**Tarea**

**Evaluación práctica de la Unidad 1: Reconociendo lo aprendido**

**Instructivo de la Tarea**

**Introducción a diversos elementos del Big Data**

Para iniciar la comprensión de cómo abordar la solución de un problema de Big Data se requiere tener claro ciertos conceptos básicos. Antes de profundizar en el manejo de grandes volúmenes de datos, vamos a realizar una actividad que cualquier proyecto de Big Data incluye, se trata de la fase ETL (Extraction, Transformation and Loading) por sus siglas en inglés) que significa Extracción, Transformación y Carga. Para lograr este propósito se ha creado un escenario hipotético para desarrollar una solución muy probable en el mundo real. Veamos cuál es el problema.

El crecimiento de la información a través de Big Data y el Internet de las Cosas requiere que el desarrollador de software esté capacitado en la gestión de datos de todo tipo y de grandes volúmenes de estos, dada su variedad actual (hojas de cálculo, audios, videos, archivos planos, sensores, bases de datos, entre otros). Es necesario que en todo proyecto el Ingeniero de Datos esté familiarizado con una serie de elementos fundamentales, entre ellos:

* Gobernanza de datos o gobierno de datos
* Identificar las fuentes y tipos de datos que tengan contexto con los objetivos del proyecto
* Administrar el sistema de bases de datos asegurando la operatividad, integridad, disponibilidad y confidencialidad de la información almacenada de acuerdo con las normas internacionales y el objeto de negocio.
* Integrar los datos en una bodega de datos, con la estructura determinada.
* Diseñar los procesos de limpieza de datos y calidad de la información, aplicando los procedimientos estándar para este tipo de actividades.
* Definir los procedimientos de protección y seguridad de los datos, garantizando la confiabilidad de las transacciones e información.
* Aplicar los principios éticos del profesional en gestión de datos.
* Comprender, planificar, codificar y ejecutar procesos ETL para la solución de problemas.
* Realizar análisis de datos a través de consultas que muestren ***Insights*** relevantes para la organización

**CASO DE ESTUDIO**

**Solución para un proceso ETL Big Data de una empresa hipotética de venta de gaseosas**

El recién estrenado líder del departamento de informática, egresado del I.U. Pascual Bravo, de la nueva Franquicia de productos de bajo costo “Gaseosas Poderosas”, desea mejorar la eficiencia de varios departamentos: Mercadeo, Compras, Ventas y Seguimiento al Cliente, entre otros. Actualmente, todo se lleva en hojas de cálculo con todos los problemas que esto representa. El objetivo es migrar la información de ventas de los productos, que se encuentran en las hojas de cálculo, a una base de datos relacional para evitar redundancias y “embasuramiento” de la información. De esta forma, será más fácil y seguro construir un Cuadro de Mando Integral, Tablero de Control y/o un sistema de visualización en tiempo real del comportamiento de venta de ciertos productos en las diferentes regiones, departamentos y municipios con el fin de tomar decisiones informadas que mejoren el desempeño, los ingresos y la utilidad de la empresa. Así que este ha decido adoptar una estrategia de gestión de Big Data utilizando técnicas de “Data Munging o Wrangling” y ”Minería de Datos”; a través de un proceso ETL que extraiga, transforme y registre la información en una base de datos para que facilite un análisis y visualización de la información al Científico de Datos. A continuación, se presentan las tres (fases) de un proceso ETL:

1.- (***EXTRACCIÓN***). En una primera etapa, se utiliza un algoritmo para extraer la información de las hojas de cálculo que contiene los registros de todos las regiones, departamentos y municipios de Colombia. Cada renglón de la hoja contiene: Región, nombre departamento, nombre municipio, códigos DANE de departamento y municipio. En una segunda etapa, se extraerá la información de la venta diaria de productos por municipio y departamento. Cada sucursal tiene su propia base de datos, independiente de las otras, y los resultados de operaciones de ventas diarias se envían en una hoja de cálculo a la Sede Principal en Támesis, Antioquia. Es decir, no hay una base de datos centralizada de todas las tiendas. Por lo tanto, se debe diseñar y construir una base de datos que almacene esta información.

