**INFORME DE RESULTADOS**

**Unidad 3 - Tarea 7**

**Peso: 20%**

**Equipo/Grupo :** Grupo 6

**Estudiantes :** Juan Pablo Arango Gaviria

Maria Camila Duran Garcia

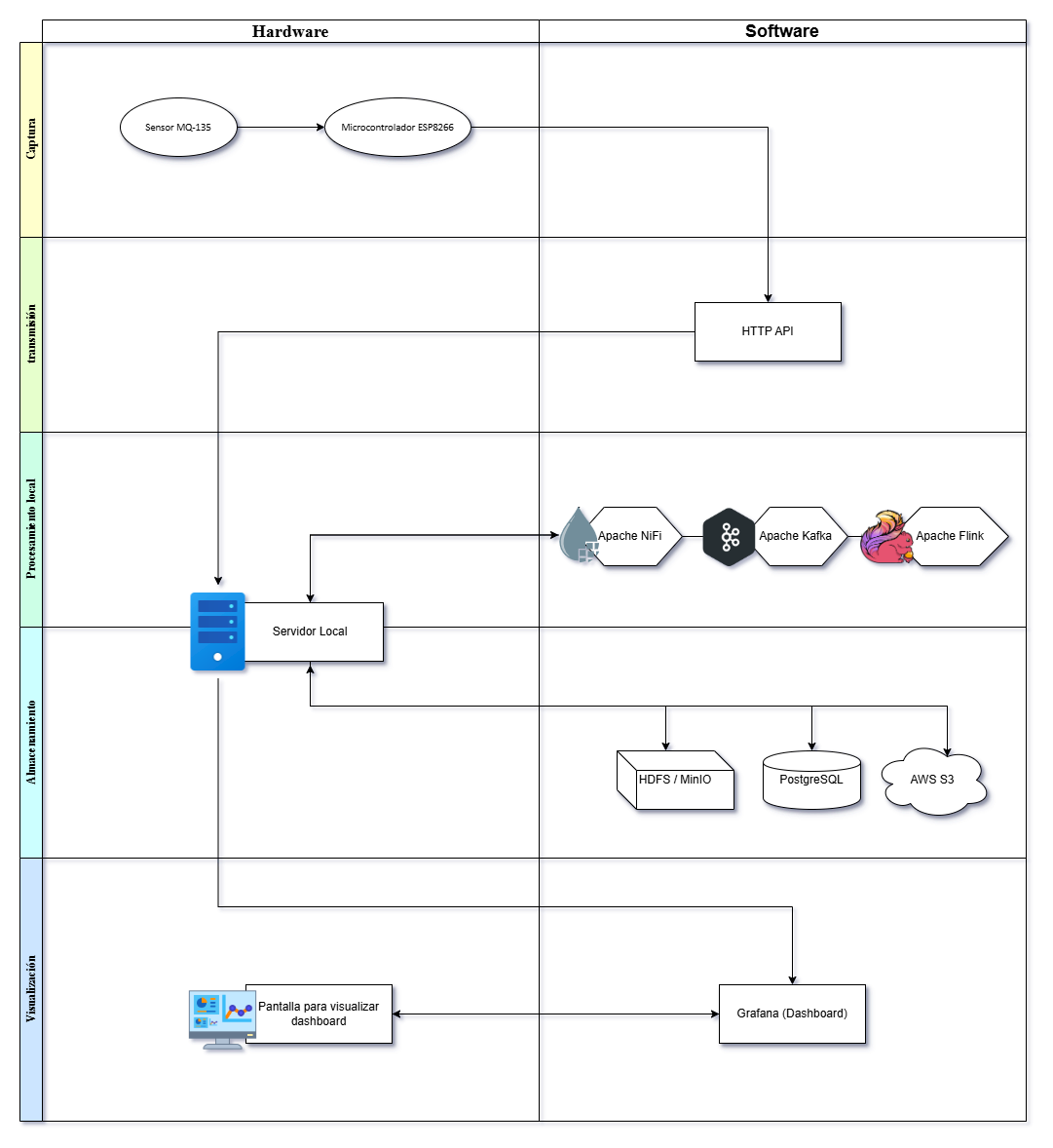
**Objetivo General**

**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)

[Diseño](https://drive.google.com/file/d/1ty6wfWQxB1wan0TFFObngqlcIhiZVu8F/view?usp=sharing)

