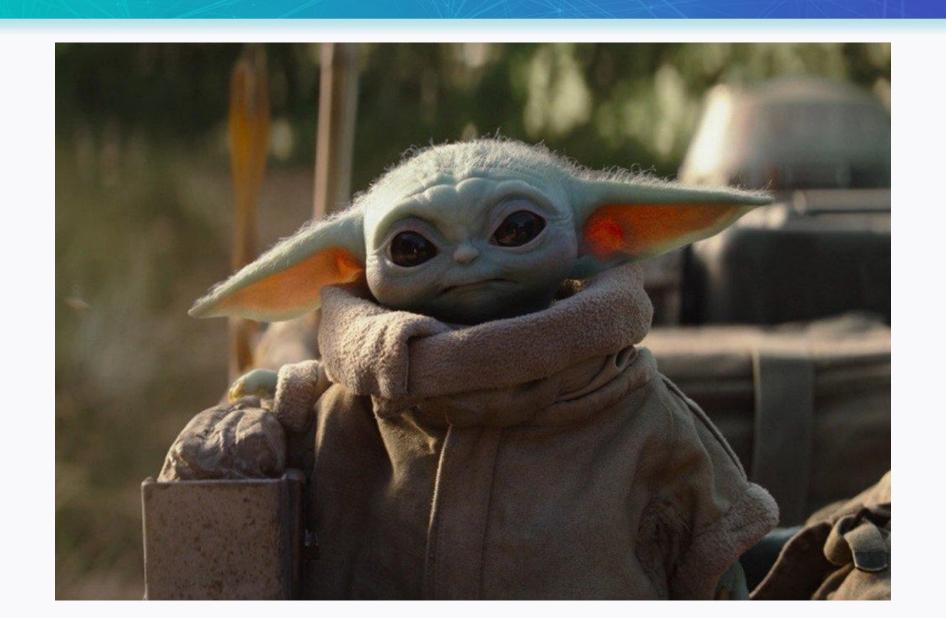
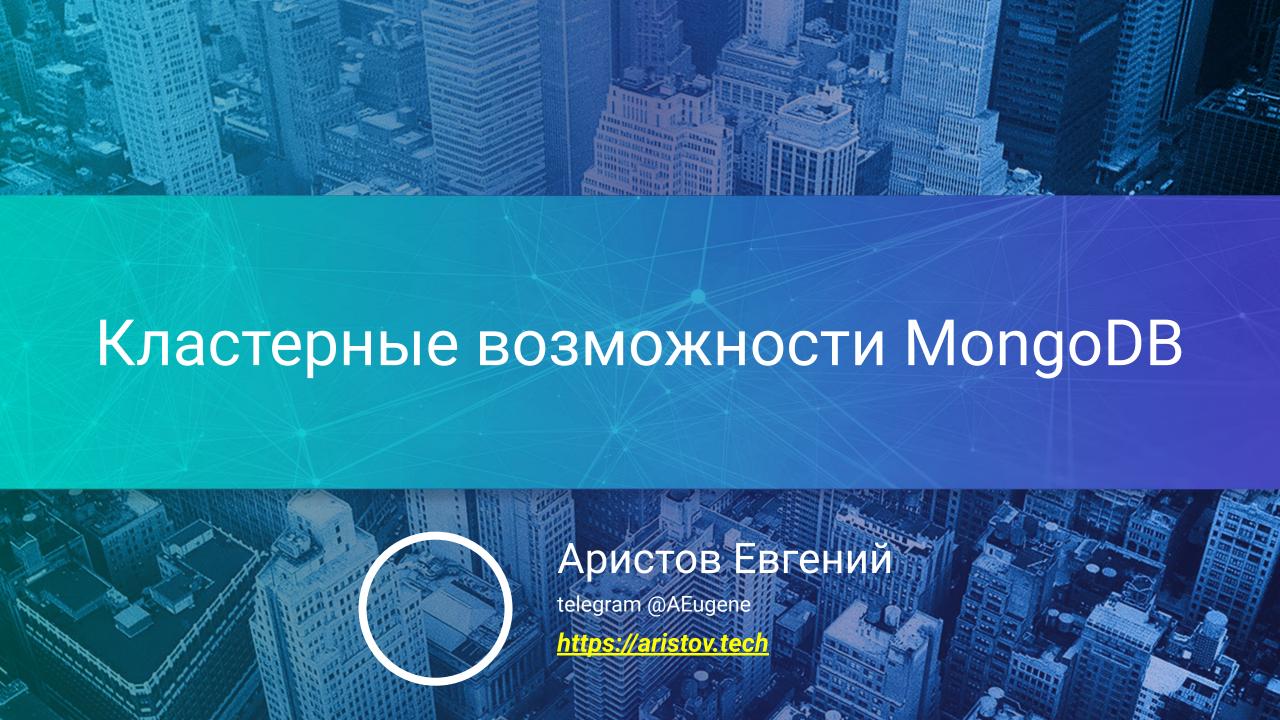


