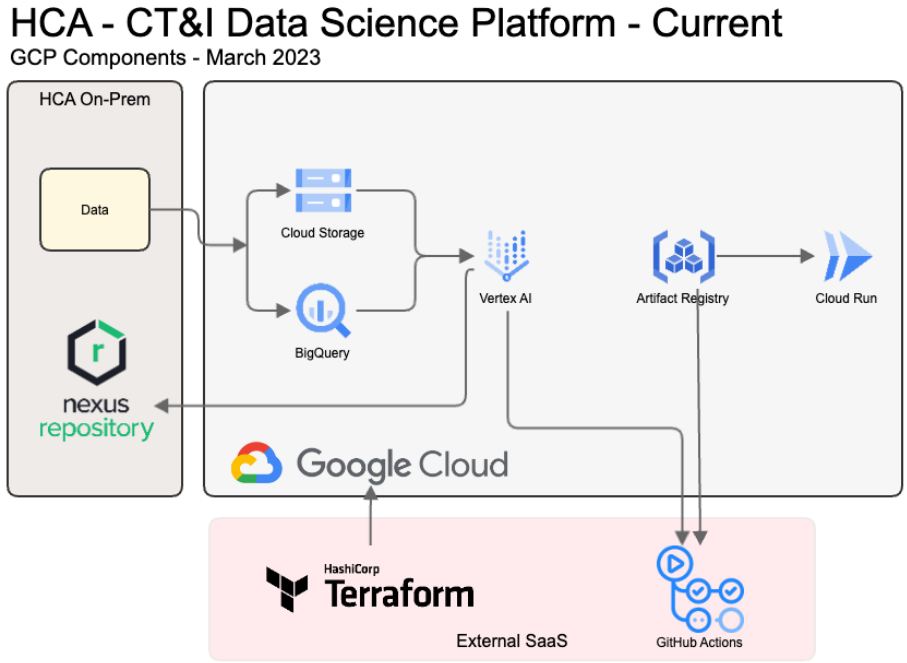
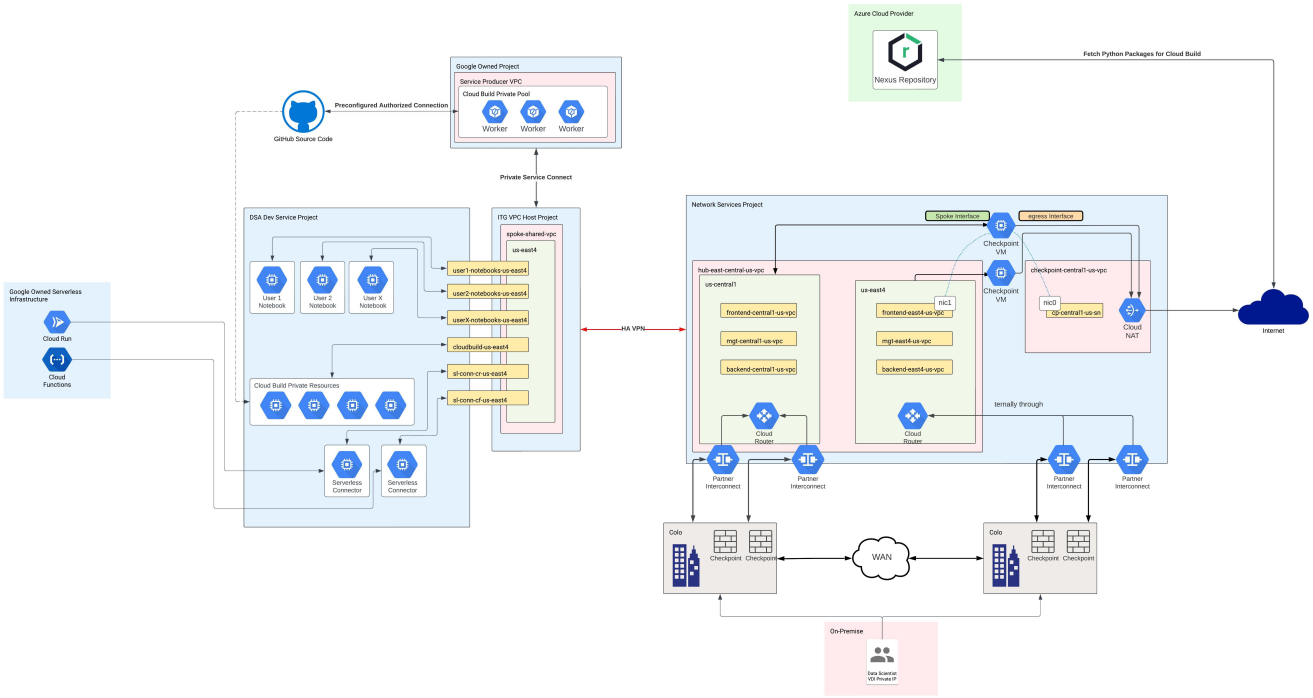
GCP ML Platform ReArch





