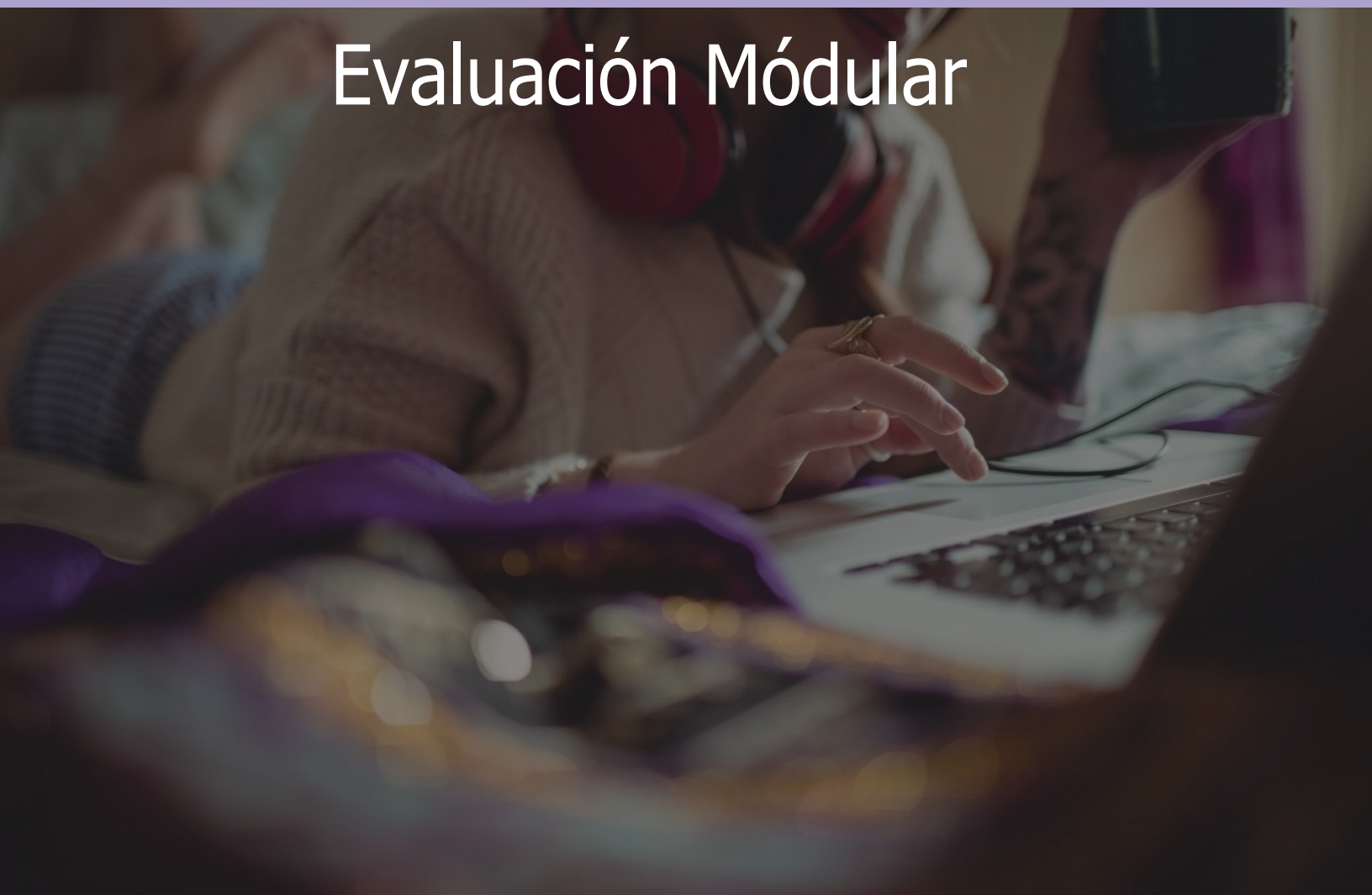
IT Academy
by KIBERNUM

Módulo 8

Evaluación Módular



ACTIVIDAD:

Automatización de Flujos de Datos con Apache NiFi y Kafka

Objetivo:
Evaluar la comprensión de los conceptos clave sobre la integración de datos, la automatización de flujos de datos, y el uso de plataformas como Apache NiFi y Apache Kafka para la ingesta de datos en tiempo real y por lotes, asegurando que el estudiante pueda aplicar de manera práctica los conocimientos adquiridos.

Contexto:
A continuación, se presentan dos tareas prácticas relacionadas con la automatización de flujos de datos. Cada tarea debe ser realizada en una plataforma que soporte el uso de Apache NiFi y Apache Kafka para la creación de flujos de datos y procesamiento de información en tiempo real o por lotes.

