AA MANIPULAR LA HERRAMIENTA INFORMÁTICA SELECCIONADA (TII)

IDENTIFICACIÓN DE LA GUÍA DE APRENDIZAJE

- Denominación del Programa de Formación: TÉCNICO PARA LA PROGRAMACIÓN PARA ANALÍTICA DE DATOS.
- Código del Programa de Formación: 228117
- Nombre del Proyecto: 1901119 APLICAR BUENAS PRACTICAS PARA PREPARAR, LIMPIAR, REFINAR Y EXPLORAR GRANDES VOLÚMENES DE DATOS EN EL SECTOR PRODUCTIVO.
- Fase del Proyecto: PREPARACIÓN DE DATOS
- Actividad de Proyecto:
  - ORGANIZAR DATA
- Competencia:
- INTEGRACIÓN DE DATOS SEGÚN TÉCNICAS DE VISUALIZACIÓN Y METODOLOGÍAS DE ANÁLISIS.
- Resultados de Aprendizaje Alcanzar:
  - REALIZAR EL PROCESO DE LIMPIEZA DE DATOS DE ACUERDO CON LA HERRAMIENTAS INFORMÁTICAS SELECCIONADA.
- Duración de la Guía: 144H (presenciales + trabajo autónomo)

2. PRESENTACIÓN

La limpieza de datos, el manejo de herramientas y el uso de técnicas permiten el éxito de un proyecto exploratorio de datos, es una invitación a ser participe como actor activo e importante en él.

3. FORMULACIÓN DE LAS ACTIVIDADES DE APRENDIZAJE

3.1 Actividades de Reflexión inicial.

Se recomienda ver el siguiente video "Análisis de datos: el futuro de las organizaciones", el cual tiene como fin dar a conocer los elementos, metodologías y procedimientos necesarios para llevar a cabo un análisis de datos efectivo dentro de las organizaciones.

3.2 Actividades de contextualización e identificación de conocimientos necesarios para el aprendizaje.

Para realizar un análisis exploratorio de datos efectivo, se debe garantizar la calidad de los datos, se recomienda ver el video “Recursos y herramientas para el análisis de datos”

Manipulando los datos

Para realizar un buen proceso de limpieza de datos es importante aprender técnicas para manipular los datos de forma correcta.
Se recomienda ver el video de "Manipulación de datos"

Situaciones para realizar limpieza en los datos:

- Datos faltantes
- Columnas irrelevantes (que no responden al problema a solucionar
- Filas repetidas
- Valores extremos o atípicos (outliers)
- Errores tipográficos.
- Formatos de fechas

Se recomienda ver los siguientes videos:

- Confidencialidad de la información
- Obtención de los datos necesarios para el análisis
- Etapas de un análisis de datos
- Limpieza de datos
- Limpieza de datos estructurados
- Limpieza de datos no estructurados
- Uso de expresiones regulares
- Programación de expresiones regulares
- Diagramas y modelos
- Cálculos estadísticos con software
- Cálculos estadísticos en lenguaje de programación
- Análisis por regresión lineal

Desarrollemos el taller "GC-F-005_LIMPIEZAPYTHON" sobre un caso propuesto por el instructor

Desarrolla el taller "GC-F-005_OPENREFINE" sobre un DATASET propuesto por el instructor.

3.3 **Actividades de transferencia del conocimiento**

Desarrolle el caso escogido en el primer trimestre y aplique las técnicas y métodos vistos.

4. ACTIVIDADES DE EVALUACIÓN

Evidencias de Aprendizaje	Criterios de Evaluación	Técnicas e Instrumentos de Evaluación
Evidencias de Conocimiento: **EV1. Responde a preguntas sobre limpieza de datos.** **Evidencias de Desempeño.** **EV2. Realiza la limpieza de datos de un caso de estudio** **Evidencias de Producto:** **EV2 Aplica técnicas y métodos a la fuente de datos escogida para la limpieza de**	CREA NUEVOS DATOS A PARTIR DE OTROS DATOS, MANEJO DE DIFERENTES FUNCIONES PARA LA TRANSFORMACIÓN.	**IEV1. Cuestionario** **IEV2. Lista de verificación** **IEV3. Lista de verificación**

datos.		

