MongoDB Essentials Training

Release 3.0

MongoDB, Inc.

October 08, 2015

Contents

1 Introduction 3
- 1.1 Warm Up . . . 3
- 1.2 MongoDB Overview . . . 4
- 1.3 MongoDB Stores Documents . . . 8
- 1.4 Storage Engines . . . 12
- 1.5 Lab: Installing and Configuring MongoDB . . . 17

2 CRUD 23
- 2.1 Creating and Deleting Documents . . . 23
- 2.2 Reading Documents . . . 29
- 2.3 Query Operators . . . 37
- 2.4 Lab: Finding Documents . . . 41
- 2.5 Updating Documents . . . 42
- 2.6 Lab: Updating Documents . . . 49

3 Indexes 51
- 3.1 Index Fundamentals . . . 51
- 3.2 Compound Indexes . . . 58
- 3.3 Lab: Optimizing an Index . . . 64
- 3.4 Multikey Indexes . . . 65
- 3.5 Hashed Indexes . . . 69
- 3.6 Lab: Finding and Addressing Slow Operations . . . 71
- 3.7 Lab: Using `explain()` . . . 71

4 Drivers 72
- 4.1 Introduction to MongoDB Drivers . . . 72
- 4.2 Lab: Driver Tutorial (Optional) . . . 74

5 Aggregation 75
- 5.1 Aggregation Tutorial . . . 75
- 5.2 Optimizing Aggregation . . . 84
- 5.3 Lab: Aggregation Framework . . . 86

6 Introduction to Schema Design 89
- 6.1 Schema Design Core Concepts . . . 89
- 6.2 Schema Evolution . . . 96
- 6.3 Common Schema Design Patterns . . . 100

7 Replica Sets 105
- 7.1 Introduction to Replica Sets . . . 105

7.2 Elections in Replica Sets . . . 109
7.3 Replica Set Roles and Configuration . . . 115
7.4 The Oplog: Statement Based Replication . . . 117
7.5 Lab: Working with the Oplog . . . 119
7.6 Write Concern . . . 121
7.7 Read Preference . . . 126
7.8 Lab: Setting up a Replica Set . . . 127

8 Sharding 132
8.1 Introduction to Sharding . . . 132
8.2 Balancing Shards . . . 140
8.3 Shard Tags . . . 143
8.4 Lab: Setting Up a Sharded Cluster . . . 145

9 Reporting Tools and Diagnostics 152
9.1 Performance Troubleshooting . . . 152

10 Backup and Recovery 160
10.1 Backup and Recovery . . . 160

11 Security 165
11.1 Security . . . 165
11.2 Lab: Creating an Admin User . . . 167
11.3 Lab: Creating a `readWrite` User . . . 169

1 Introduction

Warm Up **(page 3)** Activities to get the class started

MongoDB Overview **(page 4)** MongoDB philosophy and features

MongoDB Stores Documents **(page 8)** The structure of data in MongoDB

Storage Engines **(page 12)** MongoDB storage engines

Lab: Installing and Configuring MongoDB **(page 17)** Install MongoDB and experiment with a few operations.

1.1 Warm Up

Introductions

- Who am I?
- My role at MongoDB
- My background and prior experience

Note:

- Tell the students about yourself:
  - Your role
  - Prior experience

Getting to Know You

- Who are you?
- What role do you play in your organization?
- What is your background?
- Do you have prior experience with MongoDB?

Note:

- Ask students to go around the room and introduce themselves.
- Make sure the names match the roster of attendees.
- Ask about what roles the students play in their organization and note on attendance sheet.
- Ask what software stacks students are using.
  - With MongoDB and in general.
  - Note this informaton as well.

MongoDB Experience

- Who has never used MongoDB?
- Who has some experience?
- Who has worked with production MongoDB deployments?
- Who is more of a developer?
- Who is more of operations person?

10gen

- MongoDB was initially created in 2008 as part of a hosted application stack.
- The company was originally called 10gen.
- As part of their overarching plan to create the 10gen platform, the company built a database.
- Suddenly everybody said: "I like that! Give me that database!"

Origin of MongoDB

- 10gen became a database company.
- In 2013, the company rebranded as MongoDB, Inc.
- The founders have other startups to their credit: DoubleClick, ShopWiki, Gilt.
- The motivation for the database came from observing the following pattern with application development.
  - The user base grows.
  - The associated body of data grows.
  - Eventually the application outgrows the database.
  - Meeting performance requirements becomes difficult.

1.2 MongoDB Overview

Learning Objectives

Upon completing this module students should understand:

- MongoDB vs. relational databases and key/value stores
- Vertical vs. horizontal scaling
- The role of MongoDB in the development stack
- The structure of documents in MongoDB
- Array fields
- Embedded documents
- Fundamentals of BSON

MongoDB is a Document Database

Documents are associative arrays like:

- Python dictionaries
- Ruby hashes
- PHP arrays
- JSON objects

An Example MongoDB Document

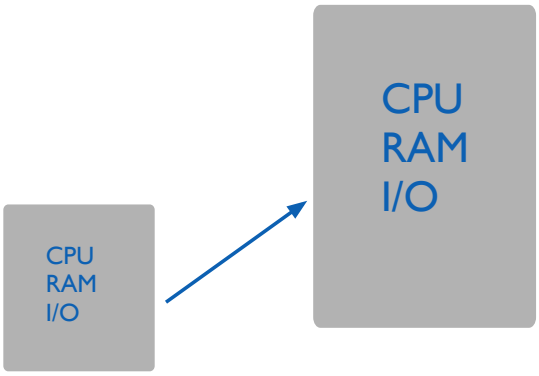
A MongoDB document expressed using JSON syntax.

```
{
  "_id" : "/apple-reports-second-quarter-revenue",
  "headline" : "Apple Reported Second Quarter Revenue Today",
  "date" : ISODate("2015-03-24T22:35:21.908Z"),
  "author" : {
    "name" : "Bob Walker",
    "title" : "Lead Business Editor"
  },
  "copy" : "Apple beat Wall St expectations by reporting ...",
  "tags" : [
    "AAPL", "Earnings", "Cupertino"
  ],
  "comments" : [
    { "name" : "Frank", "comment" : "Great Story" },
    { "name" : "Wendy", "comment" : "When can I buy an Apple Watch?" }
  ]
}
```

Note:

- How would you represent this document in a relational database? How many tables, how many queries per page load?
- What are the pros/cons to this design? (hint: 1 million comments)
- Where relational databases store rows, MongoDB stores documents.
- Documents are hierarchical data structures.
- This is a fundamental departure from relational databases where rows are flat.

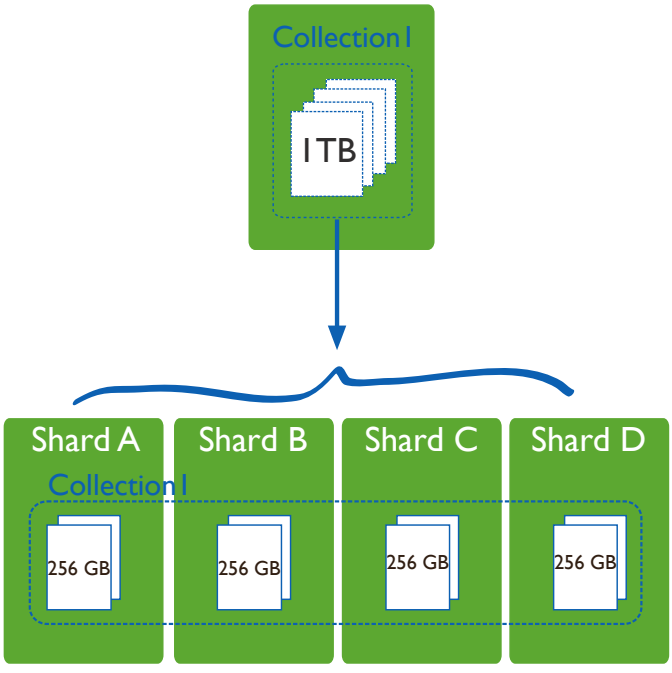
Vertical Scaling



Note: Another difference is in terms of scalability. With an RDBMS:

- If you need to support a larger workload, you buy a bigger machine.
- The problem is, machines are not priced linearly.
- The largest machines cost much more than commodity hardware.
- If your application is successful, you may find you simply cannot buy a large enough a machine to support your workload.

Scaling with MongoDB

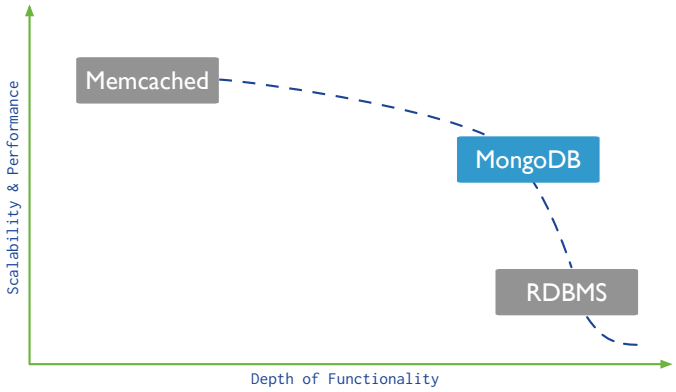


Note:

- MongoDB is designed to be horizontally scalable (linear).
- MongoDB scales by enabling you to shard your data.

- When you need more performance, you just buy another machine and add it to your cluster.
- MongoDB is highly performant on commodity hardware.

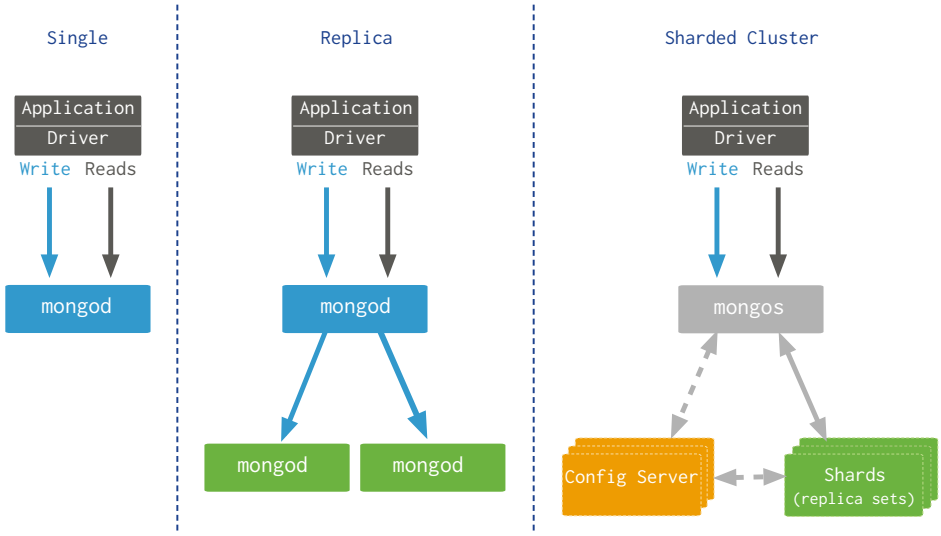
Database Landscape



Note:

- We've plotted each technology by scalability and functionality.
- At the top left, are key/value stores like memcached.
- These scale well, but lack features that make developers productive.
- At the far right we have traditional RDBMS technologies.
- These are full featured, but will not scale easily.
- Joins and transactions are difficult to run in parallel.
- MongoDB has nearly as much scalability as key-value stores.
- Gives up only the features that prevent scaling.
- We have compensating features that mitigate the impact of that design decision.

MongoDB Deployment Models



Note:

- MongoDB supports high availability through automated failover.
- Do not use a single-server deployment in production.
- Typical deployments use replica sets of 3 or more nodes.
  - The primary node will accept all writes, and possibly all reads.
  - Each secondary will replicate from another node.
  - If the primary fails, a secondary will automatically step up.
  - Replica sets provide redundancy and high availability.
- In production, you typically build a fully sharded cluster:
  - Your data is distributed across several shards.
  - The shards are themselves replica sets.
  - This provides high availability and redundancy at scale.

1.3 MongoDB Stores Documents

Learning Objectives

Upon completing this module, students should understand:

- JSON
- BSON basics
- That documents are organized into collections
- ObjectIds
- Padding Factor

JSON

- JavaScript Object Notation
- Objects are associative arrays.
- They are composed of key-value pairs.

A Simple JSON Object

```
{
    "firstname" : "Thomas",
    "lastname" : "Smith",
    "age" : 29
}
```

JSON Keys and Values

- Keys must be strings.
- Values may be any of the following:
  - string (e.g., “Thomas”)
  - number (e.g., 29, 3.7)
  - true / false
  - null
  - array (e.g., [88.5, 91.3, 67.1])
  - object
- More detail at json.org[1].

Example Field Values

```
{
  "headline" : "Apple Reported Second Quarter Revenue Today",
  "date" : ISODate("2015-03-24T22:35:21.908Z"),
  "views" : 1234,
  "author" : {
    "name" : "Bob Walker",
    "title" : "Lead Business Editor"
  },
  "tags" : [
    "AAPL",
    23,
    { "name" : "city", "value" : "Cupertino" },
    [ "Electronics", "Computers" ]
  ]
}
```

[1] http://json.org/

BSON

- MongoDB stores data as Binary JSON (BSON).
- MongoDB drivers send and receive data in this format.
- They map BSON to native data structures.
- BSON provides support for all JSON data types and several others.
- BSON was designed to be lightweight, traversable and efficient.
- See bsonspec.org[2].

Note: E.g., a BSON object will be mapped to a dictionary in Python.

BSON Hello World

```
// JSON
{ "hello" : "world" }

// BSON
"\x16\x00\x00\x00\x02hello\x00
 \x06\x00\x00\x00world\x00\x00"
```

Note:

- \x16\x00\x00\x00 (document size)
- \x02 = string (data type of field value)
- hello\x00 (key/field name, \x00 is null and delimits the end of the name)
- \x06\x00\x00\x00 (size of field value including end null)
- world\x00 (field value)
- \x00 (end of the document)

A More Complex BSON Example

```
// JSON
{ "BSON" : [ "awesome", 5.05, 1986 ] }

// BSON
"\x3b\x00\x00\x00\x04BSON\x00\x26\x00
 \x00\x00\x020\x00\x08\x00\x00
 \x00awesome\x00\x011\x00\x33\x33\x33\x33\x33\x33
 \x14\x40\x102\x00\xc2\x07\x00\x00
 \x00\x00"
```

[2] http://bsonspec.org/#/specification

Documents, Collections, and Databases

- Documents are stored in collections.
- Collections are contained in a database.
- Example:
  - Database: products
  - Collections: books, movies, music
- Each database-collection combination defines a namespace.
  - products.books
  - products.movies
  - products.music

The `_id` Field

- All documents must have an `_id` field.
- The `_id` is immutable.
- If no `_id` is specified when a document is inserted, MongoDB will add the `_id` field.
- MongoDB assigns a unique ObjectId as the value of `_id`.
- Most drivers will actually create the ObjectId if no `_id` is specified.
- The `_id` field is unique to a collection (namespace).

ObjectIds

Note:

- An ObjectId is a 12-byte value.
- The first 4 bytes are a datetime reflecting when the ObjectID was created.
- The next 3 bytes are the MAC address of the server.
- Then a 2-byte process ID
- Finally, 3 bytes that are monotonically increasing for each new ObjectId created within a collection.

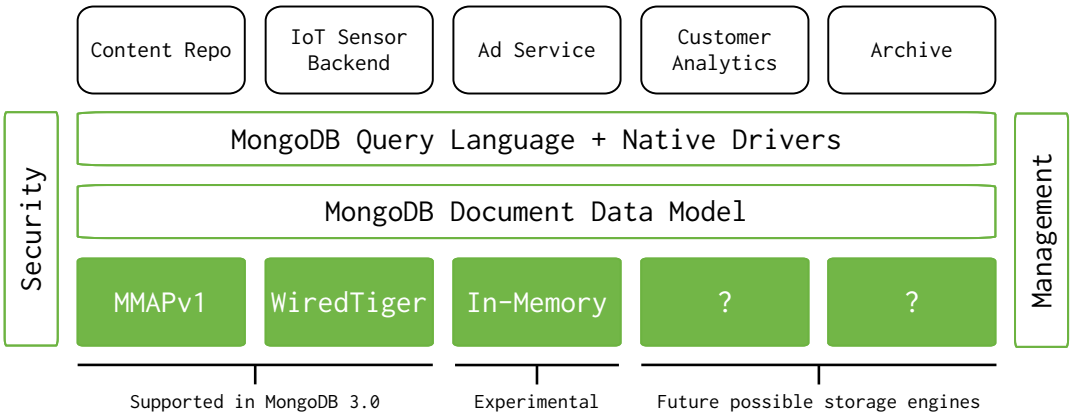
1.4 Storage Engines

Learning Objectives

Upon completing this module, students should be familiar with:

- Available storage engines in MongoDB
- MongoDB journaling mechanics
- The default storage engine for MongoDB
- Common storage engine parameters
- The storage engine API

What is a Database Storage Engine?



Note:

- A database storage engine is the underlying software component that a database management system uses to create, read, update, and delete data from a database.
- Talk through the diagram and how storage engines are used to abstract access to the data

---

How Storage Engines Affect Performance

- Writing and reading documents
- Concurrency
- Compression algorithms
- Index format and implementation
- On-disk format

Note: Can use an extreme example, such as the difference between an in-memory storage engine and mmap/wiredtiger for write performance

---

Storage Engine Journaling

- Keep track of all changes made to data files
- Stage writes sequentially before they can be committed to the data files
- Crash recovery, writes from journal can be replayed to data files in the event of a failure

MongoDB Storage Engines

With the release of MongoDB 3.0, two storage engine options are available:

- MMAPv1 (default)
- WiredTiger

Specifying a MongoDB Storage Engine

Use the `storageEngine` parameter to specify which storage engine MongoDB should use. E.g.,

```
mongod --storageEngine wiredTiger
```

Note:

- mmapv1 is used if storageEngine parameter isn't specified

Specifying a Location to Store Data Files

- Use the `dbpath` parameter

  ```
  mongod --dbpath /data/db
  ```

- Other files are also stored here. E.g.,
  - mongod.lock file
  - journal
- See the MongoDB docs for a complete list of storage options[3].

[3] http://docs.mongodb.org/manual/reference/program/mongod/#storage-options

MMAPv1 Storage Engine

- MMAPv1 is MongoDB's original storage engine and currently the default.

  ```
  mongod
  ```

- This is equivalent to the following command:

  ```
  mongod --storageEngine mmapv1
  ```

- MMAPv1 is based on memory-mapped files, which map data files on disk into virtual memory.
- As of MongoDB 3.0, MMAPv1 supports collection-level concurrency.

