Universidad Nacional Abierta y a Distancia
Vicerrectoría Académica y de Investigación
Curso: Big Data
Código: 202016911

Guía de actividades y rúbrica de evaluación – Tarea 3
Procesamiento de Datos con Apache Spark

1. Descripción de la actividad

Tipo de actividad: En grupo colaborativo	
Momento de la evaluación: Intermedio	
Puntaje máximo de la actividad: 115 puntos	
La actividad inicia el: miércoles, 2 de octubre de 2024	**La actividad finaliza el:** martes, 29 de octubre de 2024

Con esta actividad se espera conseguir los siguientes resultados de aprendizaje:

Resultado de aprendizaje 2: Diseñar e implementar soluciones de almacenamiento y procesamiento de Big Data utilizando herramientas como Hadoop, Spark y Kafka.

Resultado de aprendizaje 3: Aplicar técnicas de análisis y visualización a grandes conjuntos de datos para obtener información útil.

La actividad consiste en:

1. **Definición del problema y conjunto de datos:** Seleccionar un problema que pueda ser resuelto mediante el análisis de un conjunto de datos de gran volumen.

   - El conjunto de datos puede ser propio o provenir de alguna fuente pública (Kaggle, UCI Machine Learning Repository, datos abiertos de Colombia, etc.).

2. **Implementación en Spark:** Desarrollar aplicaciones en Spark (utilizando Python) que realicen las siguientes tareas:

- **Procesamiento en batch:**
  - Cargar el conjunto de datos seleccionados desde la fuente original.
  - Realizar operaciones de limpieza, transformación y análisis exploratorio de datos (EDA) utilizando RDDs o DataFrames.
  - Almacenar los resultados procesados.

- **Procesamiento en tiempo real (Spark Streaming & Kafka):**
  - Configurar un topic en Kafka para simular la llegada de datos en tiempo real (usar un generador de datos).
  - Implementar una aplicación Spark Streaming que consuma datos del topic de Kafka.
  - Realizar algún tipo de procesamiento o análisis sobre los datos en tiempo real (contar eventos, calcular estadísticas, etc.).
  - Visualizar los resultados del procesamiento en tiempo real.

3. **Investigación:** Realizar las siguientes actividades:

- Elaborar una tabla comparativa que resalte las diferencias clave entre Hadoop y Spark en términos de arquitectura, procesamiento, rendimiento y casos de uso
- Definir los concepto de RDD. Propiedades de los RDDs: Inmutabilidad, particionamiento. Tipos de Operaciones:
  - Transformaciones: map, filter, flatMap, groupByKey, reduceByKey.
  - Acciones: collect, count, take, reduce.
- Concepto de DataFrame. Diferencias y ventajas sobre los RDDs.
- Crear un diagrama que ilustre la arquitectura de Kafka y explicar los conceptos clave como topics, partitions, brokers, producers y consumers.

4. **Documentación y presentación:**

- **Foro de discusión:** Compartir los avances del proyecto, problemas encontrados y soluciones implementadas.

- **Repositorio de código:** Publicar el código fuente del proyecto en un repositorio Git (GitHub, GitLab, etc.) con una descripción clara de la solución y las instrucciones para su ejecución.

- **Presentación online:** Elaborar una presentación grupal que incluya:
  - Introducción al problema y conjunto de datos.
  - Arquitectura de la solución en Spark.
  - Explicación del código y las tecnologías utilizadas (RDDs, DataFrames, Spark SQL, Spark Streaming, Kafka).
  - Demostración de la ejecución de la aplicación.
  - Visualización de resultados.
  - Respuesta a las actividades de investigación
  - Conclusiones y aprendizajes obtenidos.
  - Referencias bibliográficas (Normas APA).

- **Informe escrito (Tarea3_grupo.pdf):** El documento debe contener:
  - Portada
  - Introducción
  - Objetivo General
  - Objetivos Específicos
  - Descripción del problema y conjunto de datos.
  - Diseño de la solución y arquitectura.
  - Explicación detallada del código y las tecnologías utilizadas.
  - Análisis de resultados obtenidos.
  - Enlace al repositorio del código.
  - Respuesta a las actividades de investigación.
  - Enlace a la presentación online.
  - 5 conclusiones generales del grupo.
  - Referencias Bibliográficas utilizadas con normas APA.

