

Tarea 1. Conceptos Básicos

Integrantes:

- Martínez Flores Jorge Yael
 - Cuenta: 312128726
- Sanchez Morales Rodrigo Alejandro
 - Cuenta: 312089580

1. Conceptos Generales

a. Indica las diferencias (emplea una tabla para contrastarlas) en cuanto a cómo se representan los datos entre un sistema de procesamiento de archivos y una base de datos relacional.

Sistema de procesamiento de archivos	Base de datos relacional
Existen datos que pueden repetirse en diferentes lugares o archivos, esto provoca que, teniendo esa duplicidad de datos, el almacenamiento y el costo (en recursos del sistema) de acceso sean más altos	Tiene una serie de filtros y restricciones para evitar redundancia en la base de datos
Cuando se requiere de ciertos datos diferentes de archivos diferentes, la obtención, consulta y modificación de los datos no puede hacerse directamente de forma práctica y eficiente	Permite hacer consultas para cualquier tabla en la misma base de datos desde el mismo sitio
Debido a que los datos están dispersos en varios archivos, y los archivos pueden estar en diferentes formatos	Mantiene el mismo formato de datos para las tablas en la base de datos
En muchas aplicaciones es crucial asegurar que, cuando ocurra un fallo y sea detectado, se restauren los datos a un estado de consistencia que existía antes del fallo. Es difícil asegurar esta propiedad en un sistema de archivos tradicional.	Las bases de datos permiten tener puntos de restauración en caso de que algo falle y sea necesario regresar a un estado anterior

b. ¿Qué ventajas y desventajas encuentras al trabajar con una base de datos?

Ventajas de las bases de datos:

- Obtener la información requerida de manera estructurada
- Compartir la información de manera simultánea con otros usuarios o con otras bases de datos.
- Permite controlar la duplicidad de datos (redundancia), para tener un mejor manejo de los datos.
- Rapidez para obtener información debido a la clasificación óptima de los datos.
- Independencia de datos, ya que las bases de datos son independientes de los programas y/o aplicaciones.
- Portabilidad, ya que se pueden importar y mover con mayor facilidad.
- Se tiene un mejor nivel de mantenimiento debido a que podemos modificar su estructura, tal como agregar nueva información a un registro o agregar nuevas tablas.

Desventajas de las bases de datos:

- Para su mantenimiento se necesita a personal capacitado para poder lograr un manejo óptimo de las mismas, no cualquiera es apto para manipularlas.

- Costo, algunos productos relacionados con el manejo de bases de datos suelen ser costosos.
- Si la base de datos se llega a corromper, es complejo repararla y volverla a su anterior estado
- Tamaño, espacio. Una base de datos suele requerir mucho espacio en disco, suelen volverse pesadas
- Requieren de capacitación, asesoría y acompañamiento para enseñar su manejo.
- Se requiere de una persona para que esté al tanto del mantenimiento o de fallos
- Cuando la base de datos crece demasiado puede llegar a ponerse lenta, lo que afecta las búsquedas y la recuperación de información.

c. Explica las diferencias entre independencia de datos física y lógica. ¿Cuál es más difícil de lograr? Justifica tu respuesta.

- Independencia física de datos: Es la capacidad de modificar el esquema físico sin provocar que se vuelvan a escribir los programas de aplicación. Las modificaciones en el nivel físico son ocasionalmente necesarias para mejorar el funcionamiento
- Independencia lógica de datos: Capacidad de modificar el esquema conceptual sin provocar que se vuelvan a escribir los programas de aplicación. Las modificaciones en el nivel lógico son necesarias siempre que la estructura lógica de la base de datos se altere.

Las independencias lógicas de datos son más difíciles de lograr que las independencias físicas de datos, ya que los programas de aplicación son fuertemente dependientes de la estructura lógica de los datos a los que acceden.

d. ¿Qué es el diccionario de datos y por qué es importante?

Un diccionario de datos, o repositorio de metadatos, es un repositorio centralizado de información sobre datos tales como significado, relación con otros datos, origen, uso y formato. El diccionario de datos es un listado organizado de todos los datos que pertenecen a un sistema. Es importante ya que con un diccionario de datos se puede dar precisión sobre los datos que se manejan en un sistema, evitando así malas interpretaciones o ambigüedades. Define con precisión los datos de entrada, salida, componentes de almacenes, flujos, detalles de las relaciones entre almacenes, etc. El diccionario guarda los detalles y descripciones de todos estos elementos

e. Indica las principales características de los siguientes modelos de bases de datos: jerárquico, de red, orientado a objetos.

