# UNIVERSIDAD DEL ISTMO, CAMPUS IXTEPEC

LICENCIATURA EN INFORMÁTICA

ALUMNO: JOSÉ IVÁN GARCÍA GÓMEZ

PROFESOR: FLORENTINO RUIZ AQUINO

ASIGNATURA: SISTEMAS DE INFORMACIÓN I

EJERCICIO 2: "EL DILUVIO DE DATOS: INFRAESTRUCTURA BIG DATA" (ENFOQUE: DATA LAKE & BIG DATA)

SEMESTRE: 2025-20256 A

GRUPO: 908

FECHA: 05 DE ENERO DEL 2026

**Ejercicio 2: "El Diluvio de Datos: Infraestructura Big Data" (Enfoque: Data Lake & Big Data)**

Objetivo:

Proponer una arquitectura capaz de manejar las "3 V" (Volumen, Velocidad, Variedad) para un entorno que el SQL tradicional no puede soportar.

El Caso:

Una empresa de logística llamada "TransLog" está instalando sensores IoT en sus 2,000 camiones. Estos sensores envían datos de GPS, temperatura del motor y frenadas bruscas cada 5 segundos (Velocidad y Volumen). Además, quieren analizar fotos de los choferes para detectar fatiga y comentarios en redes sociales sobre sus entregas (Variedad: datos no estructurados). Un Data Warehouse tradicional colapsaría por el costo y la rigidez.

1. **Arquitectura por capas**

- Zona Bronze/Raw

  Almacena los datos crudos tal como se generan, sin limpiarlos ni modificarlos, para conservar la información original

  Incluiría:

  - Datos de los sensores loT como GPS, temperatura de motor y frenadas bruscas.
  - Fotos de choferes
  - Comentarios en redes sociales

  Con formatos JSON, CSV, imágenes, texto plano.

- Zona Silver/Refined

  Limpiar, ordena y transforma los datos, corrigiendo errores y estandarizando formatos para que sean confiables y analizables.

  - Limpieza de datos.
  - Eliminación de registros duplicados, errores.
  - Corrección de datos incompletos

- Datos listos para realizar análisis.

- Zona Gold/Curated

  Analiza los datos y generar información útil**,** como métricas, indicadores y reportes que apoyan la toma de decisiones.

  Métricas:

  - Riesgo de conducción.
  - Fatiga del conductor.
  - Desempeño del motor.
  - Sentimiento de cliente.

  Indicadores:

  - Indicadores de mantenimiento predictivo.
  - Análisis de sentimiento en redes sociales.

2. **Selección de Tecnologías**

   - **Apache Kafka para ingestión de datos**

     Se utiliza para recibir y distribuir los datos en tiempo real provenientes de los sensores IoT de los camiones, los cuales envían información cada 5 segundos. Kafka permite manejar alta velocidad y gran cantidad de eventos sin pérdida de datos.

   - **Hadoop/Spark para procesamiento**

     Hadoop proporciona el entorno distribuido para trabajar con grandes volúmenes de datos, mientras que Apache Spark se encarga del procesamiento rápido tanto en batch como en streaming, permitiendo analizar datos de sensores, imágenes y texto de manera eficiente.

   - **AWS S3 o Azure Blob para almacenamiento**

     Se utilizan como almacenamiento del Data Lake debido a su alta escalabilidad, bajo costo y capacidad para guardar datos estructurados y no estructurados como archivos, imágenes y texto.

Un Data Lake es un sistema para guardar y analizar grandes volúmenes de datos de diferentes tipos en su forma original.

**¿Por qué SQL tradicional no sirve en este caso?**

- No maneja eficientemente datos no estructurados (imágenes, texto de redes sociales).
- Tiene limitaciones de escalabilidad frente a grandes volúmenes de datos.
- No está diseñado para procesamiento en tiempo real (streaming)**.**
- Su costo aumenta considerablemente al crecer la cantidad de datos.
- Es rígido ante cambios constantes en la estructura de los datos.

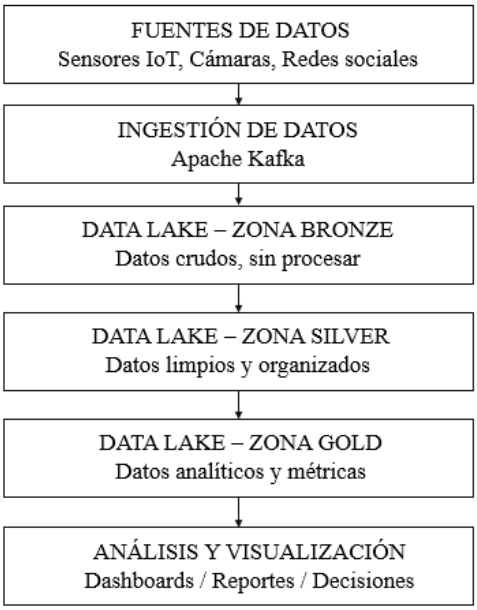
3. **Dilema Ético y Seguridad**

El uso de información biométrica**,** como las fotos de los choferes, representa un riesgo ético importante, ya que son datos personales sensibles que pueden afectar la privacidad y los derechos de los trabajadores si no se protegen adecuadamente.

Para garantizar la seguridad y la integridad de la información**,** se deben aplicar las siguientes medidas:

- Cifrar las fotos, para que nadie pueda verlas ni cambiarlas, aunque acceda al sistema.
- Limitar el acceso, de modo que solo personas autorizadas puedan ver o usar las imágenes.
- Ocultar la identidad del chofer, usando las fotos solo para análisis y no para identificar personas.
- Registrar quién entra a los datos, para detectar accesos o cambios indebidos.
- Borrar las fotos cuando ya no sean necesarias, evitando almacenarlas sin motivo.

4. **Diagrama de Arquitectura de Flujo de Datos (Pipeline).**




5. **Cuadro comparativo: ¿Por qué un Data Lake y no un Data Warehouse para este caso específico?**

| Característica | Data Lake | Data Warehouse |
|---|---|---|
| Tipo de datos | Acepta datos estructurados, semiestructurados y no estructurados (sensores, imágenes, texto) | Solo datos estructurados |
| Volumen de datos | Maneja grandes volúmenes de datos sin problema | Tiene limitaciones con grandes volúmenes |
| Velocidad | Permite trabajar con datos en tiempo real (streaming) | Procesamiento lento, generalmente por lotes |
| Flexibilidad | Muy flexible ante cambios en los datos | Rígido, requiere esquemas fijos |
| Costo | Más económico y escalable | Alto costo de almacenamiento y mantenimiento |
| Uso en IoT | Ideal para sensores IoT | No adecuado |