Tiempo estimado de desarrollo: 120 minutos.
Modalidad de desarrollo: Individual.
Formato de entrega: Word o PDF.

Instrucciones:

Tarea 1: Implementación de un flujo de datos con Apache NiFi (Ingesta Batch).
Objetivo: Crear un flujo de datos en Apache NiFi que lea archivos de datos (por ejemplo, archivos CSV) desde una fuente local, transforme los datos y cargue los datos transformados en una base de datos.

Instrucciones:

1. **Configurar NiFi en tu entorno.** Asegúrate de tener acceso a la interfaz web de NiFi y estar habilitado para crear flujos de trabajo.
2. **Crear un flujo de trabajo que realice lo siguiente:**
    a. **Extraer datos:** utiliza el procesador GetFile para leer los archivos CSV ubicados en un directorio local.
    b. **Transformar datos:** Utiliza el procesador ConvertRecord para transformar los datos a un formato adecuado (por ejemplo, convertir CSV a JSON).
    c. **Cargar datos:** Utiliza el procesador PutDatabaseRecord para cargar los datos transformados en una base de datos (puede ser una base de datos local como PostgreSQL).
3. **Configura el flujo para que se ejecute de forma periódica, por ejemplo, una vez al día.**
4. **Documenta el flujo de trabajo,** indicando cada procesador utilizado y su configuración.

Tarea 2: Implementación de un flujo de datos en tiempo real con Apache Kafka.
Objetivo: Crear un flujo de datos en tiempo real utilizando Apache Kafka para consumir eventos de un productor y procesarlos en un consumidor.

Instrucciones:

1. **Configurar un clúster de Kafka en tu entorno.** Debes asegurarte de que Kafka y Zookeeper estén correctamente instalados y configurados.
2. **Crea un productor de Kafka en Python** que simule eventos de un sistema (por ejemplo, actualizaciones de un sistema de ventas o logs de actividades). El productor enviará mensajes a un topic de Kafka.
3. **Crea un consumidor en Python** que se suscriba al topic de Kafka, lea los mensajes y procese los eventos de acuerdo a una lógica definida (por ejemplo, guardar los eventos en un archivo o base de datos).
4. **Prueba el flujo** asegurándote de que el productor envíe mensajes y el consumidor los procese correctamente en tiempo real.
5. **Documenta el flujo de trabajo,** explicando los pasos para crear el productor y consumidor, así como las configuraciones necesarias en Kafka.

Rúbrica:

Criterio / Indicador	Insuficiente (0%-20%)	Por lograrlo (21%-40%)	Medianamente logrado (41%-60%)	Logrado (61%-80%)	Sobresaliente (81%-100%)
1. Configuración y Ejecución Técnica del Flujo de Datos (40%)	El flujo de datos no se configura ni ejecuta correctamente.	El flujo de datos tiene fallos importantes que afectan su ejecución.	El flujo funciona parcialmente pero con algunos problemas menores en la configuración o ejecución.	El flujo de datos se configura correctamente y ejecuta la mayoría de los pasos sin errores.	El flujo de datos está completamente funcional y ejecutado correctamente.
2. Documentación del Flujo de Trabajo (30%)	La documentación está incompleta o falta información crucial.	La documentación cubre algunos pasos, pero carece de detalles importantes.	La documentación cubre la mayoría de los pasos pero con algunas omisiones o falta de claridad.	La documentación es clara, cubriendo todos los pasos con detalle y organización.	La documentación está perfectamente estructurada, detallada y fácilmente comprensible.
3. Funcionalidad y Pruebas de los Flujos de Datos (20%)	El flujo de datos no funciona y no pasa las pruebas.	El flujo de datos falla en varias pruebas y no procesa correctamente los mensajes.	El flujo de datos pasa algunas pruebas, pero hay errores en ciertas configuraciones o en la transmisión de datos.	El flujo de datos funciona correctamente en la mayoría de las pruebas con algunos errores menores.	El flujo de datos funciona sin errores en todas las pruebas, garantizando su funcionamiento en tiempo real.
4. Innovación y Creatividad (10%)	No se observa ninguna innovación en el enfoque o la solución.	Se presentan ideas mínimamente innovadoras, pero no aportan mejoras significativas.	Se incorporan algunos enfoques creativos, pero en su mayoría siguen enfoques convencionales.	Se muestran ideas innovadoras que mejoran significativamente el flujo de datos.	La solución propuesta es altamente creativa y aporta valor adicional con enfoques innovadores.