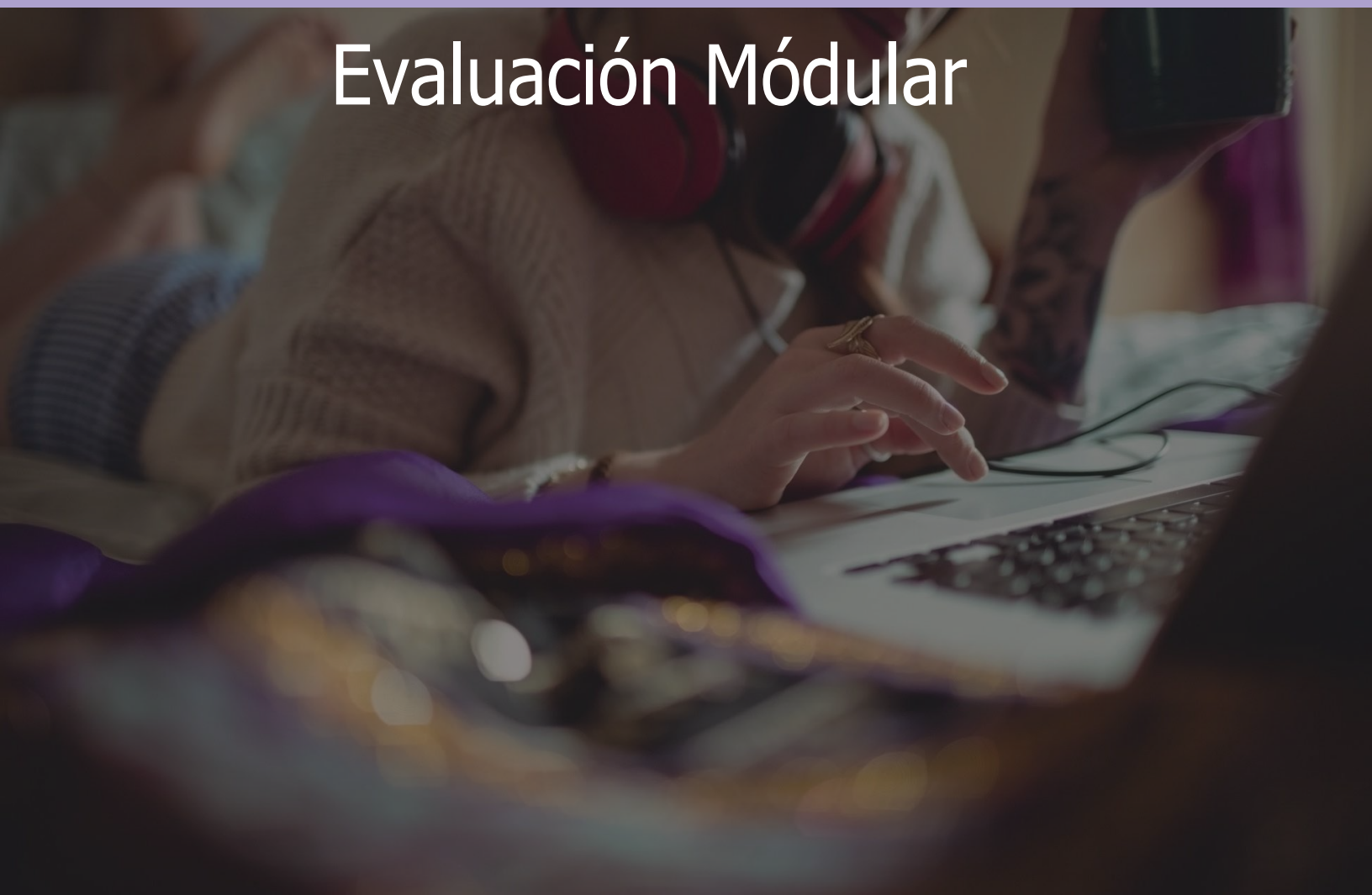
IT Academy
by KIBERNUM

Módulo 7

Evaluación Módular



ACTIVIDAD:

Desarrollo de Soluciones con Tecnologías de Procesamiento Distribuido

Objetivo:
El objetivo de esta evaluación es aplicar los conocimientos adquiridos durante el módulo sobre procesamiento distribuido y sistemas Big Data. Los estudiantes deberán demostrar su comprensión sobre conceptos clave como Big Data, Apache Spark, procesamiento en tiempo real, y el uso de tecnologías como Hadoop, MLlib, y Spark SQL a través de preguntas de respuesta corta y análisis de casos prácticos.

Contexto:
La evaluación incluirá una combinación de preguntas teóricas, análisis de casos prácticos y ejercicios aplicados en los que se utilizarán herramientas de procesamiento distribuido, SQL y machine learning. Los estudiantes deberán desarrollar soluciones y explicar su implementación en un entorno de Big Data.

