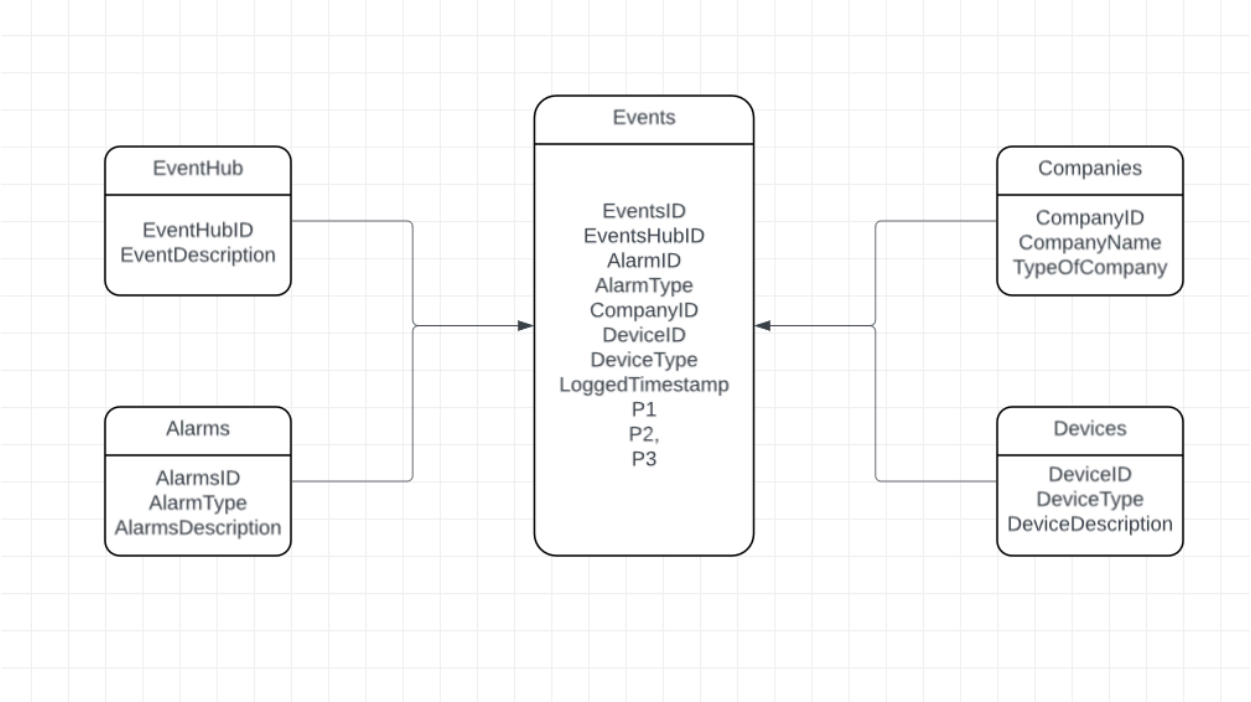
Given the sample data in the file DataModelWideTakeHome.csv, show what tables you would create in a new data warehouse off that source. Additionally, state what other data would be helpful to have and any assumptions made.

Based on the data provided, I would create 4 dimensions tables and 1 fact table in the warehouse in the staging layer. First would be an EventHub that contains the EventHub ID and description of event hubs. Secondly, I would create an Alarm table to store information about the alarmID, type of an alarm, and a description of the alarm. If applicable, we can also store if an alarm is used for vehicles, residential property, or commercial property. Similarly, the next table would be to store the company's information, such as ID, name, and type of company. In addition, I will create a table for Devices. I will store the deviceId, type of device and description of the device. Lastly, I plan to create an event fact table that will incorporate eventID, Logged data, and timestamp, eventhubID, company, deviceId, deviceType, and all the other columns included in the spreadsheet DataModelWideTakeHome.csv. I don't know what the Ps represent. I might need to create a new table to store all those information once I have more clarification. Below is the diagram that represents my model. I used Lucidchart to create a simple data flow.



Other information such as the location of the alarm, would be helpful. Creating a dimension table for dates with month, day, and year columns could be valuable during transformation. Metadata would be useful to gain more insight into data and make a decision while modeling.

Working with Data:

I used SQL to solve the data questions and uploaded files in Git Hub.

Based on the dataset, provide any other metrics that could be useful to the business.

- Based on the metrics, It could be useful to understand the unique numbers of customers that purchased each product.
- Understanding the customer location for each product could be useful for the marketing team to launch targeted campaigns.

Data Orchestration:

1. Working with ELT tools makes this process fairly easy. Previously, I have worked to extract and load CSV files from the S3 bucket to Snowflake. Therefore, I would like to share my thought process based on that experience. From my past experience, we can use tools such as Airbyte which has a connecter to set source and destination to extract and load the data from Azure Blob Storage to Snowflake. Steps to follow in Airbyte are:

- Setup new source in Airbyte (which is Azure Blob Storage) in this case
- Enter credentials to access the directory that has the CSV file.
- Once the source is verified, we can select the columns we want to extract in the Airbyte UI.
- Airbyte also makes it easy to schedule based on our preferences.
- Next, We set up our destination as Snowflake. Following a similar process of entering credentials and giving access to the warehouse.
- Specify the schema or destination where we want the files to be loaded.
- In addition, we can also choose the option for the load to be incremental or full refresh if we have appropriate timestamp columns available for example - createdOn or updateOn

2. If end users are asking for delivery of insights, we can use a few different options to make the data available to them. First of all, we need to transform and add quality checks to the data to make sure we didn't lose any during the transition. Once the data has been verified we can export those data into a different warehouse let's say Analytics Warehouse which can be used as the source of truth for Analysis. From my experience, I

have dealt with a variety of end users and they all have different requirements. Therefore I will lay out a few options below based on my experience.

a. First and foremost we need to understand the requirements of the end user. What do they want?
b. I have seen some stakeholders write a simple ad-hoc query to obtain the result they want. This is pretty rare but if that's the preference. This is the situation where end users might want to pull the data whenever they want.
c. Another option is to create a visualization board for end users to utilize. The data will be updated based on the scheduled run and they can go and get insights on topics they are interested in.