**INFORME DE RESULTADOS**

**Unidad 3 - Tarea 7**

**Peso: 20%**

**Equipo/Grupo :** “N” (cambiar por el número de grupo)

**Estudiantes :**

**-Keisy Yasiry Valoyes Jimenez**

**-**

**Objetivo General**

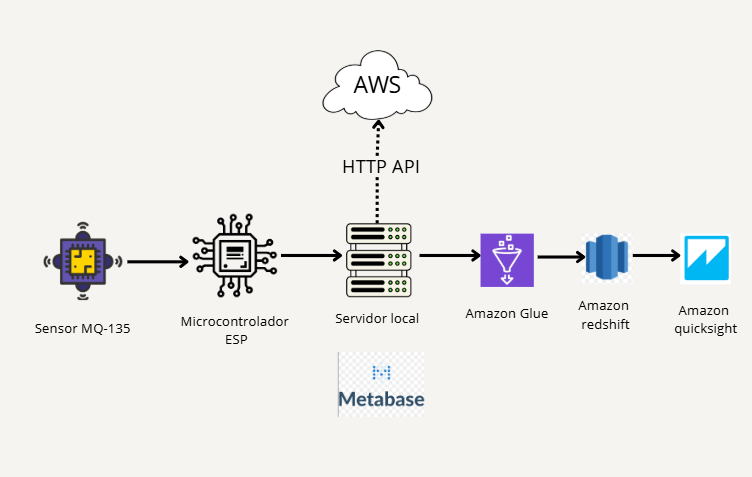
**Diseñar una Arquitectura de Sistema de Información para el procesamiento de Big Data** que esté estructurado en una serie de componentes, tanto hardware como software, que permitan la implementación de un Proceso ETL (Extracción, Transformación, Carga), Analítica, Visualización de información del monitoreo, y la generación de alarmas durante el proceso de producción y emisión de gases tóxicos de las fábricas de la empresa “Sustancias Locas”; en particular, el Benceno que representa un peligro para la Salud humana y el medio ambiente.

Adicionalmente, se plantea la posibilidad de inclusión del total de fábricas de la empresa. La cantidad de datos que se incorporan con la integración de todas la fábricas, requiere de una modificación a la Arquitectura original para incluir mayor capacidad y potencia de cómputo. En este caso, la evaluación de Hadoop como sistema de almacenamiento distribuido y escalable.

En ese mismo orden de ideas, se requiere una prueba de esfuerzo y rendimiento del sistema para llevar a cabo almacenamiento y visualización en tiempo real a través de una simulación del proceso completo: ETL, Analítica y Visualización..

**1.- Diseño Arquitectura del Sistema de Información** (para procesamiento Big Data)

*Colocar en esta sección el diseño de arquitectura del sistema de información. Gráfico con todos los componentes y elementos, tanto hardware como software, requeridos: desde la captura de emisiones de gases hasta la visualización de la información en el Dashboard.*

**

**2.- Cálculos de procesamiento y almacenamiento en la tabla lecturas**

*Colocar en esta sección los resultados de los cálculos. Nota: utilizar el tamaño de una tupla de la tabla de lecturas:*

**2.1.-Cálculos del procesamiento de lecturas**

|  |  |  |
| --- | --- | --- |
| **Tipo de Lectura** | **Cálculos** | **Total lecturas** |
| Lecturas diarias de un sensor | 6 lecturas/min \* 60min \* 24h | 8,640 |
| Lecturas de un mes de una línea de producción. Nota: un mes = 30 días. | 8,640 \* 30 | 259,200 |
| Lecturas de un año de todas las líneas de producción de todas las fábricas. Nota: un año = 365 días. | 8,640 \* 365 \* 4 líneas \* 3 fabricas | 37,843,200 |

