MODULO 09 - EJERCICIO 06-A

ALEXIS YURI M.

Diseña un flujo de procesamiento masivo de datos para una organización ficticia (por ejemplo: plataforma de streaming empresa logística o banco). Deberá decidir qué datos se procesan, con qué tecnología y cómo organizaría el cluster.

1. Definición del Escenario.

Organización Ficticia: Plataforma de Streaming de video.

Datos a Procesar: Logs de navegación de usuarios (clics, tiempo de visualización, búsquedas y ratings de contenido).

Objetivo de Procesamiento: Analizar el comportamiento masivo de los usuarios para generar un modelo de recomendación de contenido en tiempo casi real.

2. Definición del Framework.

Se elige Apache Spark ya que es el framework ideal para este caso. Su capacidad de procesamiento en memoria le permite ejecutar algoritmos iterativos (necesarios para el Machine Learning de recomendaciones) mucho más rápido que Hadoop MapReduce. Además, Spark es excelente para trabajar con grandes volúmenes de datos semi-estructurados y para consultas complejas de baja latencia, esenciales para un sistema de recomendaciones ágil.

3. Estructura del Clúster EMR.

El clúster se estructuraría para un alto rendimiento y escalabilidad, aprovechando los tres tipos de nodos de EMR:



4. Consumo y Destino de los Datos.

El EMR se integra perfectamente con el ecosistema de AWS para el origen y el destino de los datos:

Origen (Consumo): Amazon S3 (Data Lake): Los logs brutos de navegación se ingresarían y se almacenarían en un bucket de S3, particionados por fecha. El EMR leerá los datos directamente desde S3 (usando EMRFS), lo que permite separar el cómputo del almacenamiento, reduciendo costos.

Destino (Salida): Amazon DynamoDB (Salida 1 - Recomendaciones): Los resultados del modelo de Spark (las recomendaciones personalizadas) se cargarían en DynamoDB para que la aplicación las consulte con baja latencia en tiempo real.

Amazon Redshift (Salida 2 - Análisis BI): Un resumen de los datos procesados y los resultados analíticos agregados se cargarían en Redshift para análisis de Business Intelligence (BI) a largo plazo por el equipo de marketing.

5. Diagrama Simple del Flujo de Datos y Procesamiento.

Este diagrama muestra el flujo de datos desde la ingesta hasta el análisis y la aplicación final:

