**Módulo 5: Lenguajes para el procesamiento de datos y Modelos para el diseño de soluciones**

**Introducción:**

El análisis de datos estadísticos ha cobrado un papel crítico en la cultura contemporánea. Por ello no es extraño que se reclame cierto grado de alfabetización estadística en el conjunto de la sociedad, con el objetivo de aspirar a nuevas cotas de desarrollo humano (p.e., Bond, 2009; Seldmeier y Gigerenzer, 2001).

Para la ciencia, que aspira a conocer la naturaleza, y para la técnica, que anhela controlarla, la estadística se desdibuja como la herramienta clave que orienta y conduce el progreso de las comunidades humanas. Sin embargo, pese a que hoy en día disponemos de algoritmos de cómputo estadístico y que podemos implementarlos en computadoras lo suficientemente potentes como para manejar un volumen de datos considerablemente grande, parece que no somos capaces de sacar provecho de esta situación.

El enfoque es un problema muy frecuente en la enseñanza y aprendizaje de Base de Datos en el ambiente académico a nivel superior. Esto sucede porque muchas veces cuando se hace referencia a un curso sobre base de datos, se asocia al manejo de una herramienta de software como puede ser SQL Server, Oracle, Access, DB2, u otro en particular y se descuida mucho el aspecto conceptual y las áreas de aplicación de los modelos de base de datos.

Por eso, en este módulo se enfatiza la importancia que tiene modelar adecuadamente una base de datos y sus aplicaciones tan diversas en la ingeniería y los negocios. SQL es un tema que merece una mención especial, ya que es un lenguaje "casi" estándar para el mantenimiento de una base de datos, utilizado por Lenguajes de Programación, DBMS, Querys, OLAPS, así como también en la Inteligencia de Negocios y, en el análisis multidimensional de base de datos.

Es importante que el tema de base de datos se refleje en el currículo de las especialidades universitarias y técnicas, ya que todos tienen contacto con algunas bases de datos en las tareas cotidianas ya sea como usuario o administrador.

Por último, en este módulo analizaremos la aplicación de Big Data en las organizaciones. Big Data puede ser considerada como una tendencia en el avance de la tecnología que ha abierto la puerta a un nuevo enfoque para la comprensión y la toma de decisiones, que se utiliza para describir las enormes cantidades de datos (estructurados, no estructurados y semiestructurados). Así, el concepto de Big Data se aplica a toda la información que no puede ser procesada o analizada utilizando herramientas o procesos tradicionales.

**Objetivo:**

Comprender los tipos y características del lenguaje, con el fin de conocer el procesamiento de datos y analizar los modelos para el diseño de soluciones de negocio.

**5.1. Lenguaje: tipos y características**

#### Introducción:

La ciencia de datos es un campo que combina habilidades estadísticas y cuantitativas avanzadas, con capacidad de programación en el mundo real. Existen muchos lenguajes de programación en los que el científico de datos puede considerar especializarse; R es lanzado en 1995 como descendiente directo del antiguo lenguaje de programación S, de hecho, R se ha ido fortaleciendo a lo largo del tiempo; escrito en C, Fortran y en sí mismo, el proyecto cuenta actualmente con el apoyo de la R Foundation for Statistical Computing.

Todos los conceptos referentes a las bases de datos son muy claros y definidos formalmente a diferencia de los lenguajes en las bases de conocimiento. La tecnología de gestión de bases de datos se encuentra en una etapa muy madura, las bases de datos han evolucionado durante los pasados 30 años, desde sistemas de archivos rudimentarios hasta sistemas gestores de complejas estructuras de datos, los cuales ofrecen un gran número de posibilidades.

La Inteligencia Artificial tiene como principal objetivo el desarrollo de plataformas "inteligentes", que sean capaces de reproducir lo más fehacientemente posible los mecanismos intelectuales comunes de los usuarios; de forma destacada, la comprensión del lenguaje natural, el aprendizaje (acumulación de experiencia), el razonamiento lógico, la resolución de problemas y el reconocimiento de patrones.