Включил Юджин запись ли пы





Правила вебинара



Активно участвуем

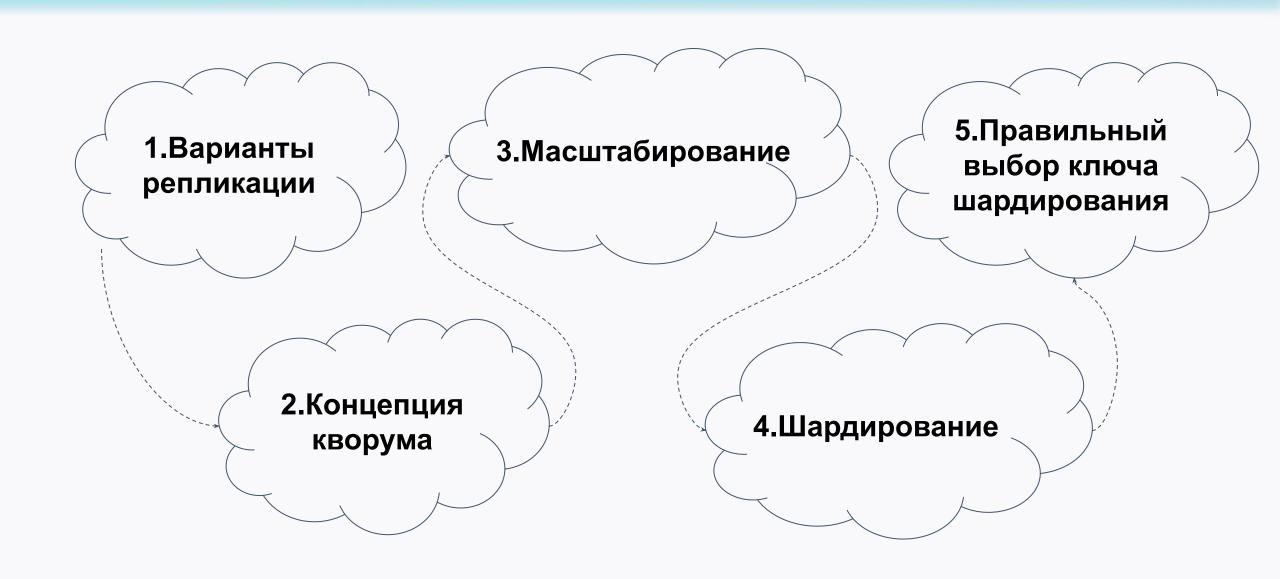


Задаем вопрос в чат



Вопросы вижу в чате, могу ответить не сразу

Маршрут вебинара



Цели вебинара После занятия вы сможете

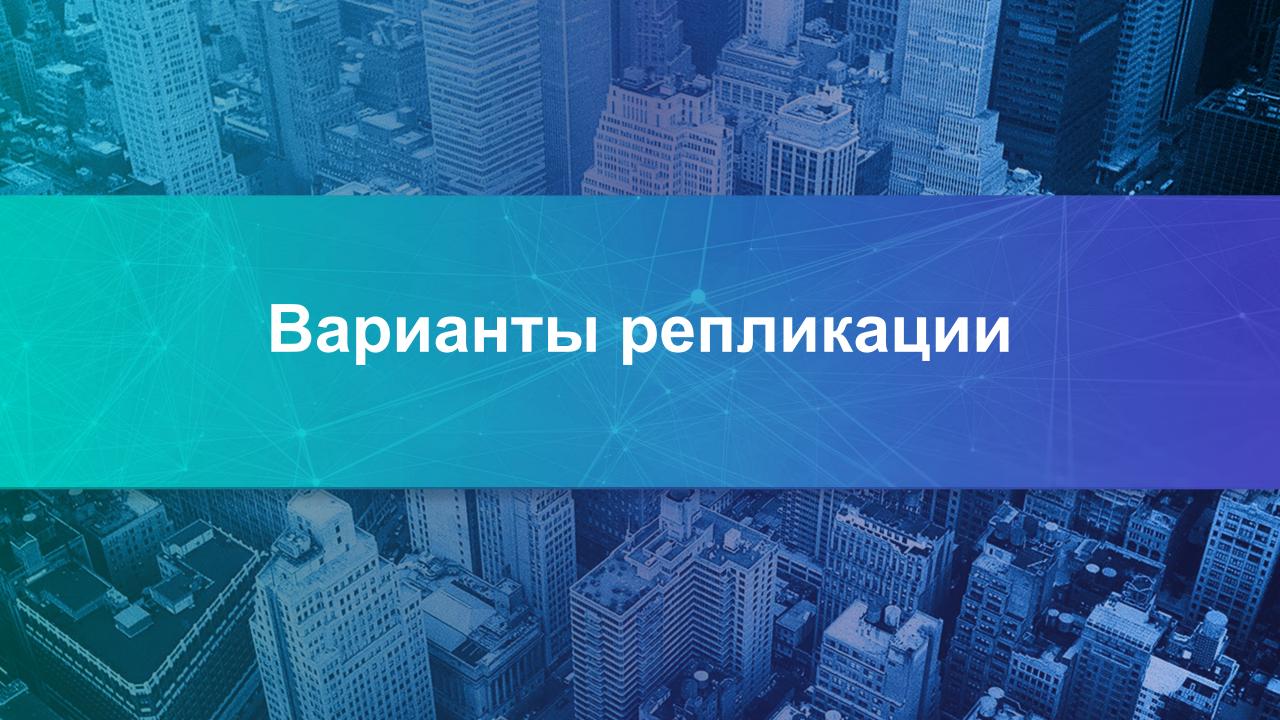
- 1 Выбирать правильный вариант репликации
- Понимать как работает концепция кворума

2 Настраивать шардирование

4 Эффективно выбирать правильный ключ шардирования

Варианты репликации





1. Высокая доступность. Бэкап это хорошо, но нужно время на его развертывание.

- 1. Высокая доступность. Бэкап это хорошо, но нужно время на его развертывание.
- 2. Что делать, когда закончились физические ядра и память у сервера? горизонтально масштабировать

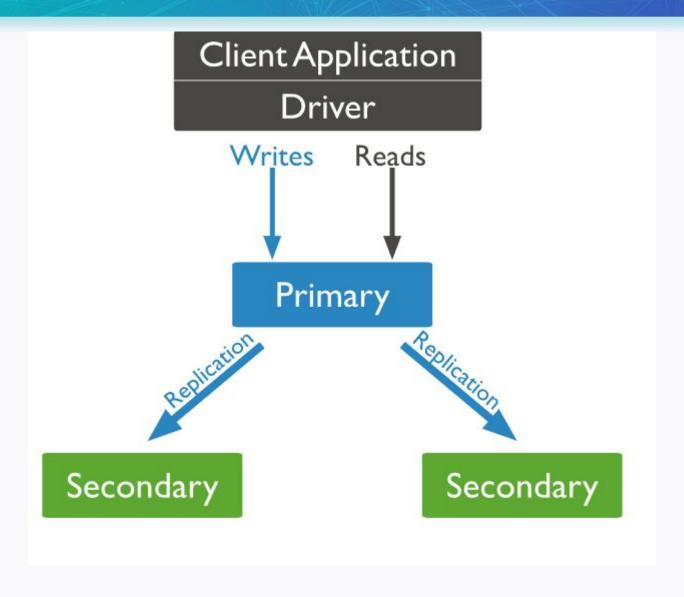
- 1. Высокая доступность. Бэкап это хорошо, но нужно время на его развертывание.
- 2. Что делать, когда закончились физические ядра и память у сервера? горизонтально масштабировать
- 3. Бэкап лучше делать с реплики, а не мастера.

