PROYECTO Big Data

Introducción al proyecto

En este proyecto, tendrán la oportunidad de aplicar los conocimientos adquiridos en el curso de Big Data y profundizar en el uso de Spark, una potente herramienta de procesamiento distribuido, junto con SparkML y los servicios de AWS (Amazon Web Services).

El objetivo de este proyecto es que los estudiantes puedan diseñar y desarrollar un sistema de procesamiento de grandes volúmenes de datos utilizando Spark, sparkML en el entorno de AWS, a partir de una problemática dada por el propio grupo de trabajo. A lo largo de la guía, se proporcionarán los pasos necesarios para completar el proyecto de manera efectiva.

El proyecto se estructura en varias etapas, comenzando por 1) la selección de la problemática, 2) la selección del dataSet (batch o streaming) 3) la configuración del entorno en AWS y la puesta en marcha de un clúster de Spark. A continuación, debe explorarán la fuente de datos para procesar y transformar estos datos utilizando las capacidades de Spark. Además, debe abordar técnicas avanzadas, como el uso de algoritmos de machine learning con Spark.

Al completar este proyecto, los estudiantes habrán adquirido habilidades prácticas en el uso de Spark en un entorno de producción basado en la nube. Además, habrán obtenido experiencia en la implementación de soluciones escalables y eficientes para el procesamiento de grandes volúmenes de datos, lo cual es fundamental en el campo del Big Data.

Desarrollo

<u>Fase 1</u>

1) Selección de la problemática:
   El primer paso de consiste en que cada grupo de estudiantes elija una problemática específica que desee abordar y seleccione el conjunto de datos adecuado para su análisis. La problemática puede estar relacionada con cualquier área de interés, como salud, finanzas, comercio electrónico, redes sociales, entre otros. Es fundamental que el dataset seleccionado proporcione la información necesaria para abordar la problemática elegida de manera efectiva. Durante este proceso, se recomienda investigar y evaluar diferentes fuentes de datos disponibles para asegurarse de elegir aquellas que sean relevantes, confiables y accesibles. La elección adecuada de la problemática y el dataset sentará las bases para el

desarrollo exitoso del proyecto y permitirá obtener resultados significativos en el análisis de datos con Spark en AWS.

2) Ingesta de datos y limpieza:
   El segundo paso del proyecto consiste en realizar la ingesta de datos y la limpieza correspondiente de los mismos. En esta etapa, se espera que adquieran los datos necesarios para su análisis y los preparen adecuadamente para su posterior procesamiento. Esto implica realizar la extracción de datos de las fuentes seleccionadas y aplicar técnicas de limpieza para asegurar la calidad y coherencia de los datos. Se debe documentar de manera clara y precisa las etapas de ingesta y limpieza realizadas, garantizando que el conjunto de datos esté listo para ser utilizado en las siguientes fases del proyecto.
3) Metodología para resolver la problemática
   Se Debe presentar los procesos a seguir utilizando técnicas con Spark ML y el cluster (no el desarrollo). Se deben describir de manera detallada los pasos específicos que se llevarán a cabo para aplicar las técnicas de machine learning utilizando Spark ML. Esto puede incluir las transformaciones, la construcción y entrenamiento de modelos, y la evaluación de su desempeño. Se debe proporcionar una explicación clara y concisa de cada paso, asegurándose de que estén relacionados con la problemática planteada. También se debe incluir cualquier consideración adicional
4) Entregable de la Fase 1
   El entregable para esta fase consiste en:
   - Documento que incluya todos los elementos descritos anteriormente, donde se encuentre el enlace al dataset a utilizar. El documento debe contener la descripción de la problemática seleccionada, la justificación de su relevancia, así como la metodología detallada para abordarla. Además, se deben presentar los pasos específicos a seguir en las técnicas con Spark ML y el clúster, incluyendo la carga de datos, la transformación, la construcción y entrenamiento de modelos, y la evaluación de su desempeño. El enlace al dataset debe estar disponible y ser accesible para su revisión.
   - Script en Python utilizado para la ingesta y limpieza de los datos. Este script debe estar correctamente comentado y documentado, de manera que sea comprensible y reproducible por otros. Es importante que se incluyan los pasos necesarios para la carga de datos y las transformaciones aplicadas, asegurando así la calidad y coherencia de los datos.
5) Items a evaluar de la Fase 1
   Para esta primera fase, la rúbrica de evaluación estará compuesta por:

- Relevancia: se evalúa la importancia y relevancia de la problemática elegida y si aborda un desafío significativo o una necesidad relevante para el contexto del curso de BigData.
- Claridad y delimitación: se verifica si la problemática está claramente definida y delimitada. Debe ser lo suficientemente específica para permitir un análisis profundo.
- Viabilidad y disponibilidad de datos: se estudia si se ha seleccionado un dataset adecuado y accesible para abordar la problemática planteada. Se verifica que los datos estén disponibles, sean relevantes y apropiados para el análisis propuesto.
- Metodología para resolver la problemática: se evalúa si se han establecido objetivos claros y una metodología adecuada para abordar la problemática. Debe contemplar la metodología los pasos necesarios para la ingesta, el analizar de los datos y el valor de los mismos al resolver la problemática.
- Ingesta y limpieza de datos. Se evalúa la correcta implementación del script en Python para la ingesta y limpieza de los datos, y que se de garantía de calidad y coherencia de los datos resultantes.

Fase 2

6) Configuración del cluster
   En este paso, se debe configurar el clúster utilizando el servicio de AWS EMR (Elastic MapReduce). Esto implica establecer la infraestructura necesaria para el procesamiento distribuido de datos utilizando Spark. Los estudiantes deberán asegurarse de asignar los recursos adecuados, como instancias EC2 y almacenamiento, para garantizar un rendimiento óptimo del clúster. El entregable esperado para este paso es la documentación detallada que describa la configuración del clúster en AWS EMR. Y donde se presente la especificación de los recursos asignados al clúster, como el número y tipo de instancias EC2 utilizadas.
7) Algoritmo para el desarrollo de la problemática
   En este paso, se debe aplicar un algoritmo de machine learning o técnica de análisis de datos para abordar la problemática seleccionada. Utilizando las capacidades de Spark ML, deberán implementar el algoritmo correspondiente y ajustar los parámetros según sea necesario. Se valora se el grupo realizar la evaluación y validación del modelo desarrollado, utilizando métricas adecuadas para medir su desempeño. Para el punto 7 los entregables son: script de la solución propuesta, (código fuente de la implemención en Spark ML en Python), donde esté debidamente documentado y explicado el código, el enfoque utilizado, los pasos seguidos y los parámetros ajustados. Resultados de salida de la implementación.

que pueden incluir gráficos, tablas u otros elementos visuales relevantes si lo considera pertinente.

8) Despliegue de la solución
   Una vez que se ha desarrollado y evaluado el modelo, los estudiantes deben proceder al despliegue de la solución. Esto implicar la implementación de los servicios de AWS. Es importante asegurarse de que la solución desplegada sea accesible y usable por los usuarios finales.
   Nota: se valora si el despliegue cuenta con mecanismo que permita utilizar y aprovechar los resultados obtenidos, ejemplo un servidor que entrega los datos, una posible interfaz gráfica.
   Para esta fase final se espera que se entregue los enlaces a la solución desplegada, como una URL un puerto, una interfaz web o cualquier otro medio utilizado, y acceso a los datos utilizados.
9) items a evaluar de la Fase 2
   - Cluster. Correcta configuración del clúster en AWS EMR y asignación adecuada de recursos, se evalúa la elección y asignación adecuada de instancias EC2, el almacenamiento utilizado.
   - Implementación y ejecución: del algoritmo se evalúa si se ha utilizado correctamente Spark ML, y si se resuelve la problemática propuesta.
   - Accesibilidad y funcionalidad: se estudia si la solución desplegada es accesible y funcional, y si cumple con los requisitos establecidos previamente.