1 - Provision

Background story

Caladan has established connections to various COVID-19 data sources for health and safety personnel. While these data sources have daily data from different geographies they are not in a format that makes it easy to analyze or create reports.
Both govenrment leaders would like policy recommendations and the ability generate their own reports to dig deeper.

To enable this long term vision, Caladan would like to establish an enterprise data lake to feed covid-19 and future infectious disease data needs. This central repository can later be leveraged for analysis, reporting and eventually - machine learning.

The datasources do not currently contain sensitive information but it is anticipated that they will in the future as other infectious disease data are ingested into the data lake. Caladan is concerned about their consituents' data, so they want to ensure such data will be protected at all times, with access limited to those who truly require it for business reasons.

Technical details

The team has the freedom during What the Hack to choose the solutions which best fit the needs of solution. However, the team must be able to explain the thought process behind the decisions to the team's coach.

At present, encryption is not a requirement for the data. However, the selected technologies must include mechanisms for controlling access to sensitive data.

Success criteria

- The team has created a repository in Github and all participants have access to the repo
- The team has selected and provisioned a storage technology for use as an enterprise data lake
- The selected storage technology must support storing both structured and unstructured data from both relational and non-relational source systems
- The team has stored the business process document and architecture document in the selected storage and within Github
- The team must explain to a coach how the selected storage technology would support restricting access to these files, such that only designated users or groups could access it

Tips

- The team does not need to **effectively set** permissions on the files; they only need to explain to a coach **how** the selected storage technology would support doing so
- A variety of storage options are available within Azure. The team should consider the features and tradeoffs between these offerings.
- In particular, the team should consider that the enterprise data lake will serve a variety of comonwealth needs and user personas. File system semantics will be extremely useful as the team advances through the challenges.

Resources

Ramp Up

- Data lakes
- Data Lakes and Data Warehouses: Why You Need Both

Choose Your Tools

- What is Azure Synapse
- Introduction to Azure Data Lake Storage Gen2
- What is Azure Blob storage?

Dive In

- Quickstart: Create an Azure Data Lake Storage Gen2 storage account
- Get started with Azure Data Lake Storage Gen1 using the Azure portal
- Create a storage account