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)

Parcial MARZO24

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.- Explica usando tus propias palabras qué son los datos no estructurados. Pon algún ejemplo de en qué situación usamos datos no estructurados. Con cuál de las 5 Vs están relacionado los datos no estructurados.

Son datos que no tienen campos estrictos ni de tamaño fijo que requieren de técnicas complejas para su interpretación.

Se usan en datos multimedia como imágenes, audio y vídeo.

Variedad

2.- ¿Qué término se usa para definir el tipo de hardware que se usa en un clúster de Big Data?

Commodity hardware

3.- Explica con tus palabras en qué consiste la "consistencia" en las bases de datos. Pon un ejemplo de alguna situación en la que una base de datos estuviera en un estado "inconsistente" en MongoDB.

Consistencia es mantener la base de datos es un estado válido, es decir, sin documentos duplicados o contradictorios.

Una réplica en un secundario que aún no se ha actualizado con su primario.

4.- Explica con tus palabras en qué consiste el teorema CAP aplicado a bases de datos distribuidas.

Consiste en que solo se pueden cumplir 2 de 3 características que son, consistencia, disponibilidad y tolerancia a la partición. Por ser distribuidas ya son tolerantes a la partición así que, según este teorema, solo pueden ser o consistentes o disponibles pero no ambas.

5.- Explica con un ejemplo de MongoDB qué significa que una base de datos tenga un estado blando o soft state.

Es posible que en un conjunto de réplicas, con la conexión configurada para que los secundarios respondan a las consultas, pero en los que no todos los secundarios estén ya actualizados, dos consultas consecutivas den respuestas distintas. Es un estado temporal, ya que la replicación llegará a todos los secundarios.

Esas dos mismas consultas ejecutadas unos segundos después ya devolverán el mismo resultado. La base de datos estuvo en un estado blando

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

6.- ¿En qué se diferencian el procesamiento distribuido del procesamiento en paralelo?

Ambos reparten el proceso en varios procesadores pero en el paralelo se comparte memoria principal y se suele dar en el mismo nodo mientras que en el procesamiento distribuido no se comparten recursos hardware y se suele dar en distintos nodos.

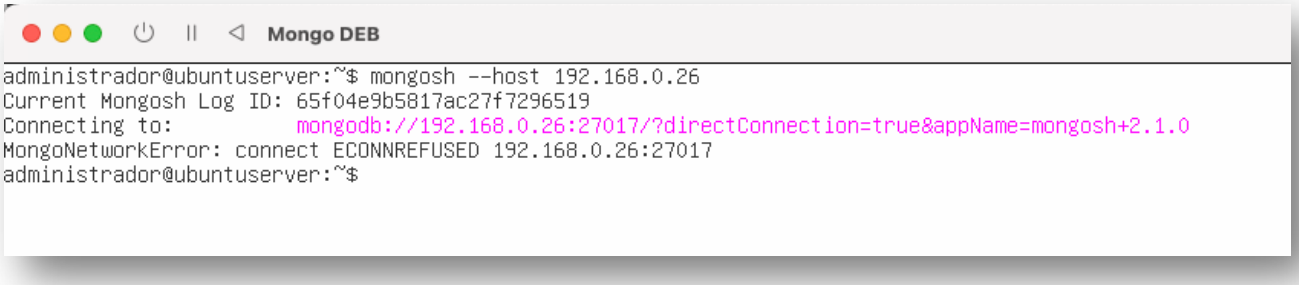
7.- Justifica con tus propias palabras qué estrategia de procesamiento de datos elegirás para el cálculo de facturas de los abonados de Vodafone.

Batch porque es una cantidad grande de datos pero su salida no es urgente y puede demorar horas.

8.- Justifica con tus propias palabras qué estrategia de procesamiento de datos elegirías para el cálculo de los datos necesarios para representar estadísticas de redes sociales.

Tiempo real porque no es fundamental dar un dato actualizado y sí tener actualizaciones rápidas.