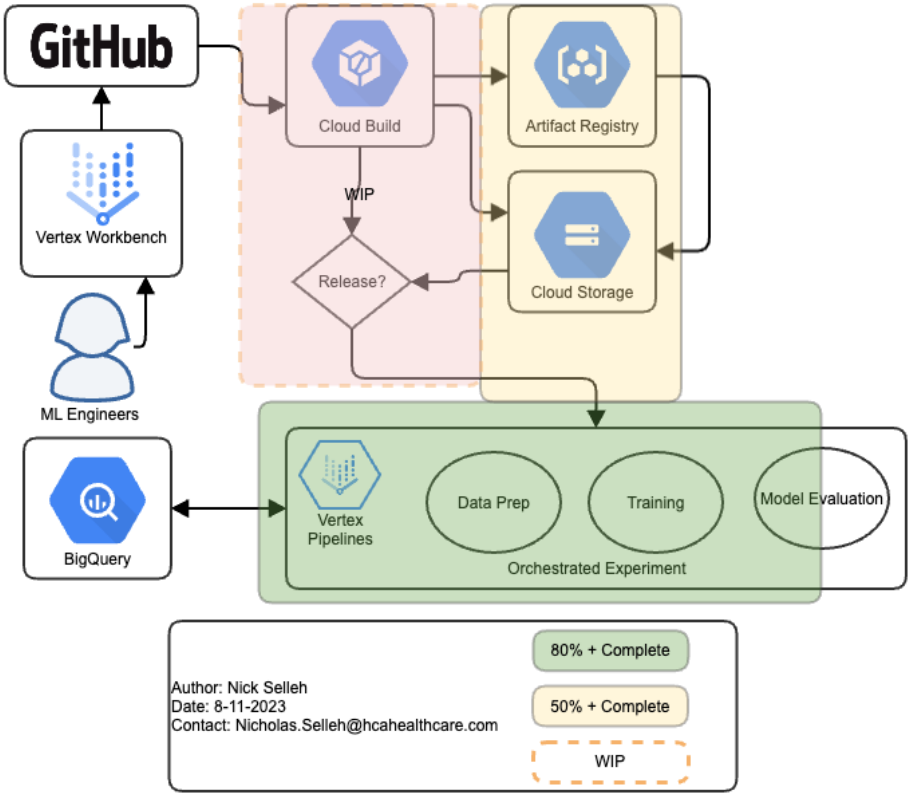
DSA Network Configuration

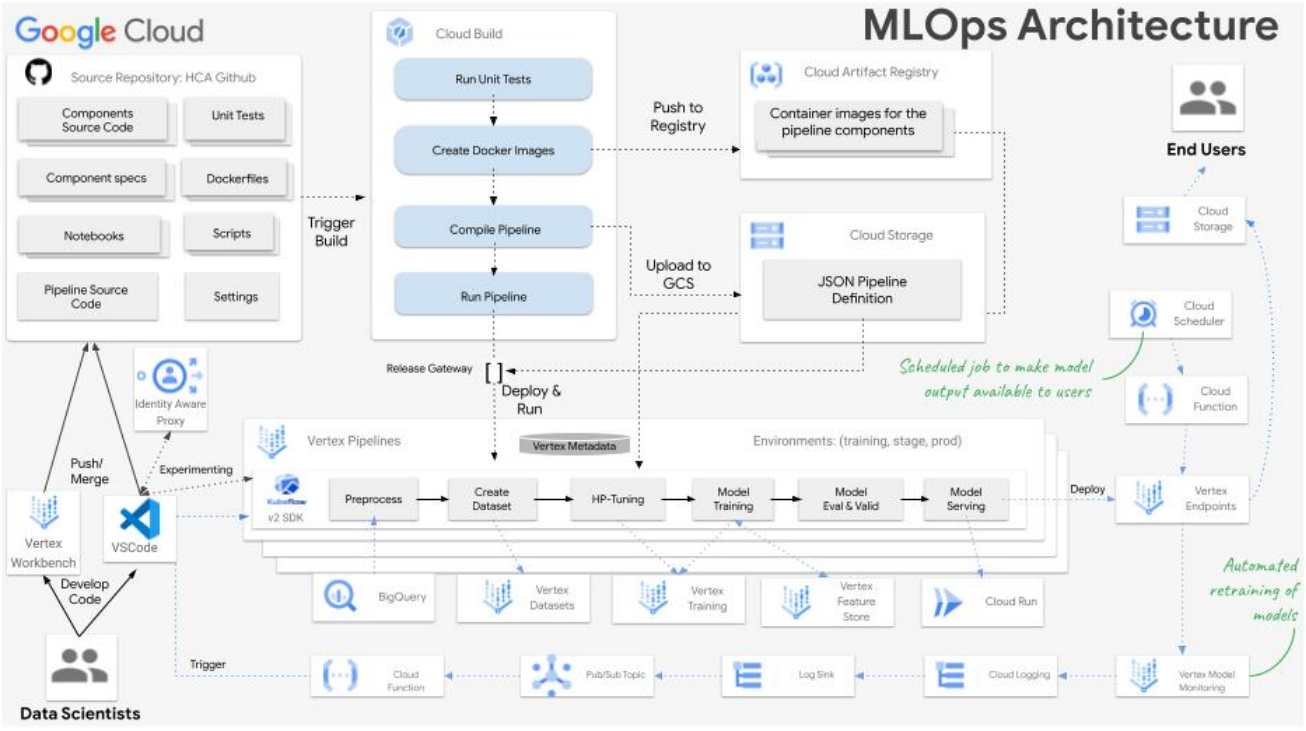


