Steps for a Successful Capstone Project

## Mock Business:

1. Name:
2. Scope (Local, State, Regional, National):
3. Description
4. Number of Locations:
5. Number of Employees:

## Model a Business Process

## Logical Data Model

1. Create an entity that represents each person, place, thing, or concept (Master Data) and the interaction between them that describes the business process (Transactional Data).
2. For each entity and its primary key, identify and define attributes that describe the entity but only those attributes needed for your business process (e.g., Customer (customer\_id, first name, last name, address, phone number, etc. or Transaction (item purchased, quantity, unit cost, sub-total (before tax), total amount (after tax), data and time the transaction took place, the employee\_id (store clerk), the customer\_id (Only of the customer participates in your companies points program.)
3. Define relationships between primary key (PK) of each entity and their corresponding foreign key (FK) in the child entity, e.g., The PK in the Employees entity (i.e., employee\_id) has a relation with the FK, i.e., the employee\_id in the Transaction entity.

## Promote the Logical Data Model to a Physical Data Model

1. Be sure to generate the DDL (data definition language – a subset of SQL) from the physical model and deploy it to an AWS RDS instance (or choice – SQL Server Express Edition, MySQL, or PostgreSQL)
2. Connect to the RDS database through a local installation of Azure Data Studio, DBeaver, DBVisualizer, DataGrip or another SQL editor of your choice.)

## Get or Generate Data to Populate Your RDS Database

1. Need to use data from at least three of the five sources:
   1. API
   2. Website (i.e., web scrapping)
   3. Other DataSet (e.g., data.gov, Kaggle, Tableau has some on their site.)
   4. Programmatically create data with Python and data generators found on the internet.
   5. Twitter/Yelp – Create a hashtag with for your business, generate 20-100 reviews from random users and collect those reviews for analysis.

## (Optional) Create a data warehouse model for Redshift (But don’t’ create a Redshift Cluster Yet)

1. At least one Fact table and a Dimension table for each master data and a time dimension.

## Ingest data from your RDS database to S3

1. Create Kinesis Data Flows (Data Stream, Firehouse, or both) for each table in your RDS database into their own folder structure. (Remember, that the folder structure will determine the names of the resulting AWS Glue tables.)

## Create an AWS Glue Crawler for each table (i.e., S3 path)

1. Verify Data with AWS Athena.
   1. Get top 10 rows of each table and ensure each table is the way you need it to be. If it isn’t, go back to #8 and re-work it until it does.

## Use Lake Formation to create a Data Lake

1. In IAM create a lake admin user with full access, log out and back into your AWS account with the lake user and add other users with varying permission in Lake Formation.
2. Connect Lake Formation to your Transaction data in the Glue Catalog with an external table.
3. Use the query editor in Lake Formation to verify your data is how you intended.

## Create a Redshift Cluster and connect it to Lake Formation and the external Transactions Table

1. Create internal tables in Redshift.
   1. Use the Redshift SQL COPY command to move your master data from S3 into your internal Redshift table.

## Start a SageMaker Notebook Instance (i.e., Jupyter notebook)

1. Once the Notebook Instance is up, create a notebook that shows the description, and/or code and out put some results for each step of your project (including diagrams).

## Create a GitHub Account (if you don’t already have one) and…

1. Create a GitHub project repo with a ReadMe with should describe your project.
2. Use Git on your machine to commit (with comments), and push your code, docs, diagrams… to the remote repo.

## Present Your Notebook to the Class

1. Make sure to include your Name, Cohort and Date at the top of the notebook.