- **Base de datos jerárquica:** Este tipo de base de datos almacena la información en una estructura jerárquica que enlaza los registros en forma de estructura de árbol (similar a un árbol visto al revés), en donde un nodo padre de información puede tener varios nodos hijo, y así sucesivamente. Esta relación jerárquica no es estrictamente obligatoria, de manera que pueden establecerse relaciones entre nodos hermanos, y en este caso, la estructura en forma de árbol se convierte en una estructura en forma de una gráfica dirigida.
- **Base de datos de red:** En este tipo de base de datos está conformada por una colección de registros, los cuales están conectados entre sí por medio de enlaces en una red. El registro es similar al de una entidad como las empleadas en el modelo relacional. Un registro es una colección o conjunto de campos (atributos), donde cada uno de ellos contiene solamente un único valor almacenado. El enlace es exclusivamente la asociación entre dos registros, así que podemos verla como una relación estrictamente binaria. Una estructura de base de datos de red, abarca más que la estructura de árbol: un nodo hijo en la estructura red puede tener más de un nodo padre.
- **Base de datos orientada a objetos:** En este tipo de base de datos la información se representa mediante objetos como los presentes en la programación orientada a objetos. Cuando se integra las características de una base de datos con las de un lenguaje de programación orientado a objetos, el resultado es un sistema gestor de base de datos orientada a objetos, (ODBMS, Object Database Management System). Un ODBMS hace que los objetos de la base de datos aparezcan como objetos de un lenguaje de programación en uno o más lenguajes de programación a los que dé soporte. Un ODBMS extiende los lenguajes con datos persistentes de forma transparente, control de concurrencia, recuperación de datos, consultas asociativas y otras capacidades, los cuales usan exactamente el mismo modelo que lenguajes de programación orientados a objetos.

f. Elabora una línea de tiempo, en dónde indiques los principales hitos en el desarrollo de las bases de datos.