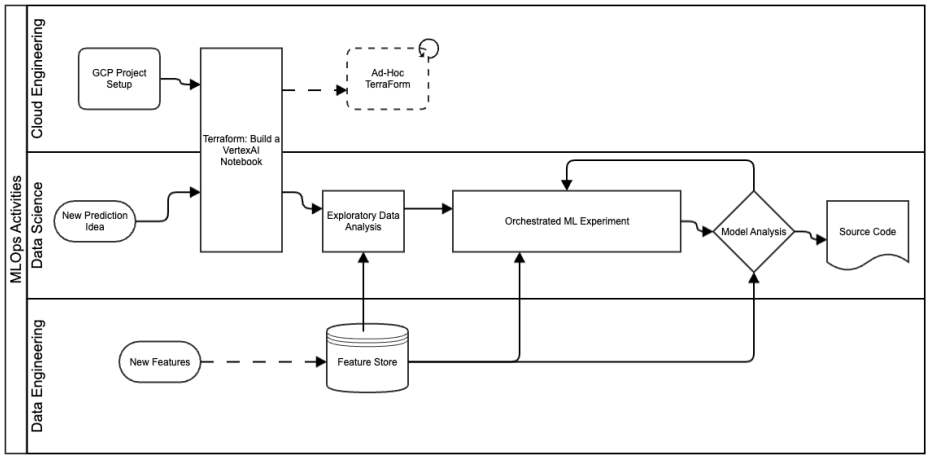
DSA MLOps

NOTE: These diagram are subject to change as we build out the pipelines.