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

Instrucciones:

Preguntas Teóricas: Responde las siguientes preguntas de manera concisa y clara. Se evaluará tanto la comprensión de los conceptos como la capacidad para aplicarlos.
Caso Práctico: Analiza el siguiente escenario práctico y propone una solución utilizando las tecnologías estudiadas (Apache Spark, Hadoop, SQL, etc.).
Ejercicio: Completa los ejercicios de desarrollo propuestos para crear una solución usando Spark SQL, RDDs, y otros módulos de Big Data.

Preguntas Teóricas

1. Explica qué son las 5V's de Big Data y por qué son fundamentales en el procesamiento de grandes volúmenes de datos.
2. Describe la diferencia principal entre procesamiento batch y procesamiento en tiempo real. Menciona las tecnologías asociadas con cada tipo de procesamiento.
3. ¿Qué es un RDD en Apache Spark? ¿Cuál es la ventaja de usar RDDs sobre bases de datos tradicionales para el procesamiento distribuido?
4. Define Apache Spark SQL. ¿Cómo se diferencia de otros motores SQL tradicionales en términos de rendimiento y escalabilidad?

Caso Práctico

Escenario: Una empresa desea analizar las transacciones bancarias en tiempo real para detectar fraudes. Se están generando miles de transacciones por segundo, y se necesita una plataforma capaz de procesar estos flujos de datos de manera continua y eficiente.
Tareas:

1. Explica cómo Apache Spark Streaming puede resolver este caso. ¿Qué componentes de Spark serían necesarios?
2. Propón una solución usando DStreams y micro-batching. Describe los pasos para configurar el sistema de procesamiento de datos en tiempo real.
3. ¿Qué beneficios tendría la integración con Apache Kafka para la ingesta de datos en este escenario?

Ejercicio

1. Implementa una consulta en Spark SQL para analizar un conjunto de datos de transacciones bancarias. La consulta debe identificar las transacciones de mayor valor por cliente, considerando los campos ClienteID, Monto, y Fecha. Explica cómo optimizarías esta consulta para manejar grandes volúmenes de datos.
2. Diseña una solución de machine learning usando MLlib para predecir la probabilidad de que una transacción sea fraudulenta, utilizando características como el Monto, Ubicación y Hora de la transacción. Menciona los pasos para entrenar y evaluar el modelo.
3. Proporciona una implementación básica para procesar un flujo de datos en tiempo real usando Structured Streaming. Explica cómo manejarías los eventos fuera de orden y qué técnicas de watermarking emplearías.

Rúbrica:

Indicador de logro / criterio	Insuficiente (0%-20%)	Por lograrlo (21%-40%)	Medianamente logrado (41%-60%)	Logrado (61%-80%)	Sobresaliente (81%-100%)
Comprensión de conceptos fundamentales del Big Data	No identifica ni explica los conceptos clave de Big Data.	Muestra comprensión parcial y poco precisa de los conceptos de Big Data.	Identifica algunos conceptos, pero con explicaciones incompletas o con errores.	Demuestra comprensión clara de la mayoría de los conceptos clave.	Explica con claridad y profundidad todos los conceptos de Big Data, integrando ejemplos relevantes.
Aplicación de tecnologías (Apache Spark, Hadoop, etc.)	No clasifica ni aplica los conceptos correctamente.	Aplica parcialmente tecnologías, pero con errores o sin justificación.	Aplica tecnologías con justificación limitada e incompleta.	Aplica correctamente las tecnologías de procesamiento distribuido con justificación.	Aplica y justifica con precisión todas las tecnologías, integrando ejemplos adecuados.
Uso de Spark SQL y procesamiento en tiempo real	No aplica ni explica correctamente el uso de Spark SQL o el procesamiento en tiempo real.	Muestra un intento parcial de explicar o aplicar el uso de Spark SQL y procesamiento en tiempo real.	Aplica conceptos de Spark SQL y procesamiento en tiempo real, pero con errores o de manera incompleta.	Utiliza Spark SQL y técnicas de procesamiento en tiempo real de manera correcta y bien justificada.	Aplica y justifica con precisión el uso de Spark SQL y procesamiento en tiempo real con ejemplos detallados.
Implementación de machine learning con MLlib	No presenta una solución de machine learning o la presenta erróneamente.	Muestra una comprensión limitada de MLlib con errores en la implementación.	Implementa una solución con MLlib, pero con justificación o explicaciones incompletas.	Implementa correctamente una solución con MLlib y la justifica adecuadamente.	Implementa una solución avanzada usando MLlib, con justificación precisa y ejemplos claros.
Manejo de flujo de datos en streaming (DStreams, Structured Streaming)	No identifica ni utiliza correctamente DStreams o Structured Streaming.	Muestra comprensión parcial de DStreams y Structured Streaming, con errores en la implementación.	Aplica conceptos de DStreams y Structured Streaming, pero de manera incompleta o errónea.	Aplica correctamente DStreams y Structured Streaming con justificación adecuada.	Aplica y justifica con precisión el uso de DStreams y Structured Streaming en escenarios reales.