**INFORME DE RESULTADOS**

**Unidad 3 - Tarea 6**

**Peso: 10%**

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

**Estudiantes :**

**-Keisy Yasiry Valoyes Jiménez**

**-**

**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.

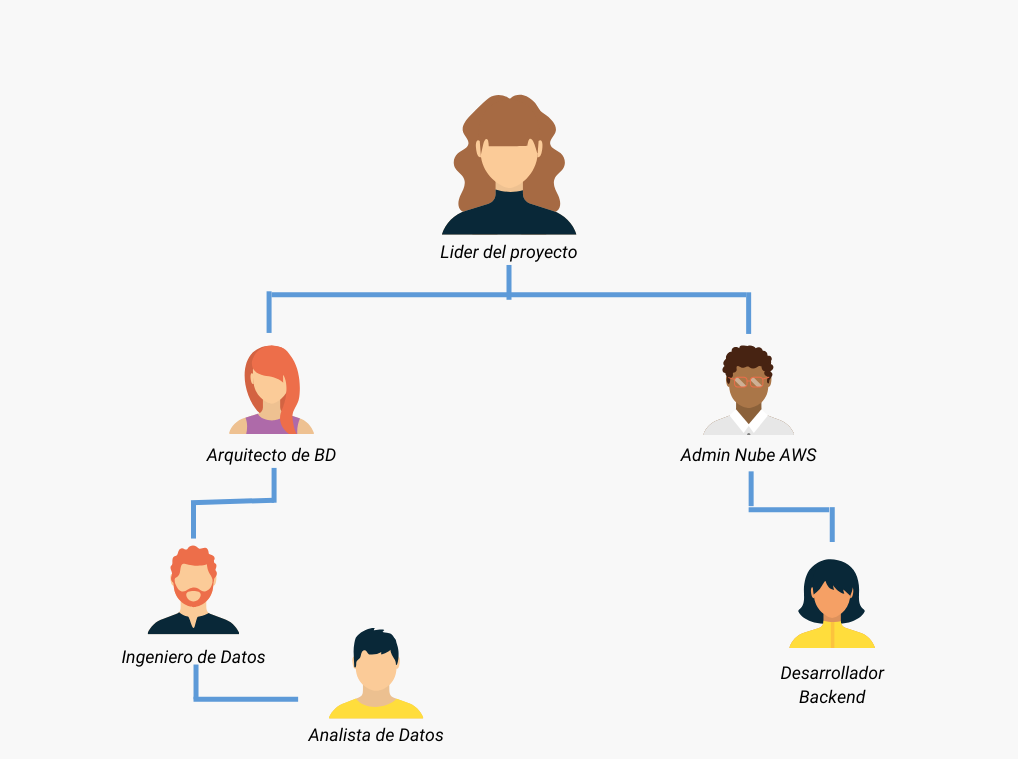
**1.-Diseño de un Gobierno de Datos**

*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** | Líder del proyecto Big Data | Responsable de coordinar y supervisar todo el proyecto | Planifica, toma decisiones estratégicas, gestiona recursos y cronograma del proyecto |
| **2** | Arquitecto de Big Data | Diseñador de la arquitectura tecnológica local y en la nube | Define como se integran sensores, bases de datos, Hadoop, Spark y AWS |
| **3** | Ingeniero de Datos | Especialista en almacenamiento y transformación de datos | Crear procesos ETL, diseña bases de datos, estructura los datos en “PostgreSQL y Hive” |
| **4** | Desarrollador Backend | Encargado de la conexión entre microcontroladores y servidores | Programa las APIs HTTP para transmitir lecturas en formato JSON |
| 5 | Analista de Datos | Profesional que interpreta y visualiza la información | Diseña dashboards en “Metabase y QuickSight”, analiza tendencias y niveles de alerta |
| 6 | Administrador de infraestructura en la Nube | Encargado de gestionar servicios en AWS | Configura EC2, S3, Glue, Redshift, QuickSight y monitorea su funcionamiento |
| **N** |  |  |  |

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



**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 MQ-135 | Detectan niveles de concentración de benceno en las líneas de producción | Son la fuente primaria de los datos |
| **2** | Microcontroladores ESP8266 | Dispositivos conectados por Wi-Fi que trasmiten datos desde sensores | Cada línea de producción tiene 20 microcontroladores |
| **3** | Hoja de calculo | Registra lecturas de cada sensor | Simulada con archivos Excel o CSV |
| **4** | Base de Datos PostgreSQL | Almacena lecturas, configuración de sensores, empleados, turnos, productos | Se implementa y conecta con Metabase para análisis |
| 5 | Apache Hadoop | Almacenamiento distribuido para grandes volúmenes de datos | Parte del diseño arquitectónico.  No implementado |
| 6 | Apache Spark | Motor de procesamiento de datos(ETL y analítica) | Incluido en la arquitectura |
| 7 | Hive | Fuente de datos con interfaz SQL sobre Hadoop | Se considera como tercera fuente |
| 8 | Metabase | Visualización local de datos de producción y alarmas | Implementado realmente con PostgreSQL |
| 9 | AWS S3 + Glue | Almacenamiento y ETL de datos en la nube | Se describe en el diseño, no se implementa |
| 10 | Amazon QuickSight | Dashboard en la nube para analizar datos consolidados | Parte del diseño conceptual |
| **N** |  |  |  |