MMAPv1 Workloads

MMAPv1 excels at workloads where documents do not outgrow their original record size:

- High-volume inserts
- Read-only workloads
- In-place updates

Note:

- None of the use cases above grow the documents (and potentially force them to move), one flaw with mmapv1

Power of 2 Sizes Allocation Strategy

- MongoDB 3.0 uses power of 2 sizes allocation as the default record allocation strategy for MMAPv1.
- With this strategy, records include the document plus extra space, or padding.
- Each record has a size in bytes that is a power of 2 (e.g. 32, 64, 128, ... 2MB).
- For documents larger than 2MB, allocation is rounded up to the nearest multiple of 2MB.
- This strategy enables MongoDB to efficiently reuse freed records to reduce fragmentation.
- In addition, the added padding gives a document room to grow without requiring a move.
  - Saves the cost of moving a document
  - Results in fewer updates to indexes

Compression in MongoDB

- Compression can significantly reduce the amount of disk space / memory required.
- The tradeoff is that compression requires more CPU.
- MMAPv1 does not support compression.
- WiredTiger does.

WiredTiger Storage Engine

- The WiredTiger storage engine excels at all workloads, especially write-heavy and update-heavy workloads.
- Notable features of the WiredTiger storage engine that do not exist in the MMAPv1 storage engine include:
  - Compression
  - Document-level concurrency
- Specify the use of the WiredTiger storage engine as follows.

```
mongod --storageEngine wiredTiger
```

WiredTiger Compression Options

- `snappy` (default): less CPU usage than `zlib`, less reduction in data size
- `zlib`: greater CPU usage than `snappy`, greater reduction in data size
- no compression

Configuring Compression in WiredTiger

Use the `wiredTigerCollectionBlockCompressor` parameter. E.g.,

```
mongod --storageEngine wiredTiger
       --wiredTigerCollectionBlockCompressor zlib
```

Configuring Memory Usage in WiredTiger

Use the `wiredTigerCacheSize` parameter to designate the amount of RAM for the WiredTiger storage engine.

- By default, this value is set to the maximum of half of physical RAM or 1GB
- If the database server shares a machine with an application server, it is now easier to designate the amount of RAM the database server can use

---

Note:

- Unlike MMAPv1, WiredTiger can be configured to use a finite amount of RAM.

Journaling in MMAPv1 vs. WiredTiger

- MMAPv1 uses write-ahead journaling to ensure consistency.
- WiredTiger uses a write-ahead transaction log in combination with checkpoints to ensure durability.
- With WiredTiger, the replication process may provide sufficient durability guarantees.

MMMAPv1 Journaling Mechanics

- Journal files in <DATA-DIR>/journal are append only
- 1GB per journal file
- Once MongoDB applies all write operations from a journal file to the database data files, it deletes the journal file (or re-uses it)
- Usually only a few journal files in the <DATA-DIR>/journal directory

MMAPv1 Journaling Mechanics (Continued)

- Data is flushed from the shared view to data files every 60 seconds (configurable)
- The operating system may force a flush at a higher frequency than 60 seconds if the system is low on free memory
- Once a journal file contains only flushed writes, it is no longer needed for recovery and can be deleted or re-used

WiredTiger Journaling Mechanics

- WiredTiger will commit a checkpoint to disk every 60 seconds or when there are 2 gigabytes of data to write.
- Between and during checkpoints the data files are always valid.
- The WiredTiger journal persists all data modifications between checkpoints.
- If MongoDB exits between checkpoints, it uses the journal to replay all data modified since the last checkpoint.
- By default, WiredTiger journal is compressed using snappy.

Storage Engine API

MongoDB 3.0 introduced a storage engine API:

- Abstracted storage engine functionality in the code base
- Easier for MongoDB to develop future storage engines
- Easier for third parties to develop their own MongoDB storage engines

Conclusion

- MongoDB 3.0 introduces pluggable storage engines.
- Current options include:
  - MMAPv1 (default)
  - WiredTiger
- WiredTiger introduces the following to MongoDB:
  - Compression
  - Document-level concurrency
- The storage engine API enables third parties to develop storage engines. Examples include:
  - RocksDB
  - An HDFS storage engine

Note:

- Good time to draw what this replica set could look like on the board and talk through even more possibilities

---

1.5 Lab: Installing and Configuring MongoDB

Learning Objectives

Upon completing this exercise students should understand:

- How MongoDB is distributed
- How to install MongoDB
- Configuration steps for setting up a simple MongoDB deployment
- How to run MongoDB
- How to run the Mongo shell

Production Releases

64-bit production releases of MongoDB are available for the following platforms.

- Windows
- OSX
- Linux
- Solaris

Installing MongoDB

- Visit http://docs.mongodb.org/manual/installation/.
- Click on the appropriate link, such as “Install on Windows” or “Install on OS X” and follow the instructions.
- Versions:
  - Even-numbered builds are production releases, e.g., 2.4.x, 2.6.x.
  - Odd-numbers indicate development releases, e.g., 2.5.x, 2.7.x.

Linux Setup

```
PATH=$PATH:<path to mongodb>/bin

sudo mkdir -p /data/db

sudo chmod -R 777 /data/db
```

Note:

- You might want to add the MongoDB bin directory to your path, e.g.
- Once installed, create the MongoDB data directory.
- Make sure you have write permission on this directory.

Install on Windows

- Download and run the .msi Windows installer from mongodb.org/downloads.
- By default, binaries will be placed in the following directory.

  ```
  C:\Program Files\MongoDB\Server\<VERSION>\bin
  ```

- It is helpful to add the location of the MongoDB binaries to your path.
- To do this, from “System Properties” select “Advanced” then “Environment Variables”

Note: Can also install Windows as a service, but not recommended since we need multiple mongod processes for future exercises

---

Create a Data Directory on Windows

- Ensure there is a directory for your MongoDB data files.
- The default location is `\data\db`.
- Create a data directory with a command such as the following.

```
md \data\db
```

Note: Optionally, talk about the –dbpath variable and specifying a different location for the data files

---

Launch a `mongod`

Explore the `mongod` command.

```
<path to mongodb>/bin/mongod --help
```

Launch a `mongod` with the MMAPv1 storage engine:

```
<path to mongodb>/bin/mongod
```

Alternatively, launch with the WiredTiger storage engine.

```
<path to mongodb>/bin/mongod --storageEngine wiredTiger
```

Specify an alternate path for data files using the `--dbpath` option. (Make sure the directory already exists.) E.g.,

```
<path to mongodb>/bin/mongod --storageEngine wiredTiger
                             --dbpath /test/mongodb/data/wt
```

Note:

- Please verify that all students have successfully installed MongoDB.
- Please verify that all can successfully launch a mongod.

The MMAPv1 Data Directory

```
ls /data/db
```

- The mongod.lock file
  - This prevents multiple mongods from using the same data directory simultaneously.
  - Each MongoDB database directory has one .lock.
  - The lock file contains the process id of the mongod that is using the directory.
- Data files
  - The names of the files correspond to available databases.
  - A single database may have multiple files.

Note: Files for a single database increase in size as follows:

- sample.0 is 64 MB
- sample.1 is 128 MB

- sample.2 is 256 MB, etc.
- This continues until sample.5, which is 2 GB
- All subsequent data files are also 2 GB.

The WiredTiger Data Directory

```
ls /data/db
```

- The mongod.lock file
  - Used in the same way as MMAPv1.
- Data files
  - Each collection and index stored in its own file.
  - Will fail to start if MMAPv1 files found

Import Exercise Data

```
cd usb_drive

unzip sampledata.zip

cd sampledata

mongoimport -d sample -c tweets twitter.json

mongoimport -d sample -c zips zips.json

cd dump

mongorestore -d sample training

mongorestore -d sample digg
```

Note: If there is an error importing data directly from a USB drive, please copy the sampledata.zip file to your local computer first.

Note:

- Import the data provided on the USB drive into the *sample* database.

Launch a Mongo Shell

Open another command shell. Then type the following to start the Mongo shell.

```
mongo
```

Display available commands.

```
help
```

Explore Databases

Display available databases.

```
show dbs
```

To use a particular database we can type the following.

```
use <database_name>
```

```
db
```

Note:

- This assigns the variable `db` to a connection object for the selected database.
- We can display the name of the database we are currently using by evaluating `db` in the mongo shell.
- Highlight the power of the Mongo shell here.
- It is a fully programmable JavaScript environment.

Exploring Collections

```
show collections
```

```
db.<COLLECTION>.help()
```

```
db.<COLLECTION>.find()
```

Note:

- Show the collections available in this database.
- Show methods on the collection with parameters and a brief explanation.
- Finally, we can query for the documents in a collection.

Admin Commands

- There are also a number of admin commands at our disposal.
- The following will shut down the mongod we are connected to through the Mongo shell.
- You can also just kill with Ctrl-C in the shell window from which you launched the mongod.

```
db.adminCommand( { shutdown : 1 } )
```

- Confirm that the mongod process has indeed stopped.
- Once you have, please restart it.

2 CRUD

Creating and Deleting Documents **(page 23)** Inserting documents into collections, deleting documents, and dropping collections

Reading Documents **(page 29)** The find() command, query documents, dot notation, and cursors

Query Operators **(page 37)** MongoDB query operators including: comparison, logical, element, and array operators

Lab: Finding Documents **(page 41)** Exercises for querying documents in MongoDB

Updating Documents **(page 42)** Using `update()` and associated operators to mutate existing documents

Lab: Updating Documents **(page 49)** Exercises for updating documents in MongoDB

2.1 Creating and Deleting Documents

Learning Objectives

Upon completing this module students should understand:

- How to insert documents into MongoDB collections.
- `_id` fields:
- How to remove documents from a collection
- How to remove a collection from a database
- How to remove a database from a MongoDB deployment

Creating New Documents

- Create documents using `insert()`.
- For example:

```
// Specify the collection name
db.<COLLECTION>.insert( { "name" : "Mongo" } )

// For example
db.people.insert( { "name" : "Mongo" } )
```

Example: Inserting a Document

Experiment with the following commands.

```
use sample

db.movies.insert( { "title" : "Jaws" } )

db.movies.find()
```

Note:

- Make sure the students are performing the operations along with you.
- Some students will have trouble starting things up, so be helpful at this stage.

Implicit `_id` Assignment

- We did not specify an `_id` in the document we inserted.
- If you do not assign one, MongoDB will create one automatically.
- The value will be of type ObjectId.

Example: Assigning _ids

Experiment with the following commands.

```
db.movies.insert( { "_id" : "Jaws", "year" : 1975 } )

db.movies.find()
```

Note:

- Note that you can assign an _id to be of almost any type.
- It does not need to be an ObjectId.

Inserts will fail if...

- There is already a document in the collection with that `_id`.
- You try to assign an array to the `_id`.
- The argument is not a well-formed document.

Example: Inserts will fail if...

```
// fails because _id can't have an array value
db.movies.insert( { "_id" : [ "Star Wars",
                              "The Empire Strikes Back",
                              "Return of the Jedi" ] } )

// succeeds
db.movies.insert( { "_id" : "Star Wars" } )

// fails because of duplicate id
db.movies.insert( { "_id" : "Star Wars" } )

// malformed document
db.movies.insert( { "Star Wars" } )
```

Note:

- The following will fail because it attempts to use an array as an `_id`.

  ```
  db.movies.insert( { "_id" : [ "Star Wars", "The Empire Strikes Back", "Return of the Jedi" ] } )
  ```

- The second insert with `_id : "Star Wars"` will fail because there is already a document with `_id` of "Star Wars" in the collection.
- The following will fail because it is a malformed document (i.e. no field name, just a value).

```
db.movies.insert( { "Star Wars" } )
```

Bulk Inserts

- MongoDB 2.6 introduced bulk inserts.
- You may bulk insert using an array of documents.
- The API has two core concepts:
  - Ordered bulk operations
  - Unordered bulk operations
- The main difference is in the way the operations are executed in bulk.

Note:

- In the case of an ordered bulk operation, every operation will be executed in the order they are added to the bulk operation.
- In the case of an unordered bulk operation however there is no guarantee what order the operations are executed.
- With an unordered bulk operation, the operations in the list may be reordered to increase performance.

Ordered Bulk Insert

- For ordered inserts MongoDB will stop processing inserts upon encountering an error.
- Meaning that only inserts occurring before an error will complete.
- The default setting for `db.<COLLECTION>.insert` is an ordered insert.
- See the next exercise for an example.

Example: Ordered Bulk Insert

Experiment with the following bulk insert.

```
db.movies.insert( [ { "_id" : "Batman", "year" : 1989 },
                    { "_id" : "Home Alone", "year" : 1990 },
                    { "_id" : "Ghostbusters", "year" : 1984 },
                    { "_id" : "Ghostbusters", "year" : 1984 } ] )
db.movies.find()
```

Note:

- This example has a duplicate key error.
- Only the first 3 documents will be inserted.

Unordered Bulk Insert

- Pass `{ ordered : false }` to insert to perform unordered inserts.
- If any given insert fails, MongoDB will still attempt the others.
- The inserts may be executed in a different order from the way in which you specified them.
- The next exercise is very similar to the previous one.
- However, we are using `{ ordered : false }`
- One insert will fail, but all the rest will succeed.

Example: Unordered Bulk Insert

Experiment with the following bulk insert.

```
db.movies.insert( [ { "_id" : "Jaws", "year" : 1975 },
                    { "_id" : "Titanic", "year" : 1997 },
                    { "_id" : "The Lion King", "year" : 1994 } ],
                    { ordered : false } )
db.movies.find()
```

The Shell is a JavaScript Interpreter

- Sometimes it is convenient to create test data using a little JavaScript.
- The mongo shell is a fully-functional JavaScript interpreter. You may:
  - Define functions
  - Use loops
  - Assign variables
  - Perform inserts

Exercise: Creating Data in the Shell

Experiment with the following commands.

```
for (i=1; i<=10000; i++) {
    db.stuff.insert( { "a" : i } )
}

db.stuff.find()
```

Deleting Documents

You may delete documents from a MongoDB deployment in several ways.

- Use `remove()` to delete documents matching a specific set of conditions.
- Drop an entire collection.
- Drop a database.

Using `remove()`

- Remove documents from a collection using `remove()`.
- This command has one required parameter, a query document.
- All documents in the collection matching the query document will be removed.
- Pass an empty document to remove all documents.
- Prior to MongoDB 2.6 calling `remove()` with no parameters would remove all documents.
- Limit `remove()` to one document using `justOne`.

Example: Removing Documents

Experiment with removing documents. Do a `find()` after each `remove()` command below.

```
for (i=1; i<=20; i++) { db.testcol.insert( { _id : i, a : i } ) }

db.testcol.remove( { a : 1 } )  // Remove the first document

// $lt is a query operator that enables us to select documents that
// are less than some value. More on operators soon.
db.testcol.remove( { a : { $lt : 5 } } )  // Remove three more

db.testcol.remove( { a : { $lt : 10 } },
                   { justOne : true } )  // Remove one more

db.testcol.remove()  // Error: requires a query document.

db.testcol.remove( { } )  // All documents removed
```

Dropping a Collection

- You can drop an entire collection with `db.<COLLECTION>.drop()`
- The collection and all documents will be deleted.
- It will also remove any metadata associated with that collection.
- Indexes are one type of metadata removed.
- More on meta data later.

Note: Mention that `drop()` is more performant than remove because of the lookup costs associated with `remove()`.

Example: Dropping a Collection

```
db.colToBeDropped.insert( { a : 1 } )
show collections  // Shows the colToBeDropped collection

db.colToBeDropped.drop()
show collections  // collection is gone
```

Dropping a Database

- You can drop an entire database with `db.dropDatabase()`
- This drops the database on which the method is called.
- It also deletes the associated data files from disk, freeing disk space.
- Beware that in the mongo shell, this does not change database context.

Example: Dropping a Database

```
use tempDB
db.testcol1.insert( { a : 1 } )
db.testcol2.insert( { a : 1 } )

show dbs  // Here they are
show collections  // Shows the two collections

db.dropDatabase()
show collections  // No collections
show dbs  // The db is gone

use sample  // take us back to the sample db
```

2.2 Reading Documents

Learning Objectives

Upon completing this module students should understand:

- The query-by-example paradigm of MongoDB
- How to query on array elements
- How to query embedded documents using dot notation
- How the mongo shell and drivers use cursors
- Projections
- Cursor methods: `.count()`, `.sort()`, `.skip()`, `.limit()`

The `find()` Method

- This is the fundamental method by which we read data from MongoDB.
- We have already used it in its basic form.
- `find()` returns a cursor that enables us to iterate through all documents matching a query.
- We will discuss cursors later.

Query by Example

- To query MongoDB, specify a document containing the key / value pairs you want to match
- You need only specify values for fields you care about.
- Other fields will not be used to exclude documents.
- The result set will include all documents in a collection that match.

Example: Querying by Example

Experiment with the following sequence of commands.

```
db.movies.drop()
db.movies.insert([ { "title" : "Jaws", "year" : 1975, "imdb_rating" : 8.1 },
                   { "title" : "Batman", "year" : 1989, "imdb_rating" : 7.6 },
                 ] )
db.movies.find()

db.movies.find( { "year" : 1975 } )

// Multiple Batman movies from different years, find the correct one
db.movies.find( { "year" : 1989, "title" : "Batman" } )
```

Note: Matching Rules:

- Any field specified in the query must be in each document returned.
- Values for returned documents must match the conditions specified in the query document.
- If multiple fields are specified, all must be present in each document returned.
- Think of it as a logical AND for all fields.

Querying Arrays

- In MongoDB you may query array fields.
- Specify a single value you expect to find in that array in desired documents.
- Alternatively, you may specify an entire array in the query document.
- As we will see later, there are also several operators that enhance our ability to query array fields.

Note: Students might find it helpful to think of an array field as having multiple values – one for each of its elements.

Example: Querying Arrays

```
db.movies.drop()
db.movies.insert(
 [{ "title" : "Batman", "category" : [ "action", "adventure" ] },
  { "title" : "Godzilla", "category" : [ "action", "adventure", "sci-fi" ] },
  { "title" : "Home Alone", "category" : [ "family", "comedy" ] }
 ])

// Match documents where "category" contains the value specified
db.movies.find( { "category" : "action" } )

// Match documents where "category" equals the value specified
db.movies.find( { "category" : [ "action", "sci-fi" ] } )  // no documents

// only the second document
db.movies.find( { "category" : [ "action", "adventure", "sci-fi" ] } )
```

Note: Later, we'll see operators that will allow us to do things like match all documents where an array field contains any of a set of values.

