**Cuestionario sobre Big Data**

1. **El uso de clústeres de máquinas es clave en los sistemas Big Data para mejorar diferentes aspectos como la disponibilidad y el rendimiento. Sin embargo, no todos los beneficios mencionados suelen obtenerse a través de clústeres. ¿Cuál de las siguientes afirmaciones no representa una ventaja clave obtenida con el uso de clústeres de máquinas?**

* Menor consumo energético
* Alta disponibilidad
* Alto rendimiento
* Escalabilidad

1. **El procesamiento paralelo permite acelerar la ejecución de tareas al distribuir el trabajo entre varias máquinas. Sin embargo, no todas las tareas pueden ser paralelizadas. ¿Cuál de las siguientes afirmaciones sobre el procesamiento paralelo es verdadera?**

* Algunas tareas no pueden paralelizarse para que se ejecuten más rápido usando varias máquinas.
* Cualquier tarea siempre tiene al menos una parte que puede ejecutarse en paralelo para ganar tiempo.
* Todas las tareas pueden paralelizarse para que se ejecuten más rápido usando varias máquinas.
* Si conseguimos paralelizar una tarea, la potencia equivalente de cómputo es la suma de las potencias de cada nodo del clúster.

1. **En la arquitectura de Big Data, la capa de colección juega un papel importante en la preparación de los datos para su procesamiento. ¿Cuál es la función principal de la capa de colección en esta arquitectura?**

* Aquella en la que se pasa de los datos que vienen de diversas fuentes a un conjunto de datos unificado y ya casi listo para ser usado.
* Aquella de la que se obtienen los datos para visualización.
* Aquella en la que se gestiona el almacenamiento de los datos.
* Aquella en que se determina para cada dato si va a ser utilizado o no para realizar analíticas.

1. **Las 5 Vs del Big Data describen los principales retos que enfrentan las organizaciones al manejar grandes volúmenes de datos. ¿Qué significa la V de "velocidad" en este contexto?**

* El bajo tiempo que transcurre desde que se genera el dato hasta que es adquirido por el sistema.
* La velocidad a la que se degrada la información.
* La alta velocidad que necesitan tener los microprocesadores con los que se tratan los datos.
* La cantidad de datos que siguen generándose constantemente.

1. **En el ámbito de Big Data, es común utilizar términos específicos para referirse a grandes conjuntos de datos. ¿Cuál de las siguientes traducciones es la más adecuada para "Big Data" en español?**

* Macrodatos
* Grandes Datos
* Megadatos
* Gran Información

1. **El principio ACID garantiza la confiabilidad de las transacciones en bases de datos. ¿Cuáles son las cuatro características que forman el principio ACID?**

* Atomicidad, Consistencia, Aislamiento y Durabilidad.
* Atomicidad, Concurrencia, Aislamiento y Durabilidad.
* Atomicidad, Consistencia, Independencia y Durabilidad.
* Aislamiento, Concurrencia, Independencia y Durabilidad.

1. **Entre las 5 Vs del Big Data, la V de volumen se refiere a la cantidad masiva de datos que las organizaciones deben manejar. ¿A qué se refiere específicamente esta V de volumen?**

* La gran cantidad de bytes de datos que es necesario almacenar y tratar.
* El elevado volumen de espacio necesario para ubicar las máquinas que conforman un clúster lo suficientemente potente.
* El hecho de que en cada posible fichero de vídeo o audio el sonido puede haberse guardado a diferente volumen.
* La complejidad de distribuir una enorme cantidad de datos en los distintos volúmenes o unidades de almacenamiento.

1. **El procesamiento paralelo y el procesamiento distribuido son dos enfoques distintos que pueden usarse en sistemas Big Data. ¿Cuál de las siguientes afirmaciones es verdadera sobre el procesamiento paralelo y distribuido?**

* Tanto procesamiento en paralelo como distribuido se realizan en varias máquinas.
* Tanto procesamiento en paralelo como distribuido se realizan dentro de la misma máquina.
* El procesamiento distribuido se realiza dentro de la misma máquina y el paralelo en varias máquinas.
* El procesamiento en paralelo se realiza dentro de la misma máquina y el distribuido en varias máquinas.

