Proyecto del curso

Profesores: Claudio Gutiérrez
Matías Toro
Auxiliares: Luis Bustos
Raúl Silva
Florencia Yáñez

El objetivo del proyecto del curso es que tengan una experiencia real con el manejo de bases de datos. Para ello deberán entregar un artefacto que consiste en una aplicación web o un estudio de datos (estadístico). Este artefacto debe utilizar a lo menos tres consultas, demostrando una mezcla de operadores de SQL (joins, consultas anidadas, agregación, etc). La base de datos que se utilizará para el proyecto será PostgreSql.

El proyecto del curso se divide en 4 hitos incrementales, donde la única entrega obligatoria es la última que corresponde a un informe final del proyecto + video de 5 minutos de presentación. Los otros 3 hitos son opcionales pero influyen positivamente en la nota final (puntos extras). En cada entrega de cada hito opcional se espera que entreguen en el informe un acumulado de todos los hitos anteriores incluyendo todo el feedback proporcionado.

Los 4 hitos son los siguientes:

Hito n°0: Grupos + fuente de datos + problema.
Fecha: 30 de Septiembre.
Acá deben entregar lo siguiente:

- Definición del grupo. Deben indicar el nombre de los integrantes de su grupo (los grupos son de 3 personas).
- Fuente de datos.
  - Deberán escoger un set de datos relativamente interesante: por lo menos 5 tablas y que existan relaciones entre ellas (no 5 tablas independientes!). Se espera que los datos se puedan cruzar entre ellos para resolver "el problema".
  - Los datos deben tener una escala razonable, es decir, al menos 10000 tuplas en total (sumando el número de tuplas de cada tabla) pero preferiblemente más.
  - La siguiente lista da algunos ejemplos de fuentes de datos:
    * `http://datos.mineduc.cl/dashboards/19731/bases-de-datos-directorio-de-establecimientos-educacionales/`
    * `https://datos.gob.cl/`
    * `https://www.kaggle.com/datasets`
    * `https://datahub.io/`
    * `https://datasetsearch.research.google.com/`

    * `https://github.com/caesar0301/awesome-public-datasets`

    No hay que usar todos los datos disponibles en un conjunto de datos; se puede usar una muestra interesante. También se pueden seleccionar datos de otra fuente no listada aquí.
  - El proyecto será más fácil si los datos ya están en un formato de tablas (como, p.ej., CSV, TSV, etc.)
  - Para los datos seleccionados deben describirlos, e indicar su origen.

- El problema.
  - Deben indicar la motivación que tienen de usar estos datos y que problema quieren resolver.
  - Esto puede ser una idea vaga, no es necesario describir precisamente en ese momento que van a hacer.
  - Ejemplos de problemas son resolver ciertas consultas, y/o mejorar algún diseño existente.

Hito nº1: Modelo Entidad Relación
Fecha: 9 de Octubre.
Acá se espera que entreguen lo siguiente:

- Modelo de datos entidad relación
- Traducción del modelo entidad relación al modelo relacional
- Similar al primer laboratorio

Hito nº2:Implementación relacional
Fecha: 6 de Noviembre.
Acá se espera que implementen el schema definido en el hito nº2:

- Para ello se les proveerá acceso a servidores virtuales donde tendrán acceso para crear su base de datos (y aplicación si correspondiese). Esta información se entregará más adelante.
- Deberán reportar en un informe el SQL de creación de las tablas y sus restricciones.
- Además se les aconseja insertar tuplas de prueba para probar restricciones y dominios.

Hito nº3:Carga de datos
Fecha: 2 de Diciembre.
Acá se espera que carguen los datos importandoles ya sea de:

- CSV, TSV,
- de papel (cargados manualmente), o
- generados sistemáticamente (generación de datos aleatorios)

Hito nº4: Entrega final.
Fecha: fecha de examen por anunciar.
Este es el único hito obligatorio y corresponde a la suma de todos los hitos anteriores más lo siguiente:

- Definición de consultas para optimizar:
  - Ilustrar consultas que necesiten índices y otras que no
  - Deben justificar la elección de los índices.
- Entrega informe final. Para estudio de datos, se deben reportar las consultas sql utilizadas.
- Video de presentación de 5 minutos. Hay que presentar:
  - el tema y la fuente de los datos
  - algunas estadísticas de los datos usados: tamaño, formato
  - el esquema relacional
  - los índices y las vistas ocupados
  - las consultas elegidas
  - una demostración de su artefacto
  - las lecciones aprendidas (qué fue difícil, fácil, interesante, etc.)