- **La antigüedad:** Se sabe que desde tiempos remotos las antiguas civilizaciones guardaban información, en varias presentaciones, desde dibujos en cuevas o hasta en grandes almacenes de libros. El ejemplo más claro son las bibliotecas donde se puede concentrar una gran cantidad de datos para poder utilizarla después.
- **1884:** Este año se dio origen a la maquina automática de perforación de tarjetas inventada por Herman Hollerith (29 de febrero de 1860—17 de noviembre de 1929), la cual fue utilizada para el censo de los Estados Unidos mejorando el proceso de terminación de siete años a sólo dos años y medio.
- **1950:** La creación de las cintas magnéticas sus principios se le atribuyen al inglés Oberlin Smith, con este mecanismo se empezó a automatizar la información referente a nómina. Utiliza un sistema de lectura secuencial y ordenada es decir si se desea buscar los datos, se debe avanzar la cinta hasta llegar a dicho punto y si desea empezar a leer la información desde 0 debe devolver la cinta hasta al principio.
- **1960:** La primera base de datos informatizada se inició en la década de los 60's, cuando el uso de las computadoras se convirtió en una opción más rentable para las organizaciones privadas. Había dos modelos de base de datos más populares en esta década: un modelo de red llamado CODASYL y un modelo jerárquico llamado IMS. Un sistema de base de datos que demostró ser un éxito comercial fue el sistema SABRE que fue utilizado por IBM para ayudar a American Airlines, la gestión de sus datos de reservas.
- **1970:** Edgar Codd publicó un importante documento para proponer el uso de un modelo de base de datos relacional, y sus ideas cambiaron la forma de pensar en las bases de datos. En su modelo, el esquema de la base de datos, o la organización lógica, se desconecta de almacenamiento de información física, y esto se convirtió en el principio de norma para los sistemas de bases de datos. Dos de los principales prototipos de relación del sistema de bases de datos fueron creadas entre los años 1974 y 1977, y fueron los de Ingres, que se desarrolló en la UBC, y del sistema de R, creado en IBM en San José. Ingres utiliza un lenguaje de consulta conocido como QUEL, y condujo a la creación de sistemas como Ingres Corporación, MS SQL Server, Sybase. Por otro lado, el sistema R utiliza el lenguaje de consulta secuela, y que ha contribuido al desarrollo de SQL / DS, DB2, Allbase, Oracle y SQL Non-Stop. Fue también en esta década que Relational Database Management System, o RDBMS, se convirtió en un término reconocido.
- **1976:** El Modelo Entidad-Relación (E-R) proporciona una herramienta para representar información del mundo real a nivel conceptual. Creado en el año de 1976 por Peter Chen, permite describir las entidades involucradas en una base de datos, así como las relaciones y restricciones de ellas.
- **1980:** El lenguaje de consulta estructurado o SQL, se convirtió en el lenguaje de consulta estándar. Los sistemas de bases de datos relacionales se convirtieron en un éxito comercial como el rápido aumento de las ventas de ordenadores de estímulo al mercado de bases de datos, y esto provocó un importante descenso en la popularidad de las redes y los modelos jerárquicos de bases de datos. DB2 se convirtió en el producto insignia de la base de datos de IBM, y la introducción de la PC de IBM como resultado en los establecimientos de muchas empresas de bases de datos nuevas y el desarrollo de productos como Paradox, 5000 RBASE, RIM, Dbase III y IV, OS / 2 de base de datos, y Watcom SQL. También se da inicio a las bases de datos orientadas a objetos.
- **1990:** Después de una sacudida industria de base de datos, la mayoría de las empresas supervivientes vende productos complejos de bases de datos a precios elevados. Alrededor de este tiempo, las nuevas herramientas de cliente para el desarrollo de aplicaciones fueron liberadas, y éstos incluyen el desarrollador de Oracle, PowerBuilder, Visual Basic, etc. Una serie de herramientas para la productividad personal, como ODBC y Excel / Access, también se desarrollaron. La llegada de la Internet condujo a un crecimiento exponencial de la industria de la base de datos. Los usuarios de escritorio comenzaron a utilizar sistemas cliente-servidor de base de datos para acceder a los sistemas informáticos.
- **Siglo XXI:** Aunque la industria de Internet experimentó un descenso en la década de los 2000, las aplicaciones de base de datos continuarán creciendo. Las nuevas aplicaciones interactivas se han desarrollado para Asistentes de Información Personal (PDA), las transacciones de punto de venta, y la consolidación de los proveedores. En la actualidad, las tres empresas líderes de base de datos en el mundo occidental son Microsoft, IBM y Oracle.

g. Explica cuál(es) de los siguientes conceptos juegan un papel importante en la representación de los datos acerca de un problema real en una base de datos y por qué: lenguaje para definición de datos, lenguaje para manipulación de datos, el modelo de datos, manejo de transacciones.

A continuación, se dará una breve explicación de cada uno de los siguientes conceptos:

- **Lenguaje de definición de datos:** Está orientado a describir de una forma abstracta las estructuras de datos y las restricciones de integridad.
- **Lenguaje para definición de datos:** Es un lenguaje proporcionado por el sistema manejador de bases de datos, que permite a los programadores de llevar a cabo las tareas de definición de las estructuras que almacenarán los datos, así como de los procedimientos o funciones que permitan consultarlos.
- **Modelo de datos:** Un modelo de datos permite describir:
 - Las estructuras de datos de la base: El tipo de los datos que hay en la base y la forma en que se relacionan
 - Las restricciones de integridad: Un conjunto de condiciones que deben cumplir los datos para reflejar la realidad deseada.
 - Operaciones de manipulación de los datos: típicamente, operaciones de agregado, borrado, modificación y recuperación de los datos de la base.
- **Manejo de transacciones:** Es un conjunto de órdenes que se ejecutan formando una unidad de trabajo, es decir, en forma indivisible o atómica.

Dadas las definiciones anteriores podemos decir que el modelo de datos juega un papel importante en la representación de los datos ya que nos permite describir los elementos de la realidad que intervienen en un problema dado y la forma en que se relacionan estos elementos entre sí.

h. Supón que una pequeña compañía desea almacenar su información en una base de datos. Desea comprar la que tenga la menor cantidad de características posibles, se desea ejecutar la aplicación en una sola computadora personal y no se planea compartir la información con nadie. Para cada una de las siguientes características explica por qué se debería o no incluir en la base de datos que se desea comprar (suponiendo que se pueden comprar por separado): Seguridad, Control de concurrencia, Recuperación en caso de fallas, lenguaje de consulta, mecanismo de vistas, manejo de transacciones.