Las diferencias entre las bases de datos y las bases de conocimiento son muchas; las bases de datos son sistemas bien establecidos, en donde todo lo que puede o no puede llevar a cabo el sistema y cómo llevarlo a cabo está perfectamente establecido. El trabajo en bases de conocimiento se encuentra aún limitado y, por lo general, se trata de sistemas específicos para necesidades específicas. La literatura en torno a las bases de conocimiento es altamente programática, intentando sentar las bases de lo que se supone que estos sistemas serán capaces de hacer.

#### Objetivo:

Comprender el lenguaje, su sintaxis, tipos y características, con el fin de entender posteriormente las prácticas para la comprensión y análisis de datos de R.

Los temas a revisar para responder al objetivo de esta sesión son:

* Lenguaje: sintaxis, tipos y características
* Prácticas para la comprensión de “R”

#### Bases de datos orientada a objetos

Los sistemas orientados a objetos deben soportar los tipos básicos de conjuntos y tuplas para construir y manipular datos complejos. Estos modelos se construyen a partir de la conjugación de tres mundos: los objetos complejos, los modelos semánticos de bases de datos y conceptos de la programación orientada a objetos.

En este sentido, los modelos de bases de datos deben ofrecer el manejo de objetos de identificadores de valores complejos, tipos, clases, métodos, jerarquías ISA, herencia y dinámico y encapsulación.

#### Esquemas

Los esquemas en base de datos están formados por signos; un esquema es una representación de alguna realidad, aunque hay otras interpretaciones de la relación entre connotación y denotación, tradicionalmente los signos en un esquema se adhieren a una interpretación aristotélica.

En una base de datos debemos ser capaces de proporcionar un procedimiento no ambiguo, para determinar si un fenómeno es un miembro de un signo o no.

#### Sintaxis sentencias DROP

Para eliminar tablas con SQL se utiliza la sentencia DROP en un sentido más amplio. La sentencia DROP se utiliza para borrar tablas, índices, procedimientos y vistas.

DROP { TABLE tabla | INDEX índice ON tabla | PROCEDURE procedimiento | VIEW

Vista }

En esta sintaxis TABLE es el nombre de la tabla que va a ser eliminada o de la cual va a ser eliminado el índice; PROCEDURE es el nombre del proceso que se va a eliminar; VIEW es el nombre de la vista asociada a eliminar e INDEX se refiere al índice de la tabla a eliminar.

#### “R”

“R es un entorno de trabajo para la ejecución de análisis estadísticos y la creación de gráficos (R Development Core Team, 2011). La interfaz gráfica del programa es una consola de comandos, es decir, que para interactuar con el mismo hay que escribir líneas de código y ejecutarlas”.

(Ruiz-Ruano, A.; Puga, J., 2016:  p.75)

R está compuesto por una estructura base que contiene funciones básicas para realizar cálculos y gráficos estadísticos.

La principal ventaja de R es la libertad y la gratuidad de su uso en comparación con otros softwares de análisis estadístico. Otras ventajas que presenta su uso es que favorece un aprendizaje básico de estadística y que puede ser ejecutado en diferentes sistemas operativos como Windows, Mac OS o Linux.

recursos

**5.2. Lenguaje para el procesamiento de datos**

**Introducción:**

Existen dos bloques de la ciencia de los datos; por un lado, se encuentran las necesidades de la infraestructura: capacidades de almacenamiento, transmisión y transformación de datos; por otro lado, se encuentran las necesidades de explotación o analítica del dato para extraer información relevante y conclusiones de negocio. Además, en ocasiones, el cliente es quien impone determinada tecnología o lenguaje.

Python es un lenguaje de programación de uso general, este cuenta con módulos específicos y soporte comunitario, de hecho, Python es un lenguaje fácil de aprender. La baja barrera de entrada lo convierte en un primer idioma, lo que es ideal para aquellos que son nuevos en programación. Paquetes como Pandas, Scikit-learn y Tensorflow hacen de Python una opción sólida para aplicaciones avanzadas de aprendizaje automático.