GCP MLOps Reference Arch







Service Level Objectives

1. Agile Data Addition - An HCA Data Scientist can
    a. add (1) new data feature to an existing MLOps pipeline
    b. perform automated training with the new data feature
    c. comparatively evaluate the new model against the previous model - in a MLOps pipeline
    d. promote the modified MLOps pipeline, with the (1) additional data feature to production
    e. **in 1 week or less**
2. New Data Scientist / MLOps Onboarding - A new to the team Data Scientist / MLOps engineer can
    a. given - their access is properly setup to the following: HCA, GCP, github, *data sources (could be DS/MLOps feature store)
    b. run automated training pipelines to build better models
    c. comparatively evaluate the new model against the previous model - in a MLOps pipeline
    d. **in 3 data or less**
3. An CT&I data scientist can
    a. instantiate a new project, complete with all the components from a reference architecture (batch, streaming, vision, TBD)
    b. with access to HCA's HIN project
    c. by utilizing a self-service portal
    d. and begin training a new model
    e. **in 4 hours or less**

ToDo: Make Swimlanes for these SLO's - NICK

Reference Architectures

Reference Architectures for Data Science are standardized designs that provide a *frame of reference* for the data science domain. They incorporate best practices, common vocabulary, and reusable designs. They are not meant to be implemented directly as solution architectures, instead they are used as constraints for more defined solution architectures. A reference architecture can be thought of as a starting template for a data science solution.

Prior Art from Google

Data Science and Analytics Platform HIN Docs - https://hcahealthcare.sharepoint.com/:f:/s/CORP-HIN/EhYQ1mVRNTtFt3yLUNHkcgIBfZhInfOQJ17csJFsH1OmsA?e=ejHCAN

Code examples from the above - https://github.com/hca-ccoe/terraform-gcp-model-training/tree/mlops

Project Overview

The aim of this project is to develop a comprehensive Data Science Platform on Google Cloud Services that will enable streaming data modeling, batch data modeling, model deployment, and a functional approach to data. This project should build on top of the existing work the DS team has done to pioneer basic GCP capabilities, and make GCP for DS easier to leverage over time. This platform will provide our business with the necessary tools to make data-driven decisions and enable us to gain a competitive edge in the market.

Objectives

The main objectives of this project are as follows:

- Develop a data science platform on Google Cloud Services that can be used by staff in CT&I Data Science and possibly other teams within HCA
- Strive for self-service for data exploration, modeling, and solution deployment
- Build repeatable reference architectures for streaming data and batch data modeling and solution deployment of those models
- Develop a functional approach to feature engineering for data science efforts
- Ensure data security and compliance with regulatory requirements
- Training and support for our team members on the use of the platform

Milestones

The following are some milestones for this project:

- Development of data science platform architecture and plan
    - to include streaming data and batch data modeling reference architectures
- Development of a functional approach to feature engineering
    - enables the reuse features across projects and stages of ML lifecycle
    - increases data quality by helping with topics like feature skew
- Implementation of self-service for:
    - instantiating the modeling reference architectures - to get a data scientist working on a problem quickly, safely, and correctly.
- Testing and quality assurance
- Training and support for team members
- Project completion

Assumptions

The following assumptions have been made for this project:

- All required resources will be available for the duration of the project
- The project team will have the necessary skills and expertise to complete the project
- The project team has access to the right environments and data

Risks

The following risks have been identified for this project:

- Changes in technology or regulatory requirements may impact project timeline and scope
- There are enough people to build out this project
- Lack of stakeholder support may impact the success of the project
- Technical challenges may arise during the development of the platform
- Non-compliance with regulations may lead to financial and legal penalties
- Using example data for engineering - this will incur technical debt downstream in the format of Data Engineering.