**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 | 1440 minutos/dia \* 6 lecturas por minuto | 8640 lecturas |
| Lecturas de un mes de una línea de producción. Nota: un mes = 30 días. | 8640 lecturas/ dia \* 20 sensores por 30 dias mes | 5.184.000 lecturas |
| Lecturas de un año de todas las líneas de producción de todas las fábricas. Nota: un año = 365 días. | Lecturas dia por linea de produccion \* 4 lineas cada fabrica\* 3 fabricas totales \*365 dias | 756.864.000 lecturas |

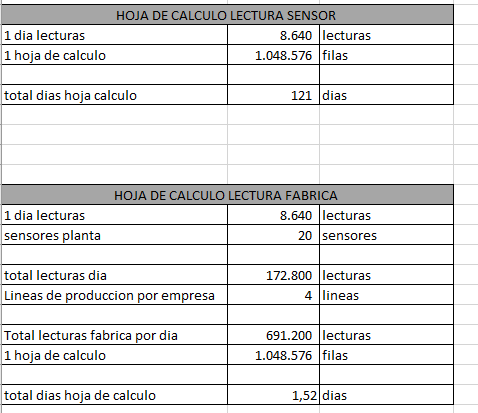
**2.2.- Almacenamiento en tabla “lecturas”**

| **Período de almacenamiento** | **Lecturas** | **Tamaño Tupla** | **Total Bytes** |
| --- | --- | --- | --- |
| Un (1) minuto | **6** | 67 | 402 |
| Un (1) hora | **360** | 67 | 24,120 |
| Un (1) día | **8640** | 67 | 578,880 |
| Un (1) mes | **259,200** | 67 | 17,366,400 |
| Un (1) año | **3,153,600** | 67 | 211,891,200 |

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

| **Lote de “lecturas”** | **Lecturas** | **Tamaño Tupla** | **Total Bytes** |
| --- | --- | --- | --- |
| Lote 1 | **3.000.000** | 67 | 201.000.000 |
| Lote 2 | **20.000.000** | 67 | 1.340.000.000 |

***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 cómo realizó los cálculos.***



En la hoja de cálculo de lector-sensor se hace el cálculo teniendo en cuenta que cada minuto se realiza 6 lecturas y un día tiene 1.440 minutos por lo cual cada día se tienen 8.640 lecturas por sensor y teniendo en cuenta que según Microsoft una hoja de cálculo de Excel tiene 1’048.576 filas nos da como resultado que se debe limpiar cada 121 días la hoja antes de sobrecargarla.

Y en la hoja de lector- fábrica se hace el cálculo teniendo en cuenta que las lecturas por día por línea de producción son 172.800 y cada fábrica tiene 4 líneas cada día se tendrían 691.200 lecturas y nuevamente al saber que la hoja de cálculo tiene 1’048.576 filas nos da como resultado que cada 1.5 días se debe limpiar la hoja de cálculo.

Referencia: Gomez, M. (2023, June 16). ¿Cuántas hojas de cálculo permite Excel? Descubre el límite - Blog - Cámara de Comercio Exterior. Blog - Cámara de Comercio Exterior. <https://blog.camaracomercioexterior.org/excel/cuantas-hojas-de-calculo-permite-excel-descubre-el-limite-ofimatica-windows-365-mac/>

**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(MENSUAL)** | **Costo Total(MENSUAL)** |
| --- | --- | --- | --- |
| Almacenamiento en Bloque Elástico (EBS) | 70GB | $0.00000000011 | $11.25 |
| Amazon Simple Storage Service (Amazon S3) | 6GB | $0.00000000011 | $0.68 |
| Amazon Aurora | 70GB | $0.00000000016 | $341.51 |

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

Las diferencias de precios se deben a:

Amazon EBS

* Baja latencia
* Requiere instancias EC2 asociadas
* Ideal para cargas de trabajo transaccionales

Amazon S3

* Mayor latencia
* Diseñado para durabilidad extrema
* Costos más bajos por volumen

Amazon Aurora

* Base de datos relacional gestionada
* Optimizado para operaciones complejas de bases de datos

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

Detalles EBS

* Región: US East (Ohio)
* Tipo de volumen: gp3
* Capacidad: 70 GB
* IOPS: 3,000
* Throughput: 125 MB/s
* Snapshots: 1 diario (1 GB retenido)
* Duración: 730 horas/mes
* Almacenamiento principal= 70 GB × $0.08/GB-mes = $5.60
* Total snapshots: $5.00

Total EBS: $11.65/mes

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

[*Calculo AWS*](https://drive.google.com/file/d/1Wvhs8wGrIzdM8JqBvLAP8-IVee5vNC6J/view?usp=sharing)

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

| **Nombre Servicio** | **Costo individual** | **Costo anual** |
| --- | --- | --- |
| EBS | 11.25 USD | 135 USD |
| S3 | 0.68 USD | 8.16 USD |
| AURORA POSTGREST | 341.5 USD | 4098 USD |