5. GLOSARIO DE TÉRMINOS

Acceso
La manera en la cual los archivos o conjunto de datos son referenciados por la computadora**.**
Archivo
Un archivo es un elemento que contiene información y que a su vez se identifica por un nombre y su extensión. Esta última comienza por un punto y determina el tipo de aplicación a la que está asociado el archivo.
Buscadores
O también llamados motores de búsqueda, son herramientas que permiten clasificar la información que existe en la red y hacerla localizable en poco tiempo según las preferencias del usuario.
Base de datos
Una colección de registros o archivos relacionados de manera lógica.
Base de datos relacional
Una colección de relaciones normalizadas en la que cada relación tiene un nombre distintivo.
Bases de datos distribuidas
Son Bases de Datos que no están almacenadas totalmente en un solo lugar físico, (están segmentadas) y se comunican por medio de enlaces de comunicaciones a través de una red de computadoras distribuidas geográficamente.
Campo
Un campo es la unidad básica de una base de datos. Un campo puede ser, por ejemplo, el nombre de una persona. Los nombres de los campos no pueden empezar con espacios en blanco y caracteres especiales. No pueden llevar puntos, ni signos de exclamación o corchetes.
Clave principal
La clave principal en una tabla de una base de datos que se selecciona para identificar de forma unívoca cada registro de la tabla. Por ejemplo, en una tabla de alumnos podría ser su número de expediente académico.
Consulta
Mediante las consultas tendrás la posibilidad de obtener toda la información contenida en las tablas añadiendo interesantes funcionalidades.
DDL
Lenguaje de definición de datos utilizado para describir todas las estructuras de información y los programas que se usan para construir, actualizar e introducir la información que contiene una base de datos.
Diseño de la base de datos
Cuando trabajamos con bases de datos relacionales es habitual distribuir la información en diferentes tablas vinculadas entre sí. Esta característica obliga a un proceso de planificación y diseño previo para obtener el resultado esperado. Piensa que deseas almacenar en la base de datos, qué datos necesitas recuperar y en definitiva, determina el propósito final del proyecto para establecer unos cimientos lo suficientemente sólidos.
DBMS
Conjunto de programas destinados a manejar la creación y todos los accesos a las bases de datos. Se compone de un lenguaje de definición de datos (DDL: Data Definition Language), de un lenguaje de manipulación de datos (DML: Data Manipulation Language) y de un lenguaje de consulta (SQL: Structured Query Language).
ELIMINACION

Es una solicitud de eliminación que se expresa de forma muy parecida a una consulta. Sin embargo, en vez de presentar tuplas al usuario, quitamos las tuplas seleccionadas de la base de datos. Sólo puede eliminar tuplas completas; no se puede eliminar únicamente valores de determinados atributos.

Facilidad de Consultas

Permitir al usuario hacer cuestiones sencillas a la base de datos. Este tipo de consultas tienen como misión proporcionar la información solicitada por el usuario de una forma correcta y rápida.

Formulario

Los formularios resultan útiles principalmente en tareas de introducción de información. Cuando se trata de incluir pocos datos podemos hacerlo directamente sobre las tablas, pero cuando el volumen es importante, este método se vuelve poco eficaz. Para resolver este problema tenemos los formularios donde la inclusión de datos se hace de forma mucho más intuitiva y sencilla.

HTML

Siglas de HyperText Markup Language (Lenguaje de Etiquetas de Hipertexto), es el lenguaje predominante para la construcción de páginas web. Se utiliza para describir la estructura y el contenido en forma de texto, así como para complementar el texto con otros objetos, como por ejemplo: imágenes. Los archivos creados en este lenguaje suelen identificarse por su extensión del tipo: "nombre_archivo.html".

Informe

Los informes tienen como objetivo proporcionar las herramientas necesarias para obtener una copia impresa de los datos existentes en una base de datos, aunque existen otras posibilidades tan interesantes como la generación de archivos en formato PDF. Habitualmente, los informes se suelen construir a partir de los resultados obtenidos de la ejecución de consultas. De esta forma combinamos la posibilidad de seleccionar sólo los datos que deseemos que nos ofrecen las consultas con la ventaja de imprimirlos que aportan los informes.

Independencia de los datos

Se refiere a la protección contra los programas de aplicaciones que pueden originar modificaciones cuando se altera la organización física y lógica de las bases de datos.

Integridad referencial

La integridad referencial es una propiedad imprescindible en cualquier base de datos. Gracias a la integridad referencial se garantiza que un conjunto de datos (registro) siempre se relacione con otros conjuntos válidos, es decir, que existen en la base de datos. Implica que en todo momento dichos datos sean correctos, sin repeticiones innecesarias, datos perdidos y relaciones mal resueltas.

JDBC

La Conectividad de Bases de Datos Java (Java Database Connectivity, JDBC) es una especificación de la interfaz de aplicación de programa (application program interface, API) para conectar los programas escritos en Java a los datos en bases de datos populares.

Lenguaje de consulta

Son los lenguajes en el que los usuarios solicitan información de la base de datos. Estos lenguajes son generalmente de más alto nivel que los lenguajes de programación. Los lenguajes de consulta pueden clasificarse como procedimentales y no procedimentales.

Modelo de base de datos orientado a objetos

Es una adaptación a los sistemas de bases de datos. Se basa en el concepto de encapsulamiento de datos y código que opera sobre estos en un objeto.

Modelos de Red

Este modelo permite la representación de muchos a muchos de una Base de Datos. El modelo de red evita redundancia en la información, a través de la incorporación de un tipo de registro denominado el conector.

Nivel lógico

Definición de las estructuras de datos que constituyen la base de datos.