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Elemento** | **Estado** | **Tiempo** |
| **1** | Diseño del Gobierno de Datos | Completado | 1 hora |
| **2** | Diseño del Proyecto Big Data | Completado | 1 hora |
| **3** | Diseño de la estructura de hoja de cálculo por sensor | Completado | 1 hora |
| **4** | Simulación de datos en la hoja de cálculo | Completado | 1 hora |
| 5 | Diseño y creación de la base de datos PostgreSQL | Completado | 2 horas |
| 6 | Inserción de datos y pruebas en PostgreSQL | Completado | 2 horas |
| 7 | Conexión y creación de dashboard en Metabase | Pendiente  (tarea7) | 2 horas |
| 8 | Redacción del informe en la plantilla | Completado | 2-3 horas |
| 9 | Diseño conceptual del flujo en la nube | Completado (teórico) | 1.5 horas |
| 10 | Grabación del video de sustentación | En proceso | 30 mins |
| **N** |  |  |  |

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

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Componente** | **Descripción (breve cita bajo Norma APA)** | **Referencia (APA)** |
| **1** | Arquitectura de Sistema de Información | Es una estructura de conjuntos que se encarga de planificar, diseñar y organizar todos los componentes necesarios para que un sistema de información cumpla con los objetivos fundamentales, operativos y estratégicos. (cecropia, s.f.) | *cecropia*. (s.f.). Obtenido de https://cecropia.co/arquitectura-de-sistemas/ |
| **2** | Hadoop | Es un framework de software de código abierto diseñado para almacenar y procesar grandes volúmenes de datos de manera distribuida y eficiente.  (wikipedia, s.f.) | *wikipedia*. (s.f.). Obtenido de https://es.wikipedia.org/wiki/Apache\_Hadoop |
| **3** | Spark | Es un framework de código abierto que procesa datos a gran escala diseñado para ejecutarse en clústeres y ofrecer un rendimiento muy superior al de sistemas tradicionales como Hadoop.  (ibm, s.f.) | *ibm*. (s.f.). Obtenido de https://www.ibm.com/mx-es/topics/apache-spark |
| **4** | PostgreSQL | Es un sistema de gestión de bases de datos relacional y orientado a objetos, de código abierto, reconocido por su estabilidad, flexibilidad y cumplimiento de estándares SQL  (microsoft, s.f.) | *microsoft*. (s.f.). Obtenido de https://azure.microsoft.com/es-es/resources/cloud-computing-dictionary/what-is-postgresql |
| **5** | Amazon Kinesis FireHose | Es un servicio totalmente gestionado de AWS que permite capturar, transformar y entregar datos de streaming en tiempo real de manera sencilla y fiable a diversos destinos.(aws, s.f.) | *aws*. (s.f.). Obtenido de https://docs.aws.amazon.com/es\_es/firehose/latest/dev/what-is-this-service.html |
| **6** | Amazon QuickSight | Es un servicio de inteligencia empresarial nativo en la nube que permite a las organizaciones conectar, combinar y analizar datos provenientes de múltiples fuentes, tanto en la nube como locales. (aws, s.f.) | *aws*. (s.f.). Obtenido de https://docs.aws.amazon.com/es\_es/quicksight/latest/user/welcome.html |
| **7** | Amazon Glue | Es un servicio sin servidor de integración de datos que facilita la detección, preparación, transformación, movimiento e integración de datos provenientes de múltiples orígenes para análisis. (aws, s.f.) | *aws*. (s.f.). Obtenido de https://docs.aws.amazon.com/es\_es/glue/latest/dg/what-is-glue.html |
| **8** | HTTP API | Es un tipo de API RESTful que utiliza el protocolo HTTP para permitir la comunicación sin estado entre clientes y servidores mediante estándar como GET, POST PUT, PATCH.  (aws, s.f.) | *aws*. (s.f.). Obtenido de https://docs.aws.amazon.com/es\_es/apigateway/latest/developerguide/welcome.html |
| **9** | Boto3 | Es el SDK oficial de AWS para Python que facilita la integración y gestión de los servicios de AWS desde aplicaciones, scripts o bibliotecas en Python.  (platzi, s.f.) | *platzi*. (s.f.). Obtenido de https://platzi.com/cursos/big-data/como-usar-boto3/ |
| **10** | Amazon RedShift | Es un servicio de almacenamiento de datos en la nube, diseñado para el análisis de grandes volúmenes de datos con alta velocidad y escalabilidad.  (aws, s.f.) | *aws*. (s.f.). Obtenido de https://docs.aws.amazon.com/es\_es/redshift/latest/mgmt/welcome.html |
| **11** | Amazon Athena | Es un servidor de consultas interactivo y sin servidor que permite analizar datos directamente almacenados en Amazon S3 usando SQL estándar, sin necesidad de administrar infraestructura ni realizar cargas previas.  (aws, s.f.) | *aws*. (s.f.). Obtenido de https://docs.aws.amazon.com/es\_es/athena/latest/ug/what-is.html |
| **12** | Amazon CloudWatch | Es un servidor de monitoreo y gestión que recopila datos operativos y métricas en tiempo real de recursos, aplicaciones y servicios tanto en AWS con en entornos híbridos o multicloud.  (aws, s.f.) | *aws*. (s.f.). Obtenido de https://docs.aws.amazon.com/es\_es/AmazonCloudWatch/latest/monitoring/WhatIsCloudWatch.html |
| **13** | Boto3 | Es el SDK oficial de AWS para Python que facilita la integración y gestión de los servicios de AWS desde aplicaciones, scripts o bibliotecas en Python.  (platzi, s.f.) | *platzi*. (s.f.). Obtenido de https://platzi.com/cursos/big-data/como-usar-boto3/ |
| **14** | Formato de datos JSON | Es un formato de texto ligero y legible para humanos que se utiliza para almacenar y transmitir datos estructurados entre sistemas de manera eficiente.  (oracle, s.f.) | *oracle*. (s.f.). Obtenido de https://www.oracle.com/co/database/what-is-json/ |
| **15** | Formato de datos CSV | Es un formato de texto plano utilizado para almacenar datos tabulares de manera sencilla y estructurada.  (adobe, s.f.) | *adobe*. (s.f.). Obtenido de https://www.adobe.com/es/acrobat/resources/document-files/text-files/csv-file.html |
| **16** | Formato de datos Parquet | Es un formato de almacenamiento columnar de código abierto para almacenar y procesar grandes conjuntos de datos de manera eficiente.  (purestorage, s.f.) | *purestorage*. (s.f.). Obtenido de https://www.purestorage.com/es/knowledge/what-is-parquet-file.html |
| **17** | Microcontrolador Arduino | Es una plataforma de hardware y software libre basada en un microcontrolador, generalmente de la familia Atmel AVR, que permite crear dispositivos electrónicos programables para interactuar con el entorno mediante sensores y actuadores.  (arduino, s.f.) | *arduino*. (s.f.). Obtenido de https://arduino.cl/que-es-arduino/ |
| **18** | Microcontrolador ESP8266 | Es un microcontrolador de 32 bits con arquitectura RISC y CPU Tensilica L106 que funciona a una frecuencia de 80 MHz.  (wikipedia, s.f.) | *wikipedia*. (s.f.). Obtenido de https://es.wikipedia.org/wiki/ESP8266 |
| **19** | Sensor MQ-135 | Es un dispositivo diseñado para medir la calidad del aire mediante la detección de gases nocivos como amoniaco (NH3), óxidos de nitrógeno (NOx), alcohol, sulfuros, benceno, monóxido de carbono (CO) y humo.  (naylampmechatronics, s.f.) | *naylampmechatronics*. (s.f.). Obtenido de https://naylampmechatronics.com/sensores-gas/73-sensor-mq-135-gas-calidad-aire.html |
| **20** | Benceno | Es un hidrocarburo aromático con fórmula molecular caracterizado por un anillo hexagonal de seis átomos de carbono con enlaces resonantes entre ellos y cada carbono unido a un átomo de hidrogeno.  (prtr, s.f.) | *prtr*. (s.f.). Obtenido de https://prtr-es.es/Benceno,15649,11,2007.html |