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

[**Diseño de Arquitectura**](https://drive.google.com/file/d/128bdtFWXce4j_mq7jMDLkFyWhDSAR5Xm/view?usp=sharing)

1. **Apache Kafka (Cloud Gateway):** Nos permite recibir flujos de datos en tiempo real desde cada fábrica (vía conectores o APIs) y los canaliza hacia el clúster Hadoop/Spark. Usado como el punto de entrada en la nube.
2. **Apache Spark (Procesamiento)**: Realiza el procesamiento en tiempo real (streaming) o en batch, haciendo limpiezas, filtrando y enriqueciendo los datos que llegan desde Kafka.
3. **Hadoop HDFS:** Es el sistema de almacenamiento distribuido donde se guardará el histórico de datos crudos y procesados, preparado para análisis masivo.
4. **Apache Hive:** Nos proporciona una capa de consulta tipo SQL sobre HDFS. Aquí se pueden modelar los datos, se generan tablas, particiones y se hacen transformaciones ETL.
5. **Amazon QuickSight**: Se conecta a las tablas de Hive o almacenes derivados y genera visualizaciones de métricas, alertas

y reportes clave de emisiones por fábrica.

**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*

**No lo realizamos**

**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.*

**No lo realizamos**

**8.- Conclusiones.**

Este proyecto nos permitió diseñar e implementar un sistema completo de monitoreo de emisiones de gases, desde la captura de datos en sensores hasta su visualización en un dashboard, a través de cálculos precisos de almacenamiento y procesamiento, se determinó que el volumen de datos generado es significativo, lo que nos hace ver la magnitud de datos que se pueden llegar a tener teniendo en cuentas líneas de tiempo como meses y años donde el uso de soluciones escalables en la nube como AWS nos brinda una ayuda magnífica.

La arquitectura propuesta integra tanto componentes locales como servicios en la nube, garantizando eficiencia y escalabilidad, también la implementación de herramientas para visualización y la exploración de tecnologías como Hadoop, Kafka y Spark nos demuestra cómo este sistema puede crecer para incorporar más fábricas sin perder rendimiento.

Esta tarea reforzó nuestros conceptos clave de la asignatura, como bases de datos, procesamiento de datos masivos, arquitectura de sistemas y análisis, aplicándolos a un caso real de monitoreo industrial, lo cual nos beneficiará enormemente en nuestro campo laboral.

**Conclusiones Juan Pablo Arango**

Al inicio de la asignatura, tenía una expectativa totalmente diferente esperaba aprender sobre gestión de datos y sistemas de información, pero el proyecto superó mis expectativas al permitirme diseñar una solución completa, desde el cálculo de costos en AWS hasta la implementación de un dashboard funcional y más con el que realice en el trabajo de recuperación.

El conocimiento adquirido en arquitecturas, bases de datos en la nube y visualización de datos no solo se me hace relevante para mi formación profesional, sino que también me abre oportunidades en campos como analista de datos, ingeniero de cloud o porque no especialista en big data.

Destaco especialmente la importancia de optimizar costos en la nube y el uso de herramientas como Metabase para tomar decisiones interpretando los datos de una manera correcta. Esta asignatura me ha dado confianza para enfrentar desafíos tecnológicos en el ámbito laboral y me motiva a seguir profundizando en análisis de datos y en el big data y como lo dije antes por qué no especializarme en este campo.

**Conclusiones Maria Camila Duran Garcia**

Recuerdo que cuando inicié con el curso sentí el primer día de clase, una sensación de curiosidad indescriptible. Era un momento en donde los datos estaban más presentes en mi día a día y conocer la orientación del curso generó en mí un completo deseo de verlo y aprenderlo de principio a fin.

El curso de big data marca en mi un antes y después en no solo cómo se gestionan los datos (Incluye todos los procesos y los sistemas) sino en cómo los veo y entiendo. Fue descubrir la importancia de estos y las formas tan atrevidas y cómodas de trabajarlos.

No suficiente con conocer diferentes sistemas, pudimos tener un acercamiento con Amazon AWS, algo que estaba trabajando pero que no comprendí en su totalidad. Definitivamente este curso ha sido uno de los mejores y en el que he encontrado más que nunca, la pasión en cada paso.

Me siento mejor desarrolladora, ya no le tengo tanto temor a la data y eso es simplemente un orgullo para mi.

**9.- Video de sustentación:**

[**Video de sustentación**](https://drive.google.com/file/d/1dsoskDz8VBulMrHE1DZJgoVKSvF1t3Qx/view?usp=sharing)