Actividad Exploración y Análisis en Big Data

Adrián Yared Armas de la Nuez

Contenido

1. Objetivo de la actividad

Entender los conceptos clave en el manejo y análisis de Big Data, fomentando el pensamiento crítico y la aplicación práctica.

2. Preguntas

2.1 Pregunta 1: Uso de Metadatos en Big Data

2.1.1 Enunciado

Los metadatos son esenciales en Big Data para diversas funciones. Identifica al menos dos tipos de metadatos (por ejemplo, descriptivos, estructurales) y explica cómo cada uno apoya el proceso de análisis de Big Data. Incluye un ejemplo práctico para cada tipo. Elemento de Reflexión: ¿Cómo cambiaría tu elección de metadatos si estuvieras analizando datos de redes sociales en comparación con datos financieros?

2.1.2 Resolución

Existen tres tipos de metadatos principales, que serían; los metadatos descriptivos; los metadatos estructurales; y los metadatos administrativos. Pero adicionalmente existen los metadatos de referencia.

Los metadatos descriptivos, como su propio nombre indica proporcionan información sobre el contenido, como por ejemplo su autor, fecha o palabras clave. Este tipo de metadato apoya el análisis del Big Data ya que hace más sencilla la clasificación y búsqueda en grandes volúmenes de datos. Un ejemplo de este tipo de metadato en las redes sociales es el filtrado de publicaciones en función al autor o palabras clave (comúnmente hashtags).

El tipo de metadatos estructurales brindan información sobre la estructura de los recursos y la relación y vínculo entre ellos. Y en cuanto a este tipo, apoya en análisis Big Data debido a que ayuda a entender la estructura interna y hacer análisis cruzados. Un ejemplo de este tipo de metadato podría ser el de los datos financieros, ya que ahí detallan cómo se relacionan tablas de transacciones y clientes, ayudando en análisis complejos.

El tercer tipo de metadato principal es el administrativo, este tipo contiene información técnica sobre la gestión de los datos tales como formato, tamaño o permisos de acceso. Este tipo ayuda al análisis Big Data ya que permiten gestionar y mantener los datos de forma segura y eficiente. Y un ejemplo de esto podría ser

en análisis de redes sociales, indicando el formato de los archivos y los permisos de acceso de los datos.

Los metadatos de referencia son otro tipo de metadatos que proveen contexto adicional al describir cómo los datos se relacionan con otros datos externos (códigos estandarizados, ubicaciones geográficas). Ayudarían al análisis Big Data facilitando la interoperabilidad y comparabilidad con otros conjuntos de datos. Y un ejemplo podría ser en datos de salud, se pueden usar códigos ICD para clasificar enfermedades, facilitando la integración de datos médicos globales.

En cuanto a esta pregunta planteada: ¿Cómo cambiaría tu elección de metadatos si estuvieras analizando datos de redes sociales en comparación con datos financieros?
La respuesta es sencilla, para redes sociales, los descriptivos y administrativos son clave, ya que se necesita segmentación rápida y control de acceso. En datos financieros, los estructurales y de referencia son esenciales para analizar patrones internos y comparar con datos externos.

2.2 Pregunta 2: Veracidad y Ruido en Big Data

2.2.1 Enunciado

La veracidad es fundamental en Big Data. Describe cómo el ruido puede afectar el procesamiento inicial de los datos y el análisis posterior. Proporciona un ejemplo donde el ruido podría tener un impacto significativo en los resultados.
Tarea de Investigación: Encuentra un estudio de caso donde el ruido en los datos haya sido un desafío y discute cómo se abordó.

2.2.2 Resolución

En Big Data, la veracidad o precisión de los datos es esencial y el ruido, es decir, datos irrelevantes o incorrectos puede afectar negativamente el análisis. El ruido se introduce a menudo por errores de medición, problemas técnicos, o datos duplicados. Esto complica el procesamiento inicial, ya que el sistema debe diferenciar entre información útil y ruido. Si no se elimina el ruido, puede generar resultados poco confiables en los modelos predictivos o en el análisis general.