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

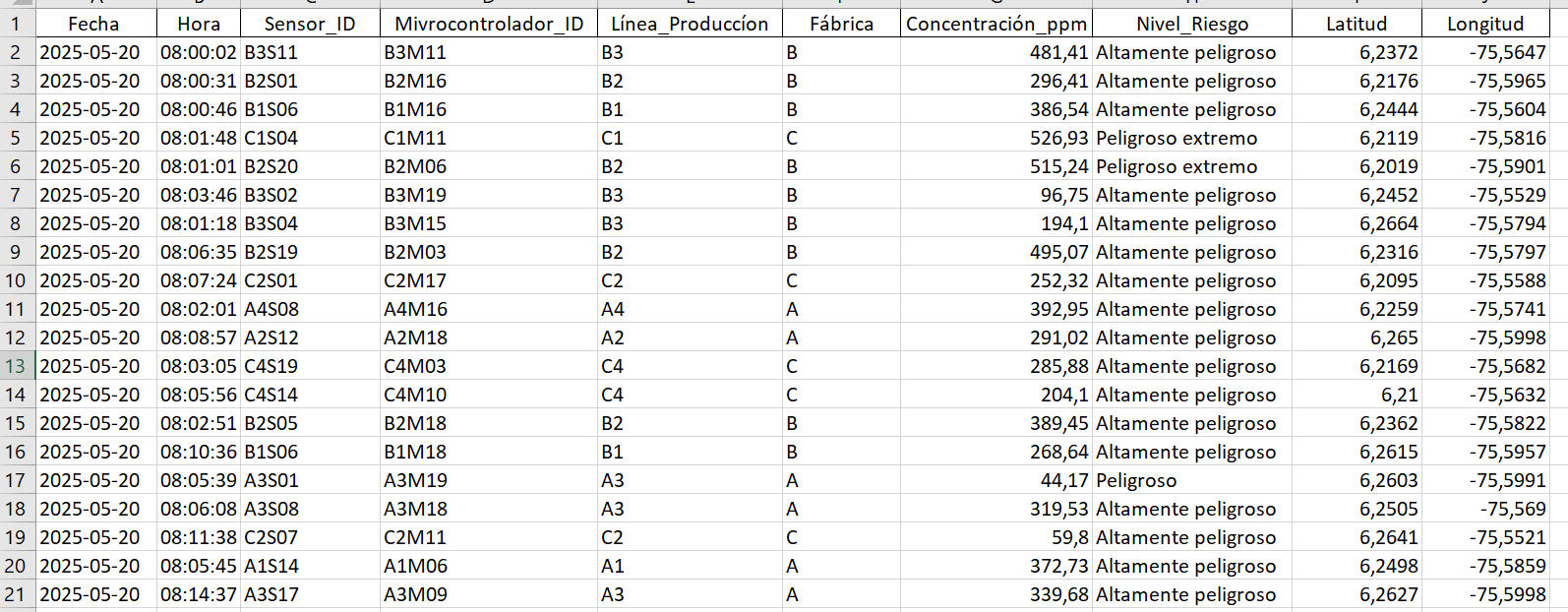
*Colocar en esta sección los datos que se generan en el sensor y se almacenan en la hoja de cálculo local “lecturas” en el microcontrolador. El mismo formato de hoja de cálculo se utiliza en el servidor local.*