Querying with Dot Notation

- Dot notation is used to query on fields in embedded documents.
- The syntax is:

  ```
  "field1.field2" : value
  ```

- Put quotes around the field name when using dot notation.

Example: Querying with Dot Notation

```
db.movies.insert(
    [ {
          "title" : "Avatar",
          "box_office" : { "gross" : 760,
                           "budget" : 237,
                           "opening_weekend" : 77
                         }
      },
      {
          "title" : "E.T.",
          "box_office" : { "gross" : 349,
                           "budget" : 10.5,
                           "opening_weekend" : 14
                         }
       }
     ] )

db.movies.find( { "box_office" : { "gross" : 760 } } ) // no values

// dot notation
db.movies.find( { "box_office.gross" : 760 } )  // expected value
```

Example: Arrays and Dot Notation

```
db.movies.insert( [
    { "title" : "E.T.",
      "filming_locations" :
         [ { "city" : "Culver City", "state" : "CA", "country" : "USA" },
           { "city" : "Los Angeles", "state" : "CA", "country" : "USA" },
           { "city" : "Cresecent City", "state" : "CA", "country" : "USA" }
         ] },
    { "title": "Star Wars",
      "filming_locations" :
         [ { "city" : "Ajim", "state" : "Jerba", "country" : "Tunisia" },
           { "city" : "Yuma", "state" : "AZ", "country" : "USA" }
         ] } ] )

 db.movies.find( { "filming_locations.country" : "USA" } ) // two documents
```

Note:

- This query finds documents where:
  - There is a `filming_locations` field.
  - The `filming_locations` field contains one or more embedded documents.
  - At least one embedded document has a field `country`.
  - The field `country` has the specified value (“USA”).
- In this collection, `filming_locations` is actually an array field.
- The embedded documents we are matching are held within these arrays.

Projections

- You may choose to have only certain fields appear in result documents.
- This is called projection.
- You specify a projection by passing a second parameter to `find()`.

Projection: Example (Setup)

```
db.movies.insert(
{
   "title" : "Forrest Gump",
   "category" : [ "drama", "romance" ],
   "imdb_rating" : 8.8,
   "filming_locations" : [
     { "city" : "Savannah", "state" : "GA", "country" : "USA" },
     { "city" : "Monument Valley", "state" : "UT", "country" : "USA" },
     { "city" : "Los Anegeles", "state" : "CA", "country" : "USA" }
   ],
   "box_office" : {
     "gross" : 557,
     "opening_weekend" : 24,
     "budget" : 55
   }
 })
```

Projection: Example

```
db.movies.findOne( { "title" : "Forrest Gump" },
                   { "title" : 1, "imdb_rating" : 1 } )
{
   "_id" : ObjectId("5515942d31117f52a5122353"),
   "title" : "Forrest Gump",
   "imdb_rating" : 8.8
}
```

Projection Documents

- Include fields with `fieldName: 1`.
  - Any field not named will be excluded
  - except _id, which must be explicitly excluded.
- Exclude fields with `fieldName: 0`.
  - Any field not named will be included.

Example: Projections

```
for (i=1; i<=20; i++) {
    db.movies.insert( { "_id" : i, "title" : i,
                        "imdb_rating" : i, "box_office" : i } )
}
db.movies.find()
// no "box_office"
db.movies.find( { "_id" : 3 }, { "title" : 1, "imdb_rating" : 1 } )
// no "imdb_rating"
db.movies.find( { "_id" : { $gte : 10 } }, { "imdb_rating" : 0 } )
// just "title"
db.movies.find( { "_id" : 4 }, { "_id" : 0, "title" : 1 } )
// just "imdb_rating", "box_office"
db.movies.find( { "_id" : 5 }, { _id : 0, "title" : 0 } )
// Can't mix inclusion/exclusion except _id
db.movies.find( { "_id" : 6 }, { "title" : 1, "imdb_rating" : 0 } )
```

The last `find()` fails because MongoDB cannot determine how to handle unnamed fields such as `c`.

Cursors

- When you use `find()`, MongoDB returns a cursor.
- A cursor is a pointer to the result set
- You can get iterate through documents in the result using `next()`.
- By default, the mongo shell will iterate through 20 documents at a time.

Example: Introducing Cursors

```
db.testcol.drop()
for (i=1; i<=10000; i++) {
    db.testcol.insert( { a : Math.floor( Math.random() * 100 + 1 ),
                         b : Math.floor( Math.random() * 100 + 1 ) } )
}
db.testcol.find()

it
it
```

Note:

- With the `find()` above, the shell iterates over the first 20 documents.
- `it` causes the shell to iterate over the next 20 documents.
- Can continue issuing `it` commands until all documents are seen.

Example: Cursor Objects in the Mongo Shell

```
// Assigns the cursor returned by find() to a variable x
var x = db.testcol.find()

// Displays the first document in the result set.
x.next()

// True because there are more documents in the result set.
x.hasNext()

// Assigns the next document in the result set to the variable y.
y = x.next()

// Return value is the value of the a field of this document.
y.a

// Displaying a cursor prints the next 20 documents in the result set.
x
```

Cursor Methods

- `count()`: Returns the number of documents in the result set.
- `limit()`: Limits the result set to the number of documents specified.
- `skip()`: Skips the number of documents specified.

Example: Using `count()`

```
db.testcol.drop()
for (i=1; i<=100; i++) { db.testcol.insert( { a : i } ) }

// all 100
db.testcol.count()

// just 41 docs
db.testcol.count( { a : { $lt : 42 } } )

// Another way of writing the same query
db.testcol.find( { a : { $lt : 42 } } ).count(  )
```

Note:

- You may pass a query document like you would to `find()`.
- `count()` will count only the documents matching the query.
- Will return the number of documents in the collection if you do not specify a query document.
- The last query in the above achieves the same result because it operates on the cursor returned by `find()`.

Example: Using `sort()`

```
db.testcol.drop()
for (i=1; i<=20; i++) {
    db.testcol.insert( { a : Math.floor( Math.random() * 10 + 1 ),
                         b : Math.floor( Math.random() * 10 + 1 ) } )
}

db.testcol.find()

// sort descending; use 1 for ascending
db.testcol.find().sort( { a : -1 } )

// sort by b, then a
db.testcol.find().sort( { b : 1, a : 1 } )

// $natural order is just the order on disk.
db.testcol.find().sort( { $natural : 1 } )
```

Note:

- Sort can be executed on a cursor until the point where the first document is actually read.
- If you never delete any documents or change their size, this will be the same order in which you inserted them.
- Sorting two or more fields breaks the convention of javascript objects that key / value pairs are unordered.
- In some drivers you may need to take special care with this.
- For example, in Python, you would usually query with a dictionary.
- But dictionaries are unordered in Python, so you would use an array of tuples instead.

The `skip()` Method

- Skips the specified number of documents in the result set.
- The returned cursor will begin at the first document beyond the number specified.
- Regardless of the order in which you specify `skip()` and `sort()` on a cursor, `sort()` happens first.

The `limit()` Method

- Limits the number of documents in a result set to the first `k`.
- Specify `k` as the argument to `limit()`
- Regardless of the order in which you specify `limit()`, `skip()`, and `sort()` on a cursor, `sort()` happens first.
- Helps reduce resources consumed by queries.

The `distinct()` Method

- Returns all values for a field found in a collection.
- Only works on one field at a time.
- Input is a string (not a document)

Example: Using `distinct()`

```
db.movie_reviews.drop()
db.movie_reviews.insert( [ { "title" : "Jaws", "rating" : 5 },
                           { "title" : "Home Alone", "rating" : 1 },
                           { "title" : "Jaws", "rating" : 7 },
                           { "title" : "Jaws", "rating" : 4 },
                           { "title" : "Jaws", "rating" : 8 } ] )
db.movie_reviews.distinct( "title" )
```

Note: Returns

```
{
    "values" : [ "Jaws", "Home Alone" ],
    "stats" : { ... },
    "ok" : 1
}
```

2.3 Query Operators

Learning Objectives

Upon completing this module students should understand the following types of MongoDB query operators:

- Comparison operators
- Logical operators
- Element query operators
- Operators on arrays

Comparison Query Operators

- `$lt`: Exists and is less than
- `$lte`: Exists and is less than or equal to
- `$gt`: Exists and is greater than
- `$gte`: Exists and is greater than or equal to
- `$ne`: Does not exist or does but is not equal to
- `$in`: Exists and is in a set
- `$nin`: Does not exist or is not in a set

Example: Comparison Operators (Setup)

```
// insert sample data
db.movies.insert( [
  {
    "title" : "Batman",
    "category" : [ "action", "adventure" ],
    "imdb_rating" : 7.6,
    "budget" : 35
  },
  {
    "title" : "Godzilla",
    "category" : [ "action",
    "adventure", "sci-fi" ],
    "imdb_rating" : 6.6
  },
  {
    "title" : "Home Alone",
    "category" : [ "family", "comedy" ],
    "imdb_rating" : 7.4
  }
] )
```

Example: Comparison Operators

```
db.movies.find()

db.movies.find( { "imdb_rating" : { $gte : 7 } } )

db.movies.find( { "category" : { $ne : "family" } } )

db.movies.find( { "title" : { $in : [ "Batman", "Godzilla" ] } } )

db.movies.find( { "title" : { $nin : [ "Batman", "Godzilla" ] } } )
```

Logical Query Operators

- `$or`: Match either of two or more values
- `$not`: Used with other operators
- `$nor`: Match neither of two or more values
- `$and`: Match both of two or more values
  - This is the default behavior for queries specifying more than one condition.
  - Use `$and` if you need to include the same operator more than once in a query.

Example: Logical Operators

```
db.movies.find( { $or : [
   { "category" : "sci-fi" }, { "imdb_rating" : { $gte : 7 } }
 ] } )

// more complex $or, really good sci-fi movie or medicore family movie
db.movies.find( { $or : [
   { "category" : "sci-fi", "imdb_rating" : { $gte : 8 } },
   { "category" : "family", "imdb_rating" : { $gte : 7 } }
 ] } )

// find bad movies
db.movies.find( { "imdb_rating" : { $not : { $gt : 7 } } } )
```

Note:

- `db.movies.find( { "imdb_rating" : { $not : { $gt : 7 } } } )` also returns everything that doesnt have an "imdb_rating"

Example: Logical Operators

```
// find movies within an imdb_rating range
db.movies.find( { "imdb_rating" : { $gt : 5 , $lte : 7 } } )  // and is implicit

// queries can be nested, why are there no results?
db.movies.find( { $and : [
   { $or : [
     { "category" : "sci-fi", "imdb_rating" : { $gte : 8 } },
     { "category" : "family", "imdb_rating" : { $gte : 7 } }
   ] } ,
   { $or : [
     { "category" : "action", "imdb_rating" : { $gte : 6 } }
   ] }
 ] } )
```

Element Query Operators

- `$exists`: Select documents based on the existence of a particular field.
- `$type`: Select documents based on their type.
- See BSON types[4] for reference on types.

[4] http://docs.mongodb.org/manual/reference/bson-types

Example: Element Operators

```
db.movies.find( { "budget" : { $exists : true } } )

// type 1 is Double
db.movies.find( { "budget" : { $type : 1 } } )

// type 3 is Object (embedded document)
db.movies.find( { "budget" : { $type : 3 } } )
```

Array Query Operators

- `$all`: Array field must contain all values listed.
- `$size`: Array must have a particular size. E.g., `$size : 2` means 2 elements in the array
- `$elemMatch`: All conditions must be matched by at least one element in the array

Example: Array Operators

```
db.movies.find( { "category" : { $all : [ "sci-fi", "action" ] } } )

db.movies.find( { "category" : { $size : 3 } } )
```

Example: $elemMatch

```
db.movies.insert( {
    "title" : "Raiders of the Lost Ark",
    "filming_locations" : [
      { "city" : "Los Angeles", "state" : "CA", "country" : "USA" },
      { "city" : "Rome", "state" : "Lazio", "country" : "Italy" },
      { "city" : "Florence", "state" : "SC", "country" : "USA" }
    ] } )

// This query is incorrect, it won't return what we want
db.movies.find( {
    "filming_locations.city" : "Florence",
    "filming_locations.country" : "Italy"
  } )

// $elemMatch is needed, now there are no results, this is expected
db.movies.find( {
    "filming_locations" : {
      $elemMatch : {
        "city" : "Florence",
        "country" : "Italy"
       } } } )
```

Note:

- Comparing the last two queries demonstrates `$elemMatch`.

2.4 Lab: Finding Documents

Exercise: Scores < 65

In the sample database, how many documents in the scores collection have a score less than 65?

Note:

- 832
- db.scores.find({ score: { $lt: 65 } })

Exercise: Exams and Quizzes

In the sample database, how many documents in the scores collection have the type “exam” or “quiz”?

Note:

- 2000
- db.scores.find({ type: { $in: [“exam”, “quiz”] } }).count()

Exercise: Highest Quiz Score

Find the highest quiz score.

Note:

- You can .limit() with .sort()
- db.scores.find({ type: “quiz” }).sort({ score: -1 }).limit(1)[0].score

---

Exercise: View Count > 1000

In the digg.stories collection, write a query to find all stories where the view count is greater than 1000.

Note:

- Requires querying into subdocuments
- db.stories.find({ “shorturl.view_count”: { $gt: 1000 } })

Exercise: Television or Videos

Find all digg stories where the topic name is "Television" or the media type is "videos". Skip the first 5 results and limit the result set to 10.

Note:

```
db.stories.find( { "$or": [ { "topic.name": "Television" },
                            { media: "news" } ] } ).skip(5).limit(10)
```

Exercise: News or Images

Query for all digg stories whose media type is either "news" or "images" and where the topic name is "Comedy". (For extra practice, construct two queries using different sets of operators to do this.)

Note:

```
db.stories.find( { media: { $in: [ "news", "images" ] },
                   "topic.name": "Comedy" })
```

2.5 Updating Documents

Learning Objectives

Upon completing this module students should understand

- The `update()` method
- The required parameters for `update()`
- Field update operators
- Array update operators
- The concept of an upsert and use cases.
- The findAndModify method

The `update()` Method

- Mutate documents in MongoDB using `update()`.
- `update()` requires two parameters:
  - A query document used to select documents to be updated
  - An update document that specifies how selected documents will change
- `update()` cannot delete a document.

Parameters to `update()`

- Keep the following in mind regarding the required parameters for `update()`
- The query parameter:
  - Use the same syntax as with `find()`.
  - By default only the first document found is updated.
- The update parameter:
  - Take care to simply modify documents if that is what you intend.
  - Replacing documents in their entirety is easy to do by mistake.

`$set` and `$unset`

- Update one or more fields using the `$set` operator.
- If the field already exists, using `$set` will change its value.
- If the field does not exist, `$set` will create it and set it to the new value.
- Any fields you do not specify will not be modified.
- You can remove a field using `$unset`.

Example: `$set` and `$unset` (Setup)

```
db.movies.insert( [
  {
    "title" : "Batman",
    "category" : [ "action", "adventure" ],
    "imdb_rating" : 7.6,
    "budget" : 35
  },
  {
    "title" : "Godzilla",
    "category" : [ "action",
    "adventure", "sci-fi" ],
    "imdb_rating" : 6.6
  },
  {
    "title" : "Home Alone",
    "category" : [ "family", "comedy" ],
    "imdb_rating" : 7.4
  }
] )
```

Example: `$set` and `$unset`

```
db.movies.update( { "title" : "Batman" }, { $set : { "imdb_rating" : 7.7 } } )

db.movies.update( { "title" : "Godzilla" }, { $set : { "budget" : 1 } } )

db.movies.update( { "title" : "Home Alone" },
                  { $set : { "budget" : 15, "imdb_rating" : 5.5 } } )

// how will this query behave?
db.movies.update( { "title" : "Batman" }, { "imdb_rating" : 7.7 } )

db.movies.update( { "title" : "Home Alone" }, { $unset :  { "budget" : 1 } } )
```

Note:

- Update will only update the first document it finds by default
- Difference between using $set and not using $set is very important

---

Example: Update Array Elements by Index

```
// add a sample document to track mentions per hour
db.movie_mentions.insert( { "title" : "E.T.",
                            "day" : ISODate("2015-03-27T00:00:00.000Z"),
                            "mentions_per_hour" : [ 0, 0, 0, 0, 0, 0, 0,
                              0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                              0, 0 ]
                          } )

// update all mentions for the fifth hour of the day
db.movie_mentions.update( { "title" : "E.T." } ,
                          { $set :  { "mentions_per_hour.5" : 2300 } } )
```

Note:

- Cool pattern for time series data
- Displaying charts is now trivial, can change granularity to by the minute, hour, day, etc.

Update Operators

- `$inc`: Increment a field's value by the specified amount.
- `$mul`: Multiply a field's value by the specified amount.
- `$rename`: Rename a field.
- `$set`: Update one or more fields.
- `$unset` Delete a field.
- `$min`: Update only if value is smaller than specified quantity
- `$max`: Update only if value is larger than specified quantity
- `$currentDate`: Set the value of a field to the current date or timestamp.

Example: Update Operators

```
db.movies.update( { "title" : "Batman" }, { $inc : { "imdb_rating" : 2 } } )

db.movies.update( { "title" : "Home Alone" }, { $inc : { "budget" : 5 } } )

db.movies.update( { "title" : "Batman" }, { $mul : { "imdb_rating" : 4 } } )

db.movies.update( { "title" : "Batman" },
                  { $rename : { "budget" : "estimated_budget" } } )

db.movies.update( { "title" : "Home Alone" }, { $min : { "budget" : 5 } } )

db.movies.update( { "title" : "Home Alone" },
                  { $currentDate : { "last_updated" : { $type : "timestamp" } } } )

// increment movie mentions by 10
db.movie_mentions.update( { "title" : "E.T." } ,
                          { $inc :  { "mentions_per_hour.5" : 10 } } )
```

`update()` Defaults to one Document

- By default, `update()` modifies the first document found that matches the query.
- The default use case is one where there is only one document that fits the query.
- This is to reduce the chances of unintended collection scans for updates.

Updating Multiple Documents

- In order to update multiple documents, we use the third (optional) parameter to `update()`.
- The third parameter is an options document.
- Specify `multi: true` as one field in this document.
- Bear in mind that without an appropriate index, you may scan every document in the collection.

