Nombres:	**Matrículas:**
Luis Fernando Lezama Araoz	2878106
Celic Gabriel Hernández Archundia	2877240
Diego Alfredo López Malerva	2999206
Iván Gutiérrez Gómez	2877087

Nombre del profesor:	**Nombre del curso:**	**Módulo**
Sergio Arturo Damián Sandoval	Fundamentos de programación para Big Data	**1**

Actividad #1	**Fecha:** 2 de febrero de 2023

Referencias bibliográficas:

[Estudiar con Manu]. (2022, Mayo 19). "Data Analytics vs Data Science vs Data Engineer ¿Qué diferencias hay?". Recopilado el 2 de febrero del 2023 de: https://www.youtube.com/watch?v=7nnl4iXjlew

¿Qué es el Data Science? (2020, 10 julio). Máster en Data Science. Recuperado 2 de febrero de 2023, de https://www.master-data-scientist.com/que-es-masters-in-data-science/

¿Qué es un data lake? | Google Cloud |. (s. f.). Google Cloud. https://cloud.google.com/learn/what-is-a-data-lake?hl=es-419

Carisio, E. (s.f). *¿Qué es Big Data y para qué sirve?* Ejemplos de uso. Recuperado de https://blog.mdcloud.es/que-es-big-data-y-para-que-sirve/

Data Warehouse: todo lo que necesitas saber sobre almacenamiento de datos. (s. f.). https://www.powerdata.es/data-warehouse

Dontha, R. (2017). *25 Big Data Terms You Must Know to Impress Your Date (or*

Hernández, E. D. K. (s. f.). *Lic. en Informática*. Recuperado 2 de febrero de 2023, de https://programas.cuaed.unam.mx/repositorio/moodle/pluginfile.php/870/mod_resource/content/5/Contenido/index.html

IBM. (s. f.). *¿Qué es Business Intelligence y cómo funciona? | IBM*. Recuperado 2 de febrero de 2023, de https://www.ibm.com/mx-es/topics/business-intelligence

Marketing KeepCoding. (2022, 20 enero). *¿Qué son los Datasets? [4 sitios donde encontrarlos]*. KeepCoding Tech School. Recuperado 2 de febrero de 2023, de https://keepcoding.io/blog/que-son-datasets/

¿Qué es un data mart? (s. f.). Oracle México. https://www.oracle.com/mx/autonomous-database/what-is-data-mart/

Power Data. (2013). *Procesos ETL: Definición, Características, Beneficios y Retos*. Recuperado de: https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/312584/procesos-etl-definici-n-caracter-sticas-beneficios-y-retos

Sánchez, E. H. (s. f.). *Lic. en Informática*. Recuperado 2 de febrero de 2023, de https://programas.cuaed.unam.mx/repositorio/moodle/pluginfile.php/1196/mod_resource/content/1/contenido/index.html

Techopedia. (2014, 1 octubre). *Data Source*. Techopedia.com. https://www.techopedia.com/definition/30323/data-source

Torres, B. (2019, 24 julio). *¿Qué es el aprendizaje automático y cómo funciona?* UNAM Global. Recuperado 2 de febrero de 2023, de https://unamglobal.unam.mx/que-es-el-aprendizaje-automatico-y-como-funciona/

What Is Data Visualization? Definition, Examples, And Learning Resources. (s. f.). Tableau. https://www.tableau.com/learn/articles/data-visualization

whomever you want). Recuperado de https://bit.ly/2T5sr2O

Investigación de conceptos

- **Algoritmo:** Conjunto detallado y lógico de pasos para alcanzar un objetivo o resolver un problema.
- **Base de datos:** Colección de datos relacionados, organizados, estructurados y almacenados de manera persistente.
- **Dataset:** Conjunto de datos tabulados en cualquier sistema de almacenamiento de datos estructurados, este término hace referencia a una única base de datos de origen, la cual se puede relacionar con otras.
- **Data Science:** Es un campo interdisciplinario que se encarga de estudiar de dónde viene la información, qué representa y cómo se puede convertir en un recurso valioso en la creación de negocios y estrategias.
- **Business Intelligence:** Es un tipo de software que se alimenta de datos de negocios y presenta reportes, paneles, tablas y gráficos de forma amigable para el usuario.
- **Machine Learning:** Es una rama específica dedicada a ayudar a las computadoras a aprender de los humanos y cómo interactuar con nosotros de una manera similar a la de los humanos.