***Nota: “Cálculos” debe mostrar el procedimiento para calcular el total de lecturas***

**2.2.- Almacenamiento en tabla “lecturas”**

|  |  |  |  |
| --- | --- | --- | --- |
| **Período de almacenamiento** | **Lecturas** | **Tamaño Tupla** | **Total Bytes** |
| Un (1) minuto | **6** | 99 | 594 |
| Un (1) hora | **360** | 99 | 35,640 |
| Un (1) día | **8,640** | 99 | 855,360 |
| Un (1) mes | **259,200** | 99 | 25,660,800 |
| Un (1) año | **37,843,200** | 99 | 3,746,476,800 |

***Nota: “Tamaño Tupla” es el tamaño en BYTES del registro de la tabla “lectura” de la base de datos “monitoreo-produccion” de PostgreSQL***

**2.3.- Almacenamiento en tabla “lecturas”**

|  |  |  |  |
| --- | --- | --- | --- |
| **Lote de “lecturas”** | **Lecturas** | **Tamaño Tupla** | **Total Bytes** |
| Lote 1 | **3.000.000** | 99 | 297,000,000 |
| Lote 2 | **20.000.000** | 99 | 1,980,000,000 |

***Nota: “Tamaño Tupla” es el tamaño en BYTES del registro de la tabla “lectura” de la base de datos “monitoreo-produccion” de PostgreSQL***

***2.4.- ¿Cada cuánto tiempo se debe limpiar la hoja de cálculo “lector-sensor” y la hoja de cálculo “lector-fabrica” antes de que se llegue al límite del máximo de registros permitidos por hoja de cálculo con formato “xlxs? Explique brevemente como realizó ellos cálculos.***

***La hoja de cálculo “lector\_sensor” puede almacenar hasta 1,048,576 registros, que es el límite máximo permitido por el formato xlxs.***

***Como cada sensor genera 8,640 lecturas por día (una cada 10 segundos), la hoja alcanzaría su límite en aproximadamente 121 días,***

***Por lo tanto, se recomienda hacer la limpieza de archivo cada 120 días como medida previa.***

***En el caso de la hoja de cálculo “lector\_fabrica”, que consolida la información de 80 sensores por fabrica, el volumen de registros es mucho mayor:***

***8,640 lecturas \* 80 sensores = 691,200 lecturas por día***

***1,048,576/ 691,200 ≈ 1.5 días***

***Por lo tanto, esta hoja debe limpiarse cada día o día y medio máximo para evitar pérdida de datos.***

**3.- Cálculo del costo de almacenamiento de en AWS de 20 millones de registros.**

*Colocar en esta sección los cálculos de los diferentes tipos de almacenamiento y el presupuesto (cotización) obtenida con la calculadora de AWS de EBS. Debe calcular el tamaño en Bytes de 20 millones de registros de la tabla “lecturas” de la base de datos “monitoreo-produccion”*

**3.1.- Cálculo de almacenamiento**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tipo Almacenamiento** | **Total Bytes** | **Costo x byte** | **Costo Total** |
| Almacenamiento en Bloque Elástico (EBS) | 1,980,000,000 | 0.08 USD/GB | 0.15 USD |
| Amazon Simple Storage Service (Amazon S3) | 1,980,000,000 | 0.023 USD/GB | 0.043 USD |
| Amazon Aurora | 1,980,000,000 | 0.10 USD/GB | 0.18 USD |

**3.2.- ¿Por qué la diferencia de costos en los diferentes tipos de almacenamiento?**

*Describir brevemente las razones por las cuáles los costos son diferentes para los tipos de almacenamiento solicitados*

*Los costos varían por el tipo de servicio que ofrece cada almacenamiento.*

*EBS está optimizado para almacenamiento en bloque y es ideal para bases de datos que requieren rendimiento estable.*

*S3 es almacenamiento de objetivos, más baratos, pero no ofrece acceso rápido a registros individuales como una base*

*de datos.*

*Aurora es un sistema de base de datos totalmente administrado y escalable, lo que lo hace más costoso por sus*