Example: Multi-Update

Use `db.testcol.find()` after each of these updates.

```
// let's start tracking the number of sequals for each movie
db.movies.update( { }, { $set : { "sequels" : 0 } } )

// we need { multi : true } to change all documents
db.movies.update( { }, { $set : { "sequels" : 0 } },
                  { multi : true } )
```

Note:

- `db.movies.update( { }, { $set : { "sequels" : 0 } } )` only updates one document.
- `db.movies.update( { }, { $set : { "sequels" : 0 } }, { multi : true } )` updates four documents.

Array Operators

- `$push`: Appends an element to the end of the array.
- `$pushAll`: Appends multiple elements to the end of the array.
- `$pop`: Removes one element from the end of the array.
- `$pull`: Removes all elements in the array that match a specified value.
- `$pullAll`: Removes all elements in the array that match any of the specified values.
- `$addToSet`: Appends an element to the array if not already present.

Note:

- These operators may be applied to array fields.

Example: Array Operators

```
db.movies.update( { "title" : "Batman" },
                  { $push : { "category" : "superhero" } } )
db.movies.update( { "title" : "Batman" },
                  { $pushAll : { "category" : [ "villain", "comicbased" ] } } )

db.movies.update( { "title" : "Batman" },
                  { $pop : { "category" : 1 } } )

db.movies.update( { "title" : "Batman" },
                  { $pull : { "category" : "action" } } )
db.movies.update( { "title" : "Batman" },
                  { $pullAll : { "category" : [ "villain", "comicbased" ] } } )

db.movies.update( { "title" : "Batman" }, { $addToSet : { "category" : "action" } } )
db.movies.update( { "title" : "Batman" }, { $addToSet : { "category" : "action" } } )
```

Note:

- Pass $pop a value of -1 to remove the first element of an array and 1 to remove the last element in an array.

The Positional $ Operator

- $[5] is a positional operator that specifies an element in an array to update.
- It acts as a placeholder for the first element that matches the query document.
- `$` replaces the element in the specified position with the value given.
- Example:

```
db.<COLLECTION>.update(
    { <array> : value ... },
    { <update operator> : { "<array>.$" : value } }
)
```

[5] http://docs.mongodb.org/manual/reference/operator/update/postional

Example: The Positional $ Operator

```
// the "action" category needs to be changed to "action-adventure"
db.movies.update( { "category": "action",  },
                  { $set: { "category.$" : "action-adventure" } },
                  { multi: true } )
db.movies.find()
```

Upserts

- By default, if no document matches an update query, the `update()` method does nothing.
- By specifying `upsert: true`, `update()` will insert a new document if no matching document exists.
- The `db.<COLLECTION>.save()` method is syntactic sugar that performs an upsert if the _id is not yet present
- Syntax:

```
db.<COLLECTION>.update( <query document>, <update document>,
                        { upsert: true } )
```

Upsert Mechanics

- Will update as usual if documents matching the query document exist.
- Will be an upsert if no documents match the query document.
  - MongoDB creates a new document using equality conditions in the query document.
  - Adds an `_id` if the query did not specify one.
  - Performs an update on the new document.

Example: Upserts

```
db.movies.update( { "title" : "Jaws" },
                  { $inc: { "budget" : 5 } },
                  { upsert: true } )

db.movies.update( { "title" : "Jaws II" },
                  { $inc: { "budget" : 5 } },
                  { upsert: true } )

db.movies.update( { "title" : "Jaws III", "category" : [ "horror" ] },
                  { $set : { "budget" : 1 } },
                  { upsert: true } )
```

Note:

```
// updates the document with "title" = "Jaws" by incrementing "budget"
db.movies.update( { "title" : "Jaws" }, { $inc: { "budget" : 5 } }, { upsert: true } )

// 1) creates a new document, 2) assigns an _id, 3) sets "title" to "Jaws II"
// 4) performs the update
db.movies.update( { "title" : "Jaws II" }, { $inc: { "budget" : 5 } }, { upsert: true } )
```

```
// 1) creates a new document, 2) sets "title" : "Jaws III",
// 3) Set budget to 1
db.movies.update( { "title" : "Jaws III" }, { "budget" : 1 }, { upsert: true } )
```

`save()`

- Updates the document if the _id is found, inserts it otherwise
- Syntax:

  ```
  db.<COLLECTION>.save( document )
  ```

Example: `save()`

- If the document does not contain an `_id` field, then the `save()` method calls the `insert()` method. During the operation, the mongo shell will create an ObjectId and assign it to the `_id` field.
- If the document contains an `_id` field, then the `save()` method is equivalent to an update with the upsert option set to true and the query predicate on the `_id` field.

```
// insert
db.movies.save( { "title" : "Beverly Hills Cops", "imdb_rating" : 7.3 })

// update with { upsert: true }
db.movies.save( { "_id" : 1234, "title" : "Spider Man", "imdb_rating" : 7.3 })
```

Note:

- A lot of users prefer to use update/insert, to have more explicit control over the operation

---

Be Careful with `save()`

Be careful that you are not modifying stale data when using `save()`. For example:

```
db.movies.drop()
db.movies.insert( { "title" : "Jaws", "imdb_rating" : 7.3 } )

db.movies.find( { "title" : "Jaws" } )

// store the complete document in the application
doc = db.movies.findOne( { "title" : "Jaws" } )

db.movies.update( { "title" : "Jaws"  }, { $inc: { "imdb_rating" : 2 } } )
db.movies.find()

doc.imdb_rating = 7.4
doc

db.movies.save(doc)  // just lost our incrementing of "imdb_rating"
db.movies.find()
```

`findAndModify()`

Modify a document and return either:

- The pre-modification document
- If "new:true" is set, the modified document

Helpful for making changes to a document and reading the document in the state before or after the update occurred.

`findAndModify()` Example

```
db.worker_queue.findAndModify({
    "query" : { "state" : "unprocessed" },
    "update" : { $set: { "worker_id" : 123, "state" : "processing" } },
})
```

2.6 Lab: Updating Documents

Exercise: Letter Grades

- Using the `sample.scores` namespace, set the proper "grade" attributes.
- For example, users with scores greater than 90 get an "A".
- Set the grade to "B" for scores falling between 80 and 90 and so on for grades "C", "D", and "F".

Note:

- `db.scores.update( { score: { $gte: 90 } }, { $set: { grade: "A" } }, { multi: true } )`
- and so on for `{ score: { $gte: 80, $lt: 90 } }` etc.
- In the shell you could write a JS script with a loop. In SQL you could do it in one statement, note the difference.

Exercise: 10 Extra-Credit Points

- You're being nice, so you decide to add 10 points to every score on every exam where the score is lower than 60.
- How do you do this update?

Note:

```
db.scores.update( { name : "exam", score : { $lt : 60 } },
                  { $inc : { score : 10 } }, { multi: true } )
```

Exercise: Updating Array Elements

Insert a document representing product metrics for a backpack:

```
db.product_metrics.insert(
   { name: "backpack",
     purchasesPast7Days: [ 0, 0, 0, 0, 0, 0, 0] })
```

Each 0 within the "purchasesPast7Days" field corresponds to a day of the week. The first element is Monday, the second element is Tuesday, etc.).

Write an update statement to increment the number of backpacks sold on Friday by 200.

Note:

- Talk about how this can be used for time series data, real-time graphs/charts
- `db.product_metrics.update( { "name" : "backpack" }, { $inc : { "purchases_past_7_days.4" : 200 } } )`

---

3 Indexes

Index Fundamentals **(page 51)** An introduction to MongoDB indexes

Compound Indexes **(page 58)** Indexes on two or more fields

Lab: Optimizing an Index **(page 64)** Lab on optimizing a compound index

Multikey Indexes **(page 65)** Indexes on array fields

Hashed Indexes **(page 69)** Hashed indexes

Lab: Finding and Addressing Slow Operations **(page 71)** Lab on finding and addressing slow queries

Lab: Using explain() **(page 71)** Lab on using the explain operation to review execution stats

3.1 Index Fundamentals

Learning Objectives

Upon completing this module students should understand:

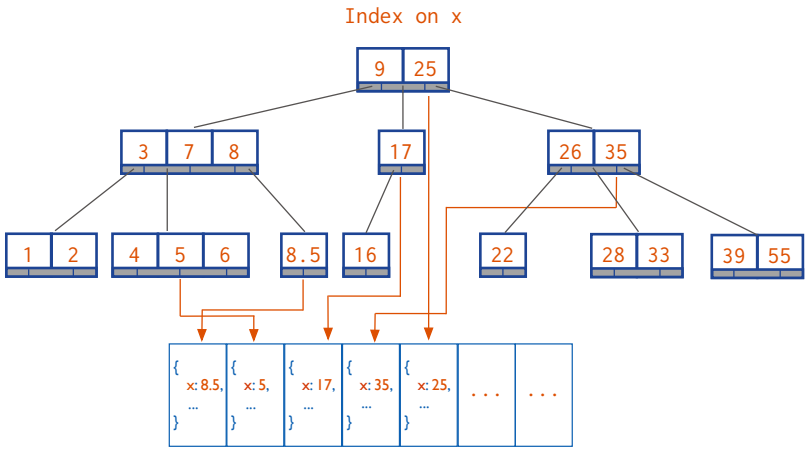
- The impact of indexing on read performance
- The impact of indexing on write performance
- How to choose effective indexes
- The utility of specific indexes for particular query patterns

---

Note:

- Ask how many people in the room are familiar with indexes in a relational database.
- If the class is already familiar with indexes, just explain that they work the same way in MongoDB.

Why Indexes?



Note:

- Without an index, in order to find all documents matching a query, MongoDB must scan every document in the collection.

- This is murder for read performance, and often write performance, too.
- If all your documents do not fit into memory, the system will page data in and out in order to scan the entire collection.
- An index enables MongoDB to locate exactly which documents match the query and where they are located on disk.
- MongoDB indexes are based on B-trees.

Types of Indexes

- Single-field indexes
- Compound indexes
- Multikey indexes
- Geospatial indexes
- Text indexes

Note:

- There are also hashed indexes and TTL indexes.
- We will discuss those elsewhere.

Exercise: Using `explain()`

Let's explore what MongoDB does for the following query by using `explain()`.

We are projecting only `user.name` so that the results are easy to read.

```
db.tweets.find( { "user.followers_count" : 1000 },
                { "_id" : 0, "user.name": 1 } )

db.tweets.find( { "user.followers_count" : 1000 } ).explain()
```

Note:

- Make sure the students are using the sample database.
- Review the structure of documents in the tweets collection by doing a find().
- We'll be looking at the user subdocument for documents in this collection.

Results of `explain()`

With the default `explain()` verbosity, you will see results similar to the following:

```
{
  "queryPlanner" : {
    "plannerVersion" : 1,
    "namespace" : "twitter.tweets",
    "indexFilterSet" : false,
    "parsedQuery" : {
      "user.followers_count" : {
        "$eq" : 1000
      }
    },
```

Results of `explain()` - Continued

```
  "winningPlan" : {
      "stage" : "COLLSCAN",
      "filter" : {
        "user.followers_count" : {
          "$eq" : 1000
        }
      },
      "direction" : "forward"
    },
    "rejectedPlans" : [ ]
  },
  ...
}
```

`explain()` Verbosity Can Be Adjusted

- **default:** determines the winning query plan but does not execute query
- **executionStats:** executes query and gathers statistics
- **allPlansExecution:** runs all candidate plans to completion and gathers statistics

---

Note:

- Default will be helpful if you're worried running the query could cause sever performance problems
- executionStats will be the most common verbosity level used
- allPlansExecution is for trying to determine WHY it is choosing the index it is (out of other candidates)

`explain("executionStats")`

```
> db.tweets.find( { "user.followers_count" : 1000 } )
  .explain("executionStats")
```

Now we have query statistics:

```
..
"executionStats" : {
  "executionSuccess" : true,
  "nReturned" : 8,
  "executionTimeMillis" : 107,
  "totalKeysExamined" : 0,
  "totalDocsExamined" : 51428,
  "executionStages" : {
    "stage" : "COLLSCAN",
    "filter" : {
      "user.followers_count" : {
        "$eq" : 1000
      }
    },
```

`explain("executionStats")` - Continued

```
    "nReturned" : 8,
    "executionTimeMillisEstimate" : 100,
    "works" : 51430,
    "advanced" : 8,
    "needTime" : 51421,
    "needFetch" : 0,
    "saveState" : 401,
    "restoreState" : 401,
    "isEOF" : 1,
    "invalidates" : 0,
    "direction" : "forward",
    "docsExamined" : 51428
  }
  ...
}
```

`explain("executionStats")` Output

- `nReturned` displays the number of documents that match the query.
- `totalDocsExamined` displays the number of documents the retrieval engine considered during the query.
- `totalKeysExamined` displays how many documents in an existing index were scanned.
- A `totalKeysExamined` or `totalDocsExamined` value much higher than `nReturned` indicates we need a different index.
- Given `totalDocsExamined`, this query will benefit from an index.

Other Operations

In addition to `find()`, we often want to use `explain()` to understand how other operations will be handled.

- `aggregate`
- `count`
- `group`
- `remove`
- `update`

`db.<COLLECTION>.explain()`

`db.<COLLECTION>.explain()` returns an ExplainableCollection.

```
> var explainable = db.tweets.explain()
> explainable.find( { "user.followers_count" : 1000 } )
```

equivalent to

```
> db.tweets.explain().find( { "user.followers_count" : 1000 } )
```

also equivalent to

```
> db.tweets.find( { "user.followers_count" : 1000 } ).explain()
```

Note:

- This will get confusing for students, may want to spend extra time here with more examples

---

Using `explain()` for Write Operations

Simulate the number of writes that would have occurred and determine the index(es) used:

```
> db.tweets.explain("executionStats").remove({ "user.followers_count" : 1000 })

> db.tweets.explain("executionStats").update({ "user.followers_count" : 1000 },
  { $set : { "large_following" : true } } )
```

Note:

- Walk through the "nWouldModify" field in the output to show how many documents would have been updated

---

Single-Field Indexes

- Single-field indexes are based on a single field of the documents in a collection.
- The field may be a top-level field.
- You may also create an index on fields in embedded documents.

Creating an Index

The following creates a single-field index on `user.followers_count`.

```
db.tweets.createIndex( { "user.followers_count" : 1 } )
db.tweets.find( { "user.followers_count" : 1000 } ).explain()
```

`explain()` indicated there will be a substantial performance improvement in handling this type of query.

Note:

- nscannedObjects should now be a much smaller number, e.g., 8.
- Operations teams are accustomed to thinking about indexes.
- With MongoDB, developers need to be more involved in the creation and use of indexes.

Listing Indexes

List indexes for a collection:

```
db.tweets.getIndexes()
```

List index keys:

```
db.tweets.getIndexKeys()
```

Indexes and Read/Write Performance

- Indexes improve read performance for queries that are supported by the index.
- Inserts will be slower when there are indexes that MongoDB must also update.
- The speed of updates may be improved because MongoDB will not need to do a collection scan to find target documents.
- An index is modified any time a document:
  - Is inserted
  - Is deleted
  - Is updated in such a way that its indexed field changes
  - If an update causes a document to move on disk

Index Limitations

- You can have up to 64 indexes per collection.
- You should NEVER be anywhere close to that upper bound.
- Write performance will degrade to unusable at somewhere between 20-30.

Use Indexes with Care

- Every query should use an index.
- Every index should be used by a query.
- Any write operation that touches an indexed field will require each index to be updated.
- Indexes require RAM.
- Be judicious about the choice of key.

Note:

- If your system has limited RAM, then using the index will force other data out of memory.
- When you need to access those documents, they will need to be paged in again.

Additional Index Options

- Sparse
- Unique
- Background

Sparse Indexes in MongoDB

Sparse indexes only contain entries for documents that have the indexed field.

```
db.<COLLECTION>.createIndex(
  { field_name : 1 },
  { sparse : true } )
```

Defining Unique Indexes

- Enforce a unique constraint on the index.
- Prevent duplicate values from being inserted into the database.
- No duplicate values may exist prior to defining the index.

```
db.<COLLECTION>.createIndex(
  { field_name : 1 },
  { unique : true } )
```

Building Indexes in the Background

- Building indexes in foreground is a blocking operation.
- Background index creation is non-blocking, however, takes longer to build.
- Initially larger, or less compact, than an index built in the foreground.

```
db.<COLLECTION>.createIndex(
  { field_name : 1 },
  { background : true } )
```

3.2 Compound Indexes

Learning Objectives

Upon completing this module students should understand:

- What a compound index is.
- How compound indexes are created.
- The importance of considering field order when creating compound indexes.
- How to efficiently handle queries involving some combination of equality matches, ranges, and sorting.
- Some limitations on compound indexes.

Introduction to Compound Indexes

- It is common to create indexes based on more than one field.
- These are called `compound indexes`.
- You may use up to 31 fields in a compound index.
- You may not use hashed index fields.

The Order of Fields Matters

Specifically we want to consider how the index will be used for:

- Equality tests, e.g.,

```
db.movies.find( { "budget" : 7, "imdb_rating" : 8 } )
```

- Range queries, e.g.,

```
db.movies.find( { "budget" : 10, "imdb_rating" : { $lt : 9 } } )
```

- Sorting, e.g.,

```
db.movies.find( { "budget" : 10, "imdb_rating" : 6 }
                ).sort( { "imdb_rating" : -1 } )
```

Note:

- The order in which the fields are specified is of critical importance.
- It is especially important to consider query patterns that require two or more of these operations.

Designing Compound Indexes

- Let's look at some guiding principles for building compound indexes.
- These will generally produce a good if not optimal index.
- You can optimize after a little experimentation.
- We will explore this in the context of a running example.

Example: A Simple Message Board

Requirements:

- Find all messages in a specified timestamp range.
- Select for whether the messages are anonymous or not.
- Sort by rating from highest to lowest.

Load the Data

```
a = [ { "timestamp" : 1, "username" : "anonymous", "rating" : 3 },
      { "timestamp" : 2, "username" : "anonymous", "rating" : 5 },
      { "timestamp" : 3, "username" : "sam", "rating" : 1 },
      { "timestamp" : 4, "username" : "anonymous", "rating" : 2 },
      { "timestamp" : 5, "username" : "martha", "rating" : 5 } ]
db.messages.insert(a)
```

Start with a Simple Index

Start by building an index on { timestamp : 1 }

```
db.messages.createIndex( { timestamp : 1 }, { name : "myindex" } )
```

Now let's query for messages with timestamp in the range 2 through 4 inclusive.

```
db.messages.find( { timestamp : { $gte : 2, $lte : 4 } } ).explain()
```

Analysis:

- Explain plan shows good performance, i.e. `totalKeysExamined` = `n`.
- However, this does not satisfy our query.
- Need to query again with `{username: "anonymous"}` as part of the query.

