Laboratorio #11
Datawarehouse y OLAP - ***PAREJAS***

I. Modalidad y fecha de entrega

a) Debe ser enviado antes de la fecha límite de entrega: martes 15 de noviembre a las 23:59
b) Luego de la fecha límite se restarán 10 puntos por cada hora de atraso en la entrega

II. Descripción de la actividad

Ejercicio 1

ETL y Datawarehousing

El objetivo de este ejercicio es levantar un pequeño *datawarehouse* en una base de datos local de PostgreSQL. Este *datawarehouse* será alimentado a partir de un proceso ETL (*extract, transform, load*) desarrollado en KNIME Analytics Platform, que tomará datos de la base de datos de su proyecto y de una base de datos local.

La base de datos centralizada contendrá únicamente una *fact table* con las reproducciones de contenido unificadas, y una *dimension table* de fechas para análisis.

A continuación los pasos a seguir:

(a) Levante una base de datos en su PostgreSQL local de nombre **lab11** y reestablezca allí el backup **test-db.sql** de Canvas, que contiene una estructura de base de datos de ejemplo basada en este esquema[1].

(b) Cree una base de datos **datawarehouse** y cree una tabla la siguiente tabla

```
reproducciones(fecha DATE, hora TIME, nombre VARCHAR(250))
```

(c) Descargue e instale KNIME Analytics Platform (486MB).

(d) Configure un proceso de extracción de datos en Knime que:

I. Consulte las reproducciones ocurridas de la base de datos de su Proyecto alojada en su PostgreSQL y extraiga de allí:
- Fecha de la reproducción (casteado a DATE)
- Hora de la reproducción (casteado a TIME)
- Nombre del contenido reproducido

II. Consulte las películas rentadas de la base de datos **lab13** mediante el siguiente query:

```
SELECT rental_date::date, rental_date::time, film.title
FROM rental
    LEFT JOIN inventory ON rental.inventory_id = inventory.inventory_id
    LEFT JOIN film ON inventory.film_id = film.film_id
;
```

III. Centralice esos datos en la tabla reproducciones de su base de datos **datawarehouse**

Muestre que la tabla creada en la base de datos **datawarehouse** contiene la suma de las reproducciones de las dos bases de datos originales.

[1] https://www.jooq.org/sakila

Ejercicio 2

Conceptos de OLAP

Utilizando la base de datos **datatawarehouse** desarrollada en el ejercicio anterior, este ejercicio tiene por objetivo desarrollar un cubo OLAP por medio de una vista materializada en PostgrSQL que permita analizar el total de reproducciones ocurridas por año, trimestre, mes, semana del año, y nombre del contenido.

Para ampliar estos conceptos puede referirse a las secciones **10.6 On-Line Analytic Processing** y **10.7 Data cubes** de Ullman, 2da edición.

A continuación, desarrolle los siguientes enunciados:

(a) En la base de datos **datawarehouse** cree y alimente una *dimension table* de fechas, que contenga la información necesaria para permitir analizar reproducciones por año, trimestre, mes y semana del año.

Se recomienda estudiar la propuesta de Nicholas Duffy documentada en la sección de **Recursos y bibliografía**.

(b) Desarrolle una consulta que unifique las reproducciones unificadas almacenadas en el **Ejercicio 1** con la tabla desarrollada en el inciso (a).

(c) Desarrolle una vista materializada que permita analizar las reproducciones por año, trimestre, mes, semana del año, género y artista. Esta vista materializada debe computar de antemano los totales de reproducciones de acuerdo con todos los posibles subconjuntos de dimensiones consideradas (utilizando el operador CUBE de Postgres).

(d) Utilizando la vista materializada responda a las siguientes preguntas:

i) ¿Cuál es el contenido más reproducido en la base de datos?
ii) ¿Cómo han evolucionado las reproducciones semanales durante el 2012?
iii) ¿Cuál ha sido nuestro mejor trimestre de reproducciones?

III. Temas a reforzar

- Consultas OLAP
- Conceptos de datawarehouse

IV. Recursos y bibliografía

1. Creating a date dimension table in PostgreSQL, Nicholas Duffy (2016): https://medium.com/@duffn/creating-a-date-dimension-table-in-postgresql-af3f8e2941ac
2. KINME Analytics Platform, https://www.knime.com/downloads/download-knime

V. Documentos a entregar

1. Backup de la base de datos PostgreSQL con el nombre **datawarehouse.bak**
2. Screencast mostrando el proceso de ETL ejecutado por KNIME, que almacena los datos en la base de datos **datawarehouse**
3. Documento PDF con los *queries* y respuestas del ejercicio 2 (d).

VI. Evaluación

- Parte 1. 50 puntos
- Parte 2. 50 puntos.

- **Total: 100 puntos**