- 1. Высокая доступность. Бэкап это хорошо, но нужно время на его развертывание.
- 2. Что делать, когда закончились физические ядра и память у сервера? горизонтально масштабировать
- 3. Бэкап лучше делать с реплики, а не мастера.
- 4. Геораспределение нагрузки.

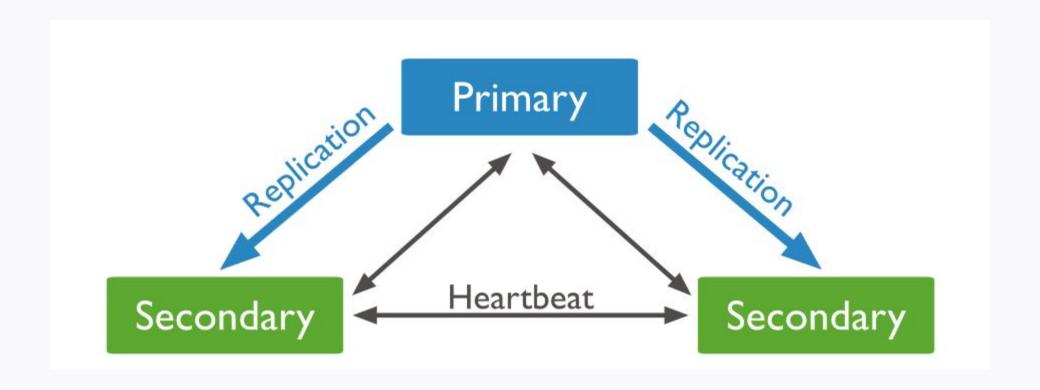
1. Replicaset

2. Master-slave 3.6

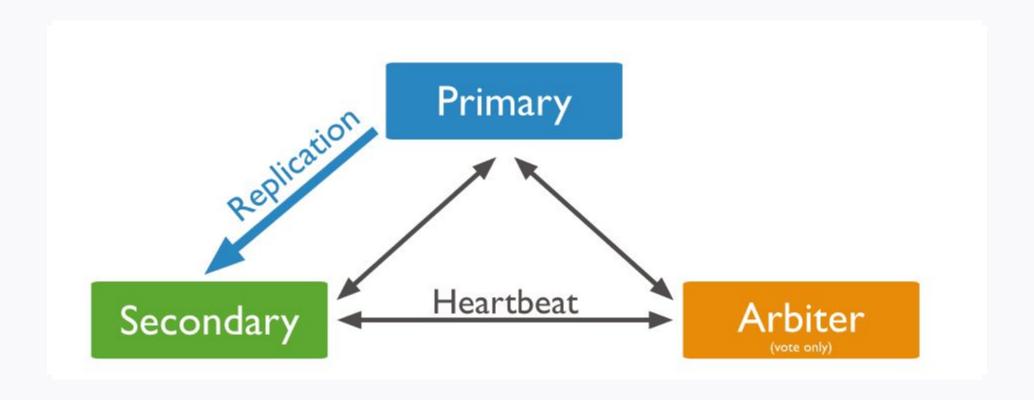
Виды репликации. Replicaset



Виды репликации. Replicaset



Виды репликации. Replicaset



Виды репликации. Основные особенности

- репликация асинхронная
- арбитр не содержит данных и поэтому очень легковесный
- primary может стать secondary и наоборот, а арбитр не может стать ни primary, ни secondary
- максимальное количество реплик 50, из них голосовать разрешено только 7
- запуск арбитра на той же машине, что и реплика не рекомендован (почему?)

Виды репликации. Практика. ReplicaSet

Можно настроить данный РепликаСет несколькими способами:

- Использовать 1 сервер и запустить 3 экземпляра самой mongoDB.
- Использовать 3 сервера с mongoDB.

Итак, для правильно работы ReplicaSet необходимо 3 запущенных экземпляра или сервера с монгой:

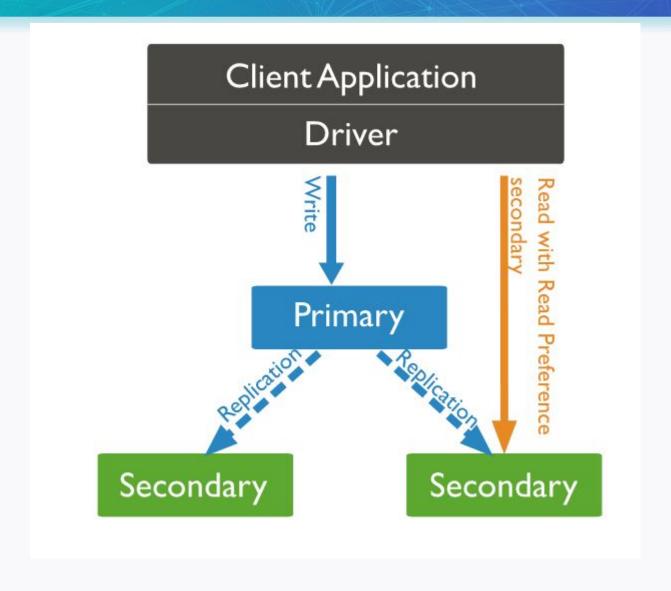
- Одна нода будет выступать как арбитр и не будет принимать на себя никакие данные. Его работа, заключается в том, чтобы выбрать того, кто будет сервером PRIMARY.
- Один сервер выступает в качестве PRIMARY сервера.
- Один сервер выступает в качестве SECONDARY сервера.
- для доступа к экземпляру mongo из cemu (не localhost) необходимо настроить встроенный файрвол и указать, например, --bind_ip_all
- <u>https://docs.mongodb.com/manual/core/security-mongodb-configuration/</u>

Виды репликации. Практика.