Query Adding `username`

Let's add the `user` field to our query.

```
db.messages.find( { timestamp : { $gte : 2, $lte : 4 },
                    username : "anonymous" } ).explain()
```

`totalKeysExamined` > `n`.

Include `username` in Our Index

```
db.messages.dropIndex( "myindex" );
db.messages.createIndex( { timestamp : 1, username : 1 },
                         { name : "myindex" } )
db.messages.find( { timestamp : { $gte : 2, $lte : 4 },
                    username : "anonymous" } ).explain()
```

`totalKeysExamined` is still > `n`. Why?

`totalKeysExamined > n`

timestamp	username
1	“anonymous”
2	“anonymous”
3	“sam”
4	“anonymous”
5	“martha”

Note:

- The index we have created stores the range values before the equality values.
- The documents with timestamp values 2, 3, and 4 were found first.
- Then the associated anonymous values had to be evaluated.

A Different Compound Index

Drop the index and build a new one with user.

```
db.messages.dropIndex( "myindex" );
db.messages.createIndex( { username : 1 , timestamp : 1 },
                         { name : "myindex" } )

db.messages.find( { timestamp : { $gte : 2, $lte : 4 },
                    username : "anonymous" } ).explain()
```

`totalKeysExamined` is 2. `n` is 2.

`totalKeysExamined == n`

username	timestamp
“anonymous”	1
“anonymous”	2
“anonymous”	4
“sam”	2
“martha”	5

Note:

- This illustrates why.
- There is a fundamental difference in the way the index is structured.
- This supports a more efficient treatment of our query.

Let Selectivity Drive Field Order

- Order fields in a compound index from most selective to least selective.
- Usually, this means equality fields before range fields.
- When dealing with multiple equality values, start with the most selective.
- If a common range query is more selective instead (rare), specify the range component first.

Adding in the Sort

Finally, let's add the sort and run the query.

```
db.messages.find( {
                    timestamp : { $gte : 2, $lte : 4 },
                    username : "anonymous"
                  } ).sort( { rating : -1 } ).explain();
```

- Note that the `winningPlan` includes a `SORT` stage.
- This means that MongoDB had to perform a sort in memory.
- In memory sorts for queries that retrieve large numbers of documents can degrade performance significantly.
- This is especially true if they are used frequently.

In-Memory Sorts

Let's modify the index again to allow the database to sort for us.

```
db.messages.dropIndex( "myindex" );
db.messages.createIndex( { username : 1 , timestamp : 1, rating : 1 },
                         { name : "myindex" } );
db.messages.find( {
                    timestamp : { $gte : 2, $lte : 4 },
                    username : "anonymous"
                  } ).sort( { rating : -1 } ).explain();
```

- The explain plan remains unchanged, because the sort field comes after the range fields.
- The index does not store entries in order by rating.
- Note that this requires us to consider a tradeoff.

Avoiding an In-Memory Sort

Rebuild the index as follows.

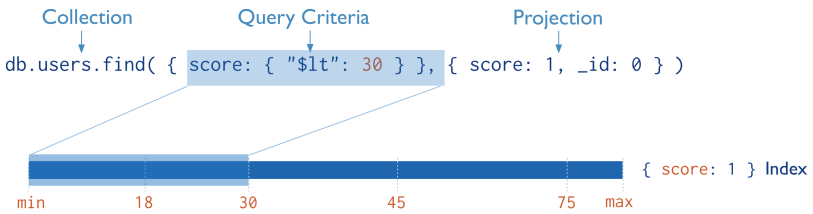
```
db.messages.dropIndex( "myindex" );
db.messages.createIndex( { username : 1, rating : 1, timestamp : 1 },
                         { name : "myindex" } );
db.messages.find( {
                    timestamp : { $gte : 2, $lte : 4 },
                    username : "anonymous"
                  } ).sort( { rating : -1 } ).explain();
```

- We no longer have an in-memory sort, but need to examine more keys.
- `totalKeysExamined` is 3 and and `n` is 2.
- This is the best we can do in this situation and this is fine.
- However, if `totalKeysExamined` is much larger than `n`, this might not be the best index.

General Rules of Thumb

- Equality before range
- Equality before sorting
- Sorting before range

Covered Queries



- When a query and projection include only the indexed fields, MongoDB will return results directly from the index.
- There is no need to scan any documents or bring documents into memory.
- These covered queries can be very efficient.

Exercise: Covered Queries

```
db.testcol.drop()
for (i=1; i<=20; i++) {
  db.testcol.insert({ "_id" : i, "title" : i, "name" : i,
                      "rating" : i, "budget" : i })
};
db.testcol.createIndex( { "title" : 1, "name" : 1, "rating" : 1 } )

// Not covered because _id is  present.
db.testcol.find( { "title" : 3 },
                 { "title" : 1, "name" : 1, "rating" : 1 }
                 ).explain("executionStats")

// Not covered because other fields may exist in matching docs.
db.testcol.find( { "title" : 3 },
                 { "_id" : 0, "budget" : 0 } ).explain("executionStats")

// Covered query!
db.testcol.find( { "title" : 3 },
                 { "_id" : 0, "title" : 1, "name" : 1, "rating" : 1 }
                 ).explain("executionStats")
```

3.3 Lab: Optimizing an Index

Exercise: What Index Do We Need?

Run the the following Javascript file from the handouts.

```
mongo --shell localhost/performance performance.js
```

In the shell that launches execute the following method

```
performance.init()
```

The method above will build a sample data set in the "sensor_readings" collection. What index is needed for this query?

```
db.sensor_readings.find( { tstamp: { $gte: ISODate("2012-08-01"),
                                     $lte: ISODate("2012-09-01") },
                           active: true } ).limit(3)
```

Note:

- Work through method of explaining query with .explain("executionStats")
- Look at differences between (timestamp, active) and (active, timestamp)

Exercise: Avoiding an In-Memory Sort

What index is needed for the following query to avoid an in-memory sort?

```
db.sensor_readings.find( { active: true } ).sort( { tstamp : -1 } )
```

Note:

- { active: 1, tstamp: 1 }

Exercise: Avoiding an In-Memory Sort, 2

What index is needed for the following query to avoid an in-memory sort?

```
db.sensor_readings.find(
    { x : { $in : [100, 200, 300, 400] } }
).sort( { tstamp : -1 })
```

Note:

- Trick question, the answer most students will give is { x: 1, tstamp: 1 }, however, the $in will require an in-memory sort
- (tstamp) or (tstamp, x) are the only indexes that will prevent an in-memory sort, but aren't selective at all

---

3.4 Multikey Indexes

Learning Objectives

Upon completing this module, students should understand:

- What a multikey index is
- When MongoDB will use a multikey index to satisfy a query
- How multikey indexes work
- How multikey indexes handle sorting
- Some limitations on multikey indexes

Introduction to Multikey Indexes

- A multikey index is an index on an array.
- An index entry is created on each value found in the array.
- Multikey indexes can support primitives, documents, or sub-arrays.
- There is nothing special that you need to do to create a multikey index.
- You create them using `createIndex()` just as you would with an ordinary single-field index.
- If there is an array as a value for an indexed field, the index will be multikey on that field.

Example: Array of Numbers

```
db.race_results.drop()
db.race_results.createIndex( { "lap_times" : 1 } )
a = [ { "lap_times" : [ 3, 5, 2, 8 ] },
      { "lap_times" : [ 1, 6, 4, 2 ] },
      { "lap_times" : [ 6, 3, 3, 8 ] } ]
db.race_results.insert( a )

// Used the index
db.race_results.find( { lap_times : 1 } ).explain()

// One document found.
// Index not used, because it is naive to position.
db.race_results.find( { "lap_times.2" : 3 } ).explain()
```

Exercise: Array of Documents, Part 1

Create a collection and add an index on the `x` field:

```
db.blog.drop()
b = [ { "comments" : [
          { "name" : "Bob", "rating" : 1 },
          { "name" : "Frank", "rating" : 5.3 },
          { "name" : "Susan", "rating" : 3 } ] },
      { "comments" : [
          { name : "Megan", "rating" : 1 } ] },
      { "comments" : [
          { "name" : "Luke", "rating" : 1.4 },
          { "name" : "Matt", "rating" : 5 },
          { "name" : "Sue", "rating" : 7 } ] }]
db.blog.insert(b)

db.blog.createIndex( { "comments" : 1 } )
// vs
db.blog.createIndex( { "comments.rating" : 1 } )

// for this query
db.blog.find( { "comments.rating" : 5 })
```

Note:

- Note: JSON is a dictionary and doesn't guarantee order, indexing the top level array (comments array) won't work

Exercise: Array of Documents, Part 2

For each of the three queries below:

- How many documents will be returned?
- Will it use our multi-key index? Why or why not?
- If a query will not use the index, which index will it use?

```
db.blog.find( { "comments" : { "name" : "Bob", "rating" : 1 } } )
db.blog.find( { "comments" : { "rating" : 1 } } )
db.blog.find( { "comments.rating" : 1 } )
```

Note:

```
// Never do this, won't give you the results expected
// JSON is a dictionary, and won't preserve ordering, second query will return no results

db.blog.find( { "comments" : { "name" : "Bob", "rating" : 1 } } )
db.blog.find( { "comments" : { "rating" : 1 } } )
```

Exercise: Array of Arrays, Part 1

Add some documents and create an index simulating a player in a game moving on an X,Y grid.

```
db.player.drop()
db.player.createIndex( { "last_moves" : 1 } )
c = [ { "last_moves" : [ [ 1, 2 ], [ 2, 3 ], [ 3, 4] ] },
      { "last_moves" : [ [ 3, 4 ], [ 4, 5 ] ] },
      { "last_moves" : [ [ 4, 5 ], [ 5, 6 ] ] },
      { "last_moves" : [ [ 3, 4 ] ] },
      { "last_moves" : [ [ 4, 5 ] ] } ]
db.player.insert(c)
db.player.find()
```

Exercise: Array of Arrays, Part 2

For each of the queries below:

- How many documents will be returned?
- Does the query use the multi-key index? Why or why not?
- If the query does not use the index, what is an index it could use?

```
db.player.find( { "last_moves" : [ 3, 4 ] } )
db.player.find( { "last_moves" : 3 } )
db.player.find( { "last_moves.1" : [ 4, 5 ] } )
db.player.find( { "last_moves.2" : [ 2, 3 ] } )
```

Note:

```
// 3 documents
db.player.find( { "last_moves" : [ 3, 4 ] } )
// Uses the multi-key index
db.player.find( { "last_moves" : [ 3, 4 ] } ).explain()

// No documents
db.player.find( { "last_moves" : 3 } )

// Does not use the multi-key index, because it is naive to position.
db.player.find( { "last_moves.1" : [ 4, 5 ] } )
db.player.find( { "last_moves.2" : [ 2, 3 ] } )
```

How Multikey Indexes Work

- Each array element is given one entry in the index.
- So an array with 17 elements will have 17 entries – one for each element.
- Multikey indexes can take up much more space than standard indexes.

Multikey Indexes and Sorting

- If you sort using a multikey index:
  - A document will appear at the first position where a value would place the document.
  - It will not appear multiple times.
- This applies to array values generally.
- It is not a specific property of multikey indexes.

Exercise: Multikey Indexes and Sorting

```
db.testcol.drop()
a = [ { x : [ 1, 11 ] }, { x : [ 2, 10 ] }, { x : [ 3 ] },
      { x : [ 4 ] }, { x : [ 5 ] } ]
db.testcol.insert(a)

db.testcol.createIndex( { x : 1 } )

// x : [ 1, 11 ] array comes first. It contains the lowest value.
db.testcol.find().sort( { x : 1 } )

// x : [ 1, 11 ] array still comes first. Contains the highest value.
db.testcol.find().sort( { x : -1 } )
```

Note:

```
// x : [ 1, 11 ] array comes first, because it contains the lowest value
db.testcol.find().sort( { x : 1 } )

// x : [ 1, 11 ] array still comes first, because it contains the highest value
db.testcol.find().sort( { x : -1 } )
```

Limitations on Multikey Indexes

- You cannot create a compound index using more than one array-valued field.
- This is because of the combinatorics.
- For a compound index on two array-valued fields you would end up with N * M entries for one document.
- You cannot have a hashed multikey index.
- You cannot have a shard key use a multikey index.
- We discuss shard keys in another module.
- The index on the `_id` field cannot become a multikey index.

Example: Multikey Indexes on Multiple Fields

```
db.testcol.drop()
db.testcol.createIndex( { x : 1, y : 1 } )

// no problems yet
db.testcol.insert( { _id : 1, x : 1, y : 1 } )

// still OK
db.testcol.insert( { _id : 2, x : [ 1, 2 ], y : 1 } )

// still OK
db.testcol.insert( { _id : 3, x : 1, y : [ 1, 2 ] } )

// Won't work
db.testcol.insert( { _id : 4, x : [ 1, 2 ], y : [ 1, 2 ] } )
```

3.5 Hashed Indexes

Learning Objectives

Upon completing this module, students should understand:

- What a hashed index is
- When to use a hashed index

What is a Hashed Index?

- Hashed indexes are based on field values like any other index.
- The difference is that the values are hashed and it is the hashed value that is indexed.
- The hashing function collapses sub-documents and computes the hash for the entire value.
- MongoDB can use the hashed index to support equality queries.
- Hashed indexes do not support multi-key indexes, i.e. indexes on array fields.
- Hashed indexes do not support range queries.

Why Hashed Indexes?

- In MongoDB, the primary use for hashed indexes is to support sharding a collection using a hashed shard key.
- In some cases, the field we would like to use to shard data would make it difficult to scale using sharding.
- Using a hashed shard key to shard a collection ensures an even distribution of data and overcomes this problem.
- See Shard a Collection Using a Hashed Shard Key[6] for more details.
- We discuss sharding in detail in another module.

Limitations

- You may not create compound indexes that have hashed index fields.
- You may not specify a unique constraint on a hashed index.
- You can create both a hashed index and a non-hashed index on the same field.

Note:

- For a field on which there is both a hashed index and a non-hashed index, MongoDB will use the non-hashed index for range queries.

Floating Point Numbers

- MongoDB hashed indexes truncate floating point numbers to 64-bit integers before hashing.
- Do not use a hashed index for floating point numbers that cannot be reliably converted to 64-bit integers.
- MongoDB hashed indexes do not support floating point values larger than 2^{53}.

[6] http://docs.mongodb.org/manual/tutorial/shard-collection-with-a-hashed-shard-key/

Creating a Hashed Index

Create a hashed index using an operation that resembles the following. This operation creates a hashed index for the active collection on the `a` field.

```
db.active.createIndex( { a: "hashed" } )
```

3.6 Lab: Finding and Addressing Slow Operations

Exercise: Determine Indexes Needed

- In a mongo shell run `performance.b()`. This will run in an infinite loop printing some output as it runs various statements against the server.
- Now imagine we have detected a performance problem and suspect there is a slow operation running.
- Find the slow operation and terminate it. Every slow operation is assumed to run for 100ms or more.
- In order to do this, open a second window (or tab) and run a second instance of the mongo shell.
- What indexes can we introduce to make the slow queries more efficient? Disregard the index created in the previous exercise.

Note:

- Look at the logs to find queries over 100ms
- { “active”: 1 }
- { “str”: 1, “x”: 1 }

3.7 Lab: Using `explain()`

Exercise: explain(“executionStats”)

Drop all indexes from previous exercises:

```
mongo performance
> db.sensor_readings.dropIndexes()
```

Create an index for the “active” field:

```
db.sensor_readings.createIndex({ "active" : 1 } )
```

How many index entries and documents are examined for the following query? How many results are returned?

```
db.sensor_readings.find(
                          { "active": false, "_id": { $gte: 99, $lte: 1000 } }
                       ).explain("executionStats")
```

4 Drivers

Introduction to MongoDB Drivers **(page 72)** An introduction to the MongoDB drivers

Lab: Driver Tutorial (Optional) **(page 74)** A quick tour through the Python driver

4.1 Introduction to MongoDB Drivers

Learning Objectives

Upon completing this module, students should understand:

- What MongoDB drivers are available
- Where to find MongoDB driver specifications
- Key driver settings

MongoDB Supported Drivers

- C[7]
- C++[8]
- C#[9]
- Java[10]
- Node.js[11]
- Perl[12]
- PHP[13]
- Python[14]
- Ruby[15]
- Scala[16]

[7] http://docs.mongodb.org/ecosystem/drivers/c
[8] http://docs.mongodb.org/ecosystem/drivers/cpp
[9] http://docs.mongodb.org/ecosystem/drivers/csharp
[10] http://docs.mongodb.org/ecosystem/drivers/java
[11] http://docs.mongodb.org/ecosystem/drivers/node-js
[12] http://docs.mongodb.org/ecosystem/drivers/perl
[13] http://docs.mongodb.org/ecosystem/drivers/php
[14] http://docs.mongodb.org/ecosystem/drivers/python
[15] http://docs.mongodb.org/ecosystem/drivers/ruby
[16] http://docs.mongodb.org/ecosystem/drivers/scala

MongoDB Community Supported Drivers

35+ different drivers for MongoDB:

Go, Erlang, Clojure, D, Delphi, F#, Groovy, Lisp, Objective C, Prolog, Smalltalk, and more

Driver Specs

To ensure drivers have a consistent functionality, series of publicly available specification documents[17] for:

- Authentication[18]
- CRUD operations[19]
- Index management[20]
- SDAM[21]
- Server Selection[22]
- Etc.

Driver Settings (Per Operation)

- Read preference
- Write concern
- Maximum operation time (maxTimeMS)
- Batch Size (batchSize)
- Exhaust cursor (exhaust)
- Etc.

Driver Settings (Per Connection)

- Connection timeout
- Connections per host
- Time that a thread will block waiting for a connection (maxWaitTime)
- Socket keep alive
- Sets the multiplier for number of threads allowed to block waiting for a connection
- Etc.

