**INFORME DE RESULTADOS**

**Unidad 3 - Tarea 7**

**Peso: 20%**

**Equipo/Grupo : 3**

**Estudiantes :**

**- *Luis Felipe Gil Gómez***

* ***Gerson Gustavo Fernández Badillo***
* ***Davidson Harley Rave Buitrango***
* ***David Pabón García***

**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)**

**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 por minuto x 60 min x 24H =8.640 lecturas/dia | 8.640 |
| Lecturas de un mes de una línea de producción. Nota: un mes = 30 días. | 8640 lecturas/día x 30 dias x 20 sensores = 5.184.000 | 5.184.000 |
| Lecturas de un año de todas las líneas de producción de todas las fábricas. Nota: un año = 365 días. | 8.640 lecturas/día x 365 dias x 3 fabricas x 4 lineas x 20 sensores = 756.864.000 | 755.136.000 |

***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 lecturas** | 65 bytes | 390 |
| Un (1) hora | **6 x 60 = 360** | 65 bytes | 23.400 |
| Un (1) día | **8.640** | 65 bytes | 561.600 |
| Un (1) mes | **259.200 (8.640 x 30)** | 65 bytes | 16.848.000 |
| Un (1) año | **3.153.600 (8.640 x 365)** | 65 bytes | 204.984.000 |

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

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Lote de “lecturas”** | **Lecturas** | **Tamaño Tupla** | **Total Bytes** |
| Lote 1 | **3.000.000** | 65 | 195.000.000 |
| Lote 2 | **20.000.000** | 65 | 1.300.000.000 |

***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.***

***1-***Una hoja de calculo XLSX tiene un limite de 1.048.576 filas

***2-***Cada sensor genera 8.640 lecturas por día.

***3-***Tiempo hasta alcanzar el limite

= 120.3 dias → Limpiar cada 120 dias (4 meses) para prevenir sobrecarga.

Para la hoja “lecturas-fabrica” que centraliza datos de 80 sensores:

***1-***80 x 8.640 = 691.200 lecturas/día

= 1.5 dias → Limpiar diariamente.

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

**3.1.- Cálculo de almacenamiento**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tipo Almacenamiento** | **Total Bytes** | **Costo x byte** | **Costo Total** |
| Almacenamiento en Bloque Elástico (EBS) | 1.21 GB | $0.00000000492 | $6.39 |
| Amazon Simple Storage Service (Amazon S3) | 1.21 GB | $0.0000000000385 | $0.05 |
| Amazon Aurora | 1.21 GB | $0.0000003503 | $455.05 |

**Nota:** Se considera que 20.000.000 registros x 65 bytes = 1.300.000.000 bytes ≈ 1.21 GB

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

En EBS esta diseñado para almacenamiento de bloques es más rápido y escalable, donde los datos son muy variables y se necesita de una respuesta más rápida a la hora de intervenir, en S3 es un almacenamiento barato para datos estáticos o archivados, estos datos no tendrán modificaciones, sino que guardaran información que requiera solo consulta, El Aurora es un servicio de gestión de base de datos con redundancia y rendimiento optimizado

La diferencia de costos entre los servicios de almacenamiento de AWS se debe principalmente al tipo de servicio, nivel de rendimiento, durabilidad, disponibilidad y características especializadas que ofrece cada uno:

**Amazon EBS (Elastic Block Store):**

Este servicio de almacenamiento es usado como discos duros virtuales. Es rápido, de baja latencia y adecuado para sistemas operativos, bases de datos y aplicaciones que requieren acceso frecuente y de alta velocidad a los datos. Su costo es favorable, mostrando su rendimiento y disponibilidad.