Característica	¿Se debería incluir o no en la base de datos?
Seguridad	Sí, ya que se tiene que asegurar algún tipo de seguridad a la información guardada ya que cualquiera podría tener acceso a la única computadora personal y modificar elementos valiosos en la base de datos.
Control de concurrencia	No, porque en un momento determinado no se debe de garantizar que haya un tipo de control para varios usuarios modificando datos al mismo tiempo
Recuperación en caso de fallas	Sí, por lo tanto se debe tener una manera de asegurar que al tener una falla hay una forma de recuperar el contenido de información que estamos recabando.
Lenguaje de consulta	Sí, ya que al estar usando una base de datos necesitamos una herramienta para hacer consultas.
Mecanismo de vistas	Sí, ya que al hacer una consulta se debe presentar como una tabla (virtual) a partir de un conjunto de tablas en una base de datos.
Manejo de transacciones	Sí, ya que una transacción es un conjunto de órdenes que se ejecutan formando una unidad de trabajo, es decir, en forma indivisible o atómica.

2. Investigación

a) **¿Qué es la Ciencia de Datos y Big Data? ¿Cómo se relacionan con las bases de datos?** La **Ciencia de Datos** es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sean datos estructurados o no estructurados, esto es una continuación de algunos campos de análisis de datos como la estadística, la minería de datos, el aprendizaje automático y la analítica predictiva.

Big Data es un concepto que se refiere a conjuntos de datos tan grandes que las aplicaciones tradicionales no son capaces de procesarlas para poder encontrar patrones repetitivos dentro de estos datos.

b) **Especifica las características más importantes de las bases de datos NoSQL, indica el modelo de datos que utilizan y principales proveedores.**

- **Facilmente escalable:** Esta nueva generación de bases de datos, están diseñados para escalar de forma limpia para tomar ventaja de nuevos nodos, y por lo general están diseñados con hardware de bajo costo.
- **Big Data:** Los volúmenes Big Data pueden ser manejados por los sistemas NoSQL.
- Las Bases de datos NoSQL son generalmente diseñadas para requerir menos gestión: reparación automática, distribución de datos y modelos de datos más simples, que existen para no depender de un administrador de base de datos.
- Las Bases de Datos NoSQL suelen usar grupos de servidores de menor costo para administrar los datos y el volumen de transacciones, mientras que lo SGBD tienden a depender de servidores costosos y sistemas de almacenamiento privativos. En otras palabras, es mucho más barato el costo por traslación o manejo de información en las bases de datos NoSQL.
- Las Bases de Datos NoSQL no dependen de un modelo en específico para ser usados

Principales proveedores:

- **Cassandra:** Su modelo de datos consiste en particionar las filas, que son reorganizadas en tablas. Las claves primarias de cada tabla tiene un primer componente que es la clave de partición. Dentro de una partición, las filas son agrupadas por las columnas restantes de la clave. Las demás columnas pueden ser indexadas por separado de la clave primaria.
- **Riak:** Se basa en el modelo orientado a objetos. Un Objeto es el elemento más grande y más pequeño de los datos. Como tal, la interacción con la base de datos es recuperando o modificando el objeto entero. No hay búsqueda parcial o actualización de los datos.
- **Redis:** El modelo de datos de Redis se basa en la estructura de datos del tipo diccionario o tabla hash que relaciona una llave a un contenido almacenado en un índice. La principal diferencia entre Redis y otros sistemas similares es que los valores no están limitados a ser de tipos string, otros tipos de datos están soportados:
 - Listas de strings
 - Conjuntos de strings
 - Hashes donde la llave y el valor son del tipo string
- **CouchDB:** No almacena los datos en sus relaciones y en tablas. Cada base de datos es una colección de documentos independientes. Cada documento mantiene sus propios datos y su esquema autocontenido.
- **MongoDB:** Utiliza un modelo basado en documentos, la forma de modelar las relaciones entre las entidades cambia. Ya que existe la ausencia de conceptos de los modelos Entidad-Relación tales como la ausencia de joins y el acceso atómico a los datos.