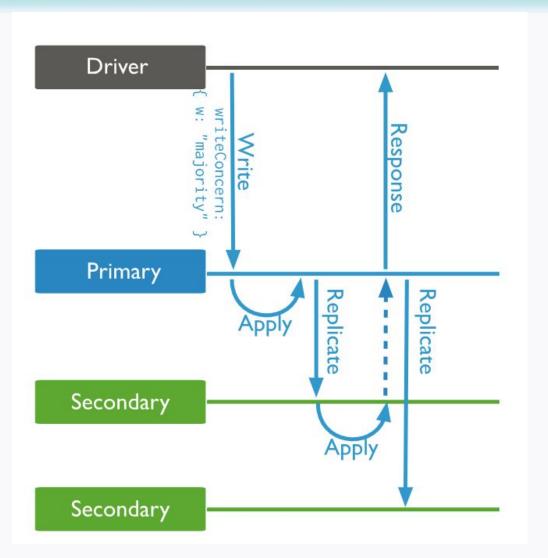
ReplicaSet



Виды репликации. Read from secondary



Виды репликации. WriteConcern/ Majority



Проблемы, которые могут возникнуть при чтении с реплики:

1. так как запись асинхронная, она может быть уже сделана на primary, но не доехать до secondary и мы прочитаем старые данные с secondary

- 1. так как запись асинхронная, она может быть уже сделана на primary, но не доехать до secondary и мы прочитаем старые данные с secondary
- 2. отсюда же вытекает что записав данные на основной сервер, мы не можем быть уверены, когда остальные получат эти данные

- 1. так как запись асинхронная, она может быть уже сделана на primary, но не доехать до secondary и мы прочитаем старые данные с secondary
- 2. отсюда же вытекает что записав данные на основной сервер, мы не можем быть уверены, когда остальные получат эти данные
- 3. опять же мы может установить приоритет на запись или чтение...

- 1. так как запись асинхронная, она может быть уже сделана на primary, но не доехать до secondary и мы прочитаем старые данные с secondary
- 2. отсюда же вытекает что записав данные на основной сервер, мы не можем быть уверены, когда остальные получат эти данные
- 3. опять же мы может установить приоритет на запись или чтение...
- 4. ну и самое веселое, когда падает основной сервер и происходят выборы %)

Варианты сделать доступной реплику для чтения:

- 1. db.slaveOk()
- 2. указывать в строке подключения драйвера нужные параметры https://docs.mongodb.com/manual/reference/connection-string/
- 3. Более точечное использование db.collection.find({}).readPref("secondary", [{ "region": "South" }])

Read Preference Mode / Description

primary

<u>Default mode</u>. All operations read from the current replica set primary.

Multi-document transactions that contain read operations must use read preference primary. All operations in a given transaction must route to the same member.

primaryPreferred

In most situations, operations read from the primary but if it is unavailable, operations read from secondary members.

secondary

All operations read from the secondary members of the replica set.

<u>secondaryPreferred</u>

In most situations, operations read from secondary members but if no secondary members are available, operations read from the primary.

nearest

Operations read from member of the replica set with the least network latency, irrespective of the member's type.

Tip

When possible, use a logical DNS hostname instead of an ip address, particularly when configuring replica set members or sharded cluster members. The use of logical DNS hostnames avoids configuration changes due to ip address changes.

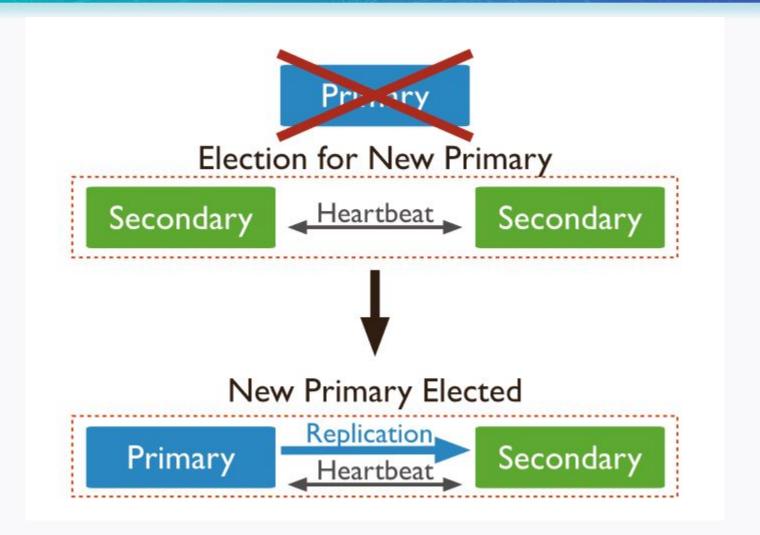
Вопросы:

Почему при возвращении 1 сервера в строй происходит его переход из secondary в primary?

Каким образом выбирается новый primary?



Концепция кворума



https://docs.mongodb.com/manual/replication/#replication-auto-failover

Концепция кворума

Особенность для 3 нод:

при конфигурации Primary-Secondary-Arbiter (P-S-A) и writeConcern: **majority** необходимо 2 ноды для согласования записи данных, соответственно при падении Primary или Secondary у нас останется только чтение и неконсистентная запись с "w:1" - подтверждение еще с 1 ноды нужно %(

при конфигурации Primary-Secondary-Secondary (P-S-S) и writeConcern : **majority**

при падении 1 ноды запись останется работать

https://docs.mongodb.com/manual/reference/write-concern/#calculating-majority-count

Концепция кворума

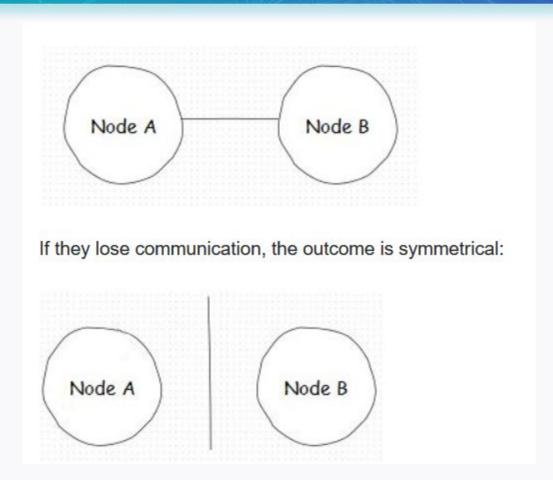
https://raft.github.io/

Концепция кворума

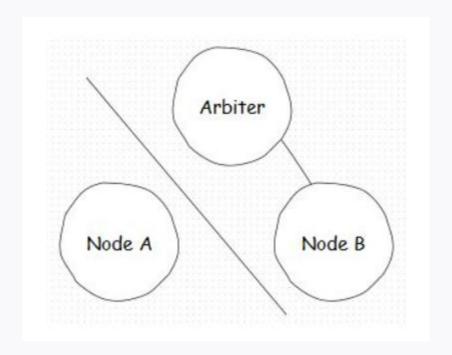
При голосовании за новую ноду, должно проголосовать округленное значение N/2 + 1, где N, это количество хостов, которым разрешено голосовать