- **Data Integration:** Proceso de combinar datos de diferentes fuentes y presentarlos en una sola vista.
- **Data Cleansing:** Proceso de limpieza de los datos, para retirar ruido, registros duplicados, corregir caracteres extraños, etcétera.
- **Data Mining:** Proceso para inferir patrones u obtener información desde grandes conjuntos de datos.
- **Data Model:** Un modelo define la estructura de los datos para el propósito de ser el puente entre el equipo funcional (conocedor del negocio) y técnico para poder mostrar los datos necesarios para los procesos del negocio que cubre tres niveles: física (tablas), lógico (entidades) y negocio (conceptos).
- **ETL:** Los procesos ETL son un término estándar que se utiliza para referirse al movimiento y transformación de datos. Se trata del proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y cargarlos en otra base de datos (denominada data mart o data warehouse) con el objeto de analizarlos.
- **ACID:** Una prueba para verificar la atomicidad, consistencia, aislamiento/resguardo y durabilidad de los datos.

- **Data Scientist:** Persona que toma los datos existentes de algún tópico y con ellos genera información nueva o predicciones para el beneficio de las empresas.
- **Data Analyst:** Persona que recibe datos tabuladores; es decir, de una base de datos. A partir de ver los datos plasmados en las herramientas de visualización del analista, este interpreta los contenidos.
- **Data Engineer:** Es la persona que se encarga de recopilar los datos. Los recolecta, organiza, hace las transformaciones necesarias, y al final los pone a disponibilidad del ingeniero en ciencia de datos.

- **Datos Estructurados:** Datos que tienen definidos su formato, tamaño y longitud, como la base de datos relacionales o Data Warehouse.
- **Datos Semi-Estructurados:** Datos almacenados según una cierta estructura flexible y con metadatos definidos, como XML y HTML, JSON, y las hojas de cálculo (CSV, Excel).
- **Datos no Estructurados:** Datos sin formato específico, como ficheros de texto (Word, PDF, correos electrónicos) o contenido multimedia (audio, vídeo, o imágenes).

- **Data Source:** es la ubicación de donde provienen los datos que se están utilizando. En un sistema de bases de datos, por ejemplo, la fuente primaria sería la base de dato (que podría ser un disco o un servidor remoto).
- **Data Visualization:** es la representación grafica de la información y los datos, empleando, por ejemplo, tablas, gráficas, mapas, entre otros.
- **Data Lake:** es un repositorio centralizado diseñado para almacenar, procesar y proteger grandes cantidades de datos estructurados, semiestructurados o no estructurados. Permite almacenar datos en su formato original ignorando los límites de tamaño.
- **Data Warehouse:** es otro tipo de repositorio centralizado, este especializado para almacenar todos los datos que recogen los diversos sistemas de una empresa. Se enfoca más que nada en la captura de datos de diversas fuentes, sobe todo para fines analíticos.
- **Data Mart:** es una forma sencilla de almacén de datos centrado en un único asunto, recogidos desde menos orígenes.

Preguntas

1. **El texto de un capítulo de un libro es considerado como un tipo de dato…**
   a. Dato estructurado
   b. Dato semi-estructurado
   c. Dato no estructurado

2. **¿Qué tipo de cargo es el responsable de la extracción, almacenamiento y mantenimiento de los datos en una organización?**
   Data Engineer

3. **El ETL es un proceso que se ejecuta durante la técnica de:**
   a. Data Integration
   b. Data Cleansing
   c. Data Mining
   d. Data Model

4. **¿Qué concepto abarca el análisis histórico de los datos con el propósito de entender el comportamiento de una organización?**
   a. Business Intelligence
   b. Machine Learning

5. **Suponga que trabaja en un banco el cual fue atacado con un virus.** Al ser el experto responsable en Big Data, se le indica que restaure los datos locales de la organización con los datos ubicados en la nube. La indicación es que restauren primero los datos de todas las transacciones de un cliente, con el que se tendrá una reunión importante en 30 minutos. Pero, al entrar a la nube, observa datos de cliente almacenados en un data lake, un data warehouse y un data mart. Indique cuáles datos son los que tendría que restaurar.
   Data mart, porque comprende un espectro de datos menor, y en este caso se busca la información de sólo un cliente en muy poco tiempo, y el data mart comprende información de un rubro en específico.

Juego