SQL define, administra y consulta bases de datos relacionales. El lenguaje apareció en 1974 y desde entonces ha sufrido muchas implementaciones, pero los principios básicos siguen siendo los mismos. SQL es utilizado en una amplia gama de aplicaciones, los módulos como [SQLAlchemy](https://www.sqlalchemy.org/" \t "_blank) hacen que la integración de SQL con otros lenguajes sea sencilla, también, la sintaxis declarativa de SQL lo hace un lenguaje muy legible. La licencia varía, ya que algunas implementaciones son gratuitas y otras son propietarias.

Los lenguajes para la ciencia de datos agilizan el proceso de gestión en torno al proceso ETL (extracción-transformación-carga). Esto hace que la descarga y la generalidad encaje perfectamente. Las librerías como Tensorflow de Google hacen de los lenguajes un aprendizaje automático, sin embargo, gran parte del proceso de la ciencia de la información depende de la longevidad y la eficiencia de las bases de datos.

**Objetivo:**

Comprender las bases del lenguaje para el procesamiento de datos, con el fin de conocer la optimización y funcionamiento de Phyton y SQL.

Los temas a revisar para responder al objetivo de esta sesión son:

* Prácticas para la comprensión de “Phyton”
* Prácticas para la comprensión de “SQL (MySQL)”

**El lenguaje SQL**

El lenguaje SQL es un lenguaje estándar completo de control e interacción con un sistema de gestión de base de datos relacional, definido por American National Standards Institute (ANSI) y la International Standards Organisation (ISO). El lenguaje administra, almacena y recupera información utilizada por la mayoría de los sistemas gestores de bases actuales, este incluye:

* DDL Lenguaje de definición de datos
* DML Lenguaje de modificación de datos
* TCL Lenguaje de control de transacciones
* DCL Lenguaje de control de datos

**Optimización de consulta**

Al hablar de la optimización en consultas nos referimos a los sistemas objeto-racional. La estimación del costo de ejecución de una consulta compleja es difícil, sobre todo si se tiene en cuenta que los usuarios pueden definir tipos de datos, operadores y funciones de datos.

Un sistema extensible de base de datos está diseñado a base de reglas; el optimizador trabaja a base de árboles y produce planes de ejecución que también lo son. Los operadores son primitivos que ofrecen el modelo de datos cuyas implementaciones específicas se denominan métodos.

**Catálogos**

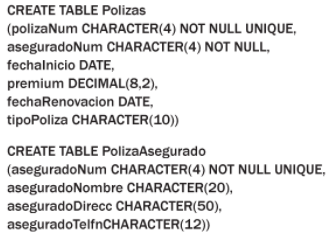
Los catálogos son un grupo de esquemas representados por un nombre, sin embargo, el estándar no define una sentencia para crear catálogos, por lo que depende de la implementación que esté utilizando. Los nombres en los esquemas deben ser únicos dentro de cada catálogo y, por lo tanto, los conflictos entre nombres se resuelven como en las tablas.

A continuación, se brindará como ejemplo una compañía de seguros, entendiendo que aquí se encuentra el principal negocio; en realidad, es un producto financiero que asegura riesgos entre el mantenedor y la póliza y algunos sucesos como la muerte, daños en edificios o accidentes de coches. Una parte básica de la base de datos de la compañía aseguradora se organizará:

Polizas(polizaNum, aseguradoNum, fechaInicio, premium, fechaRnovacion, tipoPoliza)

PolizaAsegurado(AseguradoNum, aseguradoNombre, aseguradoDirecc, aseguradoTelfn)

Podemos crear estructura de datos de SQL para la base de la siguiente forma:



recursos

**5.3. Modelos para el diseño de soluciones de negocio**

**Introducción:**

En el mundo de los empresarios, una forma de buscar la diferencia competitiva es con la excelencia de los procesos que se llevan a cabo dentro de la organización. Buscar soluciones a través de los datos, para el reconocimiento como empresa de excelencia, es un asunto de interés y atención por parte de los tomadores de decisiones, para conseguirlo se deben aplicar ciertas especificaciones y lineamientos.

