2 - Ingest from Cloud

Background story

Great job! You now have a central landing zone for your different data sources.

Caladan currently stores policy data in Azure Cosmos DB.

They also store case, death and recovery metrics for 5 sample countries in Azure SQL DB. Caladan would now like to extract the data from these systems into the data lake. This will set the stage for incorporating additional data from the on-premise country data from the Azure VM.

Technical details

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

At present, encryption and access control is not a requirement for the data.

The team will find the following resources in the OpenHack lab subscription.

Caladan Resources

Caladan has **one** Azure SQL DB with metric data from five countries and a document collection in Cosmos DB with all the policy data. The team will focus on these resources for Challenge 2.

Access keys for Cosmos DB are available from within the Azure portal.

The team's coach can provide credentials for the SQL database. They are also available on the CloudLabs homepage.

Note: Each team member should add their Client IP to the database if they are going to connect from their home machine via Azure Data Studio, etc.

Alternatively, the team may set the Active Directory Admin to one of the provided OpenHack accounts.

Success criteria

- All data from the `dbo` schema from the `covid19` Azure SQL database has been extracted and stored in the enterprise data lake.
- All data from the `covidpolicy` collection in Cosmos DB has been extracted and stored in the enterprise data lake.
- All scripts/code used to move the data to the Data Lake is persisted in Github.

Tips

- Focus on the immediate objective of landing the data in the data lake. There is no need to implement transformation, cleansing, etc. at this stage. Once the data has been landed, such processing can take place in future challenges.

- The previous challenge has mentioned setting permissions on sensitive data. It is **not required** that the team sets such permissions in this challenge, only that they have previously communicated **how they would** do so.
  - If the team does wish to set the permissions, be aware of a known issue with special characters, for which the workaround is to ensure that the object names do not contain certain special characters. For example, `dbo.Customers` will work, while `[dbo].[Customers]` will not.

When selecting a technology to ingest the datasets for this challenge, the team should consider whether that same technology may be leveraged in the future.

Resources

Ramp Up

This challenge focuses on the Extract portion of ETL and ELT workloads, and the Ingestion and Storage stages of modern data warehousing:

- ETL and ELT overviews
- Get Started with Azure Synapse Analytics
- Azure Databricks Workspace
- Introduction to Azure Data Factory

Dive In

- Azure Cosmos DB quickstart
- 5-Minute Quickstarts: Azure Data Factory
- Move data to or from Azure Blob Storage using SSIS connectors