9.- Explica con tus palabras qué harías y en qué orden en caso de recibir esta salida al ejecutar el comando “mongosh” para conectarte a una base de datos MongoDB en el host 192.168.0.26.

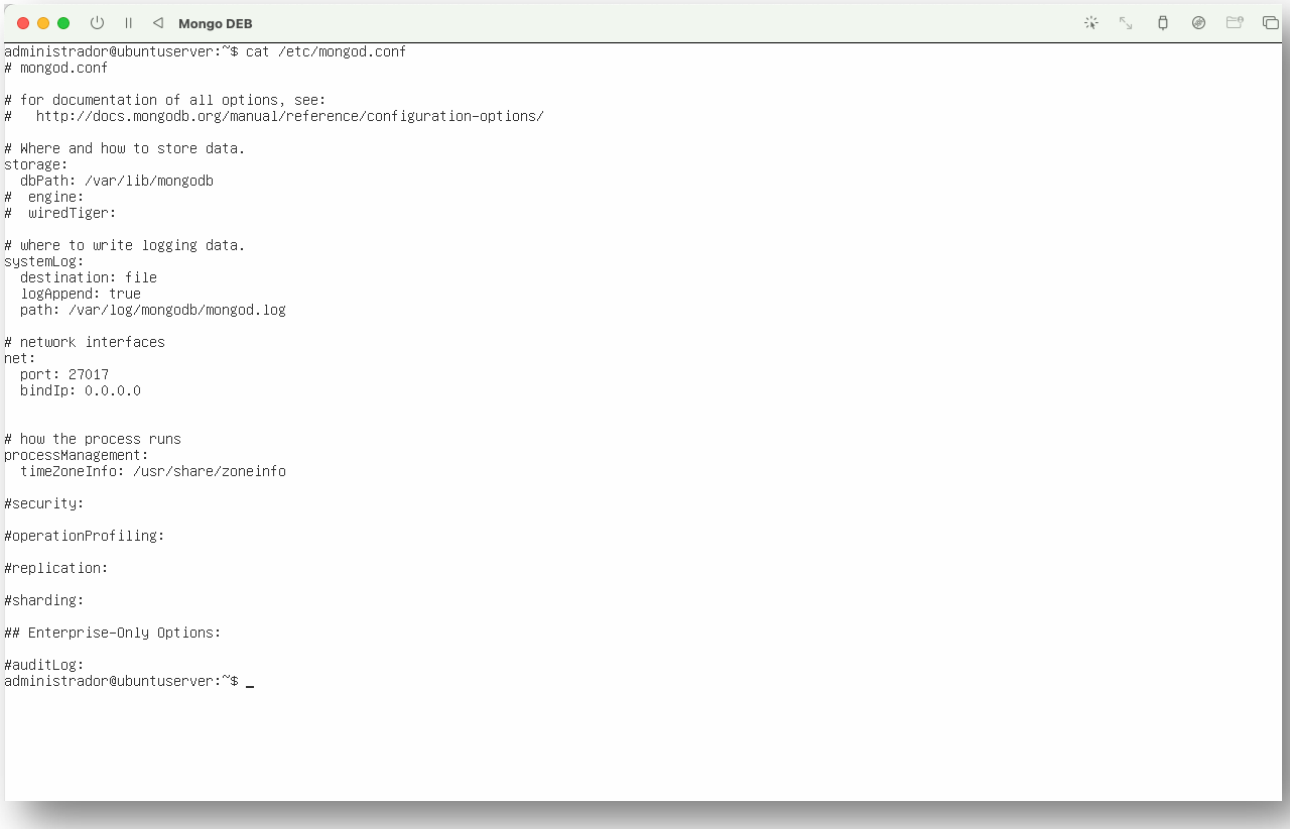


- Revisar si hay conectividad entre hosts.
- Revisar en el servidor si el proceso mondod está correctamente lanzado.
- Revisar en el servidor si archivo de configuración, en especial las direcciones IPs admitidas.