Data Nodes	N/2	Voting Minimum	Fault Tolerance	Fault Tolerance w/ Arbiter	Split Brain Possible
2	1	2	0	1	
3	1.5	2	1	2	True
4	2	3	1	2	
5	2.5	3	2	3	True
6	3	4	2	3	
7	3.5	4	3	4	True

Split brain



Split brain



Концепция кворума

Голосовать могут только ноды со статусами:

- o <u>PRIMARY</u>
- o <u>SECONDARY</u>
- o <u>STARTUP2</u>
- o <u>RECOVERING</u>
- o <u>ARBITER</u>
- o <u>ROLLBACK</u>

Концепция кворума

members[n].priority

int from 0 до 1000 for primary/secondary; 0 or 1 for arbiters.

Default: 1.0 for primary/secondary; 0 for arbiters.

Changing the balance of priority in a replica set will trigger one or more elections. If a lower priority secondary is elected over a higher priority secondary, replica set members will continue to call elections until the highest priority available member becomes primary.

- > cfg = rs.conf()
- > cfg.members[3].priority = 2
- > rs.reconfig(cfg)

members[n].votes - маскимум 7 реплик с правом голоса из 50





Проблема - все тормозит, репликасет не спасает?

Проблема - все тормозит, репликасет не спасает?

1. отключаем подтверждение записи primary большинством голосов {w:1, j:true}

Проблема - все тормозит, репликасет не спасает?

- 1. отключаем подтверждение записи primary большинством голосов {w:1, j:true}
- 2. отключаем журналирование {w:1, j:false}

Проблема - все тормозит, репликасет не спасает?

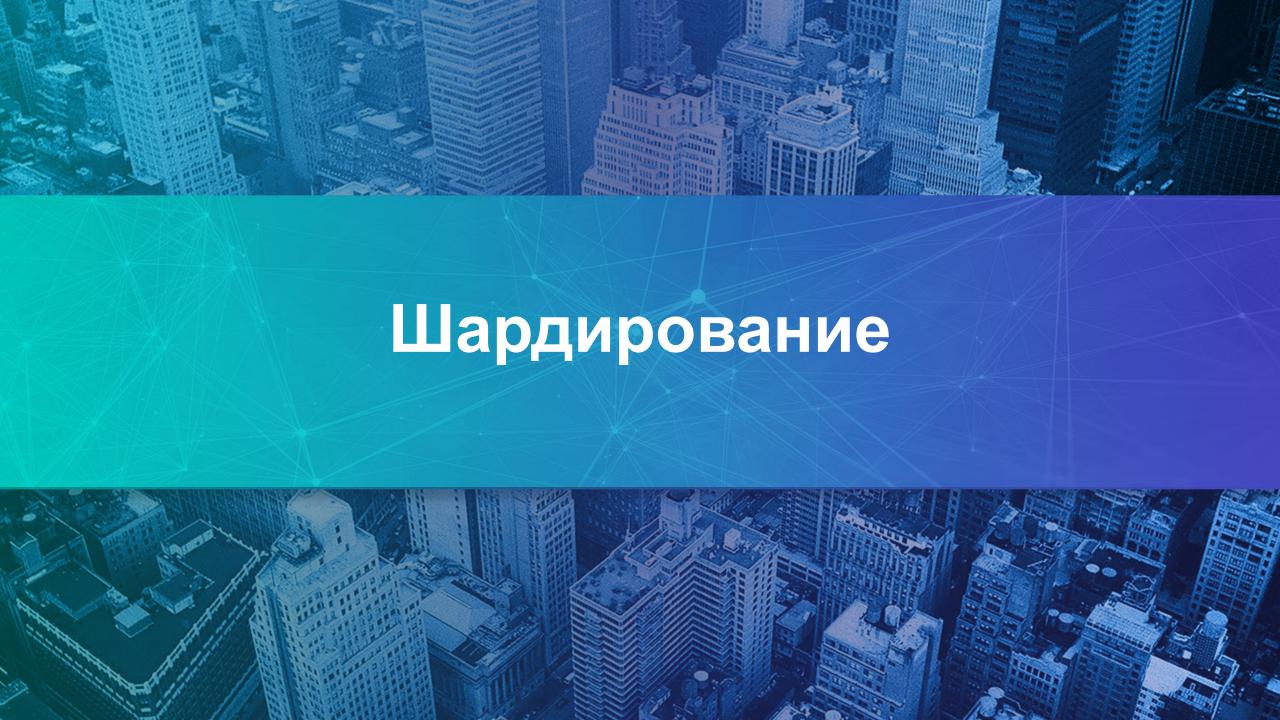
- 1. отключаем подтверждение записи primary большинством голосов {w:1, j:true}
- 2. отключаем журналирование {w:1, j:false}
- 3. отключаем любые проверки {w:0, j:false}

Проблема - все тормозит, репликасет не спасает?

- 1. отключаем подтверждение записи primary большинством голосов {w:1, j:true}
- 2. отключаем журналирование {w:1, j:false}
- 3. отключаем любые проверки {w:0, j:false}
- 4. что дальше?

https://docs.mongodb.com/manual/reference/write-concern/ https://habr.com/ru/post/335698/





Есть коллекция, мы ее хотим как-то раскидать по шардам.

Для этого MongoDB делит коллекцию на чанки с помощью ключа шардирования, пытаясь поделить их в пространстве шард-ключа на равные кусочки.

Дальше включается балансировщик, который старательно раскладывает эти чанки по шардам в кластере. Причем балансировщику все равно, сколько эти чанки весят и сколько в них документов, так как балансировка идет в чанках поштучно.