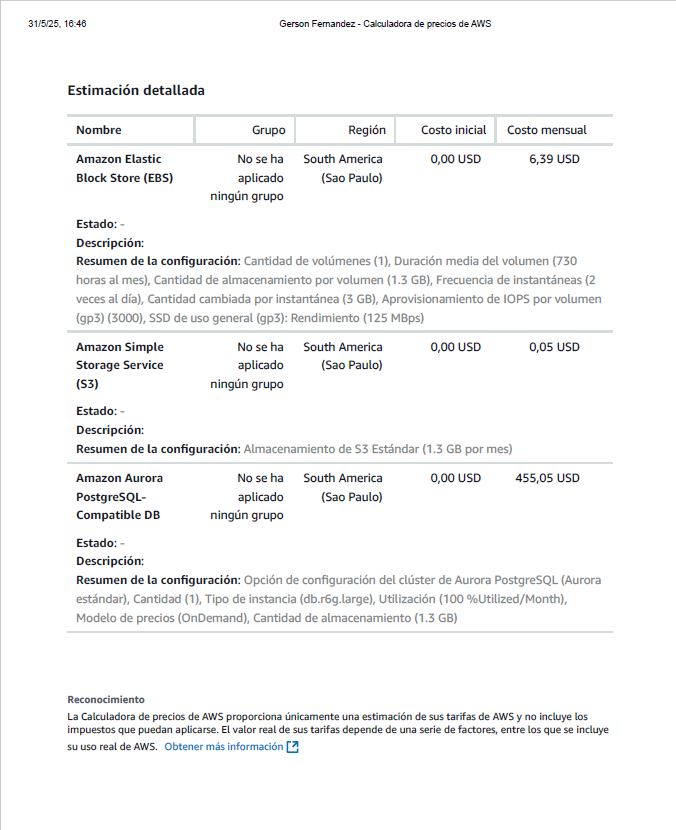
**Amazon S3 (Simple Storage Service):**

Es un servicio de almacenamiento orientada a objetos escalable y económico, ideal para almacenar grandes volúmenes de datos que no necesitan modificaciones frecuentes. Es el mas economico porque está optimizado para almacenamiento masivo con acceso menos frecuente.

**Amazon Aurora:**

Aurora es un sistema de gestión de bases de datos relacional administrado, que combina la velocidad y disponibilidad de bases de datos comerciales con la simplicidad y economía de código abierto. Su costo es considerablemente más alto porque incluye no solo almacenamiento, sino procesamiento, replicación automática, respaldos continuos, y tolerancia a fallos. Es ideal para aplicaciones críticas que requieren alta disponibilidad y rendimiento.

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



**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**

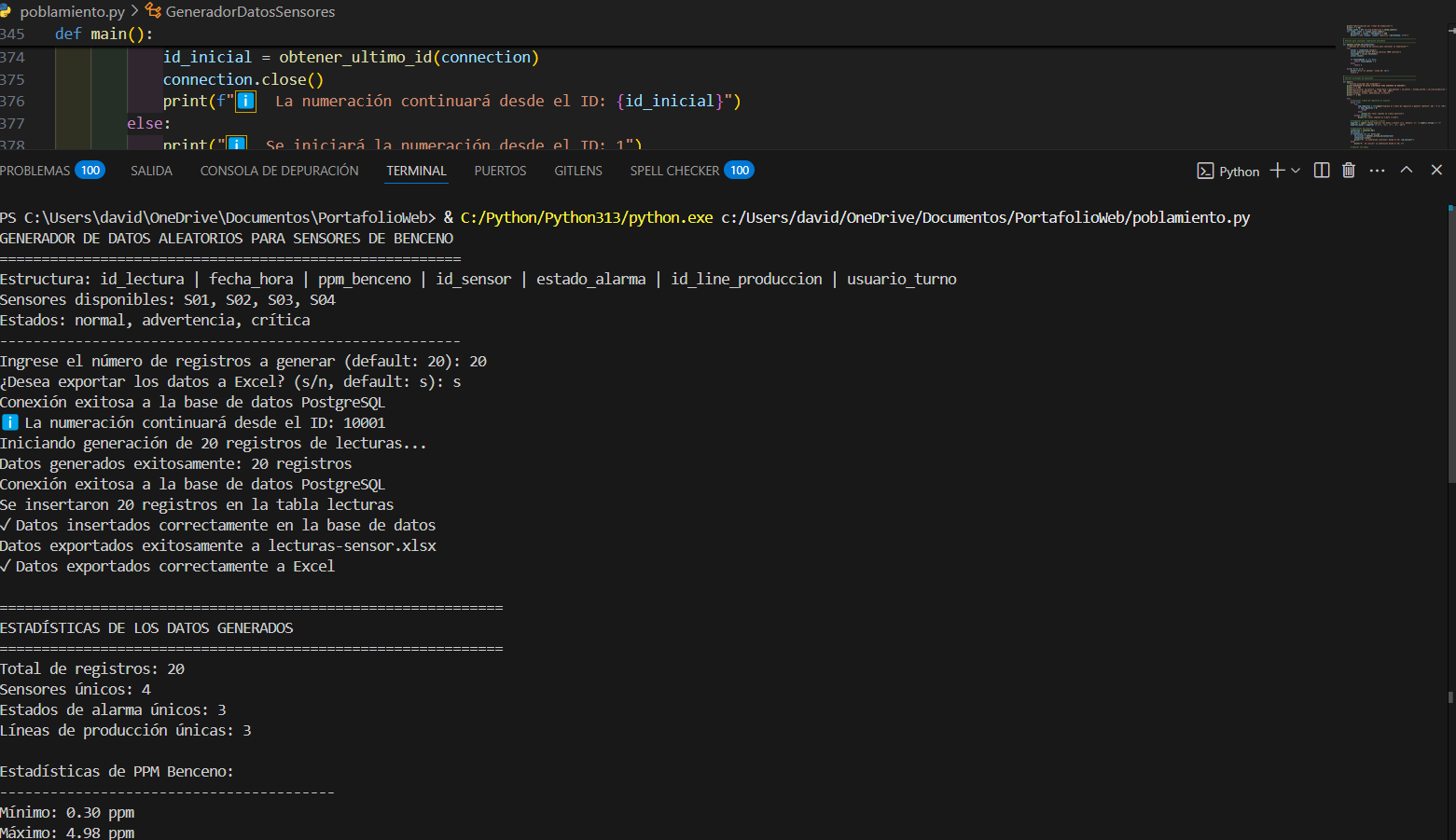
|  |  |  |
| --- | --- | --- |
| **Nombre Servicio** | **Costo individual** | **Costo anual** |
| Amazon EC2 | $29.44 | $353.28 |
| Amazon S3 | $0.05 | $0.60 |
| Amazon EBS | $6.39 | $76.68 |
| Amazon Aurora | $455.05 | $5,460.60 |
| AWS Lambda (1 millón de invocaciones/mes) | $0.20 | $2.40 |
| Amazon CloudWatch (monitoreo básico) | $3.00 | $36.00 |
| Amazon VPC (tráfico y NAT Gateway) | $10.00 | $120.00 |
| AWS Glue (procesamiento ETL esporádico) | $5.00 | $60.00 |
| Amazon QuickSight (1 usuario estándar) | $9.00 | $108.00 |
| Amazon EC2 (Instancia t3.medium, 24/7) | $29.44 | $353.28 |

