Práctico 3

Bases de datos

1. Introducción

Debido a que las nuevas tecnologías aportan cantidades significativas de datos, se hace imperiosa la necesidad de almacenar los mismos de manera que su acceso sea eficiente y útil. En general, la información de grandes relevamientos, simulaciones numéricas, etc. está organizada en *Bases de Datos*. El paradigma más utilizado para organizar bases de datos es el de Base de Datos Relacional. En este tipo de modelo la información se organiza en tablas, que se relacionan entre sí a partir de una propiedad de los datos. Cada tabla es un conjunto de registros. Existen programas o sistemas de gestión de bases de datos relacionales. Entre los más conocidos, se destacan por ejemplo MySQL, PostgreSQL, Oracle y Microsoft SQL Server. Cabe mencionar que recientemente se desarrolló el Software SciDB, que está preparado para manipular información científica (de hecho fue desarrollado para el LSST). Definiciones:

Base de datos Conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su registros.

Tabla Es un conjunto de datos con ciertas características en común.

Registro Es un objeto único de datos implícitamente estructurados en una tabla. Corresponde a una fila en las tablas.

Campo Es la mínima unidad de información a la que se puede acceder. Corresponde a una columna de una tabla.

Relación Vínculo entre los campos de distintas tablas. La información está organizada en tablas, pero puede reunir usando vínculos.

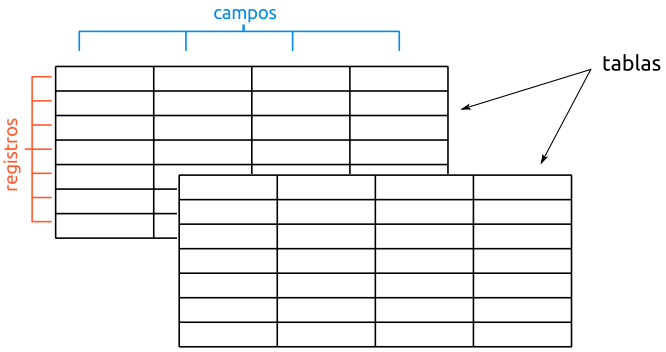

Fig. 1 *Elementos de una Base de datos relacional*

Material de lectura:

- Chilingarian et al., 2004, Astronomical Data Analysis Software and Systems (ADASS) XIII, Astronomical Society of the Pacific Conference Series.
- Bell, Hay & Szalay, 2009, Science, 323, 1297

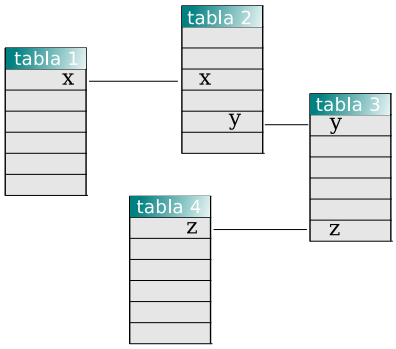

Fig. 2 *Estructura básica de una base de datos relacional*

▷ **1.** Elabore ejemplos de datos en donde es posible organizar los datos utilizando una base de datos relacional.

Elija un ejemplo y dibuje un posible esquema de la base de datos, enumerando las posibles tablas, propiedades y relaciones.

2. SQL

Uno de los lenguajes más utilizados para manipular y acceder a bases de datos es el denominado SQL (Structured Query Languaje). Con este lenguaje se pueden crear, modificar y consultar bases de datos. SQL se divide en dos partes, un lenguaje de manipulación de datos y un lenguaje de definición de datos. Para extraer informaciń de una base de datos se requiere la parte de manipulaciń de datos, cuyas instrucciones principales son:

```
SELECT, FROM, INSERT INTO, WHERE
```

Un query básico de SQL tiene la siguiente forma:

```
SELECT nombre(s)_de_columna(s)
FROM nombre_de_la_tabla
WHERE nombre_columna operador valor
```

donde "operador" puede ser AND, OR, >, <, ==, !=.

En lo que sigue se poponen dos casos de estudio para trabajar con bases de datos en observatorios virtuales: uno usando un catálogo de galaxias de la base de datos SDSS y otro usando un catálogo de exoplanetas de la base de datos exoplanets.eu

Elija uno de los dos y resuelva los ejercicios.

3. Caso de estudio: catálogo de galaxias y base de datos SDSS CasJobs

- Ingrese al sitio de CasJobs: `http://skyservice.pha.jhu.edu/casjobs/default.aspx`
- Cree una cuenta de usuario (requiere cuenta de correo electrónico)
- Identificar los elementos principales de la base de datos: tablas, campos y registros.
- Explorar la estructura de la base de datos usando *Skyserver* → *Schema Browser*
- En la pestaña *Query*, ingresar a "Sample SQL queries", elegir algunos ejemplos y ejecutarlos.
- Comprender el uso de las instrucciones "select", "from", "into", "where" y "join".
- Identificar el uso de "alias" y escribir un *query* que los utilice.