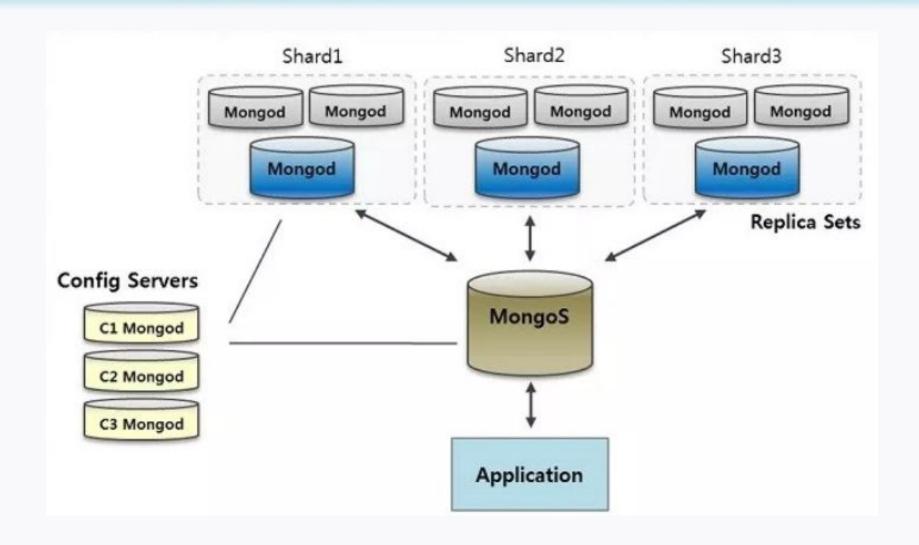
Минусы:

- С шардированием обязательно будет усложняться код во многие места придется протащить ключ шардирования. Это не всегда удобно, и не всегда возможно. Некоторые запросы пойдут либо broadcast, либо multicast, что тоже не добавляет масштабируемости. Подходите к выбору ключа по которому пойдет шардирование аккуратнее.
- **В шардированных коллекциях ломается операция count**. Она начинает возвращать число больше, чем в действительности может соврать в 2 раза. Причина лежит в процессе балансировки, когда документы переливаются с одного шарда на другой. Когда документы перелились на соседний шард, а на исходном еще не удалились count их все равно посчитает.
- Шардированный кластер гораздо тяжелее в администрировании. Девопсы перестанут с вами здороваться, потому что процедура снятия бэкапа становится радикально сложнее. Необходимо большая автоматизация работы.

Проблемы:

При включении шардинга у нас есть 2 основные группы проблем.

- 1. Шардинг коллекции не возможен, если:
 - требуется обновление полей, входящих в ключ шардирования;
- на коллекции есть несколько уникальных ключей;
- под результат запроса findAndModify попадают данные на разных шардах.
- 2. Несбалансированная нагрузка по шардам, если:
- слабая селективность ключа шардирования;
- запросы без значений ключа шардинга.



В схеме мы можем видеть config сервер, его отличие от mongod в том, что он не обрабатывает клиентские запросы, а является хранилищем метаданных — он знает физические адреса всех chunk-ов, знает какой chunk, на каком шарде искать и какой диапазон у того или иного шарда. Все эти данные он хранит в специально отведенном месте — config database.

В данной схеме также присутствует роутер запросов — mongos, на него возлагаются следующие задачи:

- 1. Кэширование данных, хранимых на config сервере
- 2. Роутинг запросов чтения и записи от драйвера маршрутизация запросов от приложений на нужные шарды, mongos точно знает где физически находится тот или иной чанк
- 3. Запуск фонового процесса "Балансер"

chunk(id)

-беск ---- + беск

```
{"name": "Max", "age": 23, "id": 23}

{"name": "John", "age": 28, "id": 15}

{"name": "Nick", "age": 19, "id": 56}

{"name": "Carl", "age": 19, "id": 78}
```

chunk(id)

-беск ---- + беск

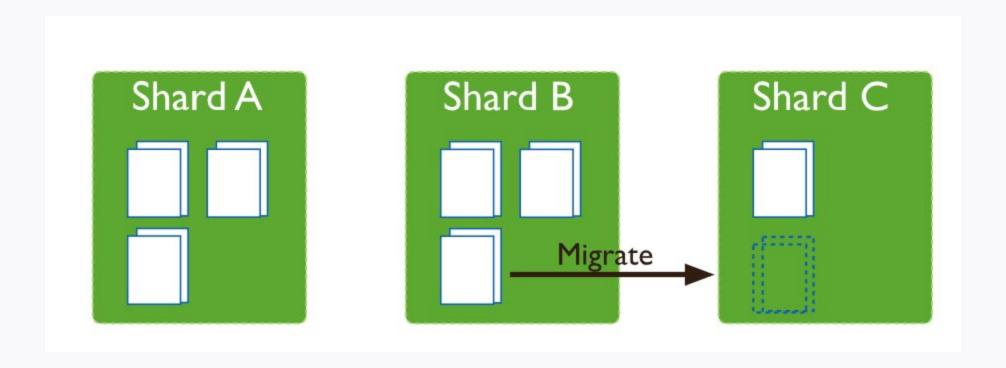
chunk(id)

-беск ---- 45

chunk(id)

45 ---- + беск

Практика



Балансировка шардов

Шардируем, запускаем, и внезапно узнаем, что балансировка не бесплатна.

Балансировка проходит несколько стадий. Балансировщик выбирает чанки и шарды, откуда и куда будет переносить. Дальнейшая работа идет в две фазы: сначала **документы копируются** с источника в цель, а потом документы, которые были скопированы, **удаляются**.

Шард у нас перегружен, в нем лежат все коллекции, но первая часть операции для него легкая. А вот вторая — удаление — совсем неприятная, потому что уложит на лопатки шард и так страдающий под нагрузкой.

Проблема усугубляется тем, что если мы балансируем много чанков, например, тысячи, то с настройками по умолчанию все эти чанки сначала копируются, а потом приходит удалятор и начинает их скопом удалять. В этот момент на процедуру уже не повлиять и приходится только грустно наблюдать за происходящим.