[17] https://github.com/mongodb/specifications
[18] https://github.com/mongodb/specifications/tree/master/source/auth
[19] https://github.com/mongodb/specifications/tree/master/source/crud
[20] https://github.com/mongodb/specifications/blob/master/source/index-management.rst
[21] https://github.com/mongodb/specifications/tree/master/source/server-discovery-and-monitoring
[22] https://github.com/mongodb/specifications/tree/master/source/server-selection

Insert a Document with the Java Driver

Connect to a MongoDB instance on localhost:

```
MongoClient mongoClient = new MongoClient();
```

Access the test database:

```
MongoDatabase db = mongoClient.getDatabase("test");
```

Insert a myDocument Document into the test.blog collection:

```
db.getCollection("blog").insertOne(myDocument);
```

Insert a Document with the Python Driver

Connect to a MongoDB instance on localhost:

```
client = MongoClient()
```

Access the test database:

```
db = client['test']
```

Insert a myDocument Document into the test.blog collection:

```
db.blog.insert_one(myDocument);
```

Insert a Document with the C++ Driver

Connect to the "test" database on localhost:

```
mongocxx::instance inst{};
mongocxx::client conn{};

auto db = conn["test"];
```

Insert a myDocument Document into the test.blog collection:

```
auto res = db["blog"].insert_one(myDocument);
```

4.2 Lab: Driver Tutorial (Optional)

Tutorial

Complete the Python driver tutorial[23] for a concise introduction to:

- Creating MongoDB documents in Python
- Connecting to a database
- Writing and reading documents
- Creating indexes

[23] http://api.mongodb.org/python/current/tutorial.html

5 Aggregation

Aggregation Tutorial **(page 75)** An introduction to the the aggregation framework, pipeline concept, and stages

Optimizing Aggregation **(page 84)** Resource management in the aggregation pipeline

Lab: Aggregation Framework **(page 86)** Aggregation labs

5.1 Aggregation Tutorial

Learning Objectives

Upon completing this module students should understand:

- The concept of the aggregation pipeline
- The stages of the aggregation pipeline
- How to use aggregation operators
- The fundamentals of using aggregation for data analysis
- Group aggregation operators
- Using the same operator in multiple stages of an aggregation pipeline

Aggregation Basics

- Use the aggregation framework to transform and analyze data in MongoDB collections.
- For those who are used to SQL, aggregation can be similar to `GROUP BY`.
- The aggregation framework is based on the concept of a pipeline.

The Aggregation Pipeline

- An aggregation pipeline in analogous to a UNIX pipeline.
- Each stage of the pipeline:
  - Receives a set of documents as input.
  - Performs an operation on those documents.
  - Produces a set of documents for use by the following stage.
- A pipeline has the following syntax:

```
db.<COLLECTION>.aggregate( [ { stage1 }, { stage2 }, ... ],
                           { options } )
```

Aggregation Stages

- `$match`: Similar to `find()`
- `$project`: Shape documents
- `$sort`: Like the cursor method of the same name
- `$skip`: Like the cursor method of the same name
- `$limit`: Like the cursor method of the same name
- `$unwind`: Used for working with arrays
- `$group`: Used to aggregate field values from multiple documents
- `$out`: Creates a new collection from the output of an aggregation pipeline

The Match Stage

- The `$match` operator works like the query phase of `find()`, `update()`, and `remove()`.
- Documents in the pipeline that match the query document will be passed to subsequent stages.
- `$match` is often the first operator used in an aggregation stage.
- Like other aggregation operators, `$match` can occur multiple times in a single pipeline.

The Project Stage

- $project allows you to shape the documents into what you need for the next stage.
  - The simplest form of shaping is using $project to select only the fields you are interested in.
  - $project can also create new fields from other fields in the input document.
    - *E.g.*, you can pull a value out of an embedded document and put it at the top level.
    - *E.g.*, you can create a ratio from the values of two fields as pass along as a single field.
- $project produces 1 output document for every input document it sees.

A Twitter Dataset

- Let's look at some examples that illustrate the MongoDB aggregation framework.
- These examples operate on a collection of tweets.
  - As with any dataset of this type, it's a snapshot in time.
  - It may not reflect the structure of Twitter feeds as they look today.

Tweets Data Model

```
{
    "text" : "Something interesting ...",
    "entities" : {
        "user_mentions" : [
            {
                "screen_name" : "somebody_else",
                ...
            }
        ],
        "urls" : [ ],
        "hashtags" : [ ]
    },
    "user" : {
        "friends_count" : 544,
        "screen_name" : "somebody",
        "followers_count" : 100,
        ...
    },
}
```

Analyzing Tweets

- Imagine the types of analyses one might want to do on tweets.
- It's common to analyze the behavior of users and the networks involved.
- Our examples will focus on this type of analysis

Note:

- We should also mention that our tweet documents actually contain many more fields.
- We are showing just those fields relevant to the aggregations we'll do.

Friends and Followers

- Let's look again at two stages we touched on earlier:
  - `$match`
  - `$project`
- In our dataset:
  - `friends` are those a user follows.
  - `followers` are others that follow a users.
- Using these operators we will write an aggregation pipeline that will:
  - Ignore anyone with no friends and no followers.
  - Calculate who has the highest followers to friends ratio.

Exercise: Friends and Followers

```
db.tweets.aggregate( [
    { $match: { "user.friends_count": { $gt: 0 },
                "user.followers_count": { $gt: 0 } } },
    { $project: { ratio: {$divide: ["$user.followers_count",
                                    "$user.friends_count"]},
                  screen_name : "$user.screen_name"} },
    { $sort: { ratio: -1 } },
    { $limit: 1 } ] )
```

Note:

- Discuss the $match stage
- Discuss the $project stage as a whole
- Remember that with project we can pull a value out of an embedded document and put it at the top level.
- Discuss the ratio projection
- Discuss screen_name projection
- Give an overview of other operators we might use in projections

---

Exercise: $match and $project

- Of the users in the "Brasilia" timezone who have tweeted 100 times or more, who has the largest number of followers?
- Time zone is found in the "time_zone" field of the user object in each tweet.
- The number of tweets for each user is found in the "statuses_count" field.
- A result document should look something like the following:

```
{ _id        : ObjectId('52fd2490bac3fa1975477702'),
  followers  : 2597,
  screen_name: 'marbles',
  tweets     : 12334
}
```

Note:

```
[ { "$match" : { "user.time_zone" : "Brasilia",
                 "user.statuses_count" : {"$gte" : 100} } },
  { "$project" : { "followers" : "$user.followers_count",
                   "tweets" : "$user.statuses_count",
                   "screen_name" : "$user.screen_name" } },
  { "$sort" : { "followers" : -1 } },
  { "$limit" : 1 } ]
```

The Group Stage

- For those coming from the relational world, `$group` is similar to the SQL `GROUP BY` statement.
- $group operations require that we specify which field to group on.
- Documents with the same identifier will be aggregated together.
- With $group, we aggregate values using accumulators[24].

Tweet Source

- The tweets in our twitter collection have a field called `source`.
- This field describes the application that was used to create the tweet.
- Let's write an aggregation pipeline that identifies the applications most frequently used to publish tweets.

Exercise: Tweet Source

```
db.tweets.aggregate( [
    { "$group" : { "_id" : "$source",
                   "count" : { "$sum" : 1 } } },
    { "$sort" : { "count" : -1 } }
] )
```

Group Aggregation Accumulators

Accumulators available in the group stage:

- $sum
- $avg
- $first
- $last
- $max
- $min
- $push
- $addToSet

[24] http://docs.mongodb.org/manual/meta/aggregation-quick-reference/#accumulators

Rank Users by Number of Tweets

- One common task is to rank users based on some metric.
- Let's look at who tweets the most.
- Earlier we did the same thing for tweet source.
  - Group together all tweets by a user for every user in our collection
  - Count the tweets for each user
  - Sort in decreasing order
- Let's add the list of tweets to the output documents.
- Need to use an accumulator that works with arrays.
- Can use either $addToSet or $push.

Exercise: Adding List of Tweets

For each user, aggregate all their tweets into a single array.

```
db.tweets.aggregate( [
   { "$group" : { "_id" : "$user.screen_name",
                  "tweet_texts" : { "$push" : "$text" },
                  "count" : { "$sum" : 1 } } },
   { "$sort" : { "count" : -1 } },
   { "$limit" : 3 }
] )
```

Note:

- $group operations require that we specify which field to group on.
- In this case, we group documents based on the user's screen name.
- With $group, we aggregate values using arithmetic or array operators.
- Here we are counting the number of documents for each screen name.
- We do that by using the $sum operator
- This will add 1 to the count field for each document produced by the $group stage.
- Note that there will be one document produced by $group for each screen name.
- The $sort stage receives these documents as input and sorts them by the value of the count field

---

The Unwind Stage

- In many situations we want to aggregate using values in an array field.
- In our tweets dataset we need to do this to answer the question:
  - "Who includes the most user mentions in their tweets?"
- User mentions are stored within an embedded document for entities.
- This embedded document also lists any urls and hashtags used in the tweet.

Example: User Mentions in a Tweet

```
...
"entities" : {
    "user_mentions" : [
        {
            "indices" : [
                28,
                44
            ],
            "screen_name" : "LatinsUnitedGSX",
            "name" : "Henry Ramirez",
            "id" : 102220662
        }
    ],
    "urls" : [ ],
    "hashtags" : [ ]
 },
 ...
```

Using $unwind

Who includes the most user mentions in their tweets?

```
db.tweets.aggregate(
    { $unwind: "$entities.user_mentions" },
    { $group: { _id: "$user.screen_name",
                count: { $sum: 1 } } },
    { $sort: { count: -1 } },
    { $limit: 1 })
```

Note:

- Many tweets contain multiple user mentions.
- We use unwind to produce one document for each user mention.
- Each of these documents is passed to the $group stage that follows.
- They will be grouped by the user who created the tweet and counted.
- As a result we will have a count of the total number of user mentions made by any one tweeter.

Data Processing Pipelines

- The aggregation framework allows you to create a data processing pipeline.
- You can include as many stages as necessary to achieve your goal.
- For each stage consider:
  - What input that stage must receive
  - What output it should produce.
- Many tasks require us to include more than one stage using a given operator.

Most Unique User Mentions

- We frequently need multiple group stages to achieve our goal.
- We just looked at a pipeline to find the tweeter that mentioned the most users.
- Let's change this so that it is more of a question about a tweeter's active network.
- We might ask which tweeter has mentioned the most unique users in their tweets.

Same Operator ($group), Multiple Stages

Which tweeter has mentioned the most unique users in their tweets?

```
db.tweets.aggregate( [
    { $unwind: "$entities.user_mentions" },
    { $group: {
        _id: "$user.screen_name",
        mset: { $addToSet: "$entities.user_mentions.screen_name"  } } },
    { $unwind: "$mset"},
    { $group: { _id: "$_id", count: { $sum: 1 } } },
    { $sort: { count: -1 } },
    { $limit: 1 }
] )
```

Note:

- We begin as we did before by unwinding user mentions.
- Instead of simple counting them, we aggregate using $addToSet.
- This produces documents that include only unique user mentions.
- We then do another unwind stage to produce a document for each unique user mention.
- And count these in a second $group stage.

The Sort Stage

- Uses the $sort operator
- Works like the `sort()` cursor method
- 1 to sort ascending; -1 to sort descending
- E.g, db.testcol.aggregate([{ $sort : { b : 1, a : -1 } }])

The Skip Stage

- Uses the $skip operator
- Works like the `skip()` cursor method.
- Value is an integer specifying the number of documents to skip.
- E.g, the following will pass all but the first 3 documents to the next stage in the pipeline.
  - db.testcol.aggregate([{ $skip : 3 }, ...])

The Limit Stage

- Used to limit the number of documents passed to the next aggregation stage.
- Works like the `limit()` cursor method.
- Value is an integer.
- E.g., the following will only pass 3 documents to the stage that comes next in the pipeline.
  - db.testcol.aggregate([{ $limit: 3 }, ...])

The Out Stage

- Used to create a new collection from the output of the aggregation pipeline.
- Can only be the last stage in the pipeline.
- If a collection by the name already exists, it replaces that collection.
- Syntax is { $out : “collection_name” }

5.2 Optimizing Aggregation

Learning Objectives

Upon completing this module students should understand:

- Aggregation pipeline options
- Key aspects of resource management during the aggregation pipeline
- How to order aggregation stages to maximize speed and minimize resource usage
- How MongoDB automatically reorders pipeline stages to improve efficiency
- Changes in the aggregation framework from MongoDB 2.4 to 2.6.

Aggregation Options

- You may pass an options document to `aggregate()`.
- Syntax:

```
db.<COLLECTION>.aggregate( [ { stage1 }, { stage2 }, ... ], { options } )
```

- Following are some of the fields that may be passed in the options document.
  - `allowDiskUse : true` - permit the use of disk for memory-intensive queries
  - `explain : true` - display how indexes are used to perform the aggregation.

Aggregation Limits

- An aggregation pipeline cannot use more than 100 MB of RAM.
- `allowDiskUse : true` allows you to get around this limit.
- The follow operators do not require the entire dataset to be in memory:
  - $match, $skip, $limit, $unwind, and $project
  - Stages for these operators are not subject to the 100 MB limit.
  - $unwind can, however, dramatically increase the amount of memory used.
- $group and $sort might require all documents in memory at once.

Limits Prior to MongoDB 2.6

- `aggregate()` returned results in a single document up to 16 MB in size.
- The upper limit on pipeline memory usage was 10% of RAM.

Optimization: Reducing Documents in the Pipeline

- These operators can reduce the number of documents in the pipeline:
  - $match
  - $skip
  - $limit:
- They should be used as early as possible in the pipeline.

Optimization: Sorting

- `$sort` can take advantages of indexes.
- Must be used before any of the following to do this:
  - `$group`
  - `$unwind`
  - `$project`
- After these stages, the fields or their values change.
- `$sort` requires a full scan of the input documents.

Automatic Optimizations

MongoDB will perform some optimizations automatically. For example:

- If a `$project` stage is used late in the pipeline it may be used to eliminate those fields earlier if possible.
- A `$sort` followed by a `$match` will be executed as a `$match` followed by a `$sort` to reduce the number of documents to be sorted.
- A `$skip` followed by a `$limit` will be executed as a `$limit` followed by a `$skip`, with the $limit parameter increased by the $skip amount to allow $sort + $limit coalescence.
- See: Aggregation Pipeline Optimation[25]

5.3 Lab: Aggregation Framework

Exercise: Working with Array Fields

Use the aggregation framework to find the name of the individual who has made the most comments on a blog.

Start by importing the necessary data if you have not already.

```
mongoimport -d blog -c posts --drop posts.json
```

To help you verify your work, the author with the fewest comments is Mariela Sherer and she commented 387 times.

Note:

```
use blog;
db.posts.aggregate([
  { "$unwind": "$comments" },
  { "$group":
    {
      _id: "$comments.author",
      num_comments: { $sum: 1 }
    }
  },
  { "$sort": { "num_comments": -1 } },
  { "$limit": 10 }
])
```

---

[25] http://docs.mongodb.org/manual/core/aggregation-pipeline-optimization/

Exercise: Repeated Aggregation Stages

Import the zips.json file from the data handouts provided:

```
mongoimport -d test -c zips --drop zips.json
```

Consider together cities in the states of California (CA) and New York (NY) with populations over 25,000. Calculate the average population of this sample of cities.

Please note:

- Different states might have the same city name.
- A city might have multiple zip codes.

Note:

```
db.zips.aggregate( [
    { $match: { state: { $in: ["CA", "NY"] } } },
    { $group: { _id: { state: "$state", city: "$city" },
                pop: {$sum: "$pop"} } },
    { $match: { pop: {$gt: 25000} } },
    { $group: { _id: null,
                pop: {$avg: "$pop"} } }
] )
```

Exercise: Projection

Calculate the total number of people who live in a zip code in the US where the city starts with a digit.

`$project` can extract the first digit from any field. E.g.,

```
db.zips.aggregate([
    {$project:
        {
            first_char: { $substr: ["$city", 0, 1] },
        }
    }
])
```

Note:

```
db.zips.aggregate( [
   { $project : {
       first_char: { $substr: [ "$city", 0, 1] },
       pop: 1,
       city: "$city",
       zip: "$_id",
       state: 1
   } },
   { $match : {
       first_char: { $in: [ '0', '1', '2', '3', '4', '5', '6', '7', '8', '9'] }
   } },
   {
      "$group" : { _id: null,
                   population: { $sum : "$pop"} }
   }
] )
```

Exercise: Descriptive Statistics

From the `grades` collection, find the class (display the `class_id`) with the best average student performance.

If you have not already done so, import the grades collection as follows.

```
mongoimport -d test -c grades --drop grades.json
```

Before you attempt this exercise, explore the `grades` collection a little to ensure you understand how it is structured.

For additional exercises, consider other statistics you might want to see with this data and how to calculate them.

6 Introduction to Schema Design

Schema Design Core Concepts **(page 89)** An introduction to schema design in MongoDB

Schema Evolution **(page 96)** Considerations for evolving a MongoDB schema design over an application's lifetime

Common Schema Design Patterns **(page 100)** Common design patterns for representing 1-1, 1-M, and M-M relationships and tree structures in MongoDB

6.1 Schema Design Core Concepts

Learning Objectives

Upon completing this module, students should understand:

- Basic schema design principles for MongoDB
- Tradeoffs for embedded documents in a schema
- Tradeoffs for linked documents in a schema
- The use of array fields as part of a schema design

What is a schema?

- Maps concepts and relationships to data
- Sets expectations for the data
- Minimizes overhead of iterative modifications
- Ensures compatibility

Example: Normalized Data Model

```
User:           Book:           Author:
- username      - title         - firstName
- firstName     - isbn          - lastName
- lastName      - language
                - createdBy
                - author
```

Example: Denormalized Version

```
User:           Book:
- username      - title
- firstName     - isbn
- lastName      - language
                - createdBy
                - author
                  - firstName
                  - lastName
```

Schema Design in MongoDB

- Schema is defined at the application-level
- Design is part of each phase in its lifetime
- There is no magic formula

Three Considerations

- The data your application needs
- Your application's read usage of the data
- Your application's write usage of the data

Case Study

- A Library Web Application
- Different schemas are possible.

Author Schema

```
{   "_id": int,
    "firstName": string,
    "lastName": string
}
```

User Schema

```
{   "_id": int,
    "username": string,
    "password": string
}
```

Book Schema

```
 {  "_id": int,
    "title": string,
    "slug": string,
    "author": int,
    "available": boolean,
    "isbn": string,
    "pages": int,
    "publisher": {
        "city": string,
        "date": date,
        "name": string
    },
    "subjects": [ string, string ],
    "language": string,
    "reviews": [ { "user": int, "text": string },
                 { "user": int, "text": string } ]
}
```

Example Documents: Author

```
{   _id: 1,
    firstName: "F. Scott",
    lastName: "Fitzgerald"
}
```

Example Documents: User

```
{   _id: 1,
    username: "emily@10gen.com",
    password: "slsjfk4odk84k209dlkdj90009283d"
}
```

Example Documents: Book

```
{   _id: 1,
    title: "The Great Gatsby",
    slug: "9781857150193-the-great-gatsby",
    author: 1,
    available: true,
    isbn: "9781857150193",
    pages: 176,
    publisher: {
        name: "Everyman's Library",
        date: ISODate("1991-09-19T00:00:00Z"),
        city: "London"
    },
    subjects: ["Love stories", "1920s", "Jazz Age"],
    language: "English",
    reviews: [
        { user: 1, text: "One of the best..." },
        { user: 2, text: "It's hard to..." }
    ]
}
```

Embedded Documents

- AKA sub-documents or embedded objects
- What advantages do they have?
- When should they be used?