*capacidades avanzadas.*

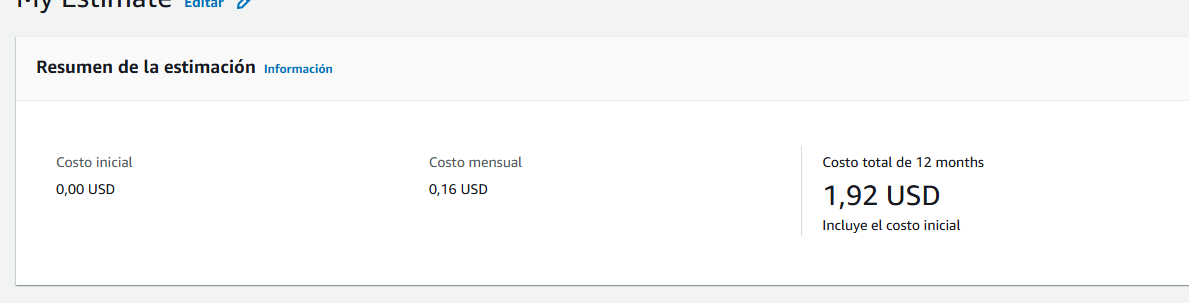
**3.3.- Cotización AWS para el tipo de almacenamiento en EBS**

*Colocar la cotización que elaboró con la calculadora de AWS para el tipo de almacenamiento EBS*

*Se realizó una simulación en la calculadora de precios de AWS para estimar el costo mensual del almacenamiento*

*requerido para 20 millones de registros en la tabla lecturas, equivalentes a aproximadamente 2 GB.*

*Se utilizó el servicio Amazon EBS en la región US East con 2 GB de almacenamiento, sin snapshost ni servicios adicionales*



**4.- Cálculo del costo por 1 año de uso de los servicios AWS**

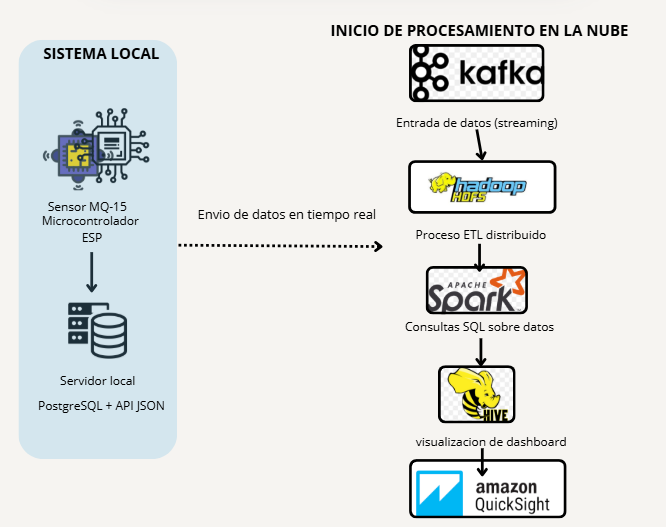
*Colocar en esta sección los cálculos del presupuesto (cotización) obtenida con la calculadora de AWS*

**4.1.- Cálculo del costo de los servicios**

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre Servicio** | **Costo individual** |  | **Costo anual** |
| Amazon EBS (2GB, tipo gp3) | $0.16 USD |  | $1.92 USD |
| Amazon S3 (2GB Standard) | $0.05 USD |  | $0.60 USD |
| AWS Glue | $1.00 USD |  | $12.00 USD |
| Amazon Redshift (nodo mínimo) | $3.00 USD |  | $36.00 USD |
| Amazon QuickSight (1 usuario) | $9.00 USD |  | $108.00 USD |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**5.- Diseño de Arquitectura de Sistema de Información Big Data con Hadoop**

*Dado que el proyecto fue exitoso, la corporación desea integrar sus 10 fábricas que capturan las emisiones en ppm en un solo ambiente. ¿Cómo integrarías Apache Kafka, Hadoop, Apache Spark o Apache Hive, Amazon QuickSight para hacer el proceso completo ETL, Analítica y Visualización? NOTA: El sistema local permanece intacto. Todo comienza a partir de Apache Kafka del lado de Computación en la Nube. En base a lo antes mencionado, proponga una Diseño de Arquitectura que integre estos elementos al procesamiento local que ya realizó. Elabore un buen diseño y explique en una leyenda al pie del Diagrama de Arquitectura los elementos del lado de la NUBE para que el lector comprenda cuál es la función de cada componente.*

