IES de Teis
Avda. de Galicia, 101
36216 – Vigo

Tfno: 886 12 04 64
e-mail: ies.teis@edu.xunta.es
http://www.iesteis.es

36018173. CSIFC90. MP5074. Sistemas de big data. 2023-2024.

(Grupo A) FINAL

Nombre completo y DNI

Leer antes de empezar.

- La duración del examen es de 2 horas con 5 minutos de cortesía.
- Al acabar el examen debes firmar en la hoja de entrega.
- Ni se permitirá la firma en la hoja de entrega ni se recogerán exámenes fuera de tiempo.
- Durante esta prueba escrita puedes utilizar libremente todo tu material escrito y digital.
- No estará permitido comunicarse con el resto de alumnado por ningún medio.
- La prueba escrita debe permanecer inmutable por lo que es obligatorio usar bolígrafo permanente y no escribir en lápiz, usar corrector o cualquier otro método que pueda poner en duda una alteración posterior.
- Es necesario usar lenguaje técnico.
- Se puede solicitar hojas extra si se necesitan.
- El espacio disponible es suficiente para contestar cada pregunta.

IES de Teis
Avda. de Galicia, 101
36216 – Vigo

Tfno: 886 12 04 64
e-mail: ies.teis@edu.xunta.es
http://www.iesteis.es

1.- Indica con tus propias palabras cuando consideras que un problema entra dentro de la categoría de Big Data y justifica tu respuesta.

RESPUESTA: cuando los datos no se pueden almacenar en un solo equipo y cuando el proceso de los mismos lleva más tiempo que el disponible (también vale traer el código a los datos) Vídeo UD1 Big Data

2.- Explica con tus propias palabras a qué hacen referencia las Vs de Velocidad, Volumen y Variedad en BigData.

RESPUESTA: Volumen hace referencia a la escala de los datos. Se mide en unidades múltiplos del byte, actualmente Petabytes. Velocidad: indica cómo de rápido se debe procesar la información (lotes o tiempo real) y Variedad: hace referencia a datos estrucutrados (tablas) y datos no semi/estructurados (multimedia). UD1 Big Data

3.- Explica con tus palabras que puede pasar con una base de datos que no cumpla con ACID y pon una situación de ejemplo para cada una de las letras. UD1 Big Data

RESPUESTA: Atomicidad: las operaciones se hacen o no se hacen. Podría quedar un registro con solo algunos campos escritos. Consistencia: la base de datos queda en un estado consistente. Podríamos insertar un registro con una clave ya existente. Aislamiento: dos operaciones cualesquiera se ejecutan secuencialmente. Dos actualizaciones simultáneas que dependan del valor de campos distintos pueden no tener sentido y ejecutarse unas sí y otras no. Durabilidad: guardar en disco los cambio. Podría fallar el equipo después de confirmar escritura pero aun no haberse guardado en disco por lo que se perderá. UD1 Big Data

4.- Justifica con tus propias palabras si está de acuerdo con la siguiente afirmación: "A las bases de datos relacionales también le es de aplicación el teorema CAP"

RESPUESTA: sí porque aquellas que no tengan tolerancia a partición siempre son disponibles y consistentes. Aquellas que tengan algún mecanismo de distribución y tolerancia a la partición necesariamente tendrán que optar por disponibilidad o consistencia. UD1 Big Data.

5.- Indica con tus palabras las diferencias que hemos visto entre procesamiento paralelo y procesamiento distribuido.

RESPUESTA: En paralelo los procesadores están en el mismo nodo y comparten memoria principal entre ellos. En distribuido procesadores y memorias están en nodos distintos y siempre será necesaria una etapa de agregación de datos para obtener el resultado final. UD1 Big Data

6.- Qué limitaciones impone el principio SCV para un clúster en el que queremos los resultados del proceso con mucha velocidad (tiempo real) y consistentes (datos exactos).

