# Technical Skill Assessment

**General Goal**

The goal of this skill assessment is to evaluate your ability some key skills that are relevant to SQL Developer/Analyst.

**Specific Skills Assessed**

* Problem Solving / ETL Design
* Dimensional Data Modeling
* Ability to clearly organize, explain, and communicate design.

**Guidelines**

* You are encouraged to ask clarifying questions (email:alex.savchenko@mexia.aero) at anytime. We will try to answer questions in a timely fashion (same day or better).
* You may use any modeling or diagramming tool you wish to prepare your design. Using a sophisticated modeling or diagramming tool is not required.
* You may present your solutions in any manner you wish (on paper, presentation, whiteboard).
* Solutions should be expressed in a platform independent manner (i.e., we should be able to adapt and implement the design in any environment).
* You may use any assistance/material or research in your development of the solutions; however, you will need to be able to explain the rationale behind your design decisions and solutions.

# Part 1: ETL Design

Develop an ETL design (process/logic/steps) to handle the case described below.

You can express the design in whichever means you find most suitable. The following formats are suggested:

* Simple Structured Flowchart
* UML Activity Diagram
* Text Narrative (step-by-step description)
* Generic/Self-invented diagram that illustrates the design.

You need to design only the ETL process/logic. You do not need to define/design any intermediate data storage or staging/temporary tables.

## Assumptions and Constraints

* The ETL process is operating in a relational database environment.
* Assume that this is a batch process that runs every day, during overnight hours while no users are using the final target table(s).
* Your design should account for the fact that occasionally a few days are missed due to source system outages.
* Extracting only new/changed data from the CRM system can be done at extraction time with adequate performance and no negative consequences on the source system.
* You can assume that the ETL solution/environment is able to keep track of the “last run/last extract date” which can be used as needed in your design logic.
* Data volume to be processed is very high and attention to optimal design is a key factor in performance.

## The Target Table: A Large Customer Dimension (Relational Table)

The target of the ETL process is a relational Customer dimension with type II history tracking and several million rows.

Type II changes are triggered on any of the following changes:

1. A change in the FSA - first three digits of the postal code (which may also cause a new Stat Canada area to be assigned).
2. A change in age group (a 10-year age band is used).

**Target Customer Dimension**

|  |  |
| --- | --- |
| ***Column*** | ***Description/Notes*** |
| Customer Surrogate Key | The assigned surrogate key for use in the data warehouse environment. |
| Customer Number | The natural key (primary key in the source system) |
| First Name | Simple identifying information (type I) This information should be maintained in sync with the source system for the current record in the dimension if changes are detected. |
| Last Name |
| Date of Birth |
| Street Address |
| City |
| Prov |
| Postal Code |
| Age | The current age of the customer. This is rounded to the year and maintained accurately for the customer’s current record. |
| Current Age Group | A 10 year grouping based on the customer’s age as of the current date (0-9, 10-19, 20 – 29, 30 – 39..etc). |
| Forward Sortation Area (FSA) | A geographic boundary based on the first three digits of postal codes in Canada. |
| Stat Canada Area | A statistical boundary (geography) to which the customer belongs based on their postal code. |
| Stat Canada Survey Year | Statistics for the area to which the customer belongs. The most recent year’s survey available is used. |
| Stat Canada Avg Household Income |
| Stat Canada Avg Education Level |
| Stat Canada Avg Household Size (# people) |
| Dimension From Date/Time | Columns to facilitate type II changes in the dimension. |
| Dimension Thru Date/Time |
| Current Record Flag |

## Source Systems / Data

**Customer Table in the CRM System**

The customer table in the CRM system contains the following information.

|  |  |
| --- | --- |
| Customer Number | Primary key |
| First Name | Simple identifying information |
| Last Name |
| Date of Birth |
| Street Address | Common/generic address information. |
| City |
| Prov |
| Postal Code |
| Insert Date/Time | Date/Time the record was created |
| Last Update Date/Time | Date/Time that a change was last applied to a record. When a record is initially created, this is the same value as the insert date/time. |

**Stats Canada Data**

This data is purchased once every 5 years and arrives in flat file format. The raw data is loaded into a staging area within the data warehouse environment. The data consists of two data sets as follows:

**Data Set 1: Stat Canada Survey Results**

Contains survey results for the past 5 years (i.e., each area may have up to 5 records dating back 5 years).

|  |  |
| --- | --- |
| Stat Canada Area ID | Unique identifier for a geographic boundary. |
| Stat Canada Area | A statistical boundary (geography) to which the customer belongs based on their postal code. |
| Stat Canada Survey Year | The year for which the survey applies. |
| Stat Canada Avg Household Income | Statistics for the area to which the customer belongs. |
| Stat Canada Avg Education Level |
| Stat Canada Avg Household Size (# people) |

Sample Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Stat Canada Area ID | Stat Canada Area | Stat Canada Survey Year | Stat Canada Avg Household Income | Stat Canada Avg Education Level | Stat Canada Avg Household Size (# people) |
| 100 | Wpg South | 2008 | 78,500 | High School | 3.2 |
| 100 | Wpg South | 2009 | 79,000 | High School | 3.1 |
| 100 | Wpg South | 2010 | 80,000 | College | 3.0 |

**Data Set 2: “Mapping” Table Connecting FSA to the Stat Canada Area**

Note: FSA is the first three digits of the Canadian postal code.

|  |  |
| --- | --- |
| Stat Canada Area ID | Defines the Stat Canada boundary |
| FSA | Identifies the FSA included in the boundary. |
| Stat Canada Survey Year | Identifies the survey year for which the FSA/boundary relationship applies to. |

**Sample Data**

|  |  |  |
| --- | --- | --- |
| Stat Canada Area ID | FSA | Stat Canada Survey Year |
| 100 | R2N | 2009 |
| 100 | R2J | 2009 |
| 100 | R2K | 2009 |
| 101 | R2M | 2009 |
| 101 | R3C | 2009 |
| 100 | R2N | 2010 |
| 100 | R2J | 2010 |
| 133 | R2K | 2010 |
| Etc…. |  |  |

# Processing Requirements

|  |  |
| --- | --- |
|  | 1. The “type I” attributes (e.g., name, address) for the current record in the dimension must be updated to stay synchronized with changes in the source CRM system. |
|  | 1. A new surrogate key must be assigned whenever a new row is added to the dimension (either due to a net new customer or a type II change). |
|  | 1. Changes in the FSA and/or age group should trigger a type II change in the dimension. If the FSA and age group were to change simultaneously (on the same overnight run), only one type II change should be triggered. |
|  | 1. The source system does not change its data when someone experiences a birthdate, so we need to consider the updating of age and age group elements of existing records in our dimension as part of our design. |
|  | 1. We also need to design an independent process for Stats Can data updates (separate from the daily run) that we can run on demand once every 5 years. This process will loads the raw Stats Canada data from flat files and updates attributes within the dimension for the all of the current customer records. |

 = key requirement

## What we are looking for in the design

* + Well-organized design with a logical, high-level breakdown of work into phases, steps, or stages that can be understood in a top-down fashion (more details at lower levels of decomposition).
  + A design that allows for some restart-ability by persisting interim results at logical/key points in the process will be considered superior.
  + Attention to performance optimization in the logical design that reduces data volume / rows which are processed/updated will be considered superior.

## Samples Illustrating the Resulting Data in Different Situations

An illustration of the expected results in the target appears below (a subset of the attributes are shown for simplicity).

A customer has moved twice over the past 6 months

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Customer Surrogate Key | Customer Number | Last Name | City | Prov | Postal Code | Forward Sortation Area (FSA) | Stat Canada Area | Dimension From Date/Time | Dimension Thru Date/Time | Current Record Flag |
| 2382 | 974 | Doe | Winnipeg | MB | R2N1M6 | R2N | 100 | 2010/01/20 | 2014/01/15 |  |
| 2479 | 974 | Doe | Winnipeg | MB | R3C1X7 | R3C | 101 | 2014/01/16 |  | Y |

A customer recently had a birthday that caused him/her to move to another age group. The source system did not indicate a change which triggered this processing (i.e., there was no change in the source record for the customer, they simply aged relative to the current date).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Customer Surrogate Key | Customer Number | Last Name | Date of Birth | Age | Age Group | Dimension From Date/Time | Dimension Thru Date/Time | Current Record Flag |
| 6475 | 3874 | Smith | 1984/03/12 | 29 | 20-29 | 2008/03/11 | 2014/03/11 |  |
| 8535 | 3874 | Smith | 1984/03/12 | 30 | 30-39 | 2014/03/12 |  | Y |

A customer coincidentally moved on the same date as their 30th birthday. Notice only one additional record version was added.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Customer Surrogate Key | Customer Number | Last Name | City | Prov | Postal Code | Date of Birth | Age | Age Group | FSA | Dimension From Date/Time | Dimension Thru Date/Time | Current Record Flag |
| 54564 | 200 | Salotia | Winnipeg | MB | R5J7L8 | 1984/03/12 | 29 | 20-29 | R5J | 2008/03/11 | 2014/03/11 |  |
| 79865 | 200 | Salotia | Brandon | MB | R7A7R2 | 1984/03/12 | 30 | 30-39 | R7A | 2014/03/12 |  | Y |

A customer had a correction in their current name and street address. Only their current record is kept in sync with the source system.

(The customer’s last name was corrected from Smyth to Smith and their street address was corrected from 668 Waverley to 678 Waverley)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Customer Surrogate Key | Customer Number | Last Name | Street | Age | Age Group | Dimension From Date/Time | Dimension Thru Date/Time | Current Record Flag |
| 2382 | 974 | Smyth | 668 Waverley | 39 | 30-39 | 2010/01/20 | 2014/01/15 |  |
| 2479 | 974 | Smith | 678 Waverley | 41 | 40-49 | 2014/01/16 |  | Y |