La creación del modelo de diseño conceptual inicia desde una etapa temprana en la fase “Análisis”, específicamente en el proceso “Desarrollo de la solución técnica”. Para esta fase “Análisis y Diseño de Sistemas”, el diseño del modelo conceptual deberá completarse y añadirle todos los aspectos de diseño que corresponden a cada una de las capas y componentes de integración que correspondan.

El instrumento sirve para valorar la forma como la organización, sus empleados aceptan el cambio y reaccionan ante el mismo. En realidad, las empresas se encuentran en posiciones muy diferentes respecto al cambio; en la primera fase, esta característica es necesaria para conocer su nivel de adaptabilidad al cambio. Este análisis de la situación se realiza a partir de criterios simples, cuya justificación es compleja ya que supone un análisis multicriterio para elaborar las tipologías de los empleados en función de su adaptabilidad*a priori*.

El modelo de la solución se enfoca en los grandes conceptos, no en detalles de la solución. La interacción básica para la definición de este modelo es con los promotores ejecutivos de negocios del cliente, quienes deben gestionar las aproximaciones a la solución diseñada por el equipo de desarrollo. El diagrama para modelar el diseño debe incluir los componentes básicos de la arquitectura de *hardware*y *software*propuesta, sin entrar en detalles.

**Objetivo:**

Conocer el desarrollo de modelos de solución a través de los datos, para definir componentes que satisfagan las necesidades de la organización y poder adecuarse a las restricciones del problema.

Los temas a revisar para responder al objetivo de esta sesión son:

* Modelos aplicables al diseño de soluciones a través de los datos
* Modelos predictivos empleando ciencia de datos

**Optimización de consultas**

El problema de optimización exacta es computacionalmente intratable y, como no se dispone de información estadística precisa sobre las bases de datos, los algoritmos de evaluación de consulta del usuario tienen una forma estándar independiente del contenido de la base de datos y se transforma de manera que mejore su evaluación.

Para cada operación se tiene una buena implementación con un costo asociado, cada una de estas secuencias es un plan de acceso.

**Semántica**

Las relaciones entre la información y los datos se encuentran dentro del campo de las semánticas con significado, la semántica estudia aquello a lo que los signos hacen referencia y supone el uso e interpretación de los signos.

En una comunicación persona a persona, esto supondría usar signos lingüísticos; el receptor del mensaje debe interpretar los signos del mensaje y la semántica se ocupa de este proceso, mediante el cual se asigna significado a los signos.

**Transacciones remotas**

Las transacciones remotas se encuentran en la etapa de petición remota, para incluir soporte de transacción y multisentencia. El usuario puede emitir una serie de peticiones, que consulta o actualiza los datos en una base de datos remota, para luego complementar o volver atrás la serie entera de sentencias como una única transacción.

El DBMS garantiza que la transacción completa tendrá éxito o fallará como una unidad, como ocurre con las transacciones sobre una base de datos local, sin embargo, todas las sentencias que forman la transacción deben referenciar una única base.

**Modelo predictivo**

Entre los modelos predictivos se encuentran:

* **Regresión logística:** instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo. Se utiliza cuando se tiene una variable dependiente dicotómica y un conjunto de variables predictoras o independientes.
* **Árboles de decisión:** proveen de una herramienta de clasificación muy potente. Su uso en el manejo de datos la hace ganar en popularidad dadas las posibilidades que brinda y la facilidad con que son comprendidos sus resultados por cualquier usuario.
* **KNN:** algoritmo basado en instancia de tipo supervisado de Machine Learning. Puede usarse para clasificar nuevas muestras (valores discretos) o para predecir (regresión, valores continuos).
* **Red Neuronal:** Las redes neuronales artificiales imitan la estructura hardware del sistema nervioso, centrándose en el funcionamiento del cerebro humano, basado en el aprendizaje a través de la experiencia, con la consiguiente extracción de conocimiento a partir de la misma.