RESPUESTA: Necesariamente el volumen de los datos tiene que ser pequeño usando solo una parte estadísticamente representactiva o aleatoria. UD1 Big Data

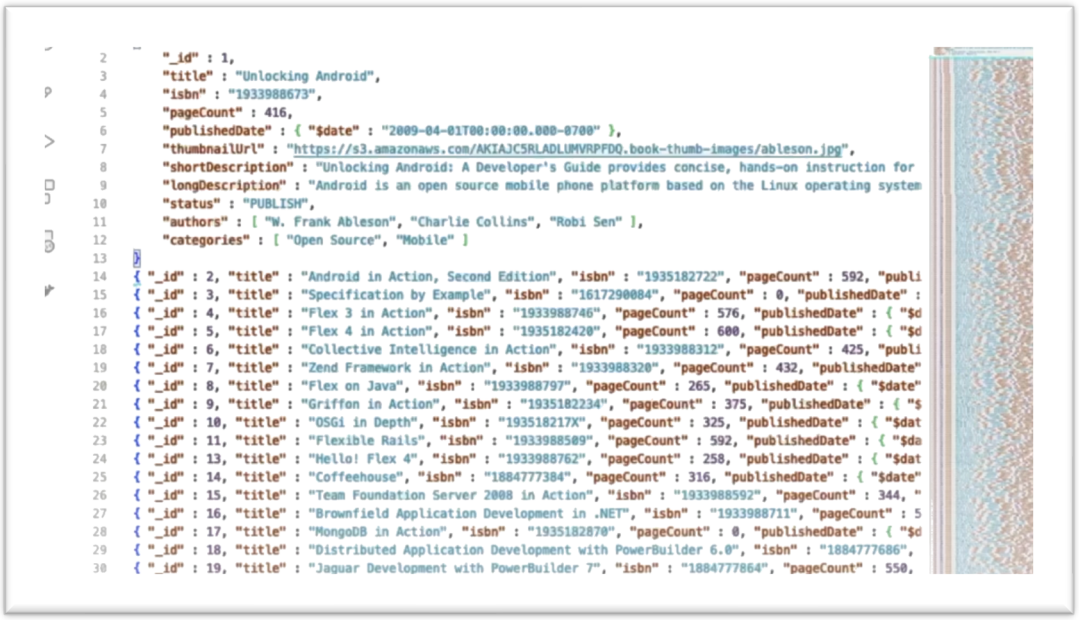
7.- Explica con tus palabras qué diferencia hay entre "mongos", "mongosh" y "mongod"

RESPUESTA: tanto mongos como mongod son los ejecutables de los servidores de mongoDB. Mongod es realmente es propio servidor de la base de datos y es imprescindible para su funcionamiento en cualquier caso. Mongos es un servidor muy ligero que solo se usa en escenarios de sharding y que hace de router entre clientes y el clúster de mongodb donde están los servidores mongod. Mongosh es el shell de la interfaz de comandos para interactuar con los servidores Mongodb o bien directamente conectando con un mongod o un mongos. Instalación personalizada MongoDB

8.- Explica con tus palabras una única ventaja de instalar MongoDB mediante "archivos", "repositorio" y "dockers".

RESPUESTA: Con archivos tenemos libertad total para instalar mongoDB en la carpeta y usuario que queramos. Con repositorio la creación de carpetas, usuario, permisos y servicio es transparente. Con docker podemos levantar mongodb para pruebas y eliminarlo cuando ya no es necesario. Instalación MongoDB desde repositorio y docker.

