Cuestionario sobre Arquitecturas de Almacenamiento

Elige la opción que consideres correcta, justificando tu respuesta y el porqué de lo inadecuado de las otras opciones.

Cuestión 1: Optimización de consultas analíticas

Una empresa necesita realizar consultas complejas sobre billones de registros de llamadas telefónicas, pero en cada consulta solo se accede a 3 o 4 atributos específicos (como la duración y el destino) de los cientos disponibles. ¿Qué formato de archivo y organización sería la más eficiente?

- **A)** Formato **Avro**, debido a su orientación a filas que permite leer el registro completo rápidamente.
- **B)** Formato **CSV**, por ser un estándar de texto que permite la interoperabilidad entre cualquier sistema.
- **C)** Formato **Parquet o ORC**, debido a su orientación a columnas y capacidad de comprimir bloques de datos.

Este formato permite acceder a atributos específicos y leer unicamente las columnas especificadas, de esta manera mejora el rendimiento. Los otros formatos, leen el registro completo ya que está orientado a filas y el otro es texto plano, por lo tanto, es poco útil para una gran cantidad de registros

Cuestión 2: Aplicación del Teorema CAP en sistemas distribuidos

Un sistema de monitorización de flotas globales requiere que el servicio esté siempre disponible para recibir datos de los sensores, incluso si hay fallos en los enlaces de red entre continentes. ¿Qué combinación del teorema CAP es la más adecuada para este caso?

- **A) CA (Consistencia y Disponibilidad)**, asegurando que todos los nodos vean lo mismo al mismo tiempo.
- **B) CP (Consistencia y Tolerancia a Particiones)**, sacrificando la disponibilidad del sistema para evitar datos erróneos.
- **C) AP (Disponibilidad y Tolerancia a Particiones)**, priorizando que el sistema siga dando servicio aunque la consistencia sea eventual.

Como el sistema tiene que estar siempre disponible, aunque haya fallos, se prioriza la disponibilidad y tolerancia a particiones. Las otras opciones no pueden ser ya que una no tolera particiones, y la otra no tiene disponibilidad

Cuestión 3: Gestión de metadatos en HDFS

En un clúster de **Apache Hadoop HDFS**, ¿cuál es la función crítica del **NameNode** primario respecto a los datos almacenados?

- **A)** Almacenar físicamente los bloques de datos y replicarlos en otros servidores del clúster.

- **B)** Gestionar exclusivamente el espacio de nombres (*namespace*) y la ubicación de los bloques sin que los datos reales pasen por él.
- **C)** Ejecutar la lógica de negocio de las aplicaciones directamente sobre los discos locales.

El NameNode gestiona el namespace y los datos reales nunca pasan por él, las otras opciones no pueden ser ya que los bloques se almacenan en el DataNode y la otra no ejecuta lógica de negocio

Cuestión 4: Evolución hacia el Data Lakehouse

¿Qué ventaja principal aporta la implementación de capas como **Delta Lake o Apache Iceberg** sobre un almacenamiento de objetos (como Amazon S3)?

- **A)** Convertir el almacenamiento de objetos en un sistema jerárquico de carpetas tradicional.
- **B)** Reducir el coste de almacenamiento eliminando la necesidad de metadatos.
- **C)** Proporcionar características **ACID**, acceso SQL y control de transacciones sobre datos no estructurados.

Las otras opciones no pueden ser ya que no convierten el almacenamiento en jerárquico y no eliminan metadatos

Cuestión 5: Selección de base de datos para relaciones complejas

Si una organización necesita detectar tramas de fraude analizando cómo se conectan conductores, médicos y abogados mediante múltiples vínculos directos e indirectos, ¿qué tipo de gestor NoSQL es el más indicado?

- **A) Clave-Valor**, por su extrema rapidez en búsquedas simples.
- **B) Grafos**, ya que representa entidades como nodos y relaciones como arcos.
- **C) Documentos**, para anidar toda la información en estructuras JSON complejas.