Поэтому если вы подходите к тому, чтобы шардировать перегруженный кластер, вам нужно планировать, так как **балансировка занимает время.** Желательно это время брать не в прайм-тайм, а в периоды низкой нагрузки. Балансировщик — отключаемая запчасть. Можно подойти к первичной балансировке в ручном режиме, отключать балансировщик в прайм-тайм, и включать, когда нагрузка снизилась, чтобы позволить себе больше.

```
sh.stopBalancer()
3.6 можно останавливать миграцию
4.2 также отключает автосплит
sh.getBalancerState()
sh.disableBalancing("students.grades")
поменять размер чанка в Мб
use config
config = db.getSiblingDB("config")
config.shards.updateOne( { "_id" : "<shard>"}, { $set : { "maxSize" : 1024 } } )
или поменять значение по умолчанию
```

db.settings.save({ _id:"chunksize", value: <sizeInMB> })

```
db.settings.update(
    {_id: "balancer" },
    { $set: { activeWindow : { start : "<start-time>", stop : "<stop-time>" } },
    { upsert: true }
)
```

Replace <start-time> and <end-time> with time values using two digit hour and minute values (i.e. HH:MM) that specify the beginning and end boundaries of the balancing window.

Необходимое количество разницы чанков между нодами для старта миграции:

Fewer than 20	2
20-79	4
80 and greater	8

если доступ по сети:

https://docs.mongodb.com/manual/core/security-mongodb-configuration/

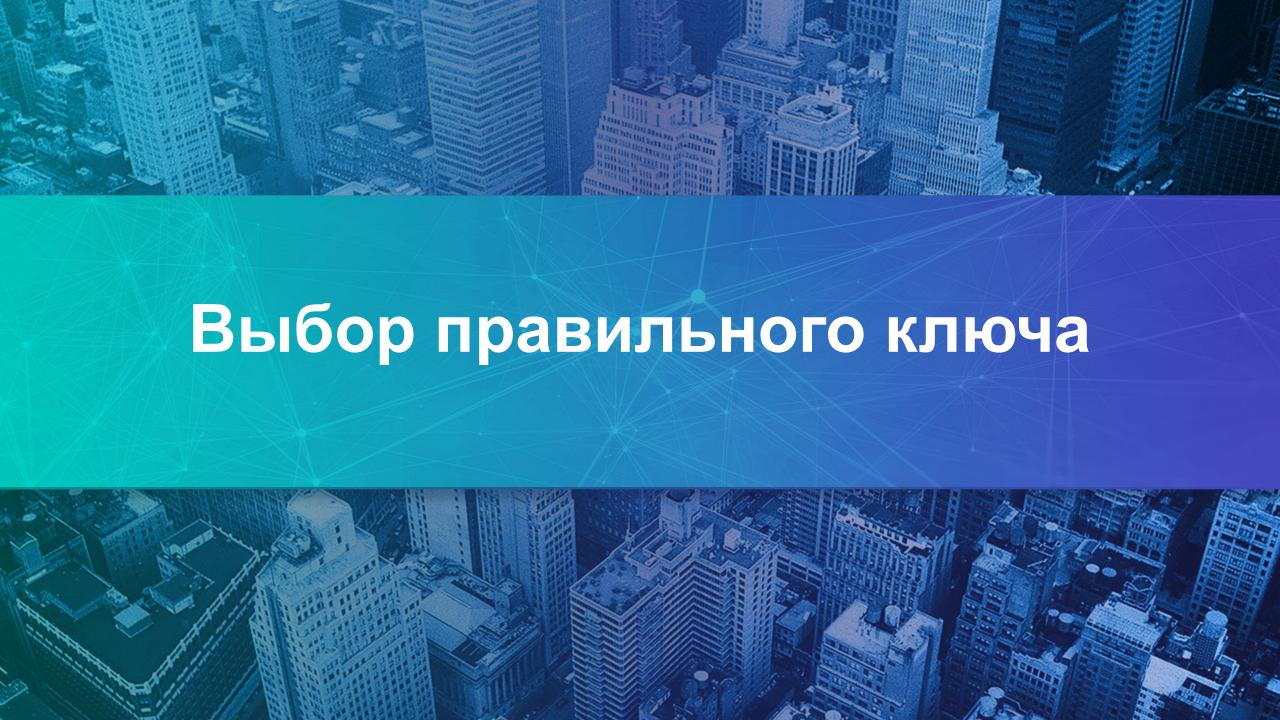
mongod --bind_ip_all

default localhost

при конфигурации конфиг сервера указываем имя, а не IP:

```
rs.initiate(
  _id: "<replSetName>",
  configsvr: true,
  members: [
   { _id : 0, host : "cfg1.example.net:27019" },
   { _id : 1, host : "cfg2.example.net:27019" },
   { _id : 2, host : "cfg3.example.net:27019" }
```





У хорошего ключа несколько параметров: **высокий cardinality**, **немутабельность** и он **хорошо ложится в частые запросы**.

Важно заметить, что при достижении каким-либо чанком неделимого диапазона, например (44, 45], деление происходить не будет и чанк будет расти свыше chunksize. Поэтому следует внимательно выбирать shard key, чтобы это была как можно наиболее случайная величина. Например, если нам надо заполнить БД всеми людьми на планете, то удачными выборами shard key были бы номер телефона, идентификационный налоговый номер, почтовый индекс. Неудачными — имя, город проживания.

Естественный выбор ключа шардирования — это первичный ключ — поле id. Если вам подходит для шардирования поле id, то лучше прямо по нему и шардировать. Это отличный выбор — у него и cardinality хороший, он и немутабельный, но а насколько хорошо ложится в частые запросы — это ваша бизнес-специфика.

Еще пример неудачного ключа шардирования. Есть коллекцию переводов — translations. В ней есть поле language, которое хранит язык. Например, коллекция поддерживает 100 языков и мы шардируем по языку. Это плохо — cardinality количество возможных значений всего 100 штук, что мало. Но это не самое плохое — может быть, для этих целей cardinality достаточно. Хуже, что как только мы пошардировали по языку, мы тут же узнаем, что у нас англоязычных пользователей в 10 раза больше, чем остальных. На несчастный шард, на котором находится английский язык, приходит в десять раз больше запросов, чем на все остальные вместе взятые.

Поэтому надо учитывать, что иногда шард-ключ естественным образом тяготеет к неравномерному распределению нагрузки.

