
PRESENTADO POR: JOHANNA RANGEL
PROYECTO INTEGRADOR
MVP - AMBIENTE DE BIG DATA CON DOCKER
Made with VISME

INTRODUCCIÓN

Me complace presentar el resultado de mi proyecto integrador, elcual es un mínimo producto viable de un ambiete de Big Data utilizando tecnologías como Hadoop, Spark, Hive, HBase MongoDB, Neo4j, Kafka y Zeppelin. Este proyecto representa un importante paso en mi dedarrollo como profesional en el campo de Data Science y será parte fundamental de mi portafolio.

ROADMAP

HDFS

SQL

HIVE

NO-SQL

ORCHESTRATION

SCALA

KAFKA

SPARK

CONTEXTO & DESAFÍOS

Para este proyecto, se me encomendó la tarea de crear un ambiente de Big Data utilizando tecnologías de código abierto, específicamente Docker, sin asignación de presupuesto para el uso de proveedores de servicios en la nube. La finalidad de este proyecto era demostrar los beneficios de las tecnologías de Big Data en un contexto de escasos recursos financieros. Durante la ejecución de este proyecto, me enfrenté a varios desafíos:

LIMITACIONES DE ESPACIO

En mi entorno de desarrollo, experimenté problemas de espacio en mi máquina virtual. Esto ralentizó el proceso, pero encontré soluciones prácticas, como la optimización de recursos y la eliminación de archivos innecesarios.

ACTIVIDADES PENDIENTES

Lamento informar que no pude completar las actividades "NoSQL Zeppelin" y "Herramientas de Orquestación". Sin embargo, considero estas tareas como pendientes y estoy comprometida a abordarlas en el futuro.

ERRORES Y COMPLEJIDAD

Algunas actividades resultaron más complejas de lo esperado y presentaron errores que consumieron tiempo adicional. Aunque estos desafíos no estaban previstos, finalmente alguno de ellos fueron resueltos, lo que contribuyó a un mayor aprendizaje y experiencia.

Colaboración en Equipo

Aunque este proyecto es presentado de manera individual, desde el inicio durante las Lectures y el Pair Programmer trabajé en estrecha colaboración con mis compañeros Gretel Sánchez y Francisco Rombini. Juntos, aprovechamos al máximo el tiempo, abordando cada actividad de manera conjunta. Cada uno de nosotros aportó ideas y soluciones muy valiosas

Gretel Sanchez

https://github.com/KGSanchezM

Johanna Rangel

https://github.com/JohannaRangel

Francisco Rombini

https://github.com/Frombini

Logros y Resultados

El proyecto alcanzó los siguientes logros:

AMBIENTE DOCKER COMPLETO

Logré configurar un entorno Docker completo que incluye Hadoop, Spark, Hive, HBase, MongoDB, Neo4J, Kafka.

CARGA DE DATOS EN HDFS

Implementé la carga de archivos CSV en HDFS de manera automatizada, facilitando el proceso para futuros usuarios.

USO DE HIVE

Creé tablas en Hive a partir de los datos en HDFS y las particioné.

INDICES EN HIVE

Agregué índices en tabla de Hive para mejorar la velocidad de consulta y demostrar las capacidades de indexación

NOSQL (HBASE Y MONGODB)

Exploré la utilización de bases de datos NoSQL, insertando y consultando datos en HBase y MongoDB.

GRAFOS CON NEO4J

Realicé operaciones de grafos en Neo4J

STREAMING CON KAFKA

Configuré y demostré el uso de Kafka para el procesamiento de datos en tiempo real.

PROCESAMIENTO CON SPARK

Utilicé Spark para cargar datos y llevar a cabo un ETL básico.

CARGA INCREMENTAL CON SPARK

Creé un script para la carga incremental de datos y lo programé para ejecutarse diariamente (ésto úultimo no lo certifiqué).

HERRAMIENTAS DE ORQUESTACIÓN

Aunque no completé esta parte del proyecto, reconocí la importancia de la orquestación y estoy dispuesto a abordarla en el futuro.

Este proyecto me permitió aprender varias lecciones valiosas

Lecciones Aprendidas

La importancia de la planificación y la asignación adecuada de recursos.

La resiliencia y la capacidad de superar desafíos técnicos.

La necesidad de un plan de contingencia para abordar problemas imprevistos.

La importancia de la documentación y la organización de ideas y soluciones.

La colaboración efectiva con compañeros de equipo.

Conclusiones

Este proyecto me brindó la oportunidad de explorar y demostrar lo que he venido aprendiendo durante el desarrollo del Bootcamp en cuanto a Big Data y tecnologías de código abierto.
Agradezco la oportunidad de trabajar en este desafío, gracias al profe **Jesus Torres** por los conocimientos trasmiitidos y su gran paciencia. Espero que esta presentación sea un reflejo fiel de mi compromiso y pasión por el campo de la ciencia de datos y el análisis de Big Data.

https://github.com/JohannaRangel/DS-M4-Herramientas_Big_Data/tree/main/Evidencia

Los archivos que recopilé como evidencias de la realización del proyecto se encuentran alojadas en GitHub

Ver Evidencias

GRACIAS

OCTUBRE, 2023