**Instructions: Task 1-5**

**Project Instructions**

To demonstrate the things you have learned in this course, you will perform the tasks outlined below. As you complete each task, you'll collect deliverables for your project submission.

**Project Checklist**

As you move through each task below, take a moment to check off the deliverables you'll need to collect from each one:





**Task 1: Create your Azure resources**

* Create an Azure Database for PostgreSQL.
* Create an Azure Synapse workspace. Note that if you've previously created a Synapse Workspace, you do not need to create a second one specifically for the project.
* Use the built-in serverless SQL pool and database within the Synapse workspace

In the cloud lab Azure environment, you will only be able to use the built-in serverless SQL Pool.

**Task 2: Design a star schema**

You are being provided a relational schema that describes the data as it exists in PostgreSQL. In addition, you have been given a set of business requirements related to the data warehouse. You are being asked to design a star schema using fact and dimension tables.

**Task 3: Create the data in PostgreSQL**

To prepare your environment for this project, you first must create the data in PostgreSQL. This will simulate the production environment where the data is being used in the OLTP system. This can be done using the Python script provided for you in [Github: ProjectDataToPostgres.py](https://github.com/udacity/Azure-Data-Warehouse-Project/tree/main/starter)

1. Download the script file and place it in a folder where you can run a Python script
2. [Download the data files](https://video.udacity-data.com/topher/2022/March/622a5fc6_azure-data-warehouse-projectdatafiles/azure-data-warehouse-projectdatafiles.zip) from the classroom resources
3. Open the script file in VS Code and add the host, username, and password information for your PostgreSQL database
4. Run the script and verify that all four data files are copied/uploaded into PostgreSQL

You can verify this data exists by using pgAdmin or a similar PostgreSQL data tool.

**Task 4: EXTRACT the data from PostgreSQL**

In your Azure Synapse workspace, you will use the ingest wizard to create a one-time pipeline that ingests the data from PostgreSQL into Azure Blob Storage. This will result in all four tables being represented as text files in Blob Storage, ready for loading into the data warehouse.

**Task 5: LOAD the data into external tables in the data warehouse**

Once in Blob storage, the files will be shown in the data lake node in the Synapse Workspace. From here, you can use the script-generating function to load the data from blob storage into external staging tables in the data warehouse you created using the serverless SQL Pool.

**Helpful Hints**

* When you use the ingest wizard, it uses the copy tool to EXTRACT into Blob storage. During this process, Azure Synapse automatically creates links for the data lake. When you start the SQL script wizard to LOAD data into external tables, start the wizard from the data lake node, not the blob storage node.
* When using the external table wizard, you may need to modify the script to put dates into a varchar field in staging rather than using the datetime data type. You can convert them during the transform step.
* When using the external table wizard, if you rename the columns in your script, it will help you when writing transform scripts. By default, they are named [C1], [C2], etc. which are not useful column names in staging.