****

**6.- Algoritmo de “poblamiento” de la tabla “lecturas” y la hoja de cálculo “lecturas-sensor”**

*Elabore un algoritmo que genere de manera aleatorio los datos necesarios para poblar la tabla “lecturas” de la base de datos y la hoja del cálculo del sensor “lecturas-sensor”. Se sugiere tomar como punto de partida el algoritmo ETL y el algoritmo de cálculo de tiempo y almacenamiento de la primera Tarea; y adaptar el programa Python al nuevo requerimiento.*

*Coloque un pantallazo en esta sección y entregue el código fuente con los otros productos solicitados*

import random

import pandas as pd

from datetime import datetime, timedelta

*def* generar\_datos\_lecturas(*n*=20):

    datos = []

    fecha\_base = datetime.now()

    sensores = ["A1S01", "A2S05", "B3S09", "C1S02"]

    microcontroladores = ["A1M01", "A2M05", "B3M09", "C1M02"]

    lineas = ["A1", "A2", "B3", "C1"]

    fabricas = ["A", "B", "C"]

    for i in range(*n*):

        fecha = (fecha\_base - timedelta(*seconds*=i\*10)).date().isoformat()

        hora = (fecha\_base - timedelta(*seconds*=i\*10)).time().strftime("%H:%M:%S")

        sensor\_id = random.choice(sensores)

        micro\_id = random.choice(microcontroladores)

        linea = random.choice(lineas)

        fabrica = random.choice(fabricas)

        concentracion = round(random.uniform(0.1, 600.0), 2)

        if concentracion <= 0.5:

            nivel = "Exposición leve"

        elif concentracion <= 1:

            nivel = "Exposición permitida"

        elif concentracion <= 10:

            nivel = "Moderada toxicidad"

        elif concentracion <= 50:

            nivel = "Peligroso"

        elif concentracion <= 500:

            nivel = "Altamente peligroso"

        else:

            nivel = "Peligro extremo"

        latitud = round(random.uniform(6.2, 6.3), 6)

        longitud = round(random.uniform(-75.6, -75.5), 6)

        datos.append([fecha, hora, sensor\_id, micro\_id, linea, fabrica,

                      concentracion, nivel, latitud, longitud])

    df = pd.DataFrame(datos, *columns*=[

        "Fecha", "Hora", "Sensor\_ID", "Microcontrolador\_ID", "Linea\_Produccion",

        "Fabrica", "Concentracion\_ppm", "Nivel\_Riesgo", "Latitud", "Longitud"