Example: Embedded Documents

```
{   ...
    publisher: {
        name: "Everyman's Library",
        date: ISODate("1991-09-19T00:00:00Z"),
        city: "London"
    },
    subjects: ["Love stories", "1920s", "Jazz Age"],
    language: "English",
    reviews: [
        { user: 1, text: "One of the best..." },
        { user: 2, text: "It's hard to..." }
    ]
}
```

Embedded Documents: Pros and Cons

- Great for read performance
- One seek to find the document
- At most, one sequential read to retrieve from disk
- Writes can be slow if constantly adding to objects

Linked Documents

- What advantages does this approach have?
- When should they be used?

Example: Linked Documents

```
{    ...
     author: 1,
     reviews: [
         { user: 1, text: "One of the best..." },
         { user: 2, text: "It's hard to..." }
     ]
}
```

Linked Documents: Pros and Cons

- More, smaller documents
- Can make queries by ID very simple
- Accessing linked documents requires extra seeks + reads.
- What effect does this have on the system?

Arrays

- Array of scalars
- Array of documents

Array of Scalars

```
{   ...
    subjects: ["Love stories", "1920s", "Jazz Age"],
}
```

Array of Documents

```
{   ...
    reviews: [
        { user: 1, text: "One of the best..." },
        { user: 2, text: "It's hard to..." }
    ]
}
```

Exercise: Users and Book Reviews

Design a schema for users and their book reviews. Usernames are immutable.

- Users
  - username (string)
  - email (string)
- Reviews
  - text (string)
  - rating (integer)
  - created_at (date)

Solution A: Users and Book Reviews

Reviews may be queried by user or book

```
// db.users (one document per user)
{   _id: ObjectId("..."),
    username: "bob",
    email: "bob@example.com"
}

// db.reviews (one document per review)
{   _id: ObjectId("..."),
    user: ObjectId("..."),
    book: ObjectId("..."),
    rating: 5,
    text: "This book is excellent!",
    created_at: ISODate("2012-10-10T21:14:07.096Z")
}
```

Solution B: Users and Book Reviews

Optimized to retrieve reviews by user

```
// db.users, one document per user with all reviews
{   _id: ObjectId("..."),
    username: "bob",
    email: "bob@example.com",
    reviews: [
      {   book: ObjectId("..."),
          rating: 5,
          text: "This book is excellent!",
          created_at: ISODate("2012-10-10T21:14:07.096Z")
      }
    ]
}
```

Solution C: Users and Book Reviews

Optimized to retrieve reviews by book

```
// db.users (one document per user)
{   _id: ObjectId("..."),
    username: "bob",
    email: "bob@example.com"
}

// db.books, one document per book with all reviews
{   _id: ObjectId("..."),
    // Other book fields...
    reviews: [
      {   user: ObjectId("..."),
          rating: 5,
          text: "This book is excellent!",
          created_at: ISODate("2014-11-10T21:14:07.096Z")
      }
    ]
}
```

Store Binary Files in MongoDB with GridFS

- Application may have a requirement for binary file storage
- GridFS is a specification for storing files larger than 16MB in MongoDB
- Handled automatically by most drivers
- “mongofiles” is the command line tool for working with GridFS

How GridFS Works

- Files are split into chunks
- Default chunk size is 255k
- fs.files collection stores meta data for the file (name, size, etc.)
- fs.chunks collection stores chunks for binary file

Schema Design Use Cases with GridsFS

- Store large video files and stream chunks to a user
- Enterprise assets, replicated across data centers
- Medical record attachments (x-rays, reports, etc.)

6.2 Schema Evolution

Learning Objectives

Upon completing this module, students should understand the basic philosophy of evolving a MongoDB schema during an application's lifetime:

- Development Phase
- Production Phase
- Iterative Modifications

Development Phase

Support basic CRUD functionality:

- Inserts for authors and books
- Find authors by name
- Find books by basics of title, subject, etc.

Development Phase: Known Query Patterns

```
// Find authors by last name.
db.authors.createIndex({ "lastName": 1 })

// Find books by slug for detail view
db.books.createIndex({ "slug": 1 })

// Find books by subject (multi-key)
db.books.createIndex({ "subjects": 1 })

// Find books by publisher (index on embedded doc)
db.books.createIndex({ "publisher.name": 1 })
```

Production Phase

Evolve the schema to meet the application's read and write patterns.

Production Phase: Read Patterns

List books by author last name

```
authors = db.authors.find({ lastName: /^f.*/i }, { _id: 1 });

authorIds = authors.map(function(x) { return x._id; });

db.books.find({author: { $in: authorIds }});
```

Addressing List Books by Last Name

"Cache" the author name in an embedded document.

```
{
    _id: 1,
    title: "The Great Gatsby",
    author: {
        firstName: "F. Scott",
        lastName: "Fitzgerald"
    }
    // Other fields follow...
}
```

Queries are now one step

```
db.books.find({ "author.firstName": /^f.*/i })
```

Production Phase: Write Patterns

Users can review a book.

```
review = {
    user: 1,
    text: "I thought this book was great!",
    rating: 5
};

db.books.update(
    { _id: 3 },
    { $push: { reviews: review }}
);
```

Caveats:

- Document size limit (16MB)
- Storage fragmentation after many updates/deletes

Exercise: Recent Reviews

- Display the 10 most recent reviews by a user.
- Make efficient use of memory and disk seeks.

Solution: Recent Reviews, Schema

Store users' reviews in monthly buckets.

```
// db.reviews (one document per user per month)
{   _id: "bob-201412",
    reviews: [
        {   _id: ObjectId("..."),
            rating: 5,
            text: "This book is excellent!",
            created_at: ISODate("2014-12-10T21:14:07.096Z")
        },
        {   _id: ObjectId("..."),
            rating: 2,
            text: "I didn't really enjoy this book.",
            created_at: ISODate("2014-12-11T20:12:50.594Z")
        }
    ]
}
```

Solution: Recent Reviews, Update

Adding a new review to the appropriate bucket

```
myReview = {
    _id: ObjectId("..."),
    rating: 3,
    text: "An average read.",
    created_at: ISODate("2012-10-13T12:26:11.502Z")
};

db.reviews.update(
      { _id: "bob-201210" },
      { $push: { reviews: myReview }}
);
```

Solution: Recent Reviews, Read

Display the 10 most recent reviews by a user

```
cursor = db.reviews.find(
    { _id: /^bob-/ },
    { reviews: { $slice: -10 }}
).sort({ _id: -1 }).batchSize(5);

num = 0;

while (cursor.hasNext() && num < 10) {
    doc = cursor.next();

    for (var i = 0; i < doc.reviews.length && num < 10; ++i, ++num) {
        printjson(doc.reviews[i]);
    }
}
```

Solution: Recent Reviews, Delete

Deleting a review

```
cursor = db.reviews.update(
    { _id: "bob-201210" },
    { $pull: { reviews: { _id: ObjectId("...") }}}
);
```

6.3 Common Schema Design Patterns

Learning Objectives

Upon completing this module students should understand common design patterns for modeling:

- One-to-One Relationships
- One-to-Many Relationships
- Many-to-Many Relationships
- Tree Structures

One-to-One Relationship

Let's pretend that authors only write one book.

One-to-One: Linking

Either side, or both, can track the relationship.

```
db.books.findOne()
{
    _id: 1,
    title: "The Great Gatsby",
    slug: "9781857150193-the-great-gatsby",
    author: 1,
    // Other fields follow...
}

db.authors.findOne({ _id: 1 })
{
    _id: 1,
    firstName: "F. Scott",
    lastName: "Fitzgerald"
    book: 1,
}
```

One-to-One: Embedding

```
db.books.findOne()
{
    _id: 1,
    title: "The Great Gatsby",
    slug: "9781857150193-the-great-gatsby",
    author: {
        firstName: "F. Scott",
        lastName: "Fitzgerald"
    }
    // Other fields follow...
}
```

One-to-Many Relationship

In reality, authors may write multiple books.

One-to-Many: Array of IDs

The “one” side tracks the relationship.

- Flexible and space-efficient
- Additional query needed for non-ID lookups

```
db.authors.findOne()
{
    _id: 1,
    firstName: "F. Scott",
    lastName: "Fitzgerald",
    books: [1, 3, 20]
}
```

One-to-Many: Single Field with ID

The “many” side tracks the relationship.

```
db.books.find({ author: 1 })
{
    _id: 1,
    title: "The Great Gatsby",
    slug: "9781857150193-the-great-gatsby",
    author: 1,
    // Other fields follow...
}

{
    _id: 3,
    title: "This Side of Paradise",
    slug: "9780679447238-this-side-of-paradise",
    author: 1,
    // Other fields follow...
}
```

One-to-Many: Array of Documents

```
db.authors.findOne()
{
    _id: 1,
    firstName: "F. Scott",
    lastName: "Fitzgerald",
    books: [
        { _id: 1, title: "The Great Gatsby" },
        { _id: 3, title: "This Side of Paradise" }
    ]
    // Other fields follow...
}
```

Many-to-Many Relationship

Some books may also have co-authors.

Many-to-Many: Array of IDs on Both Sides

```
db.books.findOne()
{
    _id: 1,
    title: "The Great Gatsby",
    authors: [1, 5]
    // Other fields follow...
}

db.authors.findOne()
{
    _id: 1,
    firstName: "F. Scott",
    lastName: "Fitzgerald",
    books: [1, 3, 20]
}
```

Many-to-Many: Array of IDs on Both Sides

Query for all books by a given author.

```
db.books.find({ authors: 1 });
```

Query for all authors of a given book.

```
db.authors.find({ books: 1 });
```

Many-to-Many: Array of IDs on One Side

```
db.books.findOne()
{
    _id: 1,
    title: "The Great Gatsby",
    authors: [1, 5]
    // Other fields follow...
}

db.authors.find({ _id: { $in: [1, 5] }})
{
    _id: 1,
    firstName: "F. Scott",
    lastName: "Fitzgerald"
}
{
    _id: 5,
    firstName: "Unknown",
    lastName: "Co-author"
}
```

Many-to-Many: Array of IDs on One Side

Query for all books by a given author.

```
db.books.find({ authors: 1 });
```

Query for all authors of a given book.

```
book = db.books.findOne(
    { title: "The Great Gatsby" },
    { authors: 1 }
);

db.authors.find({ _id: { $in: book.authors }});
```

Tree Structures

E.g., modeling a subject hierarchy.

Allow users to browse by subject

```
db.subjects.findOne()
{
    _id: 1,
    name: "American Literature",
    sub_category: {
         name: "1920s",
         sub_category: { name: "Jazz Age" }
   }
}
```

- How can you search this collection?
- Be aware of document size limitations
- Benefit from hierarchy being in same document

Alternative: Parents and Ancestors

```
db.subjects.find()
{   _id: "American Literature" }

{   _id : "1920s",
    ancestors: ["American Literature"],
    parent: "American Literature"
}

{   _id: "Jazz Age",
    ancestors: ["American Literature", "1920s"],
    parent: "1920s"
}

{   _id: "Jazz Age in New York",
    ancestors: ["American Literature", "1920s", "Jazz Age"],
    parent: "Jazz Age"
}
```

Find Sub-Categories

```
db.subjects.find({ ancestors: "1920s" })
{
    _id: "Jazz Age",
    ancestors: ["American Literature", "1920s"],
    parent: "1920s"
}

{
    _id: "Jazz Age in New York",
    ancestors: ["American Literature", "1920s", "Jazz Age"],
    parent: "Jazz Age"
}
```

Summary

- Schema design is different in MongoDB.
- Basic data design principles apply.
- It's about your application.
- It's about your data and how it's used.
- It's about the entire lifetime of your application.

7 Replica Sets

Introduction to Replica Sets **(page 105)** An introduction to replication and replica sets

Elections in Replica Sets **(page 109)** The process of electing a new primary (automated failover) in replica sets

Replica Set Roles and Configuration **(page 115)** Configuring replica set members for common use cases

The Oplog: Statement Based Replication **(page 117)** The process of replicating data from one node of a replica set to another

Lab: Working with the Oplog **(page 119)** A brief lab that illustrates how the oplog works

Write Concern **(page 121)** Balancing performance and durability of writes

Read Preference **(page 126)** Configuring clients to read from specific members of a replica set

Lab: Setting up a Replica Set **(page 127)** Launching members, configuring, and initiating a replica set

7.1 Introduction to Replica Sets

Learning Objectives

Upon completing this module, students should understand:

- Striking the right balance between cost and redundancy
- The many scenarios replication addresses and why
- How to avoid downtime and data loss using replication

Use Cases for Replication

- High Availability
- Disaster Recovery
- Functional Segregation

High Availability (HA)

- Data still available following:
  - Equipment failure (e.g. server, network switch)
  - Datacenter failure
- This is achieved through automatic failover.

Note: If we lose a server and MongoDB is correctly configured:

- Our database system can still service reads and writes, but by default not during failover period when the election takes place and there is no primary.
- Without manual intervention as long as there is still a majority of nodes available.

Disaster Recovery (DR)

- We can duplicate data across:
  - Multiple database servers
  - Storage backends
  - Datacenters
- Can restore data from another node following:
  - Hardware failure
  - Service interruption

Functional Segregation

There are opportunities to exploit the topology of a replica set:

- Based on physical location (e.g. rack or datacenter location)
- For analytics, reporting, data discovery, system tasks, etc.
- For backups

Note:

- You may direct particular queries to specific nodes (which may have different indexes or hardware) to increase overall performance.
- Backup data from secondaries to avoid performance penalties on the primary, especially when using tools like `mongodump` which are I/O intensive and evict the working set from memory (significant when data size is larger than RAM and disks are slow).
- Dedicate secondaries for other purposes such as analytics jobs.

Large Replica Sets

Functional segregation can be further exploited by using large replica sets.

- 50 node replica set limit
- Useful for deployments with a large number of data centers or offices
- Read only workloads can position secondaries in data centers around the world (closer to application servers)

Note:

- Sample use case: bank reference data distributed to 20+ data centers around the world, then consumed by the local application server

---

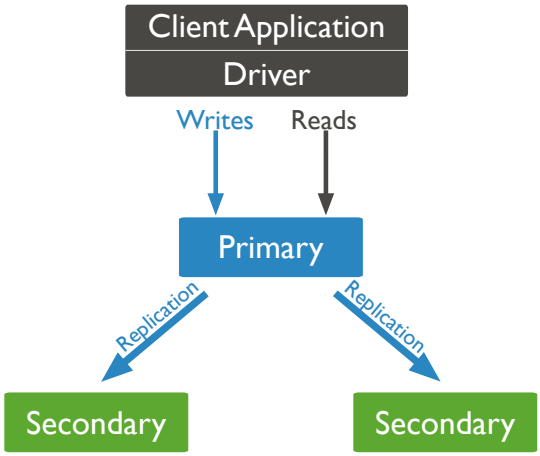
Replication is Not Designed for Scaling

- Can be used for scaling reads, but generally not recommended.
- Drawbacks include:
  - Eventual consistency
  - Not scaling writes
  - Potential system overload when secondaries are unavailable
- Consider sharding for scaling reads and writes.

Note:

- Eventual consistency: This is often tricky to handle as most developers are used to the strong consistency that most databases (and MongoDB) have by default. It also raises the question of how stale the data can be, and what to do when it crosses the threshold (e.g. fall back to reading from the primary).
- Potential system overload: For example, consider a 3 data node replica set using secondaries to scale reads. Each node is serving reads at 70% of its capacity, which is a reasonable utilization rate. What happens if one of the secondaries fail or is intentionally taken down for maintenance (e.g. upgrading the OS or MongoDB)? Even if the load splits evenly between the 2 remaining nodes, they will be at `70+(70/2)= 105%` capacity.

Replica Sets



Note:

- MongoDB implements replication in the form of replica sets. Don't use the term master-slave as that is what we had before replica sets. It still exists for some corner cases (e.g. > 12 replicas) but should otherwise be avoided.
- A replica set consists of one or more `mongod` servers. Maximum 50 nodes in total and up to 7 with votes.
- There is at most one `mongod` that is "primary" at any one time (though there are edge cases/bugs when there is more than one).
- There are usually two or more other `mongod` instances that are secondaries.
- Secondaries may become primary if there is a failover event of some kind.
- Failover is automatic when correctly configured and a majority of nodes remain.

- The secondaries elect a new primary automatically. A primary may also voluntarily step down, like when it can no longer reach the majority of nodes to avoid a potential split brain scenario.

Primary Server

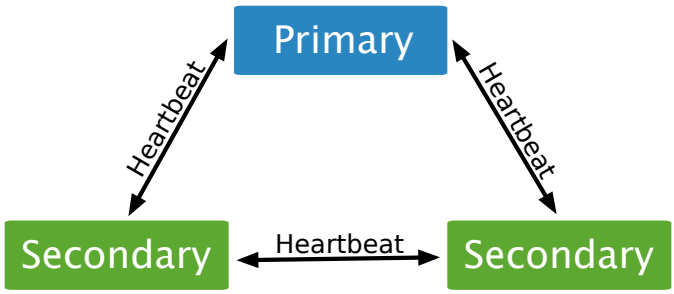
- Clients send writes the primary only.
- MongoDB, Inc. maintains client drivers in many programming languages like Java, C#, Python, Ruby, and PHP.
- MongoDB drivers are replica set aware.

Note: If the primary for a replica set changes from one node to another, the driver will automatically route writes to the correct `mongod`.

Secondaries

- A secondary replicates operations from another node in the replica set.
- Secondaries usually replicate from the primary.
- Secondaries may also replicate from other secondaries. This is called replication chaining.
- A secondary may become primary as a result of a failover scenario.

Heartbeats



Note:

- The members of a replica set use heartbeats to determine if they can reach every other node.
- The heartbeats are sent every two seconds.
- If a node is unreachable, this may indicate server failure, a network partition, or simply too slow to respond. The heartbeat will timeout and retried several times before the state is updated.

The Oplog

- The operations log, or oplog, is a special capped collection that is the basis for replication.
- The oplog maintains one entry for each document affected by every write operation.
- Secondaries copy operations from the oplog of their sync source.