Un ejemplo del ruido afectando a los datos podría ser el uso de sensores en hospitales para monitorear signos vitales, como la frecuencia cardíaca. Si los sensores recogen datos incorrectos por problemas técnicos, como interferencias o batería baja, los datos ruidosos pueden provocar una falsa alarma médica, haciendo creer a los doctores que el paciente está en riesgo cuando en realidad no lo está.

En cuanto al caso de estudio que haya tenido el ruido como un desafío podría ser el análisis de redes sociales, como Twitter, para entender la opinión de los consumidores. Las empresas enfrentan ruido en estos datos debido a mensajes irrelevantes, sarcasmo, bots, o publicaciones repetitivas. En un estudio, los analistas usaron varios métodos para reducir el ruido, como filtrar datos irrelevantes, ya que detectaron y eliminaron publicaciones de cuentas sospechosas o irrelevantes, además, mejoraron los modelos de lenguaje ajustando los modelos para entender mejor el sarcasmo y el contexto y etiquetaron manualmente algunos datos para entrenar mejor a los modelos.

Gracias a estos pasos, el equipo redujo el ruido y obtuvo una visión más clara sobre la opinión real de los consumidores.

2.3 Pregunta 3: Beneficios de Clusters en Big Data

2.3.1 Enunciado

Los clusters ofrecen varias ventajas para el procesamiento de Big Data. Explica cómo aspectos como alto rendimiento, alta disponibilidad, equilibrado de carga y escalabilidad benefician específicamente a los procesos de Big Data. Aplicación Práctica: Considera un escenario hipotético de análisis de grandes volúmenes de datos de tráfico urbano. Describe cómo un cluster podría mejorar el procesamiento de estos datos en comparación con un solo ordenador.

2.3.2 Resolución

Los aspectos de alto rendimiento, alta disponibilidad, balanceo de carga y escalabilidad son cruciales para el procesamiento eficiente de Big Data. El alto rendimiento se refiere a la capacidad de procesar grandes volúmenes de datos de manera rápida y eficiente, lo que permite obtener resultados casi en tiempo real. En Big Data, esto es fundamental, ya que los datos suelen ser masivos y deben ser analizados rápidamente para obtener información valiosa.

La alta disponibilidad asegura que el sistema continúe funcionando incluso si un nodo falla, lo que es esencial para evitar interrupciones en el procesamiento de datos. En sistemas de Big Data, la continua disponibilidad del servicio es vital para tomar decisiones inmediatas, como en el análisis de datos de tráfico o sistemas de monitoreo de salud.

El balanceo de carga distribuye las tareas de procesamiento entre los nodos de manera eficiente, evitando que algunos nodos se sobrecarguen mientras otros están

inactivos. Esto optimiza el uso de los recursos y asegura un procesamiento más rápido y equilibrado, lo cual es clave cuando se manejan grandes volúmenes de datos que podrían de otra forma generar cuellos de botella.

Finalmente, la escalabilidad permite que el sistema crezca conforme aumentan los datos, añadiendo más nodos al *cluster* sin interrumpir el servicio. Esta característica es esencial en Big Data, donde los volúmenes de datos pueden crecer rápidamente, y se necesita una infraestructura que se pueda ampliar para seguir funcionando sin perder eficiencia.

En un escenario de análisis de tráfico urbano, estos aspectos permiten gestionar de manera eficiente la enorme cantidad de datos generados por sensores, cámaras y vehículos conectados. Un *cluster* con alto rendimiento podría procesar en paralelo los datos de diferentes zonas de la ciudad, analizando patrones de tráfico en tiempo real. La alta disponibilidad garantizaría que el sistema no se detenga ante fallos, lo cual es crítico para la toma de decisiones rápidas en situaciones de congestión o accidentes. El balanceo de carga distribuiría el análisis de datos entre varios nodos, evitando que algunos se sobrecarguen, mientras que la escalabilidad permitiría agregar más nodos si aumentan los datos generados por nuevas cámaras o sensores, asegurando que el sistema se mantenga eficiente a medida que crece la ciudad.

