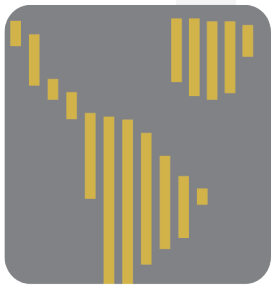

PR0MiDAT

IBEROAMERICANO

Programa Iberoamericano de Formación en Minería de Datos

Manipulación y Preparación de Datos

(506) 2268.8823 - (506) 8708.9091

info@promidat.com

facebook.com/oldemarrodriguez

www.promidat.com

Tutor: El curso será impartido por M.Sc. Fabio Fernández Senior Manager Risk Reporting en Scotiabank, Toronto, Canadá. Es Máster en Matemática Aplicada de la Universidad de Costa Rica y tiene un Bachillerato como Ingeniero en Computación del Instituto Tecnológico de Costa Rica. Además fue Analista de Modelación Matemática, Banco Nacional de Costa Rica y profesor de la Escuela de Matemática de la Universidad de Costa Rica.

Duración: Cuatro semanas.

Descripción:

En este curso se presentarán los fundamentos del lenguaje SQL. El énfasis principal del curso será examinar diversos componentes del lenguaje, como lo son declaraciones, expresiones, tipos de datos, entre otros. Se le dará especial importancia al uso del lenguaje como herramienta de análisis exploratorio de datos, como punto de partida para el desarrollo de aplicaciones de minería de datos. Para esto se utilizarán diversos motores de bases de datos como *SQL Server* (utilizando T-SQL) y MySQL, así como paquetes en R para manipulación de datos.

Objetivos:

En este curso el estudiante será capaz de:

1. Entender la estructura básica del lenguaje SQL como herramienta para consultar bases de datos.
2. Utilizar el lenguaje como mecanismo de extracción de datos e información a partir de repositorios con grandes volúmenes de datos.
3. Hacer uso correcto del lenguaje para construir consultas complejas que permitan obtener información de distintas tablas simultáneamente.
4. Entender el lenguaje SQL desde el punto de vista de teoría de conjuntos y lógica de predicados, permitiendo realizar operaciones usuales como lo son uniones, intersecciones, diferencias, entre otros.
5. Realizar análisis descriptivos de por medio de cláusulas, expresiones y funciones del lenguaje.
6. Utilizar SQL Server y MySQL como motores de bases de datos basados en SQL.

7. Explotar las ventajas del lenguaje desde plataformas conocidas como R y MS Excel.

Metodología:

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas.

- Una vídeo conferencia semanal, las cuales quedarán grabadas en Webex, para que los alumnos la puedan acceder en cualquier momento.
- Trabajos prácticos semanales.
- Foros para plantear dudas al tutor y compañeros.
- Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar proyectos de Minería de Datos que involucren alta manipulación de datos utilizando el leguaje SQL.

Contenido:

1- **Entendiendo los conceptos:** fundamentos de bases de datos
   a. Bases de datos, tablas, columnas y filas
   b. Uso de ambientes de desarrollo (IDEs)
   c. Un vistazo a SQL y T-SQL
   d. Acceso a bases de datos desde R (RODBC y SQLite)

2- **Conocer los datos**: Exploración de tablas de datos (MySQL y SQLite)
   a. Explorar tablas completas o subconjuntos (sentencia **SELECT**)
      - Selección de variables/columnas
      - Selección de individuos/filas: filtros y operadores
   b. Derivación/Creación de nuevas columnas
      - Operadores
      - Sentencias condicionales
      - Alias
      - Funciones de fecha, hora, y para manipulación de textos
   c. Ordenamiento de datos (**ORDER BY**)

3- **Potenciar fuentes de datos:** Combinando tablas de datos (MySQL y SQLite)
    a. Cómo unir una o varias tablas (*Joins*)
    b. Renombrar tablas
    c. Concatenar conjuntos de datos (**UNION**)
    d. Creación de conjuntos complejos de datos con *sub-queries*
4- **Transformación de la información:** Limpieza de los datos
    a. Búsqueda de valores nulos
    b. Búsqueda de duplicados
    c. Eliminar datos (individuos)
    d. Modificar variables (a uno o múltiples individuos)
    e. Agregar nuevos individuos
5- **Análisis descriptivos:** Resumir datos / Creación de grupos
    a. Agrupación de individuos (**GROUP BY**)
    b. Funciones sobre grupos (**COUNT**, **MAX**, **MIN**, etc.)
    c. Filtros y limpieza de datos utilizando elementos agrupados (*outliers*)
6- **Ampliar fuentes de datos:** Nuevas estructuras
    a. Crear nuevas tablas
    b. Poblar tablas a partir de tablas existentes
7- **R y Bases de Datos:** acceder a bases de datos directamente desde R
    a. Conexión a fuentes de datos
    b. Creación de estructuras (data frames) en R
    c. Uso de estructuras
    d. Almacenar datos en la Base de Datos
8- **Manipulación de datos en R:** uso de paquetes sqldf, dplyr y tidyr
    a. Manipulación de datos en R por medio de sqldf
    b. Uso de dplyr y tidyr para limpieza, procesamiento y manipulación de datos en R
        a. Filtros
        b. Agregaciones
        c. Transformaciones

Bibliografía:

1. Alexander M. Decker J. and Wehbe B. “Business Intelligence Tools for Excel Analysts”. Wiley, 2014.
2. Beaulieu A. “Learning SQL”, O’Reilly, 2009.
3. Ben-Gan I. “SQL Server 2012 T-SQL Fundamentals”. O’Reilly, 2012.
4. Date C.J. “SQL and Relational Theory: Hot to Write Accurate SQL Code”. O’Reilly, 2012.
5. DuBois P. “MySQL Developer’s Library”. Addison-Wesley, 2013.
6. Harrington J.L. “SQL Clearly Explained”. Morgan Kaufmann, 2010.
7. Linoff G. “Data Analysis Using SQL and Excel”. Wiley, 2008.
8. Mistry R, Misner S. “Introduction Microsoft SQL Server 2008 R2”. Microsoft Press, 2010.
9. Seyed M.M, “Saied” T. and Hugh E.W. “Learning MySQL”, O’Reilly, 2007.
10. Williams, G. “Data Mining with Rattle and R”, Springer, 2011.
11. Winston W. “Microsoft Excel 2013 Data Analysis and Business Modeling”, O’Reilly, 2014.
12. Zapawa T. “Excel® Advanced Report Development”. Wiley, 2005.