Obtener una lista de galaxias con las siguientes propiedades:

- clasificación como elíptica/espiral
- magnitudes Petrosian en las bandas u, g, y r
- redshift

► **2.** Estudie la distribución de índices de color g-r y u-g para galaxias elípticas y espirales. Determinar si para ambos casos las distribuciones con consistentes. Discuta la validez de la forma de la distribución propuesta para el estadístico de la prueba.

► **3.** Estudie la distribución de tipos morfológicos y determine si la misma es consistente con una distribución uniforme.

► **4.** Grafique las magnitudes aparentes de galaxias en la banda r en función de las magnitudes en la banda g, y obtenga un ajuste para la relación entre ambas.

► **5.** Calcule la magnitud absoluta para cada galaxia, usando la aproximación:

$$M = m - 25 - 5 \cdot log_{10}\left(\frac{c \cdot z}{H}\right)$$

donde c es la velocidad de la luz y $H = 75\ km\,s^{-1}/Mpc$. Grafique la magnitud absoluta vs. el *redshift* para todas las galaxias con $m_r < 17.5$, y obtenga un ajuste para la envolvente de los puntos.

Discuta en el informe el origen de la forma de los puntos y el procedimiento para el ajuste del modelo.

4. Caso de estudio: exoplanetas y base de datos exoplanet.eu

- Ingrese al sitio de exoplanets.eu: `http://exoplanet.eu/`
- No es necesario crear una cuenta de usuario
- Identificar los elementos principales de la base de datos: tablas, campos y registros.
- Instalar el paquete `pyvo`
- Leer las instrucciones para el uso de la API para python en `http://exoplanet.eu/API/`
- Comprender el uso de las instrucciones "select", "from", "into", "where" y "join".
- Correr un query y analyzar la estructura de los datos que devuelve.

Obtener una lista de exoplanetas con las siguientes propiedades:

- masa
- periodo orbital
- tipo de estrella del sistema

► **6.** Estudie la distribución de índices de masas y periodos orbitales para planetas descubiertos con diferentes técnicas observacionales. Determine si existe un sesgo en las propiedades de los exoplanetas en función del tipo de técnica utilizada.

► **7.** Estudie la distribución de distancias a las estrellas con exoplanetas y determine si la misma es consistente con una distribución gaussiana.

► **8.** Grafique un gráfico de las masas y los radios de los planetas, proponga un modelo y realice un ajuste de ese modelo. Discuta el procedimiento para el ajuste del modelo.

5. Ejercicios adicionales

▷ **9.** Deduzca las fórmulas para los coeficientes del ajuste lineal a un conjunto de puntos (x, y). Discuta las diferencias de asumir errores en los valores x o y.

▷ **10.** Sean θ_1 y θ_2 dos estimadores insesgados para el parámetro θ, y sea α una constante. Demuestre que $\theta = \alpha\theta_1 + (1 - \alpha)\theta_2$ es también un estimador insesgado para θ.

▷ **11.** Dada una muestra aleatoria de tamaño n de una población Poisson con parámetro $\lambda > 0$, use el método de máxima verosimilitud para encontrar un estimador del parámetro λ.

▷ **12.** Se sabe que la vida en horas de un foco de 100 watts de cierta marca tiene una distribución aproximadamente normal con desviación estándar de 30 horas. Para una muestra al azar de 50 focos y resultó que la vida media fue de 1550 horas. Construya un intervalo de confianza del 95% para el verdadero promedio de vida de estos focos.

▷ **13.** Las mediciones del número de cigarros fumados al día por un grupo de diez fumadores es el siguiente: 5, 10, 3, 4, 5, 8, 20, 4, 1, 10. Realice la prueba de hipótesis $H_0 : \mu = 10$ vs $H_1 : \mu < 10$, suponiendo que los datos provienen de una muestra tomada al azar de una población normal con $\sigma = 1.2$. Use un nivel de significancia del 95%.

Deberá utilizar un repositorio de git y colocar allí los códigos utilizados para resilver los ejercicios marcados con "►". En el repositorio deberá incluir la resolución de los ejercicios, junto con los códigos empleados y los gráficos correspondientes, acompañados de documentación en html que incluya una breve introducción y conclusiones.

Fecha límite de entrega de informes: 28 de octubre, a las 23:59 horas, indicando el link del repositorio por medio del Aula Virtual.