1. **En una arquitectura típica de Big Data, varias capas gestionan el flujo de datos desde su adquisición hasta su análisis. ¿Cuál de las siguientes no es una capa comúnmente presente en estas arquitecturas?**

* Capa de consolidación
* Capa de almacenamiento
* Capa de visualización
* Capa de ingestión

1. **El término "data lake" hace referencia a un sistema de almacenamiento utilizado en Big Data. ¿Cuál es la característica principal de un "data lake"?**

* Permite almacenar datos en su forma bruta y sin procesar.
* Organiza los datos estructurados para análisis detallados.
* Almacena solo datos estructurados para una fácil consulta.
* Es un sistema dedicado solo a la captura de datos en tiempo real.

1. **El término "data warehouse" se refiere a un sistema de almacenamiento optimizado para el análisis de datos. ¿Qué tipo de datos suelen almacenarse en un "data warehouse"?**

* Datos transformados y estructurados para análisis específicos.
* Datos en su forma bruta, sin ninguna transformación previa.
* Solo datos no estructurados que provienen de diversas fuentes.
* Datos en tiempo real obtenidos de sensores y redes sociales.

1. **En sistemas Big Data, el procesamiento en tiempo real es crucial para ciertas aplicaciones. ¿Qué es el procesamiento en tiempo real en el contexto de Big Data?**

* Es el procesamiento de datos en cuanto son generados para obtener resultados instantáneamente.
* Es el almacenamiento de grandes volúmenes de datos para ser procesados posteriormente.
* Es el análisis de grandes conjuntos de datos almacenados históricamente.
* Es la sincronización de los datos para que estén disponibles en diferentes ubicaciones.

1. **El término "MapReduce" es un modelo de programación ampliamente utilizado en Big Data. ¿Cuál es la función principal de "MapReduce"?**

* Dividir una tarea en partes más pequeñas que pueden procesarse en paralelo y luego combinar los resultados.
* Organizar datos en un formato adecuado para el almacenamiento.
* Procesar datos en tiempo real a medida que son generados.
* Transformar datos no estructurados en datos estructurados listos para su análisis.

1. **En Big Data, es importante manejar grandes volúmenes de datos de diversas fuentes. ¿Cuál es la principal ventaja de utilizar un sistema distribuido en lugar de uno centralizado?**

* Permite el procesamiento y almacenamiento a gran escala al repartir las tareas entre varias máquinas.
* Reduce el costo de almacenamiento porque no necesita tanta infraestructura.
* Elimina la necesidad de mantenimiento porque cada nodo se gestiona automáticamente.
* Permite que las aplicaciones corran más rápido en una única máquina debido a la optimización.

1. **El modelo CAP describe las tres propiedades clave que un sistema distribuido puede ofrecer. Sin embargo, no es posible garantizar todas las propiedades simultáneamente. ¿Cuáles son estas tres propiedades del modelo CAP?**

* Consistencia, Disponibilidad, Tolerancia a particiones.
* Confiabilidad, Disponibilidad, Consistencia.
* Consistencia, Adaptabilidad, Tolerancia a fallos.
* Disponibilidad, Tolerancia a fallos, Rendimiento.

1. **Hadoop es una de las tecnologías clave en el ecosistema de Big Data. ¿Cuál de las siguientes afirmaciones sobre Hadoop es correcta?**

* Es una plataforma que permite el procesamiento distribuido de grandes volúmenes de datos en múltiples nodos.
* Es una base de datos relacional utilizada para almacenar grandes volúmenes de datos estructurados.
* Se utiliza exclusivamente para el procesamiento en tiempo real de datos masivos.
* Se utiliza principalmente para almacenar datos en tiempo real provenientes de sensores.

1. **El procesamiento de datos por lotes es una de las formas en que se puede manejar grandes volúmenes de información. ¿Cuál es la principal característica del procesamiento por lotes?**

* Los datos se procesan en grupos acumulados en lugar de inmediatamente después de ser generados.
* Los datos se procesan en tiempo real, en cuanto son generados.
* Los datos se procesan individualmente de manera secuencial en lugar de ser agrupados.
* Los datos se almacenan temporalmente para luego ser descartados tras su uso.