Los grafos modelan entidades que son ideales para detectar conexiones. Las otras opciones no modelan entidades o no sin eficientes para conexiones múltiples

Cuestión 6: Identificación de los estados del dato

Un banco decide mover todos los registros de transacciones de hace más de cinco años, que legalmente debe conservar pero que raramente consulta, a un sistema de cintas de respaldo. Según la fuente, ¿en qué estado se encuentran estos datos?

- **A) Datos en tránsito (*data in motion*)**, ya que se están moviendo hacia el respaldo.
- **B) Datos en reposo (*data at rest*)**, ya que se encuentran fuera del acceso habitual y son inmutables.
- **C) Datos en uso (*data in use*)**, porque siguen siendo consultables bajo solicitud.

Son datos en reposo ya que no se suelen consultar, son inmutables y están almacenados en cintas. Las otras opciones son incorrectas ya que no están en tránsito y no son de uso activo

Cuestión 7: Limitaciones del modelo de escritura en HDFS

Un equipo de desarrollo intenta implementar una aplicación que requiere actualizar constantemente registros específicos (modificar una línea en medio del archivo) dentro de un fichero de 5 TB almacenado en **HDFS**. ¿Es esta una arquitectura adecuada?

- **A) Sí**, HDFS permite el acceso aleatorio y la edición de cualquier bloque del archivo de forma eficiente.
- **B) No**, HDFS utiliza un modelo **WORM** (*Write Once Read Many*), donde los archivos no pueden ser actualizados una vez creados, soportando solo el anexo al final.
- **C) Sí**, siempre que el **NameNode** coordine la reescritura de los metadatos del bloque afectado.

Es la opción B ya que HDFS usa un modelo WORM y este no permite modificar bloques

Cuestión 8: Virtualización mediante Federación en RDBMS

Una analista de datos necesita cruzar una tabla de clientes en una base de datos Oracle con un archivo de logs en formato **JSON** que reside en **MongoDB** y un archivo histórico en **Parquet** en **Amazon S3**. ¿Cuál es la solución más eficiente según las capacidades modernas de los RDBMS?

- **A)** Utilizar la **Federación**, que permite acceder a fuentes heterogéneas con una única sentencia SQL como si los datos estuvieran juntos.
- **B)** Mover físicamente todos los datos a un único clúster de **Hadoop** para procesarlos.
- **C)** Convertir todos los datos a formato **CSV** e importarlos manualmente a una tabla relacional.

La solución más eficiente es la primera, ya que se pueden consultar fuentes heterogéneas con una sola onsultar sin la necesidad de mover los datos

Cuestión 9: Estructura del Almacenamiento de Objetos

¿Cuál es la diferencia fundamental en la organización de los datos entre un sistema de archivos tradicional (como HDFS o NFS) y un **Almacenamiento de Objetos**?

- **A)** El almacenamiento de objetos utiliza una **estructura jerárquica** compleja de directorios y subdirectorios.
- **B)** El almacenamiento de objetos gestiona los datos de forma **plana**, donde cada unidad es un objeto autocontenido con un identificador único y metadatos enriquecidos.
- **C)** El almacenamiento de objetos divide los archivos en bloques físicos que el sistema operativo debe ensamblar manualmente.

La diferencia es que el almacenamiento de objetos gestiona los datos de forma plana con un identificador único. Las otras opciones no son correctas ya que la primera son sistemas de archivos jerárquicos y la tercera habla de almacenamiento por bloques

Cuestión 10: Elección de base de datos NoSQL para sesiones

Una plataforma de comercio electrónico necesita almacenar los "carritos de la compra" y las sesiones de usuario. La prioridad es que la recuperación sea extremadamente ágil utilizando únicamente el ID de la sesión. ¿Qué categoría de NoSQL es la más recomendada?

- **A) Series temporales**, para registrar cada clic del usuario cronológicamente.
- **B) Clave-Valor**, por su simplicidad y rapidez en búsquedas simples recuperando todo el valor.
- **C) Documentos**, para poder realizar consultas complejas sobre los productos dentro del carrito.

La mejor opción es usar "Clave-Valor" ya que es la más rápida y recupera el valor completo. Las otras opciones usan series temporales las cuales no sirven para esto y la otra opción es más compleja

Cuestión 11: Consistencia en el *Data Lakehouse*

Una organización utiliza un almacén de objetos en la nube para su data lake, pero experimenta problemas de inconsistencia cuando varios procesos intentan modificar los mismos datos simultáneamente. ¿Cuál es la solución tecnológica recomendada en las fuentes para resolver esto sin abandonar el almacenamiento de objetos?

- **A)** Migrar todos los datos a un sistema de archivos distribuido como HDFS para forzar el modelo WORM.
- **B)** Implementar una capa de gestión de metadatos como **Delta Lake, Apache Iceberg o Apache Hudi**.
- **C)** Aumentar el número de réplicas del objeto en diferentes regiones geográficas.

La mejor opción es implementar una capa de gestión de metadatos ya que dan consistencia sin abandonar el almacenamiento de objetos

Cuestión 12: Disponibilidad y "Rack Awareness" en HDFS

En la configuración por defecto de **HDFS**, cuando un archivo tiene un factor de replicación de 3, ¿cómo distribuye el **NameNode** los bloques para balancear la disponibilidad y el rendimiento?

- **A)** Los tres bloques se colocan en nodos aleatorios de tres bastidores (*racks*) diferentes.
- **B)** Las tres copias se guardan en el mismo nodo para minimizar la transferencia de datos por red.
- **C)** Dos copias se crean en el mismo bastidor (en nodos distintos) y la tercera en un bastidor diferente.

El NameNode distribuye los bloques en dos copias en el mismo rack, en nodos distintos y el tercero en un rack diferente para así optimizar la disponibilidad y el ancho de banda

Cuestión 13: Escalado de sistemas en Big Data

Un administrador de base de datos nota que el servidor actual ha llegado al límite de su capacidad de CPU y RAM. Según las fuentes, ¿cuál es el enfoque de crecimiento preferible para un entorno de Big Data?

- **A) Escalado Vertical**, añadiendo más recursos (CPU, RAM, disco) al servidor existente.
- **B) Escalado Horizontal**, añadiendo más servidores (nodos) al clúster y fragmentando los datos (*sharding*).
- **C)** Reemplazar el hardware por sistemas de almacenamiento de bloques (SAN) de alto rendimiento.

La mejor opción para un entorno de Big Data es un escalado horizontal el cual consiste en añadir más nodos

Cuestión 14: Flexibilidad de esquema en NoSQL

Una aplicación de comercio electrónico necesita añadir nuevos atributos a sus productos (como "color", "talla" o "voltaje") de forma dinámica y diferente para cada artículo. ¿Por qué una base de datos de **Documentos** es más apta que una **Relacional**?

- **A)** Porque el modelo relacional es rígido y requiere que todas las filas tengan las mismas columnas, mientras que NoSQL permite libertad de esquema (*schemaless*).
- **B)** Porque las bases relacionales no permiten almacenar datos en formato JSON.
- **C)** Porque las bases de documentos eliminan la necesidad de realizar copias de seguridad.

Una base de datos de Documentos es más apta ya que la relacional necesita un esquema rígido, en cambio la de documentos permite un esquema flexible

Cuestión 15: Diferencia entre Tiempo Real y Series Temporales

¿Cuál es la diferencia principal en el enfoque de uso entre una base de datos de **Tiempo Real** y una de **Series Temporales** según su taxonomía NoSQL?

- **A)** No hay diferencia; son dos nombres para la misma tecnología.
- **B)** Las de tiempo real solo almacenan datos de menos de una hora, mientras que las de series temporales son para datos de años.
- **C)** Las de tiempo real se enfocan en una alta velocidad de procesamiento y sistemas de alerta, mientras que las de series temporales priorizan el análisis retrospectivo y el pronóstico.

Big Data

Las principales diferencias es que las de tiempo real tienen una alta velocidad y procesamiento inmediato, mientras que las de series temporales tienen un análisis histórico, agregación y predicción