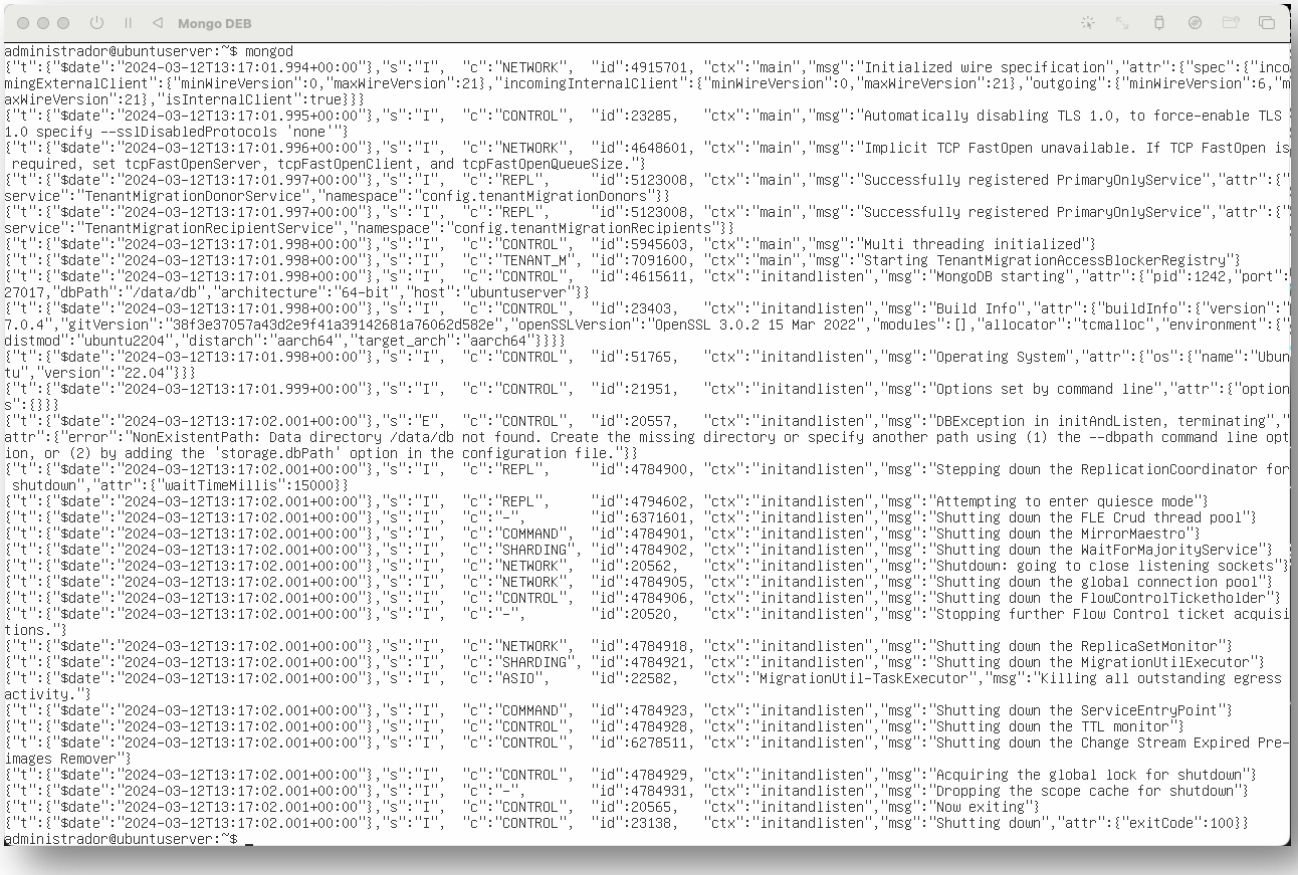
10.- En un servidor acaban de instalar MongoDB desde un repositorio oficial pero no son capaces de conectar usando varios clientes distintos tanto en local como en remoto. Los datos del archivo de configuración “/etc/mongod.conf” en el servidor son:

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



Al ejecutar el comando “mongod” obtenemos la siguiente salida:



Explica con tus propias palabras como solucionarías esta situación indicando qué harías, en qué orden y con qué objetivo.