Reglas de Integridad
Son restricciones que definen los estados de consistencias de las bases de datos.
Registro
Un registro es el conjunto de información referida a una misma unidad.
Relación
El objetivo de estas relaciones sería principalmente evitar la duplicidad de información y en consecuencia, optimizar el rendimiento de la base de datos.
Recuperación
Proporcionar como mínimo el mismo nivel de recuperación que los sistemas de bases de datos actuales. De forma que, tanto en caso de fallo de hardware como de fallo de software, el sistema pueda retroceder hasta un estado coherente de los datos.
Sistema de Administración de Base de Dato
Es el software que controla la organización, almacenamiento, recuperación, seguridad e integridad de los datos en una base de datos.
SISTEMA GESTOR DE BASE DE DATOS
Es un conjunto de programas que permiten crear y mantener una base de datos, asegurando su integridad, confidencialidad y seguridad.
Software
Es un sistema manejador de bases de datos que permite al usuario accesar con facilidad a los datos almacenados o que ande ser almacenados
Tabla
Unidad donde crearemos el conjunto de datos de nuestra base de datos. Estos datos estarán ordenados en columnas verticales. En ella se definen los campos y sus características.
Transacción
Es una secuencia de operaciones de acceso a la base de datos que constituye una unidad lógica de ejecución.
Transacciones compartidas
Las transacciones compartidas soportan grupos de usuarios en estaciones de trabajo, los cuales desean coordinar sus esfuerzos en tiempo real, los usuarios pueden compartir los resultados intermedios de una base de datos. La transacción compartida permite que varias personas intervengan en una sola transacción.
Tupla
También se denomina de este modo a un registro o fila de una tabla.
Usuario final
Es quien accesa a las bases de datos por medio de un lenguaje de consulta o de programas de aplicación.
Valor nulo
Representa un valor para un atributo que es actualmente desconocido o no es aplicable para ese registro.
Vista
El resultado dinámico de una o más operaciones relacionales que operan sobre las relaciones base para producir otra relación. Una vista es una relación virtual que no tiene por qué existir necesariamente en la base de datos, sino que puede producirse cuando se solicite por parte de un usuario concreto, generándose en el momento de la solicitud.

6. REFERENTES BIBLIOGRÁFICOS

- Microsoft (2014). Libros en pantalla de SQL Server. Disponible en: https://technet.microsoft.com/es-es/library/ms130214(v=sql.105).aspx
- ADORACION, Miguel. Fundamentos y modelos de base de datos. Colombia: Alfaomega, México: Alfaomega, 2009.

- CUADRA, Dolores. Desarrollo de bases de datos: Casos prácticos desde el análisis a la implementación. Editorial Alfaomega. 2008.
- MANNINO, Michael V. Administración de base de datos: Diseño y desarrollo de aplicaciones. Editorial McGraw-Hill. México. 2007.
- SILBERSCHATZ, Abraham. KORTH, Henry. Fundamentos de base de datos. España; McGraw-Hill, 2006. 953p.
- CASTAÑO, Adoración. PIATTINI, Mario. Diseño de Bases de Datos Relacionales. Editorial Alfaomega. Colombia. 2006. 550p.
- KENDALL, Kenneth. Análisis y diseño de sistemas. México: Pearson, 2005. 726p.
- PIATTINI, Velthuis. Mario G. Análisis y diseño de aplicaciones informáticas de gestión. Editorial Alfaomega. México. 2004. 710p.
- RICARDO, Catherine. Bases de datos. Editorial McGraw-Hill. México. 2004. 642p.
- BUYENS, Jim. Aprenda desarrollo de base de datos web Ya.!.Editorial McGraw-Hill. Madrid. 2001. 549p.
- DATE, C.J. Introducción a los sistemas de base de datos. 7a.ed. Editorial Pearson. México 2001
- SENN, James. Análisis y diseño de sistemas de información. Editorial McGraw-Hill. México. 1992. 942p.
- COVADONGA Fernández Baizán, El Modelo relacional de datos, de los fundamentos a los modelos deductivos. Base de datos Dialnet.
- RIVERO CORNELIO, Enrique, FUENTES Luis Martínez, MARTÍNEZ Israel Alonso. Bases de datos relacionales, fundamentos y diseño lógico. Universidad Pontificia Comillas, 2005. Base de datos Dialnet.
- CASELLA, Dante. Como empezar la Normalización. [Publicado 14/06/2007]. [En línea], Disponible en YouTube: Diseño de BD parte 1 http://www.youtube.com/watch?v=IPKI19SbiYQ Diseño de BD parte 2 http://www.youtube.com/watch?v=AIn8inZInAQ [2011, Marzo 21].

7. CONTROL DEL DOCUMENTO

	Nombre	Cargo	Dependencia	Fecha
Autor (es)	**JOSE FERNANDO GALINDO SUAREZ**	**INSTRUCTOR**	**CGMLTI**	**20/01/2023**

8. CONTROL DE CAMBIOS (diligenciar únicamente si realiza ajustes a la guía)

	Nombre	Cargo	Dependencia	Fecha	Razón del Cambio
Autor (es)					