03-11-2017

DWH Candidate Offsite Challenge

Introduction:

This challenge tests your understanding of DWH modeling using Data Vault 2.0 methodology and your ability to write PostgreSQL procedural language and Python scrips to load the DWH.

The sample data in the orders Excel file is your source data.

Exercise 1 – Theory

a. Name and shortly describe the three **core** entities of the raw layer of Data Vault 2.0.
b. Name and shortly describe two advance entities of Data Vault 2.0 that can be used to enhance performance.

Exercise 2 – Design DWH using Data Vault

Our stakeholders want to have an easy reporting model to build quick ad-hoc reports. They want to track Order information for Geo analysis. Additional Geo information should be added to the DWH from an external API (see Exercise 3 c for more details). Use the additional excel file as your source data from operation DB to answer the following:

a. Design a **physical ERD** for the raw layer based on Data Vault 2.0 methodology.
b. Design a Star-Schema Data Mart ERD using your entities from the previous steps.

Exercise 3: Implementation and coding

a. Create PostgreSQL functions to load **two different types of core entities** of Data Vault 2.0 based on your design in Exercise 2a.
b. Create PostgreSQL function that accepts one text parameter and returns Boolean flag if the passed value is numeric or not.
c. Write a Python script that adds additional Geo information to the zip codes in the sample Order file. The stakeholders are interested in the following additional attributes:
    - Latitude
    - Longitude
    - State
    - City

    Note: please use the following site to get the location information http://www.zipcodeapi.com/API#zipToLoc. They have free use API for 50 zip codes per hour.