9.- Indica como importarías los siguientes datos de un archivo JSON a una base de datos MongoDB.



```
    "_id" : 1,
    "title" : "Unlocking Android",
    "isbn" : "1933988673",
    "pageCount" : 416,
    "publishedDate" : { "$date" : "2009-04-01T00:00:00.000-0700" },
    "thumbnailUrl" : "https://s3.amazonaws.com/AKIAJC5RLADLUMVRPFDQ.book-thumb-images/ableson.jpg",
    "shortDescription" : "Unlocking Android: A Developer's Guide provides concise, hands-on instruction for
    "longDescription" : "Android is an open source mobile phone platform based on the Linux operating system
    "status" : "PUBLISH",
    "authors" : [ "W. Frank Ableson", "Charlie Collins", "Robi Sen" ],
    "categories" : [ "Open Source", "Mobile" ]
}
{ "_id" : 2, "title" : "Android in Action, Second Edition", "isbn" : "1935182722", "pageCount" : 592, "publi
{ "_id" : 3, "title" : "Specification by Example", "isbn" : "1617290084", "pageCount" : 0, "publishedDate" :
{ "_id" : 4, "title" : "Flex 3 in Action", "isbn" : "1933988746", "pageCount" : 576, "publishedDate" : { "$d
{ "_id" : 5, "title" : "Flex 4 in Action", "isbn" : "1935182420", "pageCount" : 600, "publishedDate" : { "$d
{ "_id" : 6, "title" : "Collective Intelligence in Action", "isbn" : "1933988312", "pageCount" : 425, "publi
{ "_id" : 7, "title" : "Zend Framework in Action", "isbn" : "1933988320", "pageCount" : 432, "publishedDate"
{ "_id" : 8, "title" : "Flex on Java", "isbn" : "1933988797", "pageCount" : 265, "publishedDate" : { "$date"
{ "_id" : 9, "title" : "Griffon in Action", "isbn" : "1935182234", "pageCount" : 375, "publishedDate" : { "$
{ "_id" : 10, "title" : "OSGi in Depth", "isbn" : "193518217X", "pageCount" : 325, "publishedDate" : { "$dat
{ "_id" : 11, "title" : "Flexible Rails", "isbn" : "1933988509", "pageCount" : 592, "publishedDate" : { "$da
{ "_id" : 13, "title" : "Hello! Flex 4", "isbn" : "1933988762", "pageCount" : 258, "publishedDate" : { "$dat
{ "_id" : 14, "title" : "Coffeehouse", "isbn" : "1884777384", "pageCount" : 316, "publishedDate" : { "$date"
{ "_id" : 15, "title" : "Team Foundation Server 2008 in Action", "isbn" : "1933988592", "pageCount" : 344, "
{ "_id" : 16, "title" : "Brownfield Application Development in .NET", "isbn" : "1933988711", "pageCount" : 5
{ "_id" : 17, "title" : "MongoDB in Action", "isbn" : "1935182870", "pageCount" : 0, "publishedDate" : { "$d
{ "_id" : 18, "title" : "Distributed Application Development with PowerBuilder 6.0", "isbn" : "1884777686",
{ "_id" : 19, "title" : "Jaguar Development with PowerBuilder 7", "isbn" : "1884777864", "pageCount" : 550,
```

RESPUESTA: Usando el comando mongoimport pasándole los datos de la conexión en –db y –collection y también la ruta del archivo json en –file. Importar documentos en MongoDB usando mongoimport

10.- Una empresa tiene un conjunto de réplica repartido por sus sedes. En Vigo tienen 2 nodos, en Santiago tienen otros 2 nodos y en Dublín tienen 1 solo nodo, todos forman un único conjunto de réplica ya configurado. Ahora mismo el primario está en uno de los nodos de la sede de Vigo. Indica qué pasaría con las lecturas y escrituras de la base de datos en los siguientes escenarios.

A) Cae el nodo primario de Vigo
B) Cae toda la sede de Vigo
C) Caen las sedes de Santiago y Dublín.

RESPUESTA: Mientras haya mayoría de 3 se puede elegir otro primario. Si no hay mayoría el primario (si lo hubiera) se degrada a secundario y la base de datos queda como solo lectura. A-otro primario, B-otro primario, C-solo lectura Práctica mongodb

11.- La base de datos de alumnado del IES Teis está creciendo demasiado y necesitamos hacer sharding con ella para poder distribuir sus datos en un clúster dedicado. Si la base de datos contiene los ”datos

personales" típicos, indica con tus palabras y justifica qué clase de sharding elegirías de entre "Provincia" y "Mes_nacimiento".

RESPUESTA: el objetivo de la clase de sharding es determinar a qué nodo va cada registro de la manera más equitativa posible. En este caso la mayoría de los registros serán de "Pontevedra" por lo que irían todos al mismo shard. Mejor es elegir "Mes_nacimiento" porque esto garantiza una distribución bastante uniforme de los registros en distintos nodos. Sharding en MongoDB

12.-Indica con tus propias palabras qué diferencia hay entre "transacciones" y "analítica" en el contexto de bases de datos.