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

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Nombre Columna** | **Descripción** | **Observaciones** |
| **1** | Fecha | Fecha de la lectura del sensor | Formato: AAAA-MM-DD |
| **2** | Hora | Hora exacta de la lectura | Formato: hh:mm:ss |
| **3** | Sensor\_ID | Identificador del sensor conectado al microcontrolador | A1S01, B3S0S |
| **4** | Mricrocontrolador\_ID | Código del microcontrolador que trasmite la lectura | A1M01 |
| **5** | Línea\_Producción | Línea donde está instalado el sensor | A1, B4 |
| **6** | Fabrica | Fabrica donde se ubica el sensor | A, B o C |
| **7** | Concentración\_ppm | Nivel de concentración de benceno en partes por millón | Valor numérico |
| **8** | Nivel\_Riesgo | Clasificación del nivel según ppm | Leve, permitida |
| **9** | Latitud | Coordenada geográfica del sensor | simulada |
| **10** | Longitud | Coordenada geográfica del sensor | simulada |
| **N** |  |  |  |

**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*

**

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

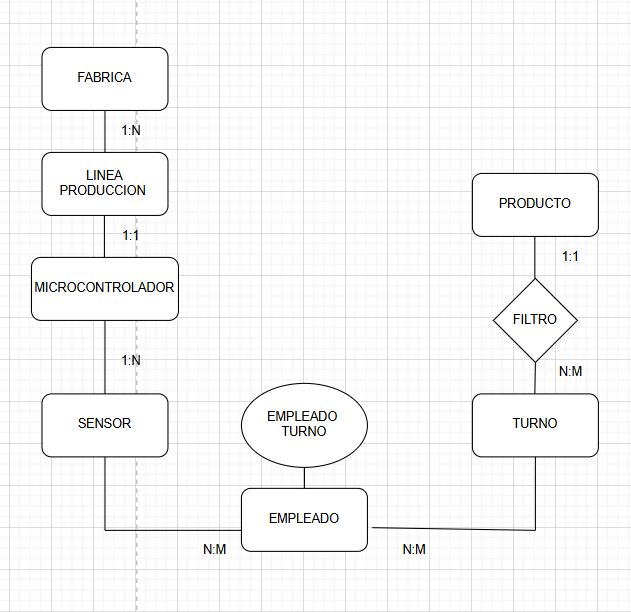
*Colocar en esta sección la estructura de la tabla de “lecturas” dónde se almacenan los datos de las diferentes hojas de cálculo que provienen de los sensores. Esta tabla se encuentra dentro de la base de datos “monitoreo-produccion”*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Tabla** | **lecturas** | |
| **#** | **Nombre Campo** | **Tipo de dato** | **Tamaño (Bytes)** |
| **1** | fecha | DATE | 4 |
| **2** | hora | TIME | 8 |
| **3** | Sensor\_id | VARCHAR(10) | 12 |
| **4** | Microcontrolador\_id | VARCHAR(10) | 12 |
| **5** | Línea\_produccion | VARCHAR(5) | 6 |
| **6** | fabrica | CHAR(1) | 1 |
| **7** | Concetrancion\_ppm | NUMERIC(6,2) | 8 |
| **8** | Nivel\_riesgo | VARCHAR(30) | 32 |
| **9** | latitud | NUMERIC(8,6) | 8 |
| **10** | longitud | NUMERIC(9,6) | 8 |
| **N** |  |  |  |
| **Tamaño de un (1) registro en bytes** | | | 99 |

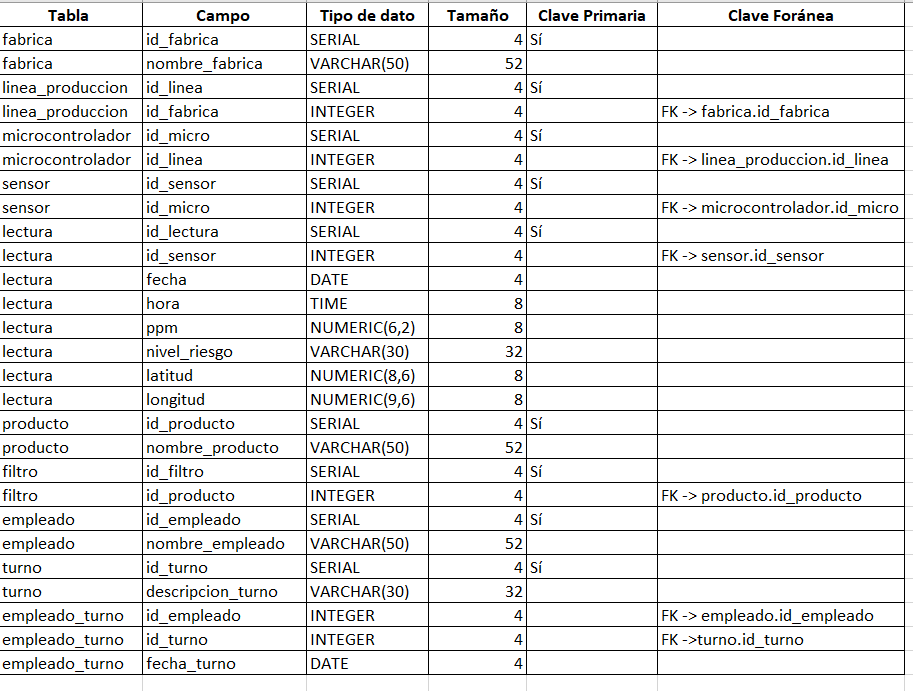
**6.- Diseño y creación de la Base de Datos “monitoreo-produccion”**