Para el desarrollo de la actividad tenga en cuenta que:

En el entorno de Información inicial debe:

Revisar la agenda del curso y ajustar sus actividades para realizar

la entrega del producto final de acuerdo con la fecha establecida.

En el entorno de Aprendizaje debe:

- Realizar las lecturas recomendadas para cada tema y participar activamente en las actividades propuestas.
- Participar en el foro de la tarea 3 – Procesamiento de Datos con Apache Spark, publicando los avances de la actividad y la presentación online con los siguientes elementos (Introducción al problema y conjunto de datos. Arquitectura de la solución en Spark. Explicación del código y las tecnologías utilizadas (RDDs, DataFrames, Spark SQL, Spark Streaming, Kafka). Demostración de la ejecución de la aplicación. Visualización de resultados. Respuesta a las actividades de investigación . Conclusiones y aprendizajes obtenidos).
- Revisar el trabajo final para que esté en conformidad con lo solicitado en la guía y rúbrica de la actividad.

En el entorno de Evaluación debe:

- Realizar la entrega del documento final en formato PDF.

Evidencias de trabajo independiente:

Las evidencias de trabajo independiente para entregar son:

1. **Participación en el foro:** Realizar mínimo 2 aportes de calidad, mostrando un análisis profundo del tema.
2. **Desarrollo de código:** Implementar al menos una parte específica del proyecto (ej: procesamiento en batch o en tiempo real).
3. **Pruebas y depuración:** Asegurar la funcionalidad del código desarrollado.
4. **Conclusiones individuales:** Redactar dos (2) conclusiones individuales sobre la Unidad 2.

Nota: Recuerde que su trabajo será evaluado con la herramienta turnitin, por lo tanto, es muy importante citar correctamente las referencias en el ensayo.

Evidencias de trabajo grupal:

Las evidencias de trabajo grupal a entregar son:

1. **Integración del código:** Unificar el código desarrollado por cada integrante en un repositorio común.
2. **Presentación unificada:** Asegurar que la presentación online sea clara, concisa y bien organizada.
3. **Entrega del documento final en formato PDF (Tarea3_grupo.pdf):** Un solo documento por grupo con la información completa. El documento debe contener:

    - Portada
    - Introducción
    - Objetivo General
    - Objetivos Específicos
    - Descripción del problema y conjunto de datos.
    - Diseño de la solución y arquitectura.
    - Explicación detallada del código y las tecnologías utilizadas.
    - Análisis de resultados obtenidos.
    - Enlace al repositorio del código.
    - Respuesta a las actividades de investigación
    - Enlace a la presentación online.
    - 5 conclusiones generales del grupo.
    - Referencias Bibliográficas utilizadas con normas APA.

Aspectos a tener en cuenta

- **Funcionalidad:** La aplicación desarrollada debe funcionar correctamente y cumplir con los objetivos del proyecto.
- **Eficiencia:** Se valorará la eficiencia del código y la optimización de los procesos.
- **Organización:** Mantener una estructura de código clara y organizada.
- **Claridad en la comunicación:** La información proporcionada en el foro, presentación e informe debe ser fácil de entender.

Formato:

- Seguir las mejores prácticas de desarrollo de software.
- Comentar el código de forma adecuada.
- Utilizar nombres de variables y funciones descriptivos.

2. Lineamientos generales para la elaboración de las evidencias de aprendizaje a entregar.

Para evidencias elaboradas **en grupo colaborativamente**, tenga en cuenta las siguientes orientaciones

1. Realice un reconocimiento general del curso y de cada uno de los entornos antes de abordar el desarrollo de las actividades.

2. Identifique y lea los recursos y los referentes de la unidad 1 que corresponden a la actividad.

3. Intervenga en el foro de discusión aplicando las normas de Netiqueta Virtual, evidenciando siempre respeto por las ideas de sus compañeros y del cuerpo docente.

4. Antes de entregar el producto solicitado revise que cumpla con todos los requerimientos que se señalaron en esta guía de actividades, rúbrica de evaluación y por parte del tutor en el foro de discusión.

5. No cometa fraudes, ni plagios ni actos que atenten contra el normal desarrollo académico de las actividades.

6. Todos los integrantes del grupo deben participar con sus aportes en el desarrollo de la actividad.

7. En cada grupo deben elegir un solo integrante que se encargará de entregar el producto solicitado en el entorno que haya señalado el docente.

8. Solo se deben incluir como autores del producto entregado, a los integrantes del grupo que hayan participado con aportes durante el tiempo destinado para la actividad.
Tenga en cuenta que todos los productos escritos individuales o grupales deben cumplir con las normas de ortografía y con las condiciones de presentación que se hayan definido. En cuanto al uso de referencias considere que el producto de esta actividad debe cumplir con las normas **APA**
En cualquier caso, cumpla con las normas de referenciación y evite el plagio académico, para ello puede apoyarse revisando sus productos escritos mediante la herramienta Turnitin que encuentra en el campus virtual. Considere que en el acuerdo 029 del 13 de diciembre de 2013, artículo 99, se considera como faltas que atentan contra el orden académico, entre otras, las siguientes: literal e) "El plagiar, es decir, presentar como de su propia autoría la totalidad o parte de una obra, trabajo, documento o invención realizado por otra persona. Implica también el uso de citas o referencias faltas, o proponer citad donde no haya coincidencia entre ella y la referencia" y liberal f) "El reproducir, o copiar con fines de lucro, materiales educativos o resultados de productos de investigación, que cuentan con derechos intelectuales reservados para la Universidad" Las sanciones académicas a las que se enfrentará el estudiante son las siguientes: a) En los casos de fraude académico demostrado en el trabajo académico o evaluación respectiva, la calificación que se impondrá será de cero puntos sin perjuicio de la sanción disciplinaria correspondiente. b) En los casos relacionados con plagio demostrado en el trabajo académico cualquiera sea su naturaleza, la calificación que se impondrá será de cero puntos, sin perjuicio de la sanción disciplinaria correspondiente.

3. Formato de Rúbrica de evaluación

Tipo de actividad: En grupo colaborativo	
Momento de la evaluación: Intermedio	
La máxima puntuación posible es de 115 puntos	
Primer criterio de evaluación: Cumple con la funcionalidad del proyecto **Este criterio representa 40 puntos del total de 115 puntos de la actividad.**	**Nivel alto:** La aplicación cumple con todos los requisitos especificados, procesa los datos correctamente y genera los resultados esperados. **Si su trabajo se encuentra en este nivel puede obtener entre 21 puntos y 40 puntos** **Nivel Medio:** La aplicación presenta errores o no completa todas las funcionalidades. **Si su trabajo se encuentra en este nivel puede obtener entre 6 puntos y 20 puntos** **Nivel bajo:** La aplicación no funciona o no cumple con los requisitos básicos. **Si su trabajo se encuentra en este nivel puede obtener entre 0 puntos y 5 puntos**
Segundo criterio de evaluación: Desarrolla código de calidad y usa las tecnologías solicitadas **Este criterio representa 35 puntos del total de 115 puntos de la actividad**	**Nivel alto:** El código es eficiente, organizado, bien documentado y utiliza las estructuras de datos y funciones adecuadas de Spark (RDDs, DataFrames, Spark SQL, Spark Streaming, Kafka). **Si su trabajo se encuentra en este nivel puede obtener entre 16 puntos y 35 puntos** **Nivel Medio:** El código presenta errores, es difícil de entender o no se utilizan las tecnologías de Spark de forma óptima. **Si su trabajo se encuentra en este nivel puede obtener entre 6 puntos y 15 puntos** **Nivel bajo:** El código es deficiente, poco legible o no se utilizan las tecnologías de Spark correctamente. **Si su trabajo se encuentra en este nivel puede obtener entre 0 puntos y 5 puntos**
Tercer criterio de evaluación:	**Nivel alto:** La presentación online es clara, atractiva y demuestra dominio del tema. El informe escrito es completo, bien redactado y cumple con los requisitos. Se evidencia un trabajo en equipo equitativo y de calidad.

Elabora la presentación y el informe escrito **Este criterio representa 40 puntos del total de 115 puntos de la actividad**	**Si su trabajo se encuentra en este nivel puede obtener entre 21 puntos y 40 puntos** **Nivel Medio:** La presentación o el informe presentan errores, falta de organización o se evidencia una participación desigual en el grupo. **Si su trabajo se encuentra en este nivel puede obtener entre 6 puntos y 20 puntos** **Nivel bajo:** La presentación o el informe son deficientes o se evidencia falta de trabajo en equipo. **Si su trabajo se encuentra en este nivel puede obtener entre 0 puntos y 5 puntos**