Data Engineering exercise

A company needs to pull vehicles’ data and sensor measurements from two respectively external REST JSON APIs once a day and perform analytical queries on this data.
Calling any of the APIs takes up to 5 minutes and brings several GB of daily data.

Sensor data from the sensors API should be loaded into the **sensor_data** table that has the following structure:

- timestamp - timestamp when a sensor value was measured (Unix epoch)
- vehicle_id - unique id of a vehicle
- sensor_value - sensor value at this timestamp (can be null)

The vehicles data from the vehicles API should be loaded into **vehicles_data** table that has the following structure:

- vehicle_id - unique vehicle id
- vehicle_name - name of the vehicle
- vehicle_type - type of the vehicle (can be truck, bus or car)

Please write answers for the following questions/requests:

1. Propose a reliable and cost effective architecture for storing and analyzing the described data. Explain what tools will you use (storage, database, compute resources etc.), in what format will you store the data, how will you trigger the loading.

2. How can you ensure that the analytical query runs only after both vehicles and sensor data are loaded into the tables?

3. The company management wants to know what was the average time (in seconds) between subsequent sensor measurements for vehicles of the same vehicle_type.

Write an SQL query that returns vehicle_type and average time between measurements (measurements with null values should be filtered out).

Good luck 🙂