Bibliografía

- Slideshare, Consideraciones para elegir un buen DBMS [<https://es.slideshare.net/evavivez/consideraciones-para-elegir-un-buen-dbms>] Consultado el día domingo 20 Agosto 2017.
- Computacion fciencias, 03ModeloE-R.pdf
[<http://computacion.fciencias.unam.mx/~gar/fbd/material/uploaden/Presentaciones/03ModeloE-R.pdf>] Consultado el día

lunes 21 de agosto 2017.

- El pensante, Ventajas y desventajas de las bases de datos [<https://educacion.elpensante.com/ventajas-y-desventajas-de-las-bases-de-datos/>] Consultado el día miércoles 23 Agosto 2017.
- Wordpress, 1.4.1.2 Independencia lógica y física de los datos [<https://tombasededatos.wordpress.com/2010/08/21/1-4-1-2-independencia-logica-y-fisica-de-los-datos/>] Consultado el día miércoles 23 Agosto 2017.
- Wikipedia, Data dictionary [https://en.m.wikipedia.org/wiki/Data_dictionary] Consultado el día miércoles 23 Agosto 2017.
- Wikipedia, Hierarchical database model [https://en.m.wikipedia.org/wiki/Hierarchical_database_model] Consultado el día miércoles 23 Agosto 2017.
- Wikipedia, Network model [https://en.m.wikipedia.org/wiki/Network_model] Consultado el día miércoles 23 Agosto 2017.
- Wikipedia, Object database [https://en.m.wikipedia.org/wiki/Object_database] Consultado el día miércoles 23 Agosto 2017.
- Wikipedia, Lenguaje de manipulación de datos [https://es.m.wikipedia.org/wiki/Lenguaje_de_manipulaci%C3%B3n_de_datos] Consultado el día miércoles 23 Agosto 2017.
- Wikipedia, Lenguaje de definición de datos [https://es.m.wikipedia.org/wiki/Lenguaje_de_definici%C3%B3n_de_datos] Consultado el día miércoles 23 Agosto 2017.
- Prezi, Línea de tiempo: Historia de las bases de Datos [<https://prezi.com/m/6fpkailpgadb/linea-de-tiempo-historia-de-las-bases-de-datos/>] Consultado el día jueves 24 Agosto 2017.
- Sistemas de Bases de Datos vs. Sistemas de Archivos [<https://uvfdatabases.wordpress.com/2009/02/02/sistemas-de-bases-de-datos-vs-sistemas-de-archivos/>] Consultado el día 24 de agosto 2017.
- Ciencia de Datos - Wikipedia [https://es.wikipedia.org/wiki/Ciencia_de_datos] Consultado el día 20 de agosto 2017.
- El científico de datos: una novedosa y necesaria profesión [<http://noticias.universia.es/ciencia-tecnologia/noticia/2017/05/10/1095994/cientifico-datos-novedosa-necesaria-profesion.html>] Consultado el día 20 de agosto 2017.
- Big Data -Wikipedia [https://es.wikipedia.org/wiki/Big_data] Consultado el día 20 de agosto 2017.
- Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad [<http://www.powerdata.es/big-data>] Consultado el día 20 de agosto 2017.
- Características De Las Bases De Datos NoSQL [<http://blog.hostdime.com.co/10-caracteristicas-que-debe-conocer-de-las-bases-de-datos-nosql/>] Consultado el día 20 de agosto 2017.
- Cassandra, la dama de las bases de datos NoSQL [<https://www.paradigmadigital.com/dev/cassandra-la-dama-de-las-bases-de-datos-nosql/>] Consultado el día 21 de agosto 2017.
- Riak - Wikipedia [<https://en.wikipedia.org/wiki/Riak>] Consultado el día 21 de agosto 2017.
- Redis - Wikipedia [<https://es.wikipedia.org/wiki/Redis>] Consultado el día 22 de agosto 2017.
- CouchDB - Wikipedia [<https://es.wikipedia.org/wiki/CouchDB>] Consultado el día 22 de agosto 2017.
- Modelado de Datos en MongoDB [<http://www.manualweb.net/mongodb/modelado-de-datos-en-mongodb/>] Consultado el día 22 de agosto 2017.