2.4 Pregunta 4: Commodity Hardware y Big Data

2.4.1 Enunciado

El uso de commodity hardware es común en sistemas de Big Data. Explica los beneficios de utilizar este tipo de hardware y discute si es posible y práctico montar un cluster con ordenadores reciclados. Justifica tu respuesta con argumentos técnicos y económicos. ¿Cuáles serían las limitaciones y los riesgos de usar hardware reciclado en un entorno de Big Data? Proporciona ejemplos.

2.4.2 Resolución

El uso de commodity hardware, que es el hardware básico y de bajo costo, en sistemas de Big Data es una práctica común para escalar infraestructuras de procesamiento de datos de forma económica y eficiente. A continuación, exploraremos los beneficios de este enfoque, la viabilidad de montar un clúster con ordenadores reciclados, así como las limitaciones y riesgos de esta alternativa:

Trae consigo una reducción de costes en comparación con hardware especializado. En sistemas distribuidos de Big Data como Hadoop o Spark, las cargas de trabajo

se pueden dividir y distribuir en múltiples nodos, permitiendo utilizar equipos menos costosos y, si falla un nodo, otros pueden compensar la carga. Además trae consigo otra ventaja como la escalabilidad horizontal, ya que es más económico y sencillo agregar nodos adicionales de *commodity hardware* para aumentar la capacidad, en lugar de invertir en costosos servidores más potentes (escalabilidad vertical). Otra ventaja a destacar está la flexibilidad y modularidad, esto quiere decir que se pueden mezclar diferentes tipos de máquinas dentro de un clúster, adaptándose a presupuestos y necesidades variables. Esto permite que las empresas puedan renovar o actualizar partes del clúster sin tener que reemplazar todo el hardware de una sola vez. Una ventaja adicional podría ser la interoperabilidad, debido a que el hardware comercial se adhiere a estándares ampliamente aceptados, puede funcionar sin problemas con diferentes programas y sistemas. Esta estandarización reduce los problemas de compatibilidad y garantiza un funcionamiento sin problemas en distintos entornos. La penúltima ventaja a destacar es la menor dependencia del proveedor, la estandarización y la amplia disponibilidad de los componentes implican que las organizaciones no están atadas a un único proveedor. Y la última ventaja que considero destacar es la fiabilidad mediante la redundancia, aunque los componentes individuales pueden no ser tan robustos como sus contrapartes especializadas, la asequibilidad permite a las organizaciones utilizar estrategias de redundancia. Al replicar el hardware, pueden crear sistemas tolerantes a fallas con menor tiempo de inactividad y mayor fiabilidad. Todas estas ventajas son importantes pero las más importantes a destacar y tener en cuenta son las tres primeras.

En cuanto al planteamiento de montar un clúster con ordenadores reciclados la respuesta corta es que si se puede y en ciertos casos es práctico especialmente para proyectos de prueba, investigación académica, o pequeñas empresas con presupuestos ajustados. Sin embargo, la efectividad y eficiencia de esta solución depende de factores técnicos específicos y del contexto económico:

En cuanto al aspecto técnico, los sistemas distribuidos como Hadoop están diseñados para tolerar fallos en el hardware. Por lo tanto, un clúster basado en ordenadores reciclados puede funcionar siempre que los nodos puedan soportar el sistema operativo y los requisitos básicos de la plataforma de Big Data. Y en cuanto al aspecto económico, los ordenadores reciclados son de bajo o nulo costo de adquisición, lo que permite crear un clúster a partir de una inversión mínima. Esto puede ser ventajoso para proyectos de inicio o como un concepto de prueba.