1. **Uno de los componentes clave en el procesamiento distribuido es el balanceo de carga. ¿Qué es el balanceo de carga en el contexto de Big Data?**

* La distribución equitativa de tareas entre los nodos de un clúster para evitar sobrecargas en alguno de ellos.
* La replicación de los datos para evitar pérdida de información en caso de fallo en un nodo.
* La reducción de tareas y datos en un clúster para optimizar el rendimiento del sistema.
* El almacenamiento de datos distribuidos de forma redundante para asegurar la disponibilidad.

1. **Spark es una herramienta utilizada para el procesamiento de grandes volúmenes de datos, ofreciendo mejoras en comparación con Hadoop. ¿Cuál de las siguientes afirmaciones describe correctamente una ventaja de Spark sobre Hadoop?**

* Spark permite el procesamiento en memoria, lo que acelera el análisis de datos respecto a Hadoop.
* Spark es una base de datos NoSQL que almacena grandes volúmenes de datos no estructurados.
* Spark permite el procesamiento exclusivo de datos estructurados en tiempo real.
* Spark se utiliza únicamente para tareas de almacenamiento y consulta de datos a gran escala.

1. **El almacenamiento de datos en entornos de Big Data presenta múltiples desafíos. ¿Cuál es el principal desafío en el almacenamiento de datos no estructurados?**

* La falta de un esquema definido que permita la organización y el análisis directo.
* El costo asociado al almacenamiento de grandes cantidades de datos no estructurados.
* La dificultad de almacenar datos no estructurados en bases de datos relacionales.

1. **El ecosistema Hadoop incluye varias herramientas que permiten el procesamiento y almacenamiento de grandes volúmenes de datos. ¿Cuál de las siguientes herramientas de Hadoop se utiliza principalmente para el almacenamiento distribuido de datos?**

* HDFS (Hadoop Distributed File System)
* MapReduce
* YARN
* Hive

1. **Los clústeres de procesamiento en Big Data permiten el manejo de grandes volúmenes de datos distribuyendo las tareas entre múltiples nodos. ¿Qué sucede cuando un nodo en un clúster falla durante el procesamiento?**

* Otro nodo puede tomar la tarea del nodo fallido y continuar el procesamiento.
* Todo el clúster se detiene hasta que el nodo fallido se restaura.
* Los datos procesados por el nodo fallido se pierden y deben reiniciarse todas las tareas.
* El nodo fallido se repara automáticamente y continúa con su tarea una vez restablecido.

1. **Los datos en un entorno de Big Data pueden ser procesados tanto en tiempo real como en lotes. ¿Qué tipo de procesamiento es más adecuado para aplicaciones que requieren resultados inmediatos, como la detección de fraudes?**

* Procesamiento en tiempo real
* Procesamiento por lotes
* Procesamiento distribuido
* Procesamiento secuencial

1. **Hive es una herramienta utilizada en el ecosistema Hadoop que facilita la consulta y el análisis de grandes volúmenes de datos. ¿Qué tipo de lenguaje utiliza Hive para ejecutar estas consultas?**

* HiveQL, un lenguaje similar a SQL adaptado a Hadoop.
* Python, un lenguaje de programación general que se adapta a Hadoop.
* MapReduce, el lenguaje nativo de Hadoop para procesar grandes volúmenes de datos.
* NoSQL, un lenguaje específico para bases de datos no relacionales en Hadoop.

1. **La replicación de datos es una técnica comúnmente utilizada en sistemas distribuidos como Hadoop para asegurar la disponibilidad y la tolerancia a fallos. ¿Qué significa la replicación de datos en este contexto?**

* Almacenar múltiples copias de los mismos datos en diferentes nodos del clúster para asegurar su disponibilidad en caso de fallo.
* Almacenar una copia única de los datos y dividirlos en diferentes nodos para mejorar el rendimiento.
* Duplicar los datos en el mismo nodo para evitar que se pierdan durante el procesamiento.
* Copiar los datos entre clústeres diferentes para aumentar la capacidad de almacenamiento total.