Universidad Simón Bolívar
Departamento de Ciencias de la Computación
CI3391 – Taller de Bases de Datos

Proyecto 2: (Valor: 20%)

Las labores de QA (Quality Assessment) en un proyecto son muy variadas y una de ellas es generar datos *sintéticos* para pruebas que se asemejen a los datos originales pero que no revelen información personal de los usuarios. Estos datos deben:

- Comportarse estadísticamente de una manera similar a los originales. Por ejemplo, los nombres y/o apellidos no suelen distribuirse uniformemente (En países hispanoparlantes Juan o José son nombres más comunes que John o Jake)
- Ser coherentes y respetar las restricciones de integridad impuestas por el sistema
- Cuando se generan elementos como órdenes de compra, la varianza entre ellas debe ser similar a la que se observa en la vida real (es decir, algunos clientes hacen unas pocas compras al año y otros hacen varias compras a la semana). El sistema de generación debe poder indicar cuales son los parámetros estadísticos en uso.
- Los datos generados deben tener sentido aunque no existan en la vida real. Es decir, si se utilizan en una aplicación donde alguno de los dominios es restringido, deben pertenecer al mismo dominio. Por ejemplo:
    - Si una columna es “País”, los valores deben ser nombres de países.
    - En USA muchas direcciones son de la forma: <Número> <Nombre de Calle>. Es decir, aunque algo como “123 Main St” puede no existir para una ciudad aún se consideraría valida.
    - Las fechas deberían tener sentido (bajo cualquier criterio lógico)

Dado el modelo de entrega de víveres definido en el sitio web de Portobello (A Grocery Delivery Data Model), que fue estudiado al principio del curso.

1. Implementar el modelo físico (en postgresql).

2. Definir un procedimiento almacenado que genere datos sintéticos:

spCreateTestData

number_of_customers	Número de clientes a crear
number_of_orders	Número de órdenes a crear
number_of_items	Número de Items (productos) a crear
avg_items_per_order	Cantidad **promedio** de productos por orden (promedio significa que este número puede variar de una orden a otra)

3. Escribir queries que respondan estas preguntas:
    a. ¿Cuáles son los clientes que viven en las ciudades donde se compran el TOP 5% de las órdenes más costosas?
    b. ¿Cuáles son las 5 ciudades donde las órdenes tardan en promedio más tiempo en ser despachadas?
    c. Cuales son los 10 mejores clientes (en monto total de compras)
    d. ¿Cuál es el ítem que causa más retrasos en el despacho de órdenes?

Reglas:

- El/los archivos a entregar deberán ser enviados por email a jhovanny.villegas@gmail.com a más tardar el día 15/Marzo@23:59
- Los datos generados deben mostrar *variedad*.
- El código para generar los datos debe respetar principios básicos de modularidad/reutilización (DRY: Don't repeat yourself)
- La entrega debe contener:
    - Script (bash) que:
        - Crea la base de datos con nombre único
        - Carga las tablas auxiliares requeridas por el generador de datos
        - Crea el stored procedure y cualquier view/función/etc.
    - Queries (SELECT) respuesta a cada pregunta.
        - No se aceptarán queries en sintaxis previa a ANSI-99 (es decir, que no utilicen la sintaxis de JOIN)