Pero pese a este planteamiento positivo, también posee limitaciones y riesgos, estos son algunos de ellos; rendimiento y Consumo Energético, debido a la antigüedad de los equipos son menos eficientes energéticamente lo que determina una peor capacidad computacional y un mayor consumo energético, haciendo que

el procesamiento de grandes volúmenes de datos sea más lento y menos eficiente. Un ejemplo podría ser el implícito en mi explicación, a continuación lo detallo mejor, si se usan ordenadores de hace 10 años en un clúster de Hadoop, los tiempos de procesamiento pueden ser significativamente mayores en comparación con máquinas más recientes. Además, equipos antiguos no suelen optimizar el consumo de energía, generando un costo oculto a largo plazo; fiabilidad y tolerancia a fallos, debido a la antigüedad del hardware, podremos enfrentar problemas de averías o desgaste. Un ejemplo podría ser el siguiente: en un clúster de 50 nodos donde varios nodos presentan fallos frecuentes, el sistema puede saturarse en la recuperación de datos perdidos, afectando el rendimiento general y aumentando el tiempo de procesamiento de tareas; mantenimiento y administración, la gran variedad de hardware genera una mayor complejidad de configuración y esto incrementa el tiempo y esfuerzo de los administradores de sistemas. Un ejemplo podría ser: un clúster de Big Data con varios tipos de ordenadores requerirá una configuración específica y personalizada para cada tipo de equipo, lo cual es más laborioso que un clúster homogéneo. Esto puede aumentar los costos en personal de IT y en el tiempo necesario para la configuración; compatibilidad y escalabilidad limitada, los ordenadores reciclados pueden tener arquitecturas y especificaciones obsoletas que dificultan la instalación de software reciente o su escalabilidad a futuro. Un ejemplo podría ser el siguiente: Debido a los discos duros de una empresa y su baja capacidad, así como sus procesadores antigüos, se limita la capacidad de almacenaje y procesamiento. Y en caso de querer sustituirlos había un gran coste económico.

En conclusión, montar un clúster de Big Data con ordenadores reciclados es viable principalmente en ambientes de prueba académicos o de bajo presupuesto, pero para proyectos mayores o con futura escalabilidad supone un problema que derivará en un coste mayor y ralentización del proyecto.

3. Debate

En el debate he aprendido nuevos usos de los metadatos como en el caso de los descriptivos organizar música.

En la pregunta del ruido hablamos de los outliers y pusimos ejemplos como una transacción inusual por parte de un cliente, comentarios fuera de la media o los clicks involuntarios, etc.

En cuanto al clúster hablamos de las ventajas.

Y finalmente en cuanto al commodity hardware hablamos de que es una oportunidad y tiene cosas positivas, pero no compensa ya que tiene muchas cosas negativas.

4. Bibliografía

1 Información sobre metadatos:
https://arxiversvalencians.org/wp-content/uploads/2020/04/revista2009_raventos.pdf
https://keepcoding.io/blog/tipos-de-metadatos/

2 Ruido en el Big Data:
https://medium.com/@gdellamattia/el-desaf%C3%ADo-del-ruido-en-los-datos-una-mirada-cr%C3%ADtica-a-los-softwares-acumuladores-c7a3bc4505d9#:~:text=El%20ruido%20en%20los%20datos,atletas%20en%20la%20%C3%BAltima%20d%C3%A9cada.

https://www.techtarget.com/searchbusinessanalytics/definition/noisy-data

3 Beneficios de Clusters en Big Data
https://www.universidadunie.com/blog/que-es-clustering#:~:text=Gracias%20al%20clustering%2C%20uno%20es,similares%2C%20permitiendo%20estrategias%20m%C3%A1s%20efectivas

4.1 Hardware commodity
https://phoenixnap.com/glossary/commodity-hardware
4.2 Viabilidad de un Clúster con Ordenadores Reciclados
https://upcommons.upc.edu/bitstream/handle/2117/167977/143810.pdf