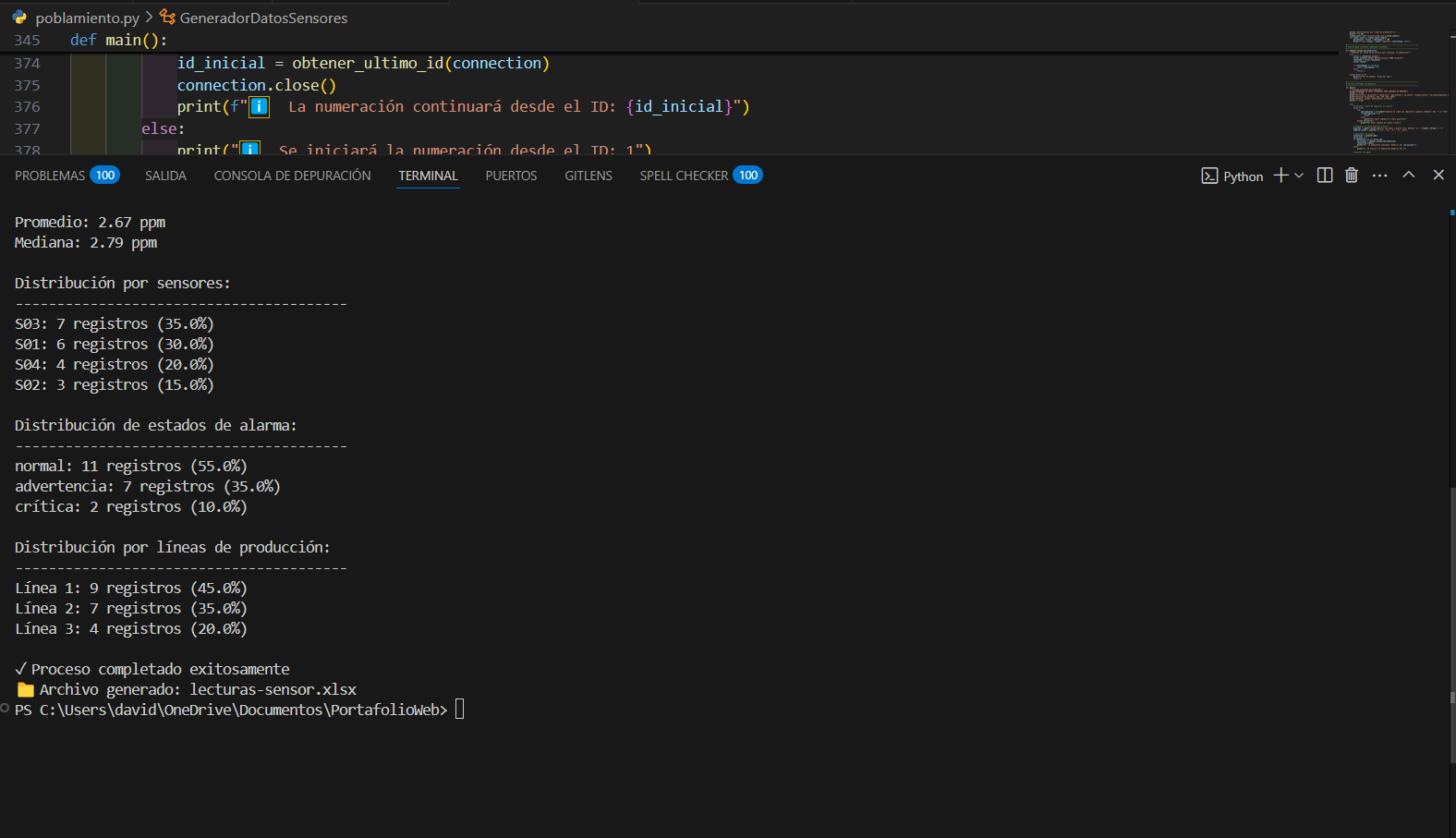
**5.- Diseño de Arquitectura de Sistema de Información Big Data con Hadoop**