El archivo de configuración está bien, el problema es que no se está haciendo referencia a ese archivo en la llamada a mongod. El mensaje de error hace referencia a una ruta por defecto que no se encuentra. La solución pasa por arrancar mongod con el parámetro –dbconfig y la ruta del archivo de configuración.

11.-En una instalación de MongoDB usando un contenedor docker llamado mongo, ¿Cómo importarías los documentos del archivo “sales.json”? Aunque partes desde el terminal del equipo anfitrión puedes suponer que el docker ya está en ejecución y que todas las herramientas/archivos que necesites ya están dentro del docker. Lo importante es el procedimiento de la importación a la base de datos MongoDB en docker.

Podría hacer un docker exec -it mongo bash para tener acceso a una terminal interactiva dentro del docker y desde allí hacer un mongoimport –db examen –collection ventas –file sales.json

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

12.- Señala y corrige los errores que encuentres en esta consulta de MongoDB.

```
db.consulta01.InsertOne([
    { _id: “Blanca”,
    nombre.completo : “Blanca Suárez”,
    “añonacimiento” : 1.988
    “altura” : 1,65,
    filmografía : { “Disco, Ibiza, Locomía”; “Me he hecho viral”; “El cuarto pasajero”}
])
```

db.consulta01.InsertOne([	-Error en el I mayúscula y en usar array con un insertone
{ _id: “Blanca”,	El valor de la clave no es apropiado
Nombre.completo : “Blanca Suárez”,	La clave está mal escrita
“añonacimiento” : 1.988	El punto es el separador decimal y falta coma
“altura” : 1,65,	La coma no es separador decimal
filmografía : { “Disco, Ibiza, Locomía”; “Me he hecho viral”; “El cuarto pasajero”}	corchete array
])	Para cerrar el documento llave

13.- Realiza una única consulta a la colección de “ventas” de MongoDB Shell (no interfaz gráfica) en la que devuelvas los campos “_id”, “correo” con los datos del email y “fecha” con los datos de la fecha de compra, de los documentos que han comprado, al menos, un portátil (“laptop”) de menos de 600 euros.

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

(En la base de datos “sample_supllies”, colección "sales” de los datos de prueba de MongoDB en Atlas, el resultado es 397 documentos)

```
{
  "_id": {"$oid": "5bd761dcae323e45a93ccfe8"},
  "saleDate": ISODate("2024-03-12"),
  "items": [
    {
      "name": "notepad",
      "tags": [ "office", "writing", "school"],
      "price": 35.29,
      "quantity": 2
    },
    {
      "name": "pens",
      "tags": ["writing", "office", "school", "stationary"],
      "price": 56.12,
      "quantity": 5
    }
  ],
  "storeLocation": "Denver",
  "customer": {
    "gender": "M",
    "age": 42,
    "email": "cauho@witwuta.sv",
    "satisfaction": 4
  },
  "couponUsed": true,
  "purchaseMethod": "Online"
}
```

Db.sales.find({ “ítems” : { $elemMatch : { “name”: “laptop”, “price” : {$lt : 600} } } } ,

{“correo”:”$customer.email”, “fecha”:”$saleDate”})

14.- Realiza una consulta a la colección de “ventas” de MongoDB Shell (no intefaz gráfica) en la que devuelvas lo documentos cuya fecha sea el 1 de enero de 2016.