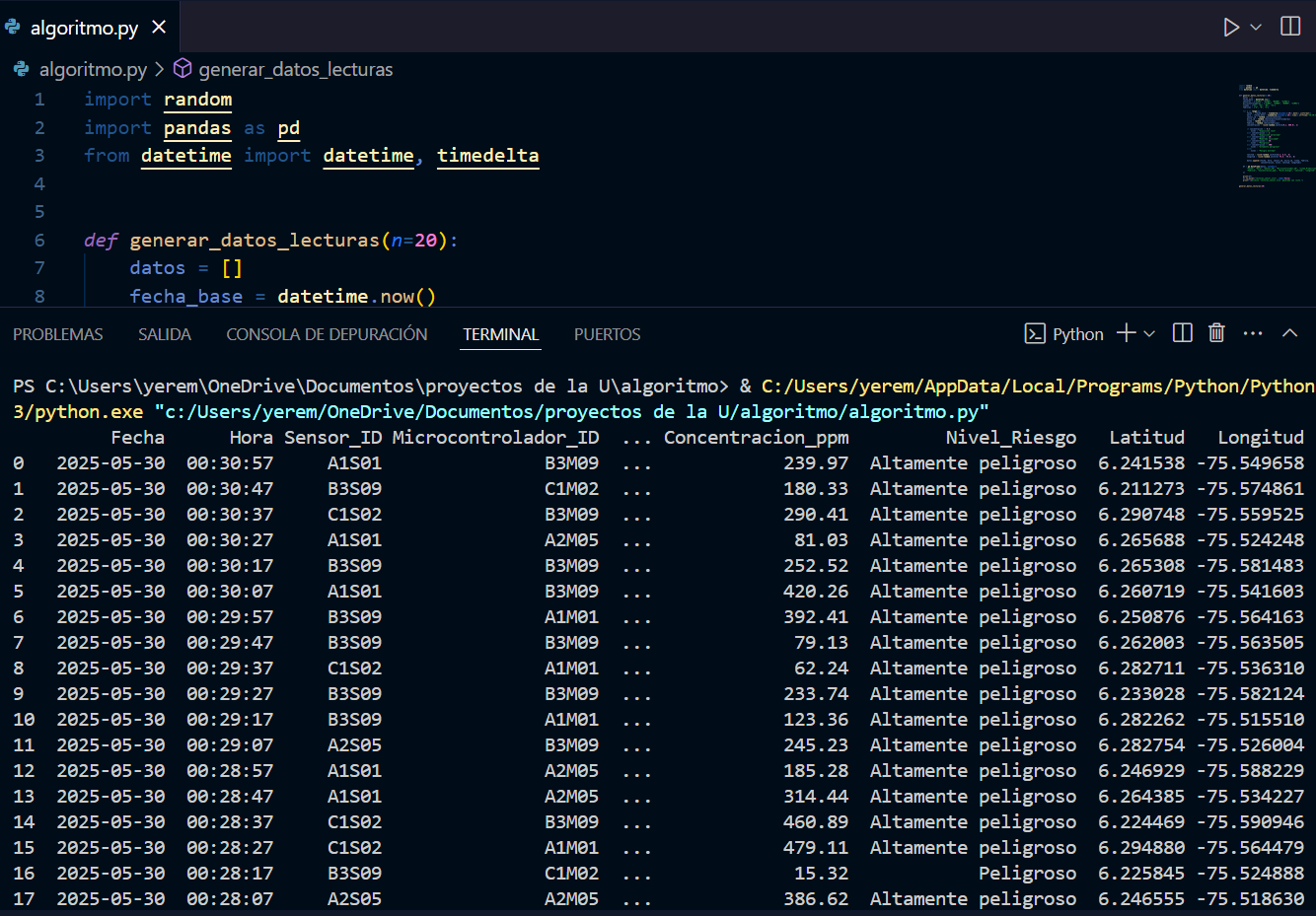
    ])

    print(df)

    df.to\_excel("lecturas\_sensor.xlsx", *index*=False)

