**INFORME DE RESULTADOS**

**Unidad 3 - Tarea 6**

**Peso: 10%**

**Equipo/Grupo : NUMERO 2**

**Estudiantes :**

* ***JUAN FELIPE BARRERA GARCIA***
* ***SOFIA RAMIREZ HIGUITA***
* ***JUAN DIEGO GLLARDO VILORIA***

**Objetivo General**

**Diseñar un Gobierno de Datos y la estructura de un Proyecto de Big Data de envergadura para enfrentar la problemática en relación a la implementación de una Arquitectura de Plataforma de control y monitoreo de concentración de gases tóxicos en las fábricas de la empresa “Sustancias Locas”.**

Se requiere la convergencia de talento humano, planificación del trabajo, contratación de diversos servicios, instalación y configuración de aplicaciones y servidores locales y en la Nube, desarrollo de software, soporte técnico, entre otros.

Por lo tanto, es necesario la conformación de un gobierno de datos y la estructuración de un Proyecto de Big Data; así como una lista de verificación del Proyecto Big Data que incluya todos los ítems necesarios y los tiempos estimados para lograr el objetivo.

**1.-Diseño de un Gobierno de Datos**

*Después de analizar el enunciado y evaluar los requerimientos, proponga la estructura y roles que debe tener el gobierno de datos que se encargara del Proyecto Big Data*

**1.1- Roles del gobierno de datos**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Departamento o persona** | **Descripción** | **Rol, responsabilidades y/o funciones** |
| **1** | **Gerente del Proyecto** | Líder del proyecto, encargado de la planificación y ejecución del sistema Big Data. | Definir los objetivos del proyecto, coordinar al equipo, asegurar el cumplimiento de los plazos, y supervisar el uso de los recursos. |
| **2** | **Data Engineer** | Responsable del diseño e implementación de la arquitectura Big Data. | Desarrollar la estructura de almacenamiento de datos, optimizar el procesamiento de grandes volúmenes de datos y gestionar los pipelines de datos (sistemas que conectan los orígenes de datos, los sistemas de procesamiento de datos, las herramientas de análisis y las aplicaciones). |
| **3** | **Data Analyst** | Analista de datos para la interpretación de las lecturas de ppm y emisión de gases. | Analizar los datos recopilados, generar informes, detectar patrones y proporcionar insights para la toma de decisiones. |
| **4** | **Especialista en seguridad de Datos** | Encargado de implementar medidas de seguridad y cumplimiento de normativas. | Asegurar la integridad y confidencialidad de los datos, implementar protocolos de seguridad y gestionar los permisos de acceso. |
| **5** | **Ingeniero de IoT** | Integración de sensores y microcontroladores con el sistema Big Data. | Configurar sensores, monitorear el rendimiento de los dispositivos IoT y asegurar la transmisión de datos en tiempo real. |
| **6** | **Administrador de la Nube** | Gestión de los servicios en la nube (AWS) para almacenamiento y visualización de datos. | Implementar soluciones en la nube, optimizar costos, y asegurar la disponibilidad del sistema. |
| **7** | **Desarrollador de Dashboard** | Creación del Tablero de Control (dashboard) para la visualización en tiempo real. | Diseñar y desarrollar el Dashboard, integrar alertas y reportes en tiempo real y asegurar la usabilidad del sistema. |
| **8** | **Responsable de Calidad** | Verificación de la precisión y confiabilidad de los datos generados. | Realizar pruebas de calidad en los datos, supervisar los niveles de ppm y asegurar el cumplimiento de los estándares. |
| **9** | **Técnico de Mantenimiento** | Supervisión y mantenimiento de los sensores y microcontroladores. | Monitorear los sensores, realizar mantenimientos preventivos y correctivos, y coordinar el recambio de filtros. |
|  |  |  |  |
| **N** |  |  |  |

***1.2.- Estructura gráfica del gobierno de datos***

Interfaz de usuario gráfica, Aplicación, Word

El contenido generado por IA puede ser incorrecto.

**2.-Diseño del Proyecto Big Data**

*Después de analizar el enunciado y evaluar los requerimientos, proponga la estructura y elementos que deben conformar el Proyecto Big Data*

**2.1- Elementos del Proyecto Big Data**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Elemento** | **Descripción** | **Observaciones** |
| **1** | **Sensores IoT** | Dispositivos que capturan los datos sobre la concentración de gases tóxicos. | Se integran a través de microcontroladores ESP8266 y sensores MQ-135. |
| **2** | **Microcontroladores** | Placas electrónicas que procesan los datos obtenidos por los sensores antes de enviarlos a la nube. | Uso de Arduino y ESP8266 para la comunicación. |
| **3** | **Servicios en la nube (AWS)** | Plataforma de almacenamiento y procesamiento en la nube para manejar grandes volúmenes de datos. | Se usarán Amazon Kinesis, Redshift, Glue y CloudWatch. |
| **4** | **Base de datos "monitoreo-produccion"** | Almacén de datos estructurados sobre las lecturas de los sensores. | Uso de PostgreSQL para consultas avanzadas. |
| **5** | **Procesamiento en tiempo real** | Sistema que analiza los datos y genera alertas en caso de niveles críticos de toxicidad. | Implementación con Apache Spark para procesamiento distribuido. |
| **6** | **Dashboard de visualización** | Plataforma web que muestra métricas y tendencias sobre la emisión de gases. | Desarrollo con Amazon QuickSight y APIs HTTP. |
| **7** | **Seguridad de datos** | Protección de la integridad y confidencialidad de la información. | Uso de protocolos de cifrado y gestión de accesos. |

**2.2- Lista de verificación del Proyecto Big Data**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Elemento** | **Estado** | **Tiempo** |
| **1** | **Instalación de sensores IoT** | Pendiente | 2 semanas |
| **2** | **Configuración de microcontroladores** | En proceso | 1 semana |
| **3** | **Implementación de base de datos** | Pendiente | 1 semana |
| **4** | **Integración con AWS** | En proceso | 3 semanas |
| **5** | **Desarrollo del dashboard** | Pendiente | 4 semanas |
| **6** | **Pruebas de seguridad y calidad** | Pendiente | 2 semanas |
| **7** | **Creación de scripts de procesamiento en tiempo real** | Pendiente | 3 semanas |
| **8** | **Configuración de protocolos de seguridad** | Pendiente | 2 semanas |
| **9** | **Desarrollo de APIs para la comunicación entre sistemas** | En proceso | 3 semanas |
| **10** | **Pruebas de rendimiento en la infraestructura en la nube** | Pendiente | 2 semanas |
| **11** | **Capacitación del equipo en el uso del sistema** | Pendiente | 1 semana |
| **12** | **Documentación técnica del sistema** | Pendiente | 2 semanas |
| **13** | **Validación de cumplimiento normativo** | Pendiente | 1 semana |
| **14** | **Implementación de monitoreo con Amazon CloudWatch** | Pendiente | 2 semanas |
| **15** | **Evaluación de mejoras en el modelo de datos** | Pendiente | 3 semanas |
| **16** | **Mantenimiento preventivo de sensores y hardware** | Pendiente | 2 semanas |

**3.- Conceptos y elementos de Big Data**

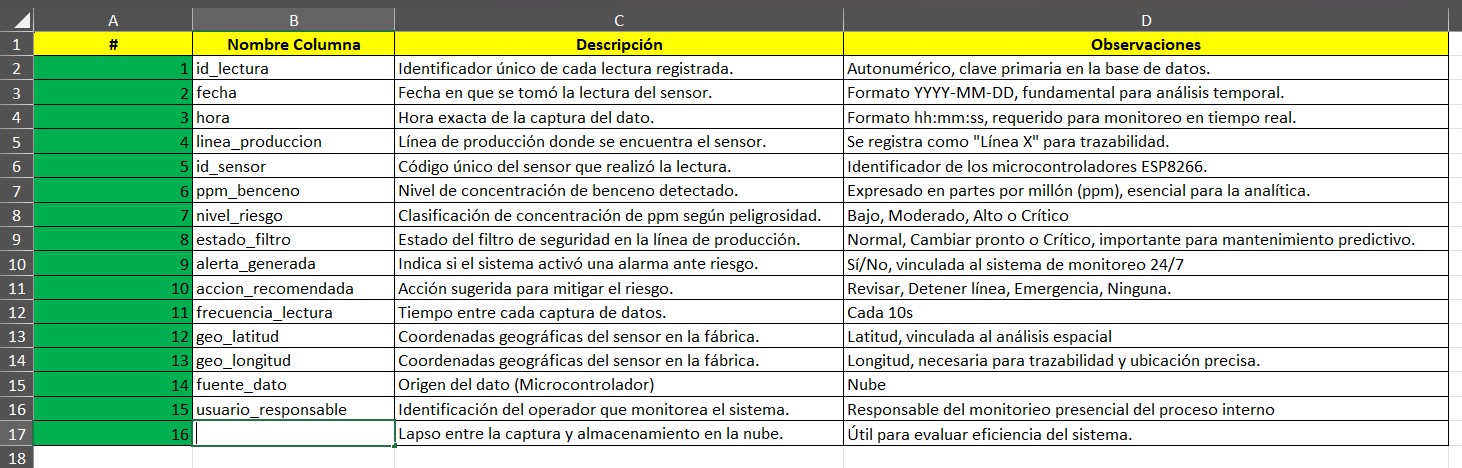
|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Componente** | **Descripción (breve cita bajo Norma APA)** | **Referencia (APA)** |
| **1** | **Arquitectura de Sistema de Información** | Permite planear y diseñar la arquitectura, el ciclo de vida, las aplicaciones, los soportes y la gestión de los sistemas de información que habilitan el cumplimiento de las funciones de una entidad pública. | <https://mintic.gov.co/arquitecturaempresarial/630/articles-237650_recurso_1.pdf> |
| **2** | **Hadoop** | Entorno de trabajo de software de código abierto que permite el procesamiento y análisis distribuido de grandes volúmenes de datos. | <https://es.wikipedia.org/wiki/Apache_Hadoop> |
| **3** | **Spark** | Motor unificado de analíticas para procesar datos a gran escala que integra módulos para SQL, streaming, aprendizaje automático y procesamiento de grafos. Spark se puede ejecutar de forma independiente o en Apache Hadoop, Apache Mesos, Kubernetes, la nube y distintas fuentes de datos. | <https://cloud.google.com/learn/what-is-apache-spark?hl=es> |
| **4** | **PostgreSQL** | Base de datos de código abierto que tiene una sólida reputación por su fiabilidad, flexibilidad y soporte de estándares técnicos abiertos. Soporta tipos de datos relacionales y no relacionales. Esto la convierte en una de las bases de datos relacionales más compatibles, estables y maduras disponibles actualmente. | <https://www.ibm.com/mx-es/topics/postgresql> |
| **5** | **Amazon Kinesis FireHose** | Servicio web de Amazon diseñado para procesar flujos de datos a gran escala de una multitud de servicios en tiempo real. Puede considerarse una especie de intermediario de mensajes. Esto significa que actúa como intermediario entre diversas fuentes generadoras de datos, permitiendo que otras aplicaciones o servicios trabajen con los datos de origen. | <https://www.upsolver.com/glossary/amazon-kinesis> |
| **6** | **Amazon QuickSight** | Servicio de inteligencia empresarial a escala de nube que se puede utilizar para ofrecer “easy-to-understand” (fácil de entender o claro) información a las personas con las que trabajas, estén donde estén. Amazon QuickSight se conecta a tus datos en la nube y combina datos de muchas fuentes diferentes. | <https://docs.aws.amazon.com/es_es/quicksight/latest/user/welcome.html> |
| **7** | **Amazon Glue** | Servicio de AWS de integración de datos sin servidor , por ejemplo,cuando tenemos multiples origenes de datos y queremos realizar tareas de carga desde diferentes origenes a diferentes destinos para su posterior explotación con diferentes herramientas, podemos hacer tareas de movimientos de datos como creación de datasets para analitica del dato, este tipo de herramientas nos ofrecen automatizar este tipo de tareas. | <https://www.linkedin.com/pulse/aws-glue-la-etl-de-amazon-web-services-jos%C3%A9-mar%C3%ADa-baena-moya-47lsf/> |
| **8** | **HTTP API** | permite que una aplicación (como una app móvil o una página web) envíe solicitudes a otra aplicación (como un servidor) y reciba respuestas. Esta comunicación ocurre mediante URLs y métodos como GET, POST, PUT, DELETE, entre otros. | <https://chatgpt.com/> |
| **9** | **Boto3** | Es una herramienta esencial para los desarrolladores de Python que trabajan con AWS, ya que simplifica la interacción con los servicios de AWS y permite crear aplicaciones y scripts para gestionar la infraestructura en la nube. | <https://www-geeksforgeeks-org.translate.goog/what-is-the-difference-between-the-aws-boto-and-boto3/?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=sge#:~:text=La%20interacci%C3%B3n%20program%C3%A1tica%20con%20estos,boto3%2C%20dos%20bibliotecas%20de%20Python.> |
| **10** | **Amazon RedShift** | Es un producto de almacenamiento de datos que forma parte de la plataforma de computación en la nube Amazon Web Services. Se utiliza para para gestionar conjuntos de datos a gran escala y migraciones de bases de datos . | <https://en.wikipedia.org/wiki/Amazon_Redshift> |
| **11** | **Amazon Athena** | Servicio de consultas interactivo que facilita el análisis de datos directamente en Amazon Simple Storage Service (Amazon S3) con SQL estándar. | <https://docs.aws.amazon.com/es_es/athena/latest/ug/what-is.html> |
| **12** | **Amazon CloudWatch** | Servicio que supervisa las aplicaciones, responde a los cambios de rendimiento, optimiza el uso de los recursos y proporciona información sobre el estado operativo. Al recopilar datos en todos los recursos de AWS, CloudWatch brinda visibilidad del rendimiento de todo el sistema y permite a los usuarios configurar alarmas, reaccionar automáticamente a los cambios y obtener una visión unificada del estado operativo. | <https://aws.amazon.com/es/cloudwatch/> |
| **13** | **Formato de datos JSON** | Formato basado en texto para almacenar e intercambiar datos de una manera que es legible por humanos y analizable por máquina. Es relativamente fácil de aprender y de solucionar problemas. Se ha convertido en un formato de datos muy capaz que simplifica el intercambio de datos en diversas plataformas y lenguajes de programación. | <https://www.oracle.com/co/database/what-is-json/> |
| **14** | **Formato de datos CSV** | "CSV" se refiere a la extensión de archivo CSV, es decir, cualquier archivo que termine en ".csv". La extensión .csv significa archivo de "valores separados por comas" y es uno de los archivos de salida más comunes para cualquier programa de hojas de cálculo.  Un archivo de valores separados por comas se refiere a un archivo donde los datos se introducen separados por comas. Una aplicación de hoja de cálculo convierte estos datos separados por comas en celdas de tablas para facilitar su lectura y edición. | <https://flatfile.com/blog/what-is-a-csv-file-guide-to-uses-and-benefits/> |
| **15** | **Formato de datos Parquet** | En lugar de agrupar filas como una hoja de cálculo de Excel o una base de datos relacional estándar, un archivo Apache Parquet agrupa columnas para un rendimiento más rápido. El parquet es un formato de almacenamiento columnar y no una base de datos en sí misma, pero el formato Parquet es común con los lagos de datos, especialmente con Hadoop. Como es un formato columnar, es popular entre el almacenamiento de datos analíticos y las consultas. | <https://www.purestorage.com/es/knowledge/what-is-parquet-file.html> |
| **16** | **Microcontrolador Arduino** | Arduino es una plataforma de creación de electrónica de código abierto, la cual está basada en hardware y software libre, flexible y fácil de utilizar para los creadores y desarrolladores. Esta plataforma permite crear diferentes tipos de microordenadores de una sola placa a los que la comunidad de creadores puede darles diferentes tipos de uso. | <https://www.xataka.com/basics/que-arduino-como-funciona-que-puedes-hacer-uno> |
| **17** | **Microcontrolador ESP8266** | Chip WiFi con MCU integrado que puede realizar la función de comunicarse con la red.ESP8266 se usa ampliamente en Internet de las cosas.Se puede conectar a un microcontrolador (como Arduino) a través de una interfaz de comunicación en serie, realizando así una variedad de escenarios de aplicación como luces inteligentes, vehículos inteligentes y casas inteligentes. | <https://www.ariat-tech.es/blog/features-and-applications-of-esp8266.html> |
| **18** | **Sensor MQ-135** | Modulo sensor de calidad del aire MQ135 para monitorear Amoníaco (NH3), Óxidos de nitrógeno (NOx), Alcohol, Sulfuros, Benceno (C6H6), Monóxido de carbono (CO), humo y otros gases nocivos. | <https://www.vistronica.com/sensores/modulosensordecalidaddelairemq135-detail.html> |
| **19** | **Benceno** | El benceno es una sustancia química líquida, incolora o de color amarillo claro cuando está a temperatura ambiente. Se utiliza principalmente como solvente en la industria química y en la industria farmacéutica, como producto inicial e intermedio en la síntesis de numerosas sustancias químicas, y en la gasolina. El benceno se produce tanto en procesos naturales como artificiales. Es un componente natural del petróleo crudo, que es la principal fuente de benceno producido en la actualidad. | <https://www.cancer.gov/espanol/cancer/causas-prevencion/riesgo/sustancias/benceno> |

**4.- Diseño de la estructura de la hoja de cálculo del microcontrolador “lecturas”**

*Colocar en esta sección los datos que se generan en el sensor y se almacenan en la hoja de cálculo local “lecturas” en el microcontrolador. El mismo formato de hoja de cálculo se utiliza en el servidor local.*

**4.1. - Diseño de la estructura de la hoja de cálculo para almacenar lecturas del microcontrolador**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Nombre Columna** | **Descripción** | **Observaciones** |
| **1** | **id\_lectura** | Identificador único de cada lectura registrada. | Autonumérico, clave primaria en la base de datos. |
| **2** | **fecha** | Fecha en que se tomó la lectura del sensor. | Formato YYYY-MM-DD, fundamental para análisis temporal. |
| **3** | **hora** | Hora exacta de la captura del dato. | Formato hh:mm:ss, requerido para monitoreo en tiempo real. |
| **4** | **linea\_produccion** | Línea de producción donde se encuentra el sensor. | Se registra como "Línea X" para trazabilidad. |
| **5** | **id\_sensor** | Código único del sensor que realizó la lectura. | Relacionado con microcontroladores ESP8266. |
| **6** | **ppm\_benceno** | Nivel de concentración de benceno detectado. | Expresado en partes por millón (ppm), esencial para la analítica. |
| **7** | **nivel\_riesgo** | Clasificación de concentración de ppm según peligrosidad. | Bajo, Moderado, Alto o Crítico |
| **8** | **estado\_filtro** | Estado del filtro de seguridad en la línea de producción. | Normal, Cambiar pronto o Crítico, importante para mantenimiento predictivo. |
| **9** | **alerta\_generada** | Indica si el sistema activó una alarma ante riesgo. | Sí/No, vinculada al sistema de monitoreo 24/7. |
| **10** | **accion\_recomendada** | Acción sugerida para mitigar el riesgo. | Revisar, Detener línea, Emergencia, Ninguna. |
| **11** | **frecuencia\_lectura** | Tiempo entre cada captura de datos. | Cada 10s |
| **12** | **geo\_latitud** | Coordenadas geográficas del sensor en la fábrica. | Latitud, vinculada al análisis espacial. |
| **13** | **geo\_longitud** | Coordenadas geográficas del sensor en la fábrica. | Longitud, necesaria para trazabilidad y ubicación precisa. |
| **14** | **fuente\_dato** | Origen del dato (Microcontrolador) | Nube |
| **15** | **usuario\_responsable** | Identificación del operador que monitorea el sistema. | Registrado solo si hay intervención humana. |
| **16** | **tiempo\_procesamiento** | Lapso entre la captura y almacenamiento en la nube. | Útil para evaluar eficiencia del sistema. |

**

**4.2. – Poblamiento de la hoja de cálculo del microcontrolador**

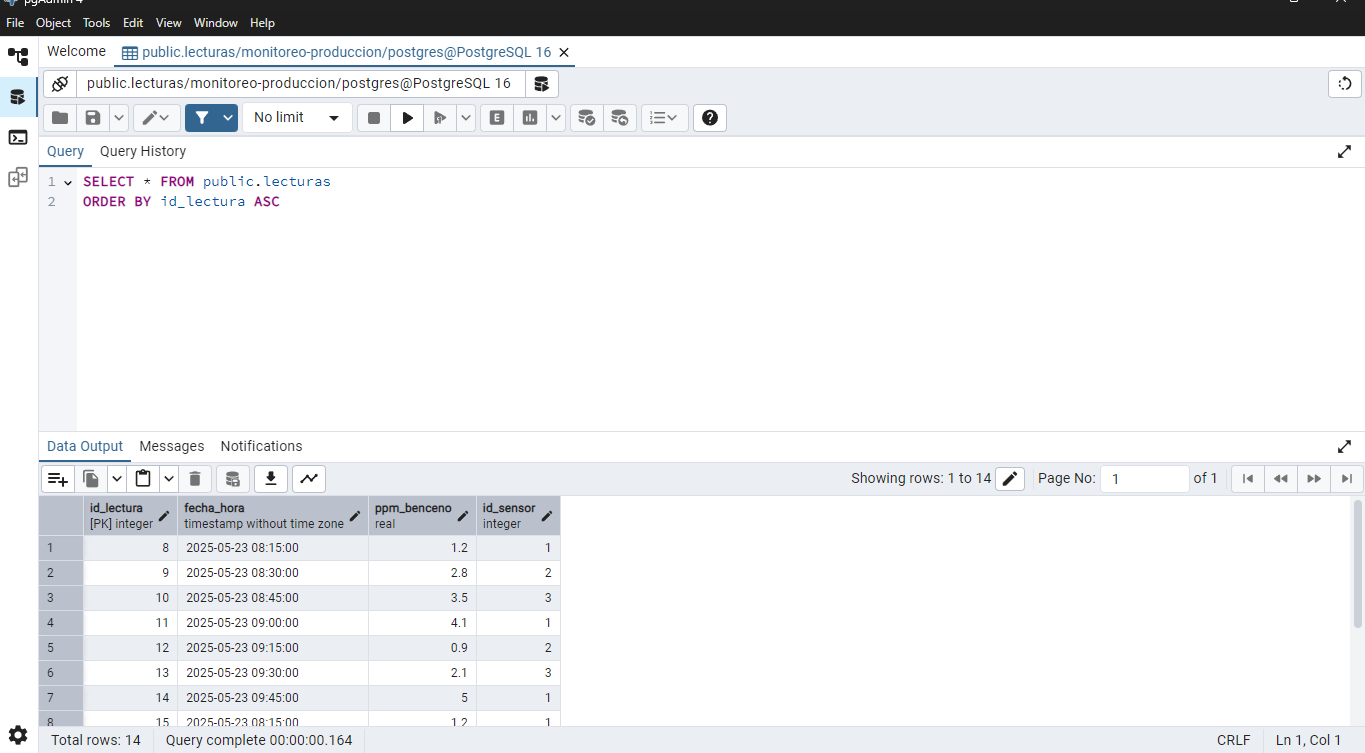
*Colocar en esta sección veinte (20) registros con datos ficticios. Nota: puede colocar el “pantallazo” de la hoja de cálculo o incrustar “copia” y “pega”. Las columnas son los datos*

*Interfaz de usuario gráfica, Aplicación, Tabla, Excel

El contenido generado por IA puede ser incorrecto.*

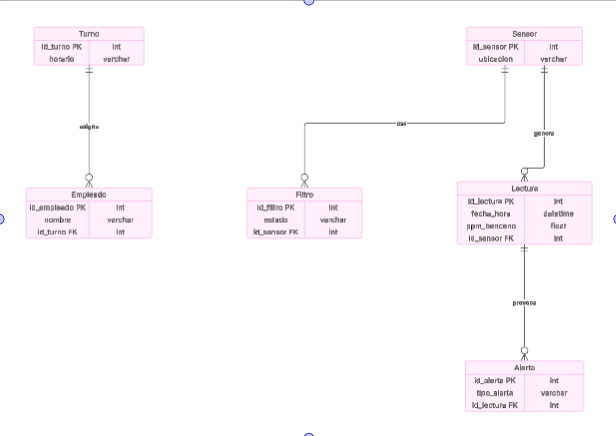
**5.- Diseño de la estructura de la tabla de “lecturas”**

*Colocar en esta sección la estructura de la tabla de “lecturas” dónde se almacenan los datos de las diferentes hojas de cálculo que provienen de los sensores. Esta tabla se encuentra dentro de la base de datos “monitoreo-produccion”*

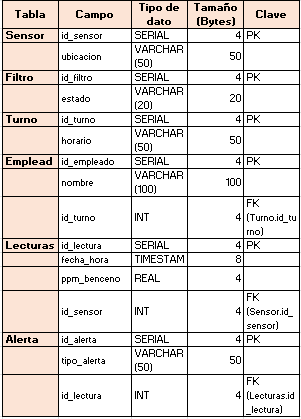


**6.- Diseño y creación de la Base de Datos “monitoreo-produccion”**

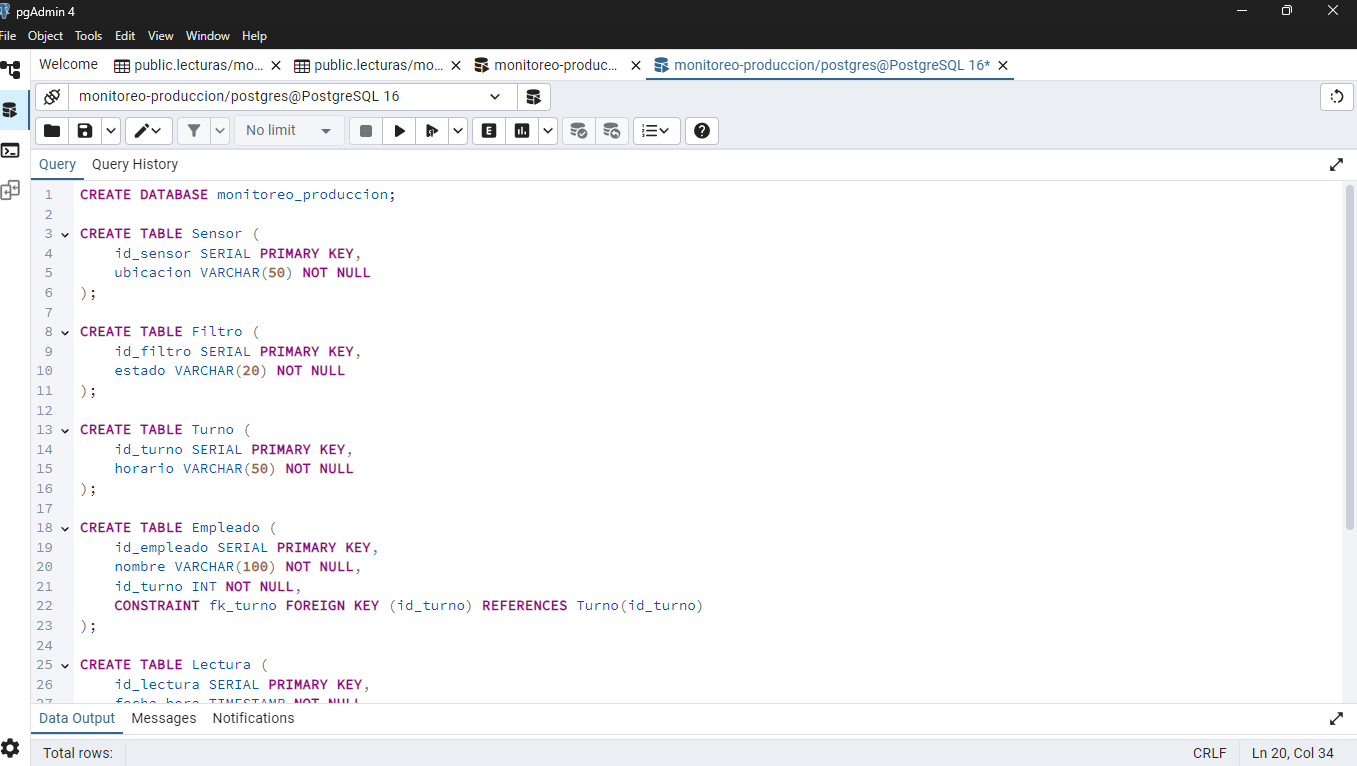
*Colocar en esta sección el diagrama de Entidad-Relación sin atributos y el Diccionario de datos de la base de datos de gestión. Esta es la base de datos que contiene, además de la tabla lecturas,*

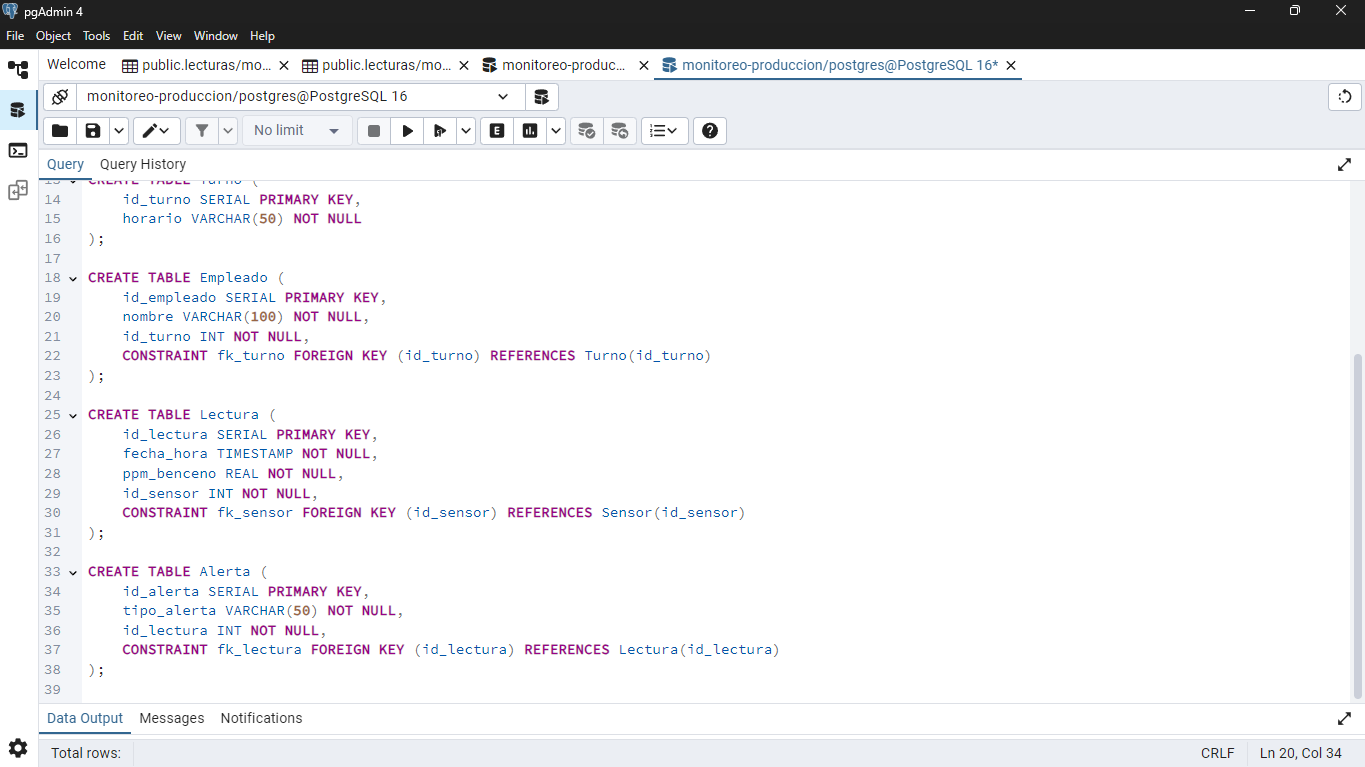
** 6.1.- Diagrama Entidad-Relación (entidades, relaciones, atributos y cardinalidades)**

**6.2- Diccionario de Datos (tipos de dato, tamaño, claves primarias y foráneas)**

****

**6.3- Scripts de creación de las tablas de la Base de Datos “monitoreo-produccion”**

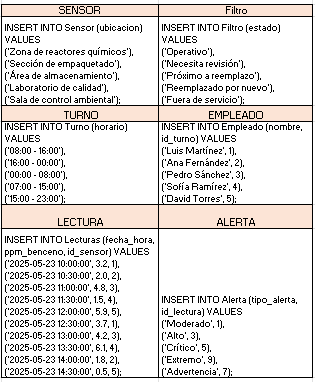
****

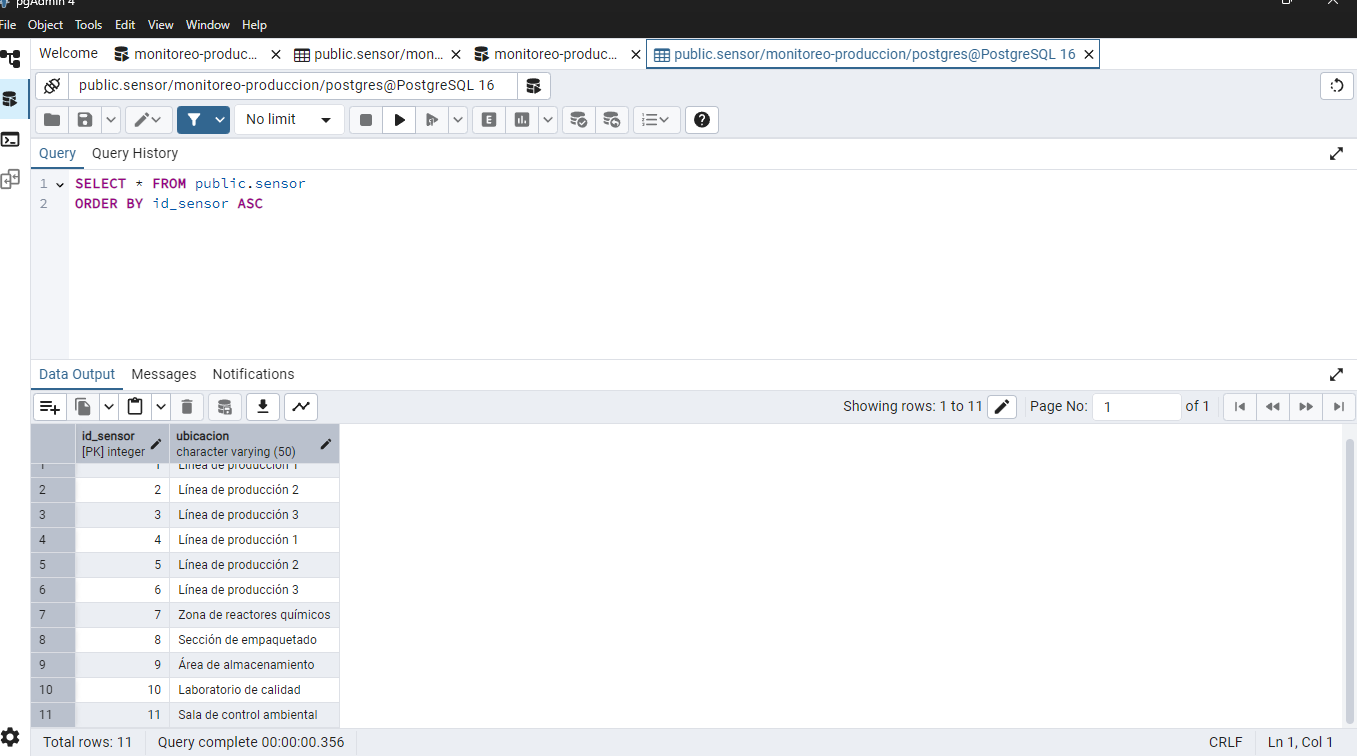
****

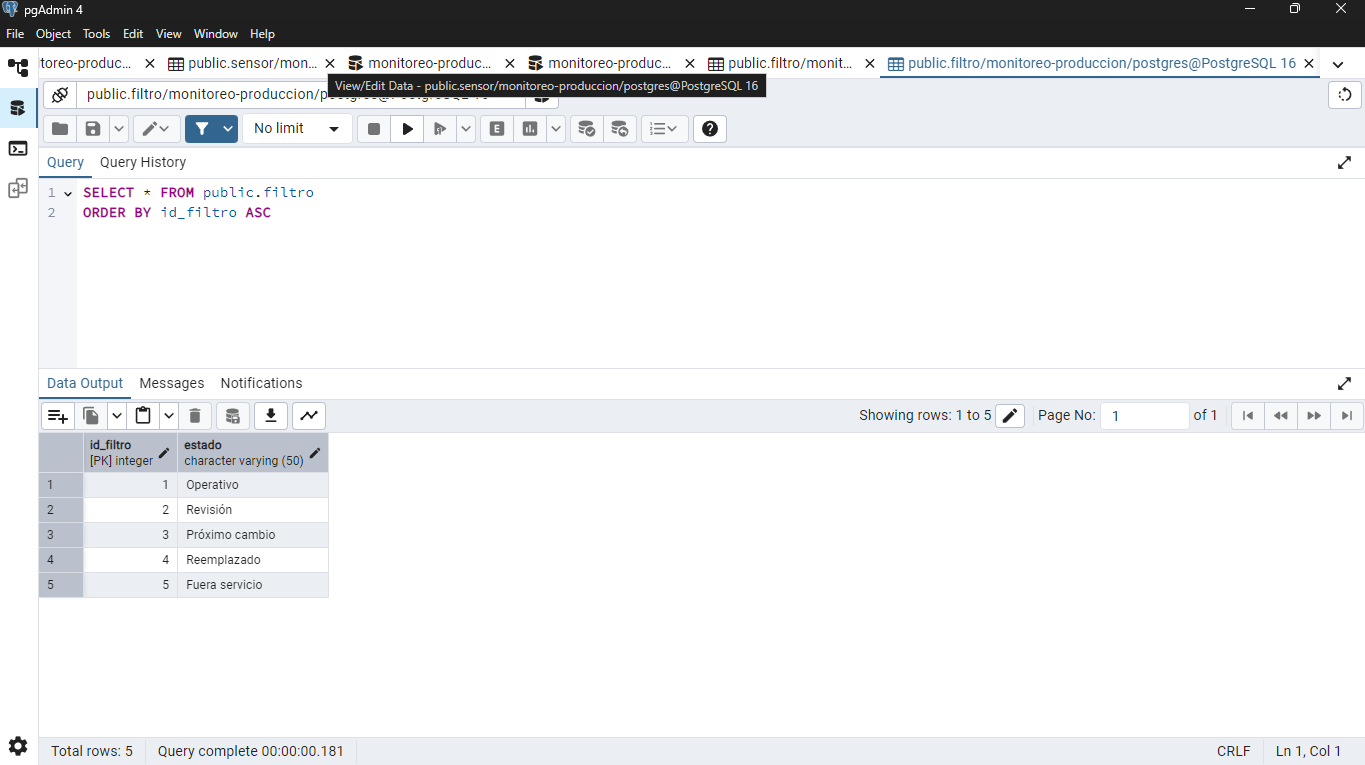
**7.- Poblamiento de la base de datos “monitoreo-produccion”**

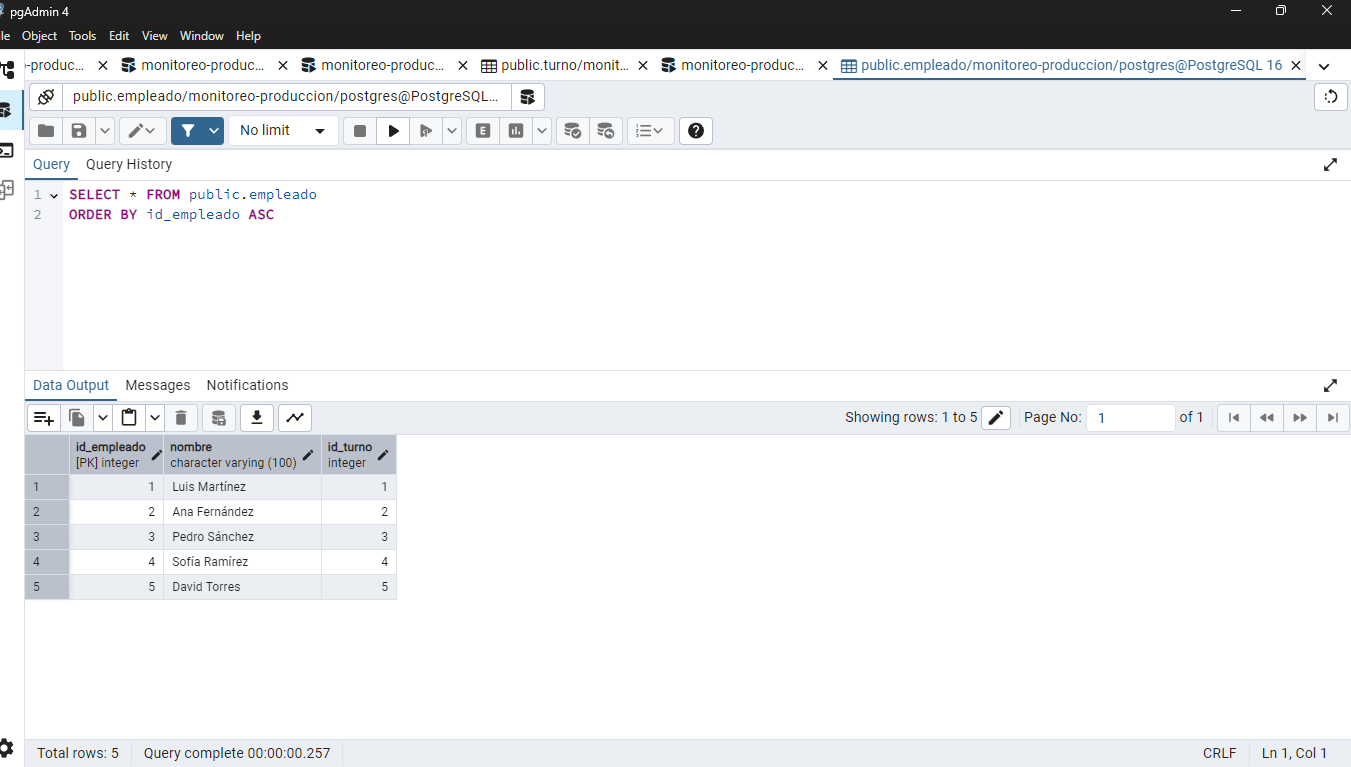
*Colocar en esta sección un pantallazo de los resultados de DML (INSERTS) en pgAdmin4 realizados en todas las tablas de la base de datos.*

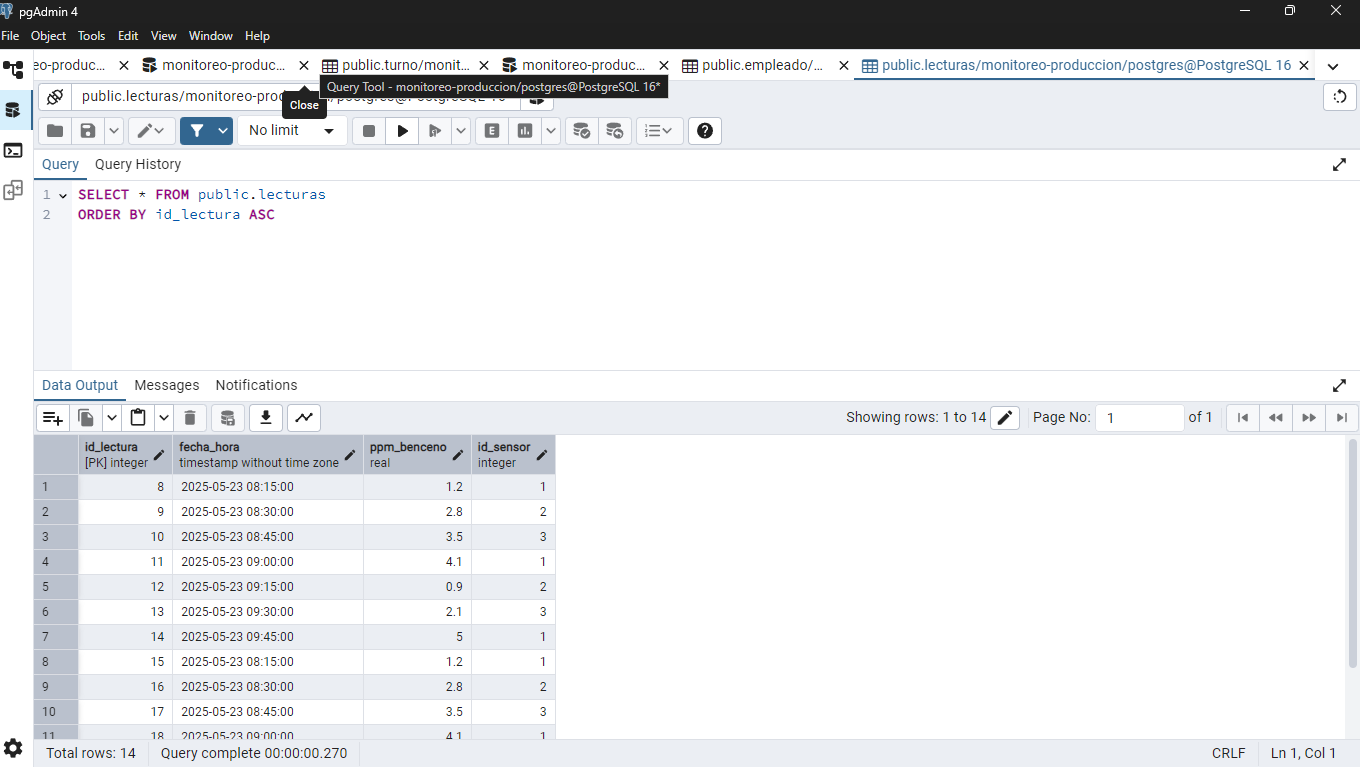
**7.1.- Scripts de inserción de registros en todas las tablas de la base datos (INSERTS)**

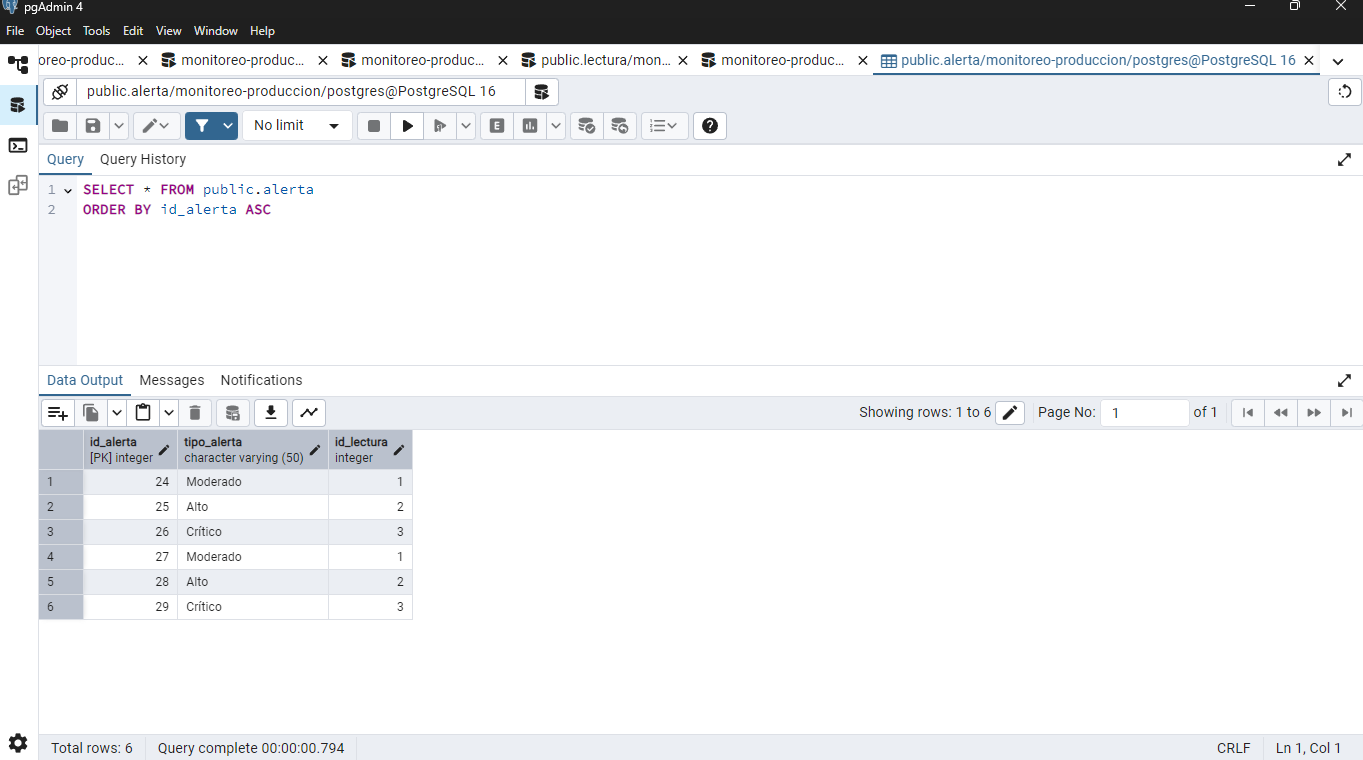
****

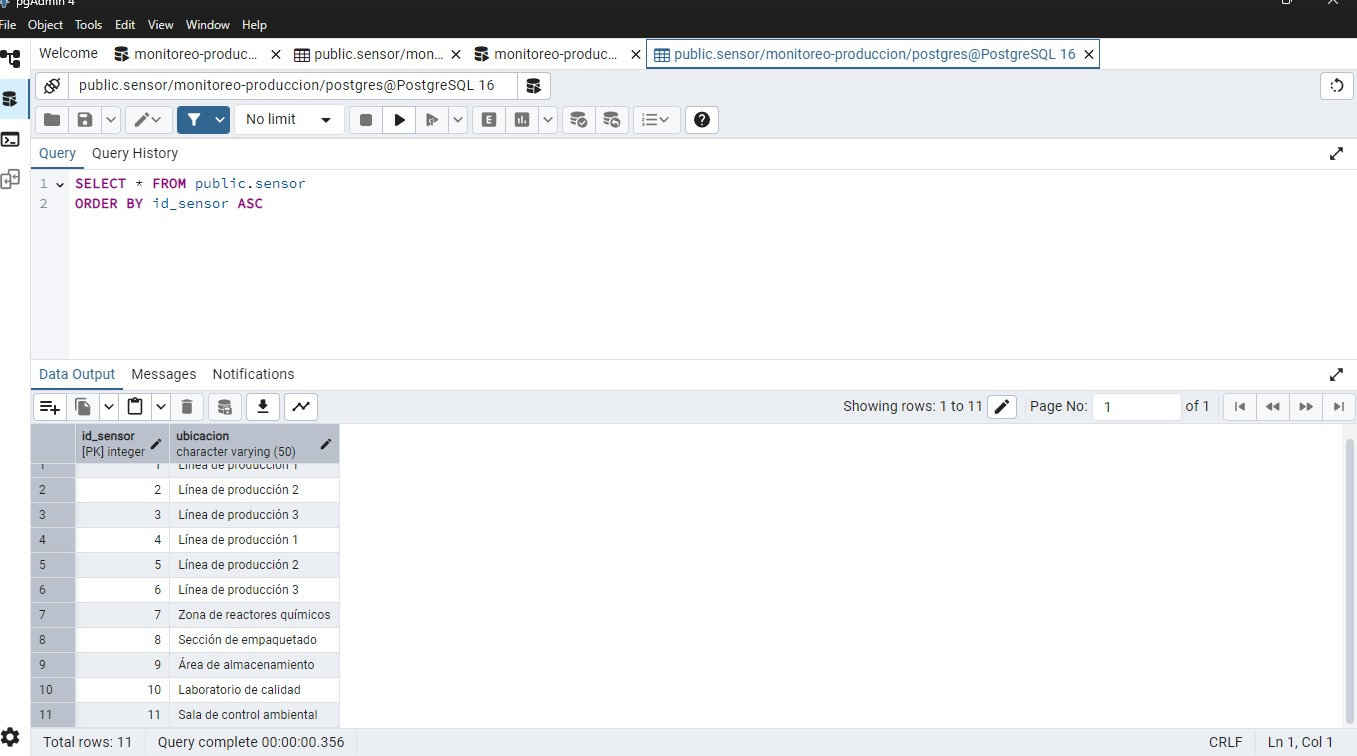
** 7.2- Pantallazos de consultas SELECT de las tablas pobladas**

****

****

****

****

**

**8.- Conclusiones.**

**SOFIA RAMIREZ**

“Esta tarea me ayudó a entender mejor cómo funciona el Big Data en la práctica. No solo aprendí conceptos teóricos, sino que también puse manos a la obra diseñando una arquitectura de monitoreo para una fábrica, conectando sensores, bases de datos y servicios en la nube. Fue un reto, pero al final logré fortalecer habilidades como la organización, el trabajo en equipo y la planificación. Esta experiencia me dejó con más confianza para enfrentar nuevos desafíos tecnológicos y me motiva a seguir aprendiendo más sobre este campo”.

**JUAN DIEGO GALLARDO VILORIA**

“A lo largo del curso y especialmente con el desarrollo de esta tarea, logré no solo comprender los conceptos, sino también ponerlos en práctica al diseñar una arquitectura de monitoreo para una fábrica, integrando sensores, bases de datos, microcontroladores y servicios en la nube.

Esta experiencia me permitió desarrollar habilidades técnicas concretas, como la creación de estructuras de datos, la implementación de proyectos en AWS y el uso de herramientas como PostgreSQL, además de reforzar mi capacidad para trabajar en equipo y planificar proyectos complejos.

Estoy convencido de que todo lo aprendido tendrá un impacto positivo en mi futuro profesional. Me siento más preparado para asumir retos tecnológicos en la industria, y creo que contar con estos conocimientos diferenciales en Big Data y análisis de datos me abrirá muchas puertas cuando obtenga mi título en el Pascual Bravo”.

**JUAN FELIPE BARRERA GARCIA**

“Este informe nos permitió aplicar de manera práctica los conceptos de Big Data, entendiendo cómo se estructura un Gobierno de Datos, cómo se diseñan procesos de almacenamiento y análisis de información, y la importancia de monitorear datos en tiempo real para mejorar la toma de decisiones en industrias como la de control de gases tóxicos.

Durante el desarrollo del trabajo, logramos integrar sensores IoT, definir procesos ETL y estructurar una base de datos eficiente para el almacenamiento y análisis de datos, asegurando que la información se pueda consultar de forma organizada y segura.

Más allá de la parte técnica, este proyecto nos dejó claro lo esencial que es trabajar en equipo y planificar bien cada etapa, porque gestionar grandes volúmenes de información no solo depende de la tecnología, sino de una buena organización.

Sin duda, todo lo aprendido nos servirá en el futuro, tanto en el ámbito profesional como en el desarrollo de proyectos que requieran estructuras de datos eficientes, escalabilidad y almacenamiento inteligente”.

**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 . Coloque el enlace en esta sección*