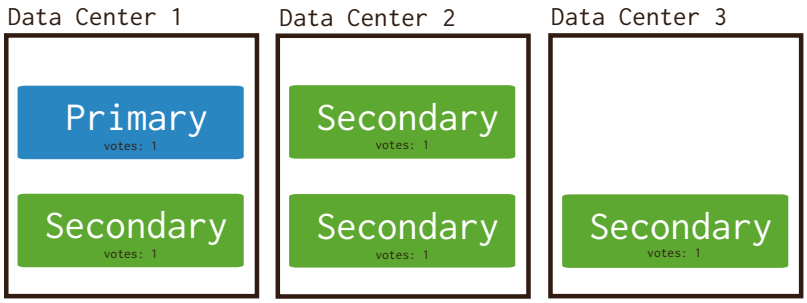
7.2 Elections in Replica Sets

Learning Objectives

Upon completing this module students should understand:

- That elections enable automated failover in replica sets
- How votes are distributed to members
- What prompts an election
- How a new primary is selected

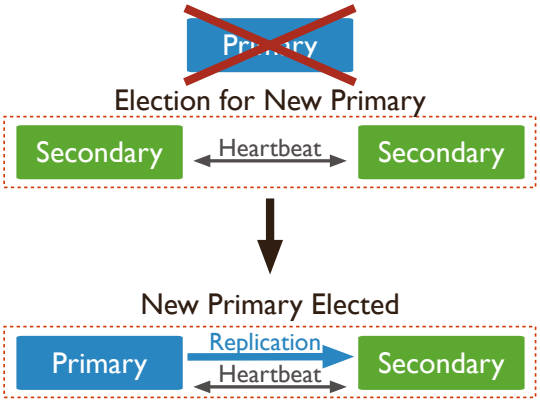
Members and Votes



Note:

- In order for writes to occur, one member of a replica set must be primary.
- In the event the current primary becomes unavailable, the remaining members elect a new primary.
- Voting members of replica set each get one vote.
- Up to seven members may be voting members.
- This enables MongoDB to ensure elections happen quickly, but enables distribution of votes to different data centers.
- In order to be elected primary a server must have a true majority of votes.
- A member must have greater than 50% of the votes in order to be elected primary.

Calling Elections



Note:

- MongoDB uses a consensus protocol to determine when an election is required.
- Essentially, an election will occur if there is no primary.
- Upon initiation of a new replica set the members will elect a primary.
- If a primary steps down the set will hold an election.
- A secondary will call for an election if it does not receive a response to a heartbeat sent to the primary after waiting for 10 seconds.
- If other members agree that the primary is not available, an election will be held.

Selecting a New Primary

Three factors are important in the selection of a primary:

- Priority
- Optime
- Connections

Priority

- The higher its priority, the more likely a member is to become primary.
- The default is 1.
- Servers with a priority of 0 will never become primary.
- Priority values are floating point numbers 0 - 1000 inclusive.

Note:

- Priority is a configuration parameter for replica set members.
- Use priority to determine where writes will be directed by default.
- And where writes will be directed in case of failover.

- Generally all identical nodes in a datacenter should have the same priority to avoid unnecessary failovers. For example, when a higher priority node rejoins the replica set after a maintenance or failure event, it will trigger a failover (during which by default there will be no reads and writes) even though it is unnecessary.
- More on this in a later module.

Optime

- Optime: Operation time, which is the timestamp of the last operation the member applied from the oplog.
- To be elected primary, a member must have the most recent optime.
- Only optimes of visible members are compared.

Connections

- Must be able to connect to a majority of the members in the replica set.
- Majority refers to the total number of votes.
- Not the total number of members.

Note: To be elected primary, a replica set member must be able to connect to a majority of the members in the replica set.

When will a primary step down?

- After receiving the `replSetStepDown` or `rs.stepDown()` command.
- If a secondary is eligible for election and has a higher priority.
- If it cannot contact a majority of the members of the replica set.

replSetStepDown Behavior

- Primary will attempt to terminate long running operations before stepping down.
- Primary will wait for electable secondary to catch up before stepping down.
- "secondaryCatchUpPeriodSecs" can be specified to limit the amount of time the primary will wait for a secondary to catch up before the primary steps down.

Note:

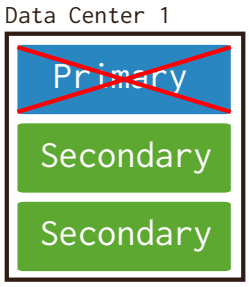
- Ask the class what the tradeoffs could be in setting secondaryCatchUpPeriodSecs to a very short amount of time (rollbacks could occur or operations not replicated)

Exercise: Elections in Failover Scenarios

- We have learned about electing a primary in replica sets.
- Let's look at some scenarios in which failover might be necessary.

Scenario A: 3 Data Nodes in 1 DC

Which secondary will become the new primary?

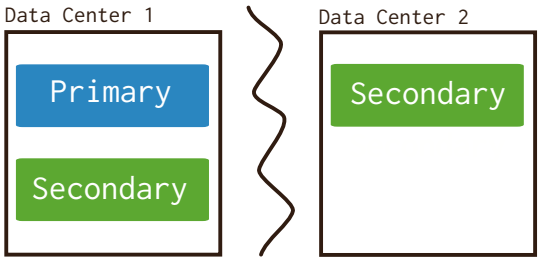


Note:

- It depends on the priorities of the secondaries.
- And on the optime.

Scenario B: 3 Data Nodes in 2 DCs

Which member will become primary following this type of network partition?

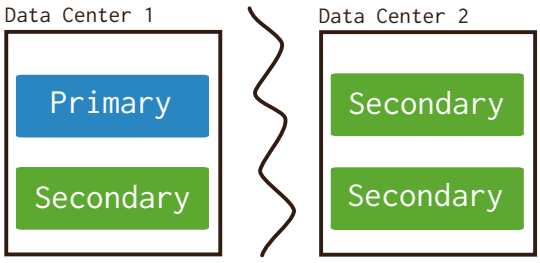


Note:

- The current primary is likely to remain primary.
- It probably has the highest priority.
- If DC2 fails, we still have a primary.
- If DC1 fails, we won't have a primary automatically. The remaining node in DC2 needs to be manually promoted by reconfiguring the replica set.

Scenario C: 4 Data Nodes in 2 DCs

What happens following this network partition?

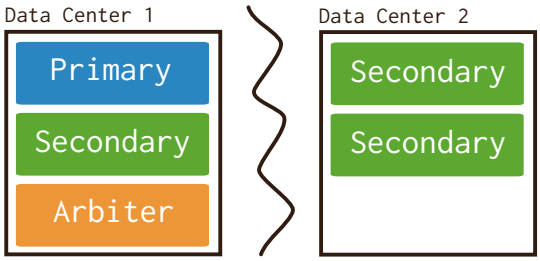


Note:

- We enter a state with no primary.
- Each side of the network partition has only 2 votes (not a majority).
- All the servers assume secondary status.
- This is avoidable.
- One solution is to add another member to the replica set.
- If another data node can not be provisioned, MongoDB has a special alternative called an arbiter that requires minimal resources.
- An arbiter is a `mongod` instance without data and performs only heartbeats, votes, and vetoes.

Scenario D: 5 Nodes in 2 DCs

The following is similar to Scenario C, but with the addition of an arbiter in Data Center 1. What happens here?

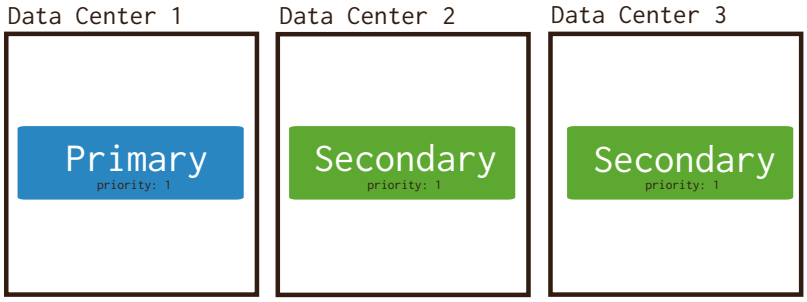


Note:

- The current primary is likely to remain primary.
- The arbiter helps ensure that the primary can reach a majority of the replica set.

Scenario E: 3 Data Nodes in 3 DCs

- What happens here if any one of the nodes/DCs fail?
- What about recovery time?

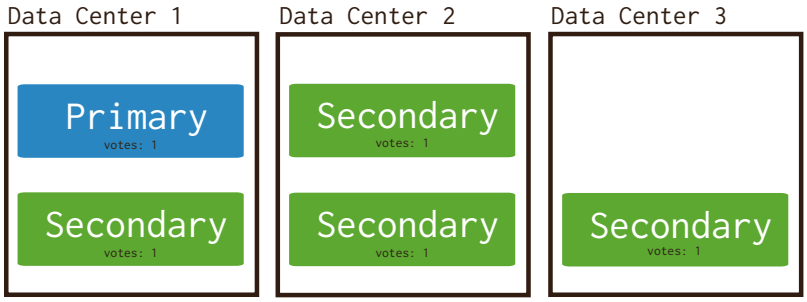


Note:

- The intent is to explain the advantage of deploying to 3 DCs - it's the minimum number of DCs in order for MongoDB to automatically failover if any one DC fails. This is generally what we recommend to customers in our consult and health check reports, though many continue to use 2 DCs due to costs and legacy reasons.
- To have automated failover in the event of single DC level failure, there must be at least 3 DCs. Otherwise the DC with the minority of nodes must be manually reconfigured.
- One of the data nodes can be replaced by an arbiter to reduce costs.

Scenario F: 5 Data Nodes in 3 DCs

What happens here if any one of the nodes/DCs fail? What about recovery time?



Note:

- Adds another data node to each "main" DC to reduce typically slow and costly cross DC network traffic if an initial sync or similar recovery is needed, as the recovering node can pull from a local replica instead.
- Depending on the data sizes, operational budget, and requirements, this can be overkill.
- The data node in DC3 can be replaced by an arbiter to reduce costs.

7.3 Replica Set Roles and Configuration

Learning Objectives

Upon completing this module students should understand:

- The use of priority to preference certain members or datacenters as primaries.
- Hidden members.
- The use of hidden secondaries for data analytics and other purposes (when secondary reads are used).
- The use of slaveDelay to protect against operator error.

Example: A Five-Member Replica Set Configuration

- For this example application, there are two datacenters.
- We name the hosts accordingly: `dc1-1`, `dc1-2`, `dc2-1`, etc.
  - This is just a clarifying convention for this example.
  - MongoDB does not care about host names except to establish connections.
- The nodes in this replica set have a variety of roles in this application.

Configuration

```
conf = {              // 5 data-bearing nodes
  _id: "mySet",
  members: [
    { _id : 0, host : "dc1-1.example.net:27017", priority : 5 },
    { _id : 1, host : "dc1-2.example.net:27017", priority : 5 },
    { _id : 2, host : "dc2-1.example.net:27017" },
    { _id : 3, host : "dc1-3.example.net:27017", hidden : true },
    { _id : 4, host : "dc2-2.example.net:27017", hidden : true,
      slaveDelay: 7200 }
  ]
}
```

Principal Data Center

```
{ _id : 0, host : "dc1-1.example.net", priority : 5 },
{ _id : 1, host : "dc1-2.example.net", priority : 5 },
```

Note:

- The objective with the priority settings for these two nodes is to prefer to DC1 for writes.
- The highest priority member that is up to date will be elected primary.
- Up to date means the member's copy of the oplog is within 10 seconds of the primary.
- If a member with higher priority than the primary is a secondary because it is is not up to date, but eventually catches up, it will force an election and win.

Data Center 2

```
{ _id : 2, host : "dc2-1.example.net:27017" },
```

Note:

- Priority is not specified, so it is at the default of 1.
- `dc2-1` could become primary, but only if both `dc1-1` and `dc1-2` are down.
- If there is a network partition and clients can only reach DC2, we can manually failover to `dc2-1`.

What about `dc1-3` and `dc2-2`?

```
// Both are hidden.
// Clients will not distribute reads to hidden members.
// We use hidden members for dedicated tasks.
{ _id : 3, host : "dc1-3.example.net:27017", hidden : true },
{ _id : 4, host : "dc2-2.example.net:27017", hidden : true,
  slaveDelay: 7200 }
```

Note:

- Will replicate writes normally.
- We would use this node to pull reports, run analytics, etc.
- We can do so without paying a performance penalty in the application for either reads or writes.

What about `dc2-2`?

```
{ _id : 4, host : "dc2-2.example.net:27017", hidden : true,
  slaveDelay : 7200 }
```

Note:

- `slaveDelay` permits us to specify a time delay (in seconds) for replication.
- In this case it is 7200 seconds or 2 hours.
- slaveDelay allows us to use a node as a short term protection against operator error:
  - Fat fingering – for example, accidentally dropping a collection in production.
  - Other examples include bugs in an application that result in corrupted data.
  - Not recommended. Use proper backups instead as there is no optimal delay value. E.g. 2 hours might be too long or too short depending on the situation.

7.4 The Oplog: Statement Based Replication

Learning Objectives

Upon completing this module students should understand:

- Binary vs. statement-based replication.
- How the oplog is used to support replication.
- How operations in MongoDB are translated into operations written to the oplog.
- Why oplog operations are idempotent.
- That the oplog is a capped collection and the implications this holds for syncing members.

Binary Replication

- MongoDB replication is statement based.
- Contrast that with binary replication.
- With binary replication we would keep track of:
  - The data files
  - The offsets
  - How many bytes were written for each change
- In short, we would keep track of actual bytes and very specific locations.
- We would simply replicate these changes across secondaries.

Tradeoffs

- The good thing is that figuring out where to write, etc. is very efficient.
- But we must have a byte-for-byte match of our data files on the primary and secondaries.
- The problem is that this couples our replica set members in ways that are inflexible.
- Binary replication may also replicate disk corruption.

Note:

- Some deployments might need to run different versions of MongoDB on different nodes.
- Different versions of MongoDB might write to different file offsets.
- We might need to run a compaction or repair on a secondary.
- In many cases we want to do these types of maintenance tasks independently of other nodes.

Statement-Based Replication

- Statement-based replication facilitates greater independence among members of a replica set.
- MongoDB stores a statement for every operation in a capped collection called the `oplog`.
- Secondaries do not simply apply exactly the operation that was issued on the primary.

Example

Suppose the following remove is issued and it deletes 100 documents:

```
db.foo.remove({ age : 30 })
```

This will be represented in the oplog with records such as the following:

```
{ "ts" : Timestamp(1407159845, 5), "h" : NumberLong("-704612487691926908"),
  "v" : 2, "op" : "d", "ns" : "bar.foo", "b" : true, "o" : { "_id" : 65 } }
{ "ts" : Timestamp(1407159845, 1), "h" : NumberLong("6014126345225019794"),
  "v" : 2, "op" : "d", "ns" : "bar.foo", "b" : true, "o" : { "_id" : 333 } }
{ "ts" : Timestamp(1407159845, 4), "h" : NumberLong("8178791764238465439"),
  "v" : 2, "op" : "d", "ns" : "bar.foo", "b" : true, "o" : { "_id" : 447 } }
{ "ts" : Timestamp(1407159845, 3), "h" : NumberLong("-1707391001705528381"),
  "v" : 2, "op" : "d", "ns" : "bar.foo", "b" : true, "o" : { "_id" : 1033 } }
{ "ts" : Timestamp(1407159845, 2), "h" : NumberLong("-6814297392442406598"),
  "v" : 2, "op" : "d", "ns" : "bar.foo", "b" : true, "o" : { "_id" : 9971 } }
```

Replication Based on the Oplog

- One statement per document affected by each write: insert, update, or delete.
- Provides a level of abstraction that enables independence among the members of a replica set:
  - With regard to MongoDB version.
  - In terms of how data is stored on disk.
  - Freedom to do maintenance without the need to bring the entire set down.

Note:

- Can do maintenance without bringing the set down because statement-based replication does not depend on all nodes running the same version of MongoDB or other restrictions that may be imposed by binary replication.
- In the next exercise, we will see that the oplog is designed so that each statement is idempotent.
- This feature has several benefits for independent operation of nodes in replica sets.

Operations in the Oplog are Idempotent

- Each operation in the oplog is idempotent.
- Whether applied once or multiple times it produces the same result.
- Necessary if you want to be able to copy data while simultaneously accepting writes.

Note: We need to be able to copy while accepting writes when:

- Doing an initial sync for a new replica set member.
- When a member rejoins a replica set after a network partition a member might end up writing operations it had already received prior to the partition.

The Oplog Window

- Oplogs are capped collections.
- Capped collections are fixed-size.
- They guarantee preservation of insertion order.
- They support high-throughput operations.
- Like circular buffers, once a collection fills its allocated space:
  - It makes room for new documents.
  - By overwriting the oldest documents in the collection.

Sizing the Oplog

- The oplog should be sized to account for latency among members.
- The default size oplog is usually sufficient.
- But you want to make sure that your oplog is large enough:
  - So that the oplog window is large enough to support replication
  - To give you a large enough history for any diagnostics you might wish to run.

7.5 Lab: Working with the Oplog

Create a Replica Set

Let's take a look at a concrete example. Launch mongo shell as follows.

```
mongo --nodb
```

Create a replica set by running the following command in the mongo shell.

```
replicaSet = new ReplSetTest( { nodes : 3 } )
```

ReplSetTest

- ReplSetTest is useful for experimenting with replica sets as a means of hands-on learning.
- It should never be used in production. Never.
- The command above will create a replica set with three members.
- It does not start the mongods, however.
- You will need to issue additional commands to do that.

Start the Replica Set

Start the mongod processes for this replica set.

```
replicaSet.startSet()
```

Issue the following command to configure replication for these mongods. You will need to issue this while output is flying by in the shell.

```
replicaSet.initiate()
```

Status Check

- You should now have three mongods running on ports 31000, 31001, and 31002.
- You will see log statements from all three printing in the current shell.
- To complete the rest of the exercise, open a new shell.

Connect to the Primary

Open a new shell, connecting to the primary.

```
mongo --port 31000
```

Create some Inventory Data

Use the `store` database:

```
use store
```

Add the following inventory:

```
inventory = [ { _id: 1, inStock: 10 }, { _id: 2, inStock: 20 },
              { _id: 3, inStock: 30 }, { _id: 4, inStock: 40 },
              { _id: 5, inStock: 50 }, { _id: 6, inStock: 60 } ]
db.products.insert(inventory)
```

Perform an Update

Issue the following update. We might issue this update after a purchase of three items.

```
db.products.update({ _id: { $in: [ 2