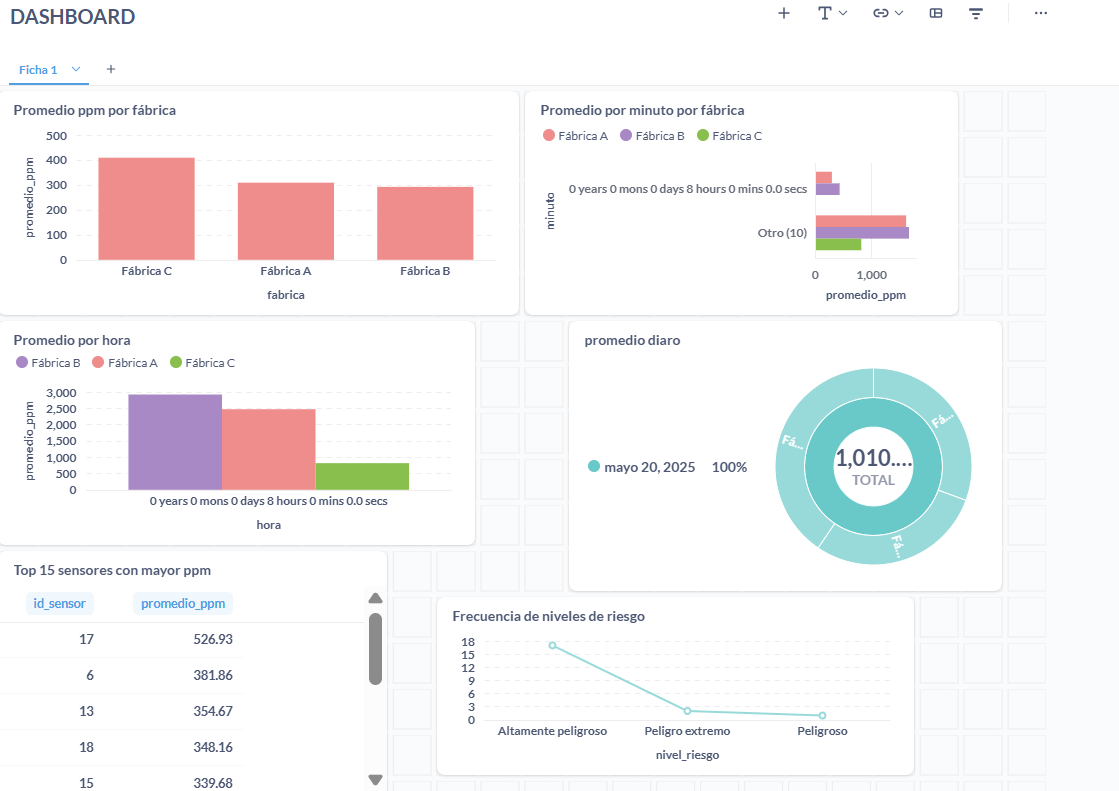
    print("\nArchivo 'lecturas\_sensor.xlsx' generado con éxito.")

generar\_datos\_lecturas(20)



**7.- Implementación de un tablero de control y monitoreo con la herramienta “Metabase”**

*Elabore un tablero de control y monitoreo que contenga los gráficos que se encuentran en los requerimientos del documento de instrucciones. Coloque un pantallazo del tablero en esta sección. NOTA: recuerde que durante la grabación del video de sustentación, debe mostrar el tablero (Dashboard) en funcionamiento.*



**8.- Conclusiones.**

*Elabore las conclusiones planteando la importancia y utilidad de esta tarea y su relación con el contenido de la asignatura. ¿Cuáles eran sus expectativas al inicio de la asignatura y en este momento final? ¿Cómo cree Ud. que el conocimiento aprendido en la asignatura afectará sus oportunidades laborales y desempeño profesional una vez obtenga su título en el Pascual Bravo?*

*Adicionalmente a estas conclusiones generales del grupo de trabajo, cada estudiante debe expresar sus propias conclusiones.*

*En comparación con las tareas anteriores, esta no me pareció tan exigente, lo cual agradecí bastante. Me gustó mucho volver a trabajar con Metabase, ya que para mí es una herramienta más sencilla y cómoda de utilizar. También fue interesante volver a poner en práctica el algoritmo de poblamiento en Python. En general, fue una buena actividad para cerrar el curso, aunque como siempre, estuve algo estresada con la fecha de entrega, ya que las tareas de Big Data suelen llevarme varios días de trabajo y algunas trasnochadas. A pesar de eso, siento que aprendí mucho, y valoro el conocimiento adquirido porque sé que será útil para mi futuro profesional como tecnóloga en desarrollo de software.*

**9.- Video de sustentación:**

*Elabore un video de sustentación con la participación de todos los integrantes (si es en equipo). Este vídeo debe informar sobre las actividades realizadas en general, pero PRINCIPALMENTE debe mostrar el diseño y la explicación detallada de los componentes de las dos arquitecturas solicitadas; así como el flujo de la información desde la captura hasta el despliegue en un Tablero de Control. Debe tener buena calidad de sonido.*

*https://github.com/Kei-2005/TIA7.git*