(En la base de datos “sample_supllies”, colección "sales” de los datos de prueba de MongoDB en Atlas, el resultado es 5 documentos)

Db.sales.find(

{ “saleDate” : { $gte : ISODate(“2016-01-01”), $lt : ISODate(“2016-01-02”) } }

)

15.- Realiza una consulta a la colección de “ventas” de MongoDB Shell (no interfaz gráfica) en la que devuelvas en un único documento, con un único campo llamado “ciudadConMasVentasDePortátiles”, la localización de la tienda que más número de portátiles ha vendido.

```
Db.sales.aggregate([
{ $match : { “ítems.name” : { $in : [“laptop”] } } },      menos docs a sig etapas, uso de índices y eficiencia
{ $unwind : “$items”},      desempaqueto el array de items
{ $match : {“ítems.name” : “laptop” } },      filtro solo los portátiles
{ $group : { “_id”: “$storeLocation”, cantidad: { $sum: “$items.quantity” } } },    agrupo por ciudad y voy sumando cant.
{ $sort : { “cantidad”:-1} },              ordeno para que mayor aparezca primero
{ $limit: 1},          solo quiero el primero
{ $project: {“ciudadConMasVentasDePortatiles” : “$_id”, “_id”:0 } }     creo nuevo campo y oculto _id
])
```

16.- Realiza una consulta a la colección de “ventas” de MongoDB Shell (no interfaz gráfica) en la que devuelvas los documentos en los que su campo email tiene el dominio de nivel superior geográfico “.es”

(En la base de datos “sample_supllies”, colección "sales” de los datos de prueba de MongoDB en Atlas, el resultado es 20 documentos)

```
Db.sales.find({“customer.email”:/.es$/})
```

17- Viendo el siguiente archivo de configuración, indica cual o cuales de las características de MongoDB se están utilizando y en qué te basas para deducirlo.

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

```
 GNU nano 6.2                    /etc/mongod.conf
# mongod.conf

# for documentation of all options, see:
#   http://docs.mongodb.org/manual/reference/configuration-options/

# Where and how to store data.
storage:
  dbPath: /home/administrador/mongodb
#  engine:
#  wiredTiger:

# where to write logging data.
systemLog:
  destination: file
  logAppend: true
  path: /home/administrador/mongod.log

# network interfaces
net:
  port: 27018
  bindIp: 0.0.0.0


# how the process runs
processManagement:
  timeZoneInfo: /usr/share/zoneinfo

#security:

#operationProfiling:

replication:
  replSetName: shard2ReplSet

sharding:
  clusterRole: shardsvr

## Enterprise-Only Options:

#auditLog:
```

Es un nodo shard, lo indica en el clusterRole, además usa el puerto típico 27018 aunque esto último no es concluyente. También forma parte de un conjunto de réplica llamado “shard2ReplSet”

18.- Una empresa sin problemas de presupuesto quiere disponer de un servidor local de bases de datos MongoDB. Te preguntan si es mejor invertir en un RAID o configurar un conjunto de réplica. Toma una decisión y argumenta con ventajas de tu opción y contras de la otra.

RAID OK -> barato, sencillo, aumenta velocidad lectura y escritura

RAID NOOK -> único nodo,

REPL OK -> ampliable, totalmente tolerante a fallo, permite más conexiones

REPL NO OK-> Escrituras lentas

19.- Una empresa tiene un conjunto de réplica repartido por sus sedes. En Vigo tienen 4 nodos, en Santiago tienen otros 2 nodos y en Dublín tienen 1 único 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

Cualquier otro toma el relevo ya que solo hace falta 4 votos y entre todos los nodos disponibles suman.

B) Cae toda la sede de Vigo

Base de datos de solo lectura, no se pueden conseguir 4 votos para la mayoría

C) Caen las sedes de Santiago y Dublín.

Sigue funcionando todo

D) Cae el primario de Vigo y 2 de los secundarios de Vigo .

Cualquier otro toma el relevo ya que solo hace falta 4 votos y entre todos los nodos disponibles suman.

20.- Explica como configurarías un conjunto de réplica y sus conexiones para tener una base de datos MongoDB donde la prioridad máxima es la velocidad de respuesta a las consultas y, a las escrituras, aun a costa de la consistencia de los datos.

Añadiría hasta 50 nodos al conjunto de réplica haciendo que todos contesten por cercanía, además configuraría el primario para que no requiera de ninguna mayoría para dar el documento por insertado.