Прейдем к практике

Алгоритм выбора может быть следующим:

- 1. Еще раз вспоминаем ограничения.
- 2. К существующим индексам добавляем другие со всеми возможными (и осмысленными) комбинациями полей.
- 3. Перечисляем всех кандидатов на ключи шардинга. По каждому допустимому индексу можно выбрать ключ по его любому префиксу. Например для индекса { a: 1, b: 1, c: 1 } можно выбрать префиксы { a: 1} и { a: 1, b: 1 } или полный ключ { a: 1, b: 1, c: 1 }.
 - Важно! Если поле в индексе, но ключ шарда построен по префиксу без этого поля, то такое поле остается изменяемым.
- 4. Для каждого кандидата оцениваем:
 - Селективность, есть ли ограничения на рост минимального чанка.
 - Возможность добавить в каждый поисковый запрос конкретные значения ключа чанка.
 - Возможность исключить обновление поля, попавшего в ключ шардинга.
 - Долю запросов, которые не изолируются на одном шарде.
- 5. Выбираем ключ с максимальной долей изолированных запросов.

Шардинг по префиксу индекса с уникальностью

Пример выбора ключа, когда пришлось остановить выбор на менее селективном ключе шардинга.

Сущности в гипотетическом проекте Smartcat:

- Аккаунт (Account) владелец документов.
- Документ (Document) документ для перевода (идентификатор глобально уникальный).
- Сегмент (Segment) предложение в документе.

Поля сегмента:

- accountld идентификатор аккаунта.
- idlnAccount уникальный идентификатор в рамках аккаунта.
- documentId идентификатор документа.
- order порядковый номер сегмента в документе.
- и другие...

Индексы:

- { documentId: 1, idInAccount: 1 }
- { documentId: 1, order: 1 }
- { accountld: 1, idInAccount: 1 }, unique = 1

Большинство запросов включают в себя documentld. Значит, выбирать будем из первых двух индексов, и это исключает шардинг по accountld. Каждый документ имеет глобально уникальный идентификатор и не может принадлежать двум аккаунтам.

idInAccount в нашем коде известен только в одном поисковом запросе, т.е. при выборе полного ключа шардинга сегменты документа могут быть разделены по разным чанкам, и поисковые запросы могут обращаться к нескольким шардам.

Рассматриваем вариант шардинга по *{ documentld: 1}* — это допустимо, т.к. он является префиксом с уникальностью. Мы оцениваем максимальный объем сегментов одного документа — это примерно 50 Mb. Значит, документ отлично входит в размер одного чанка шардинга (он у нас 64 Mb).

Результат:

- Строим индекс с уникальностью по { documentId: 1, idInAccount: 1 }
- Включаем шардинг по { documentId: 1}

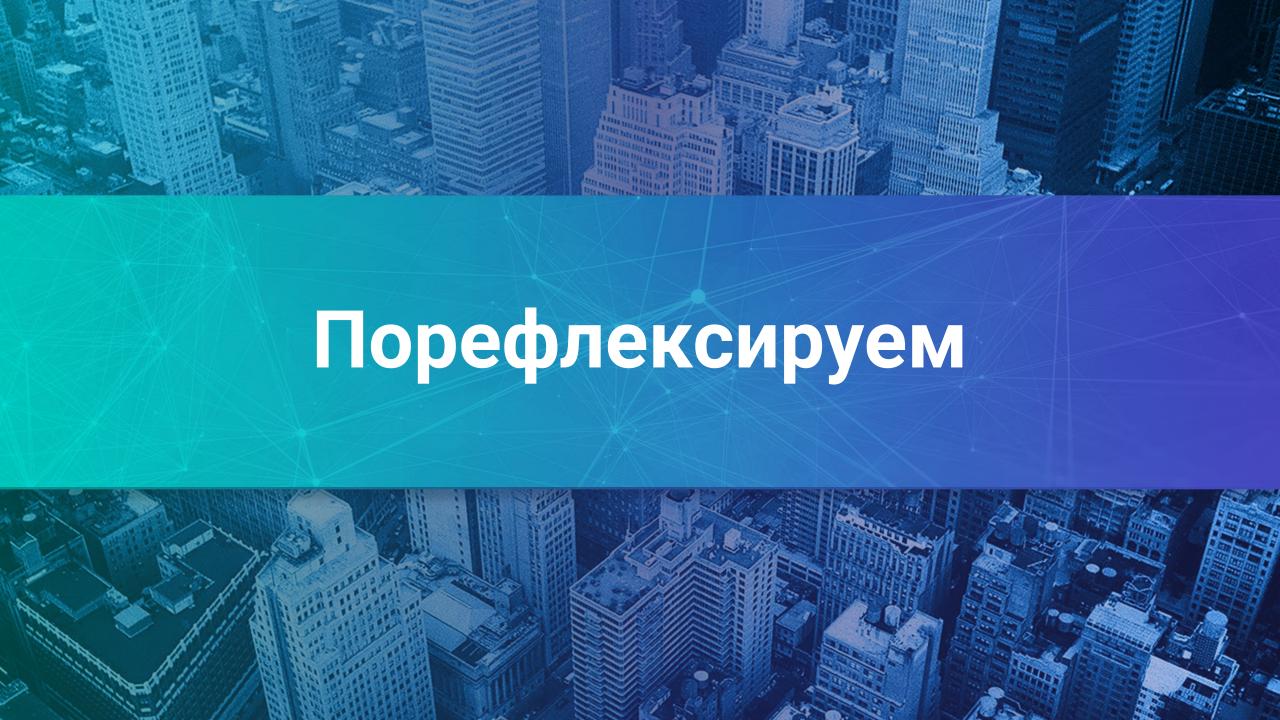
Чек-лист при планировании схемы данных и индексов коллекции:

- 1. Экономим трафик репликации минимизируем размер обновления документа.
- 2. Помним о rollback бизнес-логика должна поддерживать обрыв операций записи, так же как и внезапное выключение сервера (w:majority).
- 3. Не препятствуем шардингу на больших коллекциях не заводим больше одного индекса с уникальностью, но лучше вообще не включать уникальность, если это не критично для бизнес-логики.
- 4. Изолируем поисковые запросы одним шардом большинство поисковых запросов должно включать в себя значения ключа шардинга.
- 5. Балансировка шардов тщательно выбираем тип идентификатора, самый лучший из них GUID.



Primary shard

https://docs.mongodb.com/manual/core/sharded-cluster-shards/



Вопросы?

• Какая служба отвечает за доступ к шардированному кластеру и распределяет чанки?