RESPUESTA: las transacciones modifican la base de datos añadiendo o modificando registros. Son operaciones típicas de insertar o actualizar. La analítica no modifica datos y sí los consulta, típicamente los select. UD3 Bases de datos NoSQL

13.- Explica con un ejemplo qué quiere decir que, en las bases de datos NoSQL como MongoDB, la consistencia blanda puede implicar leer datos antiguos en algunos casos y en otros casos el dato más reciente.

RESPUESTA: en MongoDB las escrituras se hacen sobre el primario. Si hay replicación, la confirmación de escritura llega cuando la mayoría de las réplicas también han escrito el dato (pero no todas). El resto de las réplicas aun no han confirmado pero lo están haciendo. Si en ese momento hay una lectura de ese dato, dependiendo del nodo al que consulten pueden tener la versión más actual o la anterior. UD3 Bases de datos NoSQL

14.- En escala Big Data enfocando solo y exclusivamente a la "compresión de los datos" en disco, qué base de datos recomendarías, HBase o MongoDB? Justifica tu respuesta con argumentos técnicos.

RESPUESTA: HBase. Aunque MongoDB comprime los datos porque pasa de JSON a BSON, HBASE comprime mucho más los datos porque se aprovecha de que las columnas almacenan el mismo tipo de datos y eso permite una compresión mucho más eficiente que si los tipos son diversos. UD3 Bases de datos NoSQL

15.- Indica cuales son las características generales que tienen la bases de datos NoSQL. Es suficiente con el título, no es necesarios ejemplos.

RESPUESTA:

Modelo de datos flexible, escalabilidad horizontal, esquema dinámico o sin esquema, alto rendimiento, Alta disponibilidad y tolerancia a fallos, almacenamiento de datos no estructurados. UD3 Bases de datos nosql

16.- Escribe la consulta para la base de datos Neo4j en la que tenemos una base de datos genealógica en la que solo hay relaciones del tipo "Hijo de" y nodos de tipo "Persona" en la que quieres obtener un grafo con los nodos y relaciones desde un nodo con propiedad "nombre: Yo" hasta todos sus familiares de primer grado (padres e hijos).

RESPUESTA: MATCH (yo:Persona {nombre:"Yo"})-[x1:Hijo_de]->(ascendente:Persona), (descendente:Persona)-[x2:Hijo_de]->(yo:Persona {nombre:"Yo"}) RETURN *

Práctica Neo4j

17.- Explica con tus propias palabras en qué consiste el proceso ETL.

RESPUESTA: Son las siglas de Extract, Transform y Load. Hacen referencia a la cada una de las fases desde que se capturan los datos desde fuentes diversas. La parte de transformación hace referencia a adaptar los datos para para que puedan ser mostrados ya sea filtrando, calculando o modificando los datos. Por último la carga hace referencia a llevarlos a una fuente de datos estructurada para el usuario final o bien generar visualizaciones. Introducción a PowerBI

18.- En que consiste la "anulación de dinamización de columnas" en Power Query? Explícalo con tus propias palabras y pon un ejemplo muy sencillo.

RESPUESTA: Consiste en transformar los valores y cabeceras de una o varias columnas en dos nuevas columnas mostrando algo parecido a una pareja de clave-valor donde la columna clave contiene el nombre de la cabecera y en la columna valor el contenido de ese registro para la columna de la cabecera. En una tabla con dos columnas, gasolina95 y gasoil donde cada una almacena el precio por litro de cada carburante, aplicar una anulación de dinamización de columna seleccionando las dos columnas nos

IES de Teis
Avda. de Galicia, 101
36216 – Vigo

Tfno: 886 12 04 64
e-mail: ies.teis@edu.xunta.es
http://www.iesteis.es

devolvería una columna ”combustible” y otra ”precio” con los registros ”gasolina95” – 1,50 y ”gasoil” - 1,40. Práctica PowerBI