2.- (***TRANSFORMACIÓN***). La información que extrae de esta hoja de cálculo tiene que “manipularse” o “transformarse” para adecuarse a los requerimientos de la plataforma de monitoreo que tendrá como fuente de información una base de datos relacional. Uno de los requerimientos es que las claves primarias de departamentos y municipios tienen que ser números enteros y únicos. Esta situación impide que se utilicen los códigos DANE de los municipios (que vienen en texto representados por un punto). Ejemplo: al convertir el código de XXXXXX a 99.99 y el código de YYYYYY a 9.999 en enteros, ambos resultan en el mismo número “999999” lo cual rompería la restricción de unicidad de la clave primaria. Por lo tanto, se requiere construir un sistema de creación de los códigos de departamentos y municipios de manera automática y transparente al usuario. Al utilizar una base de datos relacional, se le podrá sacar provecho a la relación que se establezca con las entidades Departamento-Municipio. Después, en otra fase del proyecto, se integrarán otras entidades como Ventas Diarias de Productos.

3.- (***CARGA***). Una vez que la información se halla en el formato deseado, se procede a “cargarla” en las tablas respectivas de departamentos, municipios y productos. Los registros de operaciones de venta se encuentran ya en formato SQL (tienen algunos problemas que se deben solucionar a través de la transformación, limpieza y/o imputación de información). De esta manera, se dispondría de un sistema relacionado que está listo para incluir las ventas diarias de productos, y a través del poder de SQL, realizar consultas de todo tipo, tales como, el producto más vendido, el producto que genera más utilidad, ventas en un periodo específico de fechas, ¿Cuál es la clasificación de ventas por municipio y departamento? Ventas por producto. Toda esta información es vital para la toma de decisiones. Este es uno de los grandes propósitos de la gestión de Big Data.

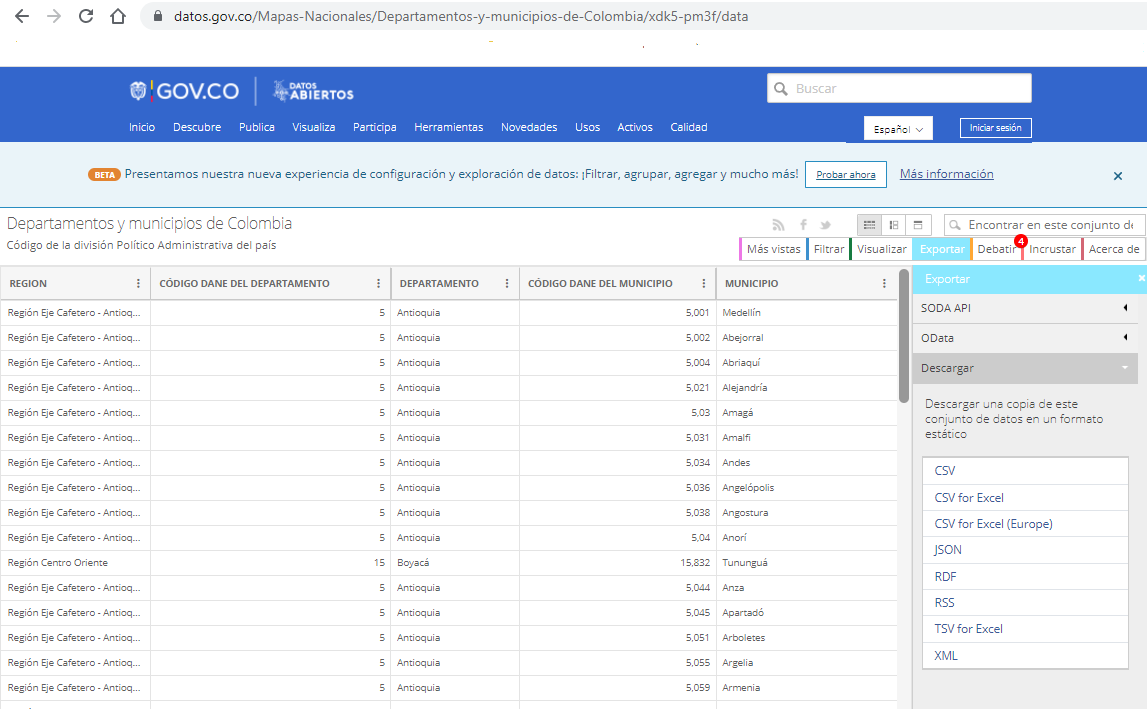
**NOTA**: Estas tres actividades representan el inicio de un proyecto de Big Data: ETL.

**Archivos provistos para realizar la tarea**

A continuación, se presentan los algoritmos, scripts y fuentes de datos que se le entregan al estudiante para que realice la tarea correctamente. Algunos de estos archivos no sufrirán modificaciones de ningún tipo y otros tendrá que ser modificados por el estudiante para resolver los problemas que se plantean.

**FUENTE DE DATOS:**

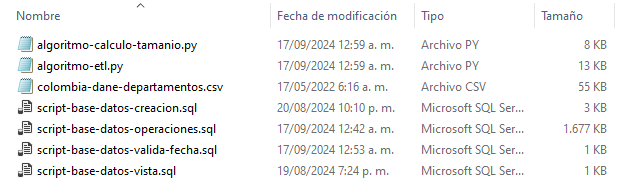
La fuente de datos que se utilizará en la tarea proviene de los archivos digitales del DANE. Se trata del Código de la División Político Administrativa del país: Departamentos y Municipios de Colombia. Esta información es fundamental para el desarrollo de la tarea ya que se trata de monitorear las ventas de un producto tanto a nivel municipal como departamental en Colombia.



**Acceso a los datos:** <https://www.datos.gov.co/Mapas-Nacionales/Departamentos-y-municipios-de-Colombia/xdk5-pm3f/data>

**Archivos adjuntos**:

Se entregan a los estudiantes los archivos (artefactos) siguientes (necesarios para la realización de la tarea)



**DESCRIPCIÓN DE LOS ARCHIVOS ENTREGADOS**

|  |  |  |
| --- | --- | --- |
| **Descripción archivos** | **Nombre físico** | **Observaciones** |
| **Script SQL** de creación de tablas de la base de datos | **script-base-datos-creacion.sql** | Script SQL para creación de las tablas de la base de datos. Incluye la carga de la tabla productos. |
| **Script SQL** de inserción de registros de operaciones de venta | **script-base-datos-operaciones.sql** | Script SQL de inserción de 1.000.000 registros de operaciones ventas de productos por municipio. ATENCIÓN: los registros van en la tabla de “operaciones” |
| **Script SQL** de creación de la vista (VIEW) de la tabla operaciones | **script-base-datos-vista.sql** | Esta vista es una consulta SQL prediseñada para utilizar en las consultas y análisis de datos. **NOTA**: Su uso es **OBLIGATORIO** |
| **Consultas SQL** de selección de registros con fechas válidas y NO válidas. | **script-base-datos-valida-fecha.sql** | Este archivo contiene las consultas para la selección de los registros con formato de fecha DIFERENTE al formato válido “AAAA-MM-DD”. Es un instrumento que sirve para detectar los registros con problemas de formato de fecha y proceder a su corrección. |
| **Fuente de Datos:** Hoja de Cálculo con la información oficial de departamentos y municipios (DANE) | **colombia-dane-departamentos.csv** | Hoja de cálculo con departamentos y municipios obtenida del repositorio del DANE |
| **Algoritmo Python.** Programa ETL para cargar departamentos y municipios | **algoritmo-etl.py** | Programa en Python que realiza un proceso ETL. Se encarga de extraer, transformar y cargar los departamentos y municipios de la hoja de cálculo en la base de datos. En este algoritmo se realiza la modificación para valorizar el campo “id\_region” en tabla operaciones. |
| **Algoritmo Python**. Programa para cargar la tabla ***“tamanio***” de manera aleatoria | **algoritm-calculo-tamanio.py** | Carga aleatoria de registros en tabla “***tamanio***” para experimentar con los tiempos de procesamiento y el crecimiento del tamaño de las tablas de una base de datos en un proceso ETL de Big Data. |
| **Plantilla Informe**. Documento para entrega de informe de resultados | **2024-2-bigdata-et0155-G0100-tia-02-informe-equipo-X.docx** | Recuerde sustituir la letra “X” por el número del grupo de estudiantes. al que pertenece y que se conformó previamente en clase. NOTA: Se hará una sola entrega de un informe por grupo de estudiantes. Debe estar bien identificado. |

**INSTRUCCIONES PARA RESOLVER LA TAREA**

Esta tarea es una perfecta introducción al mundo del Big Data. Las actividades a realizar fortalecerán sus conocimientos en base de datos, extracción de información de fuentes de datos, Técnicas ETL, limpieza de datos, almacenamiento y visualización que se abordará durante todo el curso.

La información que se provee a continuación es crucial para resolver las preguntas de la tarea. Lea detenidamente cada uno de los enunciados antes de comenzar a resolver la tarea:

1. Analizar el Caso de Estudio, la hoja de cálculo, los scripts SQL, y los algoritmos suministrados.
2. Al elaborar el diagrama de Entidad-Relación (E-R) asegúrese que es un diagrama de Chen (E-R). Para lograr esto, debe determinar cuáles son las entidades, las relaciones y los atributos. Debe definir cuáles son las claves primarias y las claves únicas. Utilice los símbolos que se usan para diseñar una diagrama de Chen (NO un diagrama “Pata de Cuervo”)
3. El diccionario de datos debe mostrar los cuadros de las “tablas” separados por un espacio. Cada tabla debe contener la clave primaria, las claves únicas y las claves foráneas; así como los tipos de dato y los tamaños de los campos.
4. El formato de dato fecha de operación de venta correcto es **AAAA-MM-DD** (año-mes-dia). En las operaciones que se cargaron, algunos registros no presentan ese formato de fecha (hay formatos de fecha diferentes). El analista de datos debe hacer una propuesta para corregir esta situación antes de migrar la información a la base de datos temporal o puede migrar la información y corregir después de cargada (ETL o ELT). Esta situación es parte del proceso relacionado como “***Transformación de datos***”. Debe hacer el esfuerzo de transformar (corregir) todos aquellos datos que sea posible (TODOS se pueden corregir). La tarea incluye las instrucciones SQL para determinar cuáles fechas están en formato incorrecto.
5. Algunos registros no contienen la información completa. ¿Qué propuesta tiene UD. para corregir esta situación? Nota: es posible que este dato no se pueda migrar en el eventual almacenamiento de servicios en la base de datos temporal. Esta actividad forma parte del proceso de “***Limpieza de Datos***”. No necesariamente significa que hay que borrar registros.
6. Algunos registros, aunque no tienen la información completa, es posible rescatarlos a través de “***Técnicas de Imputación***” (referenciando algunos de los datos presentes).
7. Investigue y practique consultas SQL dónde se utiliza ***distinct, SUM, Group By, Order by***. Tendrá que utilizarlas estos elementos durante la construcción de las consultas. Verifique siempre que las consultas que realiza cuando utilice la VIEW “***vista\_operaciones***” sean los mismos resultados que las consultas directas a la tabla operaciones.
8. En la sección de gráficos, investigue qué son, para que sirven y cómo se construyen los Diagramas de Pareto. Los resultados de las consultas tendrán que llevarlos a Excel para poder construir los gráficos.
9. **ATENCIÓN**: Revise bien la rúbrica presente en el documento de requerimientos para que sepa la manera en que se evaluará la tarea.
10. **SOLO SE RECIBIRÁN TRABJOS EN GRUPO. NO SE RECIBIRÁN ENTREGAS INDIVIDUALES**

**Ruta sugerida de desarrollo de la tarea**:

**Se sugiere a los estudiantes los siguientes pasos de ejecución de actividades para culminar la tarea. No son obligatorios. Cada uno puede establecer su estrategia:**

1. Descargar e instalar todas las aplicaciones de la caja de herramientas
2. Descargar en una carpeta de su PC todos los programas y scripts suministrados por el docente.
3. Instalar las librerías de Python para interactuar con PostgreSQL y la Hoja de Cálculo del DANE:. Debe correr los siguientes comandos en “Símbolo de Sistema” o “Terminal” de Visual Studio Code:
   1. **pip install psycopg2**
   2. **pip install csv**
4. Al instalar PostgreSQL, debe crear la base de datos “***bigdata***” y asignar como ***usuario*** “***postgres***” y como ***password*** “***postgres***”. El ***puerto*** ***5432*** viene por defecto. Debe realizar las siguientes operaciones desde el pgAdmin (utilice los scripts SQL que le fueron suministrados por el docente):
   1. Correr el script SQL de creación de las tablas y alimentación de la tabla de productos
   2. Alimentar la tabla “operaciones” con el script **script-base-datos-operaciones.sql**
   3. Crear la vista “***vista\_operaciones***”. Esta vista es fundamental para construir las consultas.
   4. Crear las funciones de validación de fechas
   5. Crear la tabla regiones
   6. Modificar la tabla operaciones para agregar el campo “id\_region”.
5. Elaborar el diagrama de flujo del algoritmo (“**algotimo-etl.py”**) para conocer mejor su funcionamiento
6. Modificar el algoritmo de Python (“**algotimo-etl.py”**) para incluir la valorización del campo “id\_region”
7. Correr el programa python “**algotimo-etl.py”.** Este programa se encarga de extraer la información de la hoja de cálculo del DANE y cargar la tabla temporal. Después de que se carga la tabla temporal, el algoritmo carga las tablas de departamentos y municipios.
8. Realizar el “Data Munging” o “Data Clenasing”. Corregir todos los problemas existentes en la tabla operaciones
9. Verificar que la VIEW vista\_operaciones arroja el mismo número de registros que la tabla operaciones. Eso certifica que no hay problemas de relacionamiento.
10. Construir las consultas SQL
11. Realizar los gráficos con los resultados de las consultas
12. Realizar el análisis y las conclusiones
13. Revisar que el INFORME que se entrega es de calidad: Presentación, Estructura, Organización, Redacción, Ideas, Resultados, entre otros.
14. Elaborar un video con las especificaciones solicitadas