**Read carefully before you begin**

* For question 1 and 2: you only need to answer them if you have not provided answers in your application screening questions.
* For questions 3 to 8:
  + Pick 4 questions that you are most comfortable with to answer. And state your assumptions if anything is not clear.
  + If you have used any large language models (e.g. ChatGPT) to assist with your answers, declare it with a note at the end of the answer, along with the LLM model and all the prompts you used to get to the final answer.

**Question 1:** Have you written any Python code in your personal GitHub repository that you can share with a link?

**Question 2:** How many years of commercial experience do you have in Snowflake and Python programming respectively?

**Question 3**

Design and create a logical data model using the star schema methodology to store the electricity outage dataset below in Snowflake. Define primary/foreign/surrogate keys for each table and the data types for each column. Note that the number of customers on a transformer may change over time, and the model and table structure must support preserving these changes during the ETL process.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Outage ID** | **Outage Suburb** | **Transformer Code** | **# of Customers on Transformer** | **Outage Start Time** | **Outage End Time** | **Outage Status** | **Outage Duration in Minutes** |
| 12345 | Ponsonby | KCN ME01 | 1,200 | 25/06/2024 8:00 | 25/06/2024 9:00 | Closed | 60 |
| 12346 | Albany | KNN CEP1 | 500 | 25/07/2024 8:30 | 25/07/2024 10:30 | Closed | 120 |
| 12347 | Remuera | REMU MK01 | 30 | 27/08/2024 8:30 | 27/08/2024 10:30 | Closed | 120 |
| 12347 | Remuera | REMU MK09 | 2,000 | 27/08/2024 8:30 | 27/08/2024 9:00 | Closed | 30 |
| 12347 | Remuera | REMU MU78 | 100 | 27/08/2024 8:30 | 27/08/2024 8:50 | Closed | 20 |
| 13349 | Ponsonby | KCN ME01 | 13 | 31/08/2024 20:00 |  | Open | 300 |
| 13350 | Takapuna | KNN CXP8 | 150 | 31/08/2024 22:10 |  | Open | 430 |

**Question 4**

On 1 Sep 2024, the daily ETL receives the following outage record from the Outage Management System (OMS):

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Outage ID** | **Outage Suburb** | **Transformer Code** | **# of Customers on Transformer** | **Outage Start Time** | **Outage End Time** | **Outage Status** | **Outage Duration in Minutes** |
| 14508 | Milford | KNN CXP9 | 1,000 | 01/09/2024 8:00 |  | Open | 60 |

However, the TRANS system—your source for transformer attributes (size, install date, etc.)—is offline and is not expected back until 6 Sep 2024. Describe an ETL strategy that:

* preserves referential integrity between the fact and transformer dimension tables,
* allows outage reporting to continue between 1 Sep and 6 Sep, and
* back-fills the correct transformer attributes once TRANS is online.

**Question 5**

Starting 1 Jul 2025, Microsoft Yammer (Viva Engage) will deprecate the OAuth2 based authentication, which means that only Azure Active Directory (AAD) token option is available. See below link for details of how Yammer AAD token authentication works.

<https://techcommunity.microsoft.com/blog/viva_engage_blog/yammer-api-with-aad-tokens-postman-collection/857923>

You are tasked with creating an automatic data pipeline that ingests messages daily from Yammer API v1. The data pipeline does not have an available web URL that you could use as callback URL when you pass parameters to get the AAD token. How would you design your data pipeline in Python or PowerShell to ingest data automatically without human intervention? Note code is not required but please explain:

* which OAuth2 grant type you would use and why
* the key values in header/body that you would pass to the authentication endpoint.
* how you would secure the credentials used in the authentication process.

**Question 6**

You are working on an AI use case to classify causes of electricity outages by using the combined information from field notes from electricity technicians and the photos they have taken on the fault sites. Snowflake’s multimodal capability (AI\_COMPLETE) only allows you to process up to 20 images at a time (see below the documentation), how would you design your AI solution in Snowflake if the field crews sometimes attach 30-40 photos per outage? Draw an architecture diagram (hand-drawn is acceptable) of the end-to-end solution. Provide enough details for a junior data & AI engineer to implement the solution.

<https://docs.snowflake.com/en/user-guide/snowflake-cortex/ai-images>

**Question 7**

Below is an example of first a few lines of an EIEP3 file, which is standard file format for half hourly electricity meter readings in NZ. Please write a SQL or Python script to load the file from SFTP into Snowflake without losing any useful information of the dataset.

*HDR,ICPHH,11,ENERGYCO,ENRG,DIST,09/06/2025,14:00:00,ID2025060914001,5,202505,E,I*

*DET,1234567890ABCD1,MTR123456789012,F,31/05/2025,1,20.35,5.12,,X,*

*DET,1234567890ABCD1,MTR123456789012,F,31/05/2025,2,19.87,4.98,,X,*

*DET,1234567890ABCD1,MTR123456789012,F,31/05/2025,3,21.10,5.34,,X,*

*DET,1234567890ABCD1,MTR123456789012,F,31/05/2025,4,22.45,5.68,,X,*

**Question 8**

You are working on an AI chatbot that can help the policy department to draft internal policies. The policy needs to quote KPIs from the company’s data warehouse, including total number of customers, total revenue, total spend by department, etc. The policy also needs to refer to the previous versions (there were about 20 of them and in PDF) published in the past 20 years. Draw an architecture diagram of the solution and describe the key points of how an AI engineer should build it if you are the technical lead of the project. Explain how you would measure the accuracy and quality of the chatbot response.