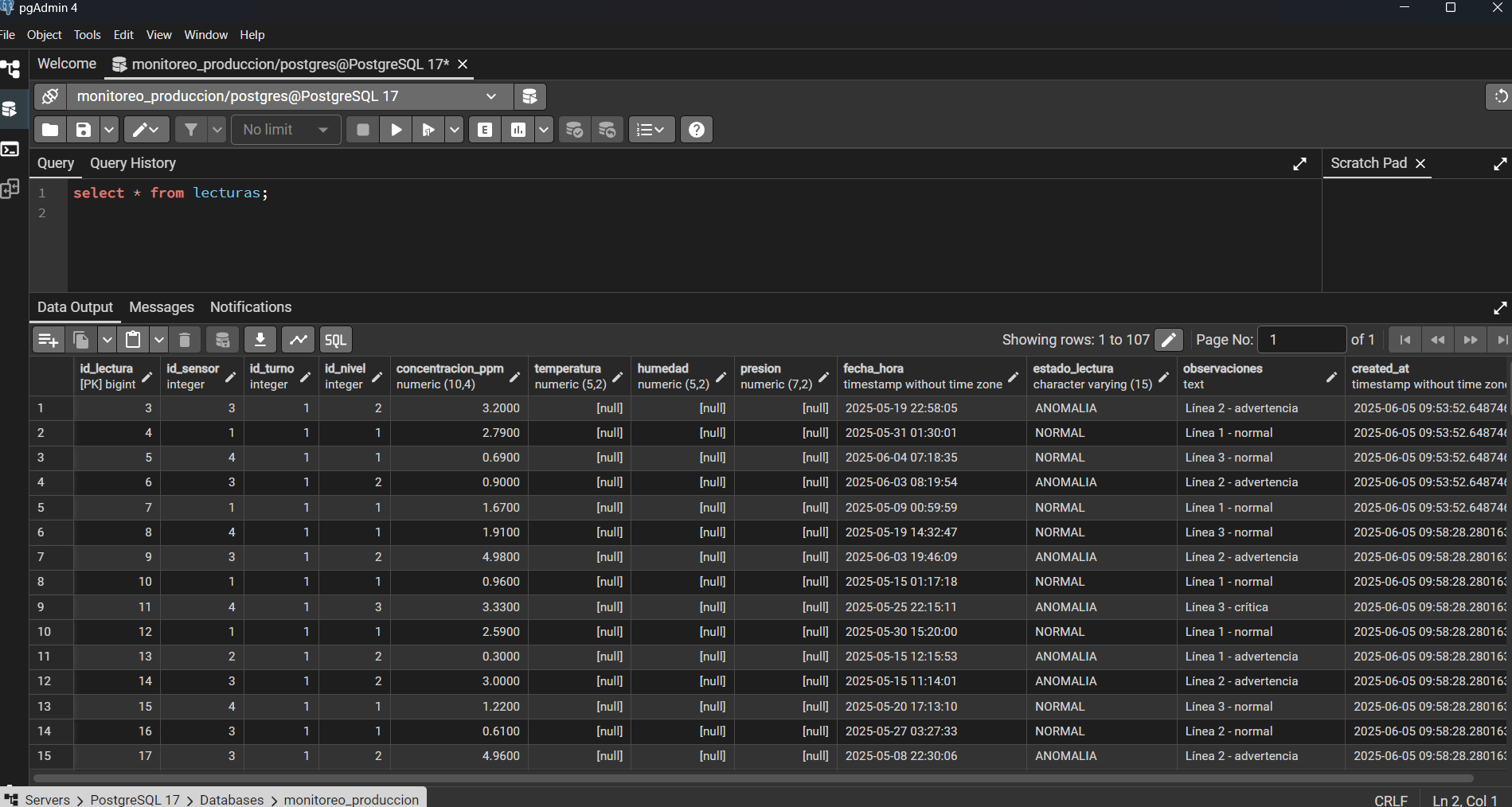
****

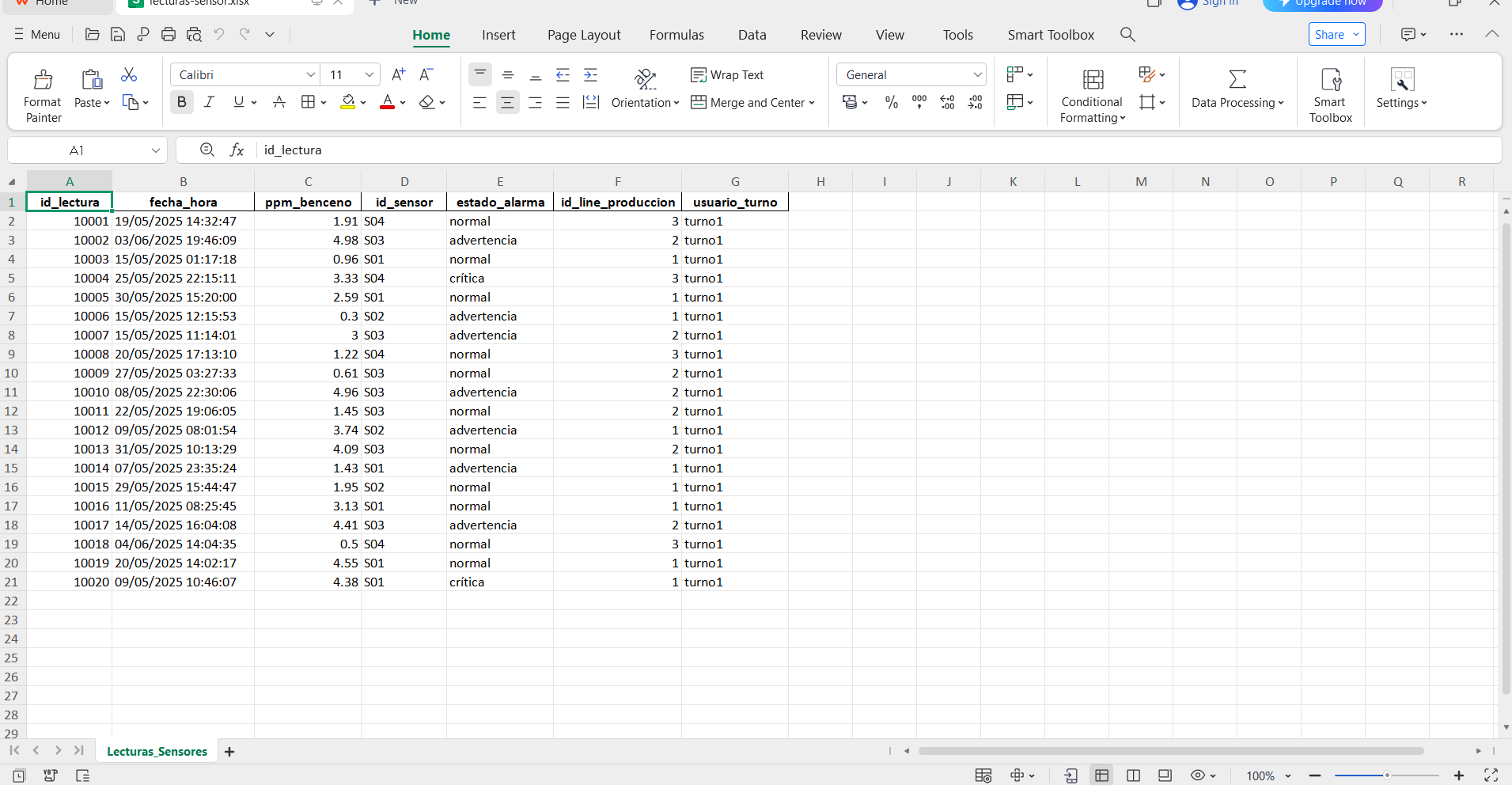
.

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

**

**

**

**

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

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Interfaz de usuario gráfica, Aplicación, Word, Teams

El contenido generado por IA puede ser incorrecto.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Gráfico

El contenido generado por IA puede ser incorrecto.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

**Gráfico

El contenido generado por IA puede ser incorrecto.**

**8.- Conclusiones.**

El desarrollo de este proyecto permitió diseñar e implementar de una arquitectura de sistemas de información basada en Big Data, aplicada al monitoreo en entorno industrial. A través de un proceso completo de captura, transformación, almacenamiento y visualización de datos de sensores, se comprendió de forma práctica el ciclo de vida de los datos en sistemas distribuidos.

La elaboración de cálculos de volumen de datos, consumo de almacenamiento, estimación de costos en plataformas como AWS y el uso de herramientas como Metabase para la visualización, fortalecieron competencias técnicas clave para enfrentar retos actuales en ciencia de datos e ingeniería de software. Adicionalmente, se evaluó la escalabilidad de la solución mediante el uso de tecnologías como Hadoop, orientadas al tratamiento eficiente de grandes volúmenes de información. El trabajo en equipo y la integración de conocimientos permitieron diseñar soluciones viables, sostenibles y alineadas con las necesidades de la empresa, permitiendo una toma de decisiones basada en datos

Gerson Gustavo Fernández Badillo:

Durante el desarrollo de esta actividad pude evidenciar la importancia de una arquitectura bien definida como base para implementar soluciones efectivas en proyectos de Big Data. El diseño de una arquitectura local clara y funcional permite establecer los componentes tecnológicos necesarios para garantizar la conectividad, el procesamiento y el almacenamiento de los datos en tiempo real. De igual manera, la elaboración de una cotización detallada para los recursos tecnológicos y servicios requeridos (puntos 3 y 4), no solo permitió estimar costos reales, sino también alinear la propuesta con la viabilidad técnica y financiera del proyecto. El diseño de la arquitectura de alto nivel en el punto 5 fue fundamental para visualizar de forma integrada el flujo de los datos y los actores involucrados, desde la recolección hasta la explotación analítica. Finalmente, la organización general del trabajo me permitió coordinar las actividades del equipo, asegurar el cumplimiento de las entregas y mantener una línea coherente en el desarrollo del proyecto. Esta experiencia reafirma que el éxito de una estrategia de datos depende tanto de la infraestructura tecnológica como del trabajo colaborativo y la planificación detallada de cada componente.

David Pabón García:

El desarrollo de este trabajo nos permitió comprender cómo funciona una arquitectura de Big data, desde la captura de datos hasta su análisis y visualización, el cálculo real de volumen de datos que se puede generar en un entorno industrial y evaluar costos para los servicios. Adquirimos conocimientos prácticos sobre herramientas de procesamiento y servicios en la nube, lo que nos permitió diseñar soluciones escalables y eficientes. En general este proyecto nos ayudó a implementar la teoría con la práctica enfocado al mundo laboral.

**9.- Video de sustentación:**

<https://youtu.be/fJdk8CpQ6Fk>