*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”**

CREATE TABLE fabrica (

id\_fabrica SERIAL PRIMARY KEY,

nombre\_fabrica VARCHAR(50)

);

CREATE TABLE linea\_produccion (

id\_linea SERIAL PRIMARY KEY,

id\_fabrica INTEGER REFERENCES fabrica(id\_fabrica)

);

CREATE TABLE microcontrolador (

id\_micro SERIAL PRIMARY KEY,

id\_linea INTEGER REFERENCES linea\_produccion(id\_linea)

);

CREATE TABLE sensor (

id\_sensor SERIAL PRIMARY KEY,

id\_micro INTEGER REFERENCES microcontrolador(id\_micro)

);

CREATE TABLE lectura (

id\_lectura SERIAL PRIMARY KEY,

id\_sensor INTEGER REFERENCES sensor(id\_sensor),

fecha DATE,

hora TIME,

ppm NUMERIC(6,2),

nivel\_riesgo VARCHAR(30),

latitud NUMERIC(8,6),

longitud NUMERIC(9,6)

);

CREATE TABLE producto (

id\_producto SERIAL PRIMARY KEY,

nombre\_producto VARCHAR(50)

);

CREATE TABLE filtro (

id\_filtro SERIAL PRIMARY KEY,

id\_producto INTEGER REFERENCES producto(id\_producto)

);

CREATE TABLE empleado (

id\_empleado SERIAL PRIMARY KEY,

nombre\_empleado VARCHAR(50)

);

CREATE TABLE turno (

id\_turno SERIAL PRIMARY KEY,

descripcion\_turno VARCHAR(30)

);

CREATE TABLE empleado\_turno (

id\_empleado INTEGER REFERENCES empleado(id\_empleado),

id\_turno INTEGER REFERENCES turno(id\_turno),

fecha\_turno DATE,

PRIMARY KEY (id\_empleado, id\_turno, fecha\_turno)

);

**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)**

**-- INSERTS fabrica**

**INSERT INTO fabrica (nombre\_fabrica) VALUES ('Fábrica A');**

**INSERT INTO fabrica (nombre\_fabrica) VALUES ('Fábrica B');**

**INSERT INTO fabrica (nombre\_fabrica) VALUES ('Fábrica C');**

**-- INSERTS linea\_produccion**

**INSERT INTO linea\_produccion (id\_fabrica) VALUES (1);**

**INSERT INTO linea\_produccion (id\_fabrica) VALUES (1);**

**-- INSERTS microcontrolador**

**INSERT INTO microcontrolador (id\_linea) VALUES (12);**

**INSERT INTO microcontrolador (id\_linea) VALUES (11);**

**-- INSERTS sensor**

**INSERT INTO sensor (id\_micro) VALUES (1);**

**INSERT INTO sensor (id\_micro) VALUES (2);**

**-- INSERT lectura**

**INSERT INTO lectura (id\_sensor, fecha, hora, ppm, nivel\_riesgo, latitud, longitud) VALUES (20, '2025-05-20', '08:00:02', 481.41, 'Altamente peligroso', 6.2372, -75.5647);**

**INSERT INTO lectura (id\_sensor, fecha, hora, ppm, nivel\_riesgo, latitud, longitud) VALUES (10, '2025-05-20', '08:00:31', 296.41, 'Altamente peligroso', 6.2176, -75.5965);**

**INSERT INTO lectura (id\_sensor, fecha, hora, ppm, nivel\_riesgo, latitud, longitud) VALUES (18, '2025-05-20', '08:00:46', 386.54, 'Altamente peligroso', 6.2444, -75.5604);**

**-- INSERTS producto**

**INSERT INTO producto (nombre\_producto) VALUES ('Benceno Refinado');**

**INSERT INTO producto (nombre\_producto) VALUES ('Tolueno Industrial');**

**-- INSERTS filtro**

**INSERT INTO filtro (id\_producto) VALUES (1);**

**INSERT INTO filtro (id\_producto) VALUES (2);**

**-- INSERTS empleado**

**INSERT INTO empleado (nombre\_empleado) VALUES ('Darwin Salazar');**

**INSERT INTO empleado (nombre\_empleado) VALUES ('Lisa Morelos');**

**-- INSERTS turno**

**INSERT INTO turno (descripcion\_turno) VALUES ('08:00-16:00');**

**INSERT INTO turno (descripcion\_turno) VALUES ('16:00-00:00');**

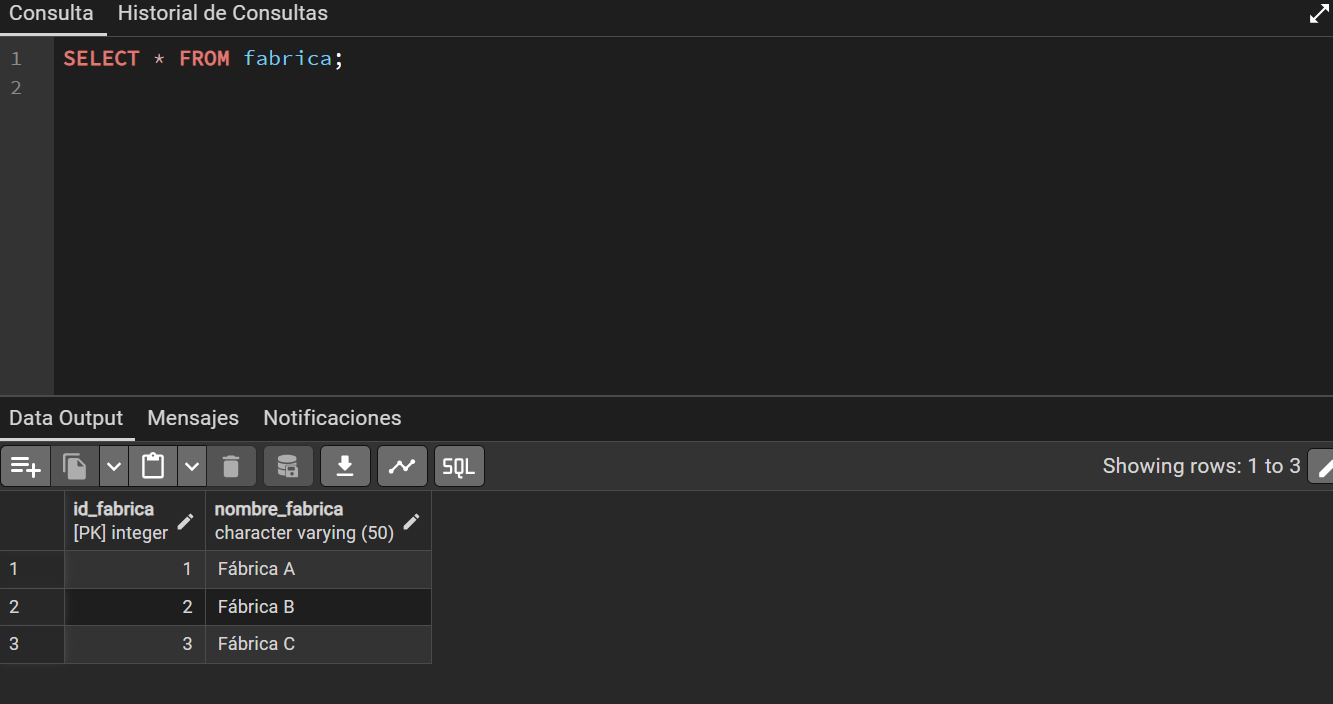
**INSERT INTO turno (descripcion\_turno) VALUES ('00:00-08:00');**

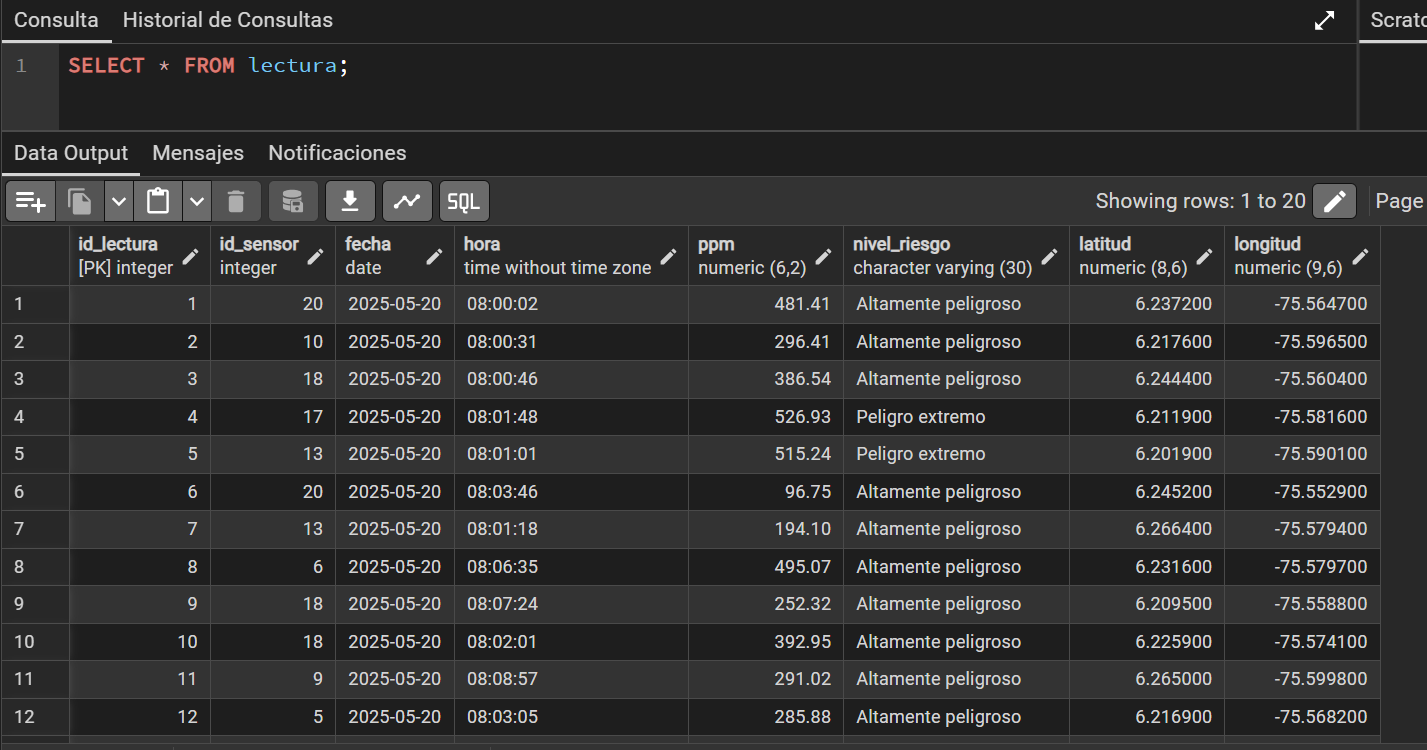
**-- INSERTS empleado\_turno**

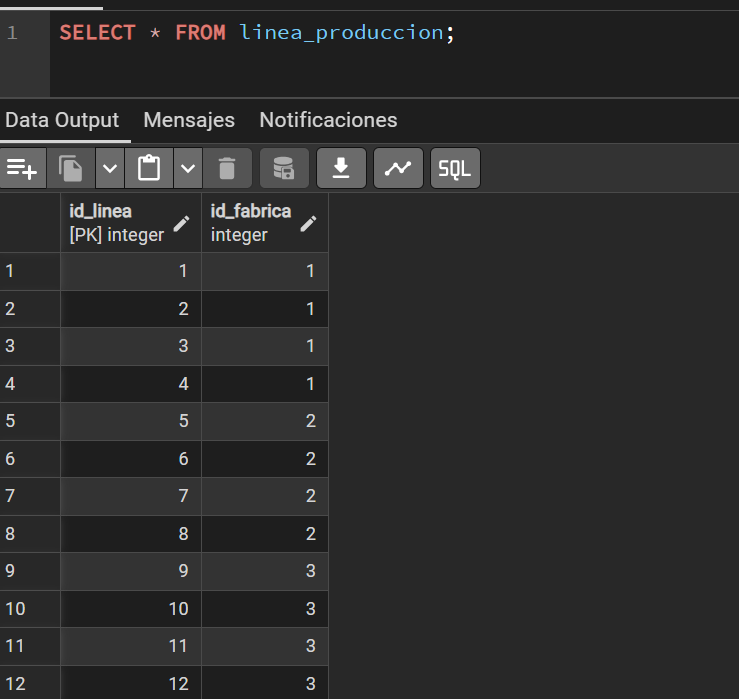
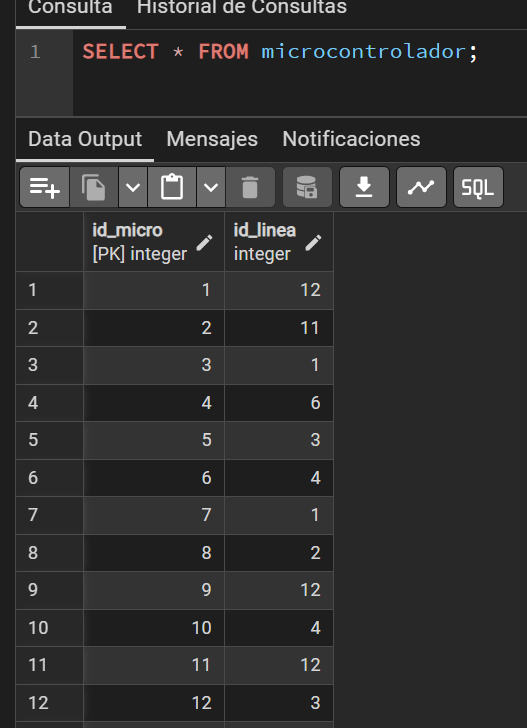
**INSERT INTO empleado\_turno (id\_empleado, id\_turno, fecha\_turno) VALUES (4, 3, '2025-02-14');**

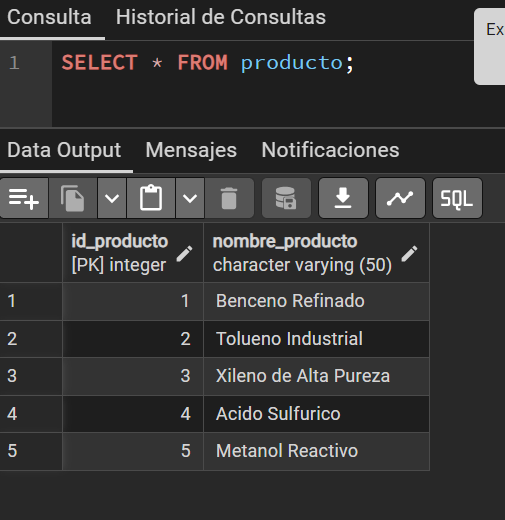
**INSERT INTO empleado\_turno (id\_empleado, id\_turno, fecha\_turno) VALUES (4, 2, '2025-03-04');**

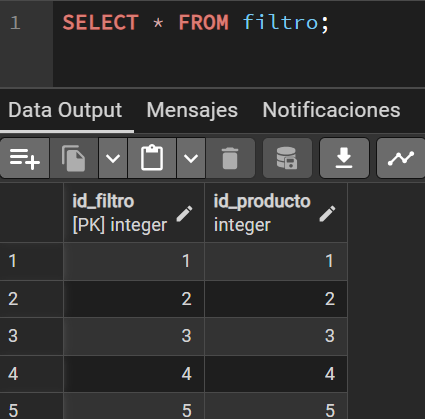
**7.2- Pantallazos de consultas SELECT de las tablas pobladas**

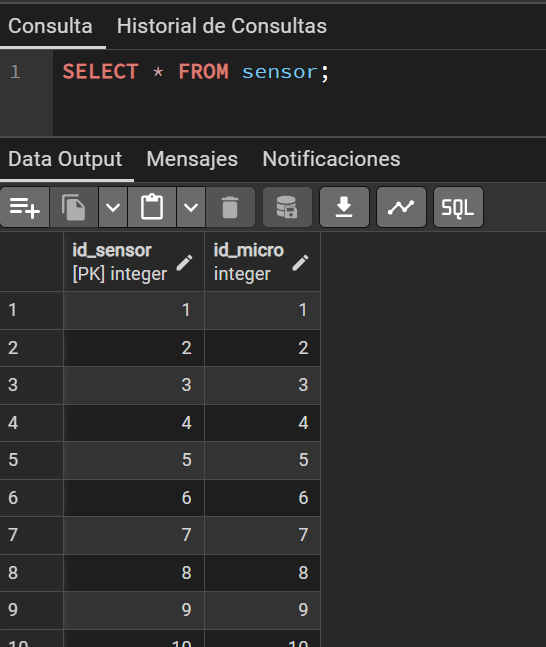


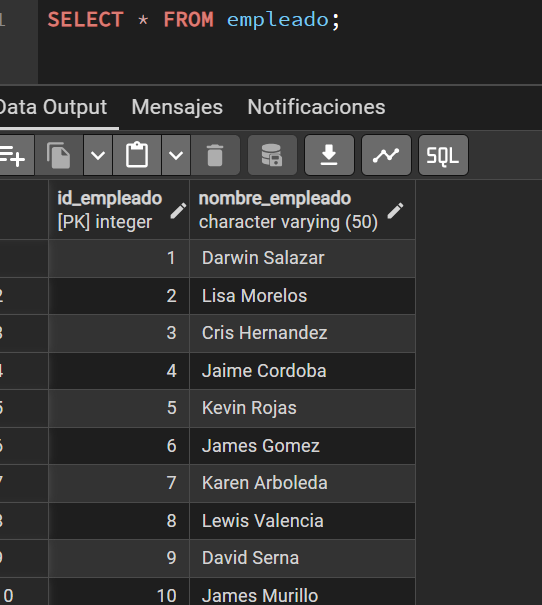
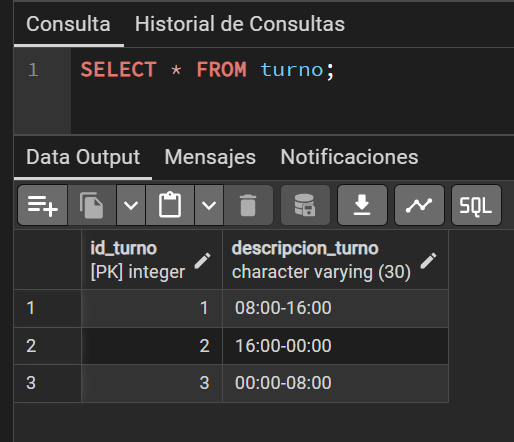


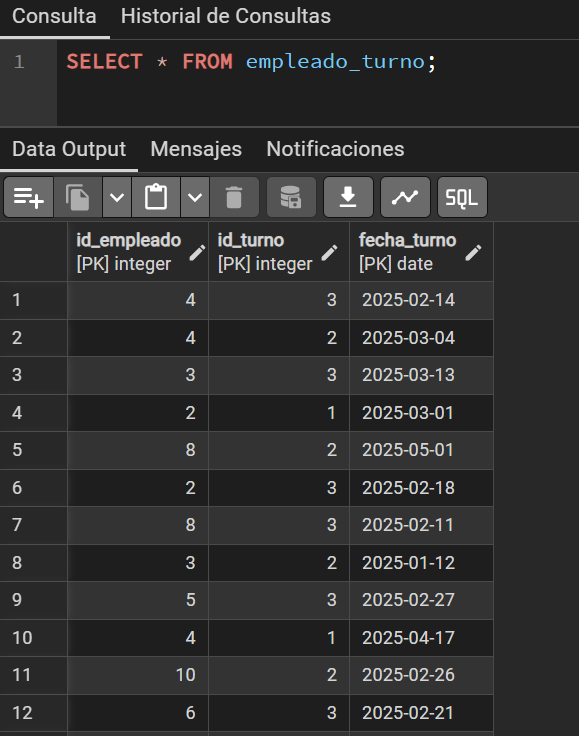












**8.- Conclusiones.**

*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.*

*Keisy Valoyes:*

*Al inicio no creí que esta tarea fuera tan extensa, parecía más simple, pero resulto ser una actividad bastante larga que exige mucha dedicación. Sin embargo, con esfuerzo y compromiso pude sacarla adelante. Algunas partes fueron confusas, pero investigando con cuidado logre comprender y resolver lo necesario. Esta experiencia me dejo aprendizajes importantes sobre la estructura de un sistema Big Data real y siento que ahora tengo más herramientas y confianza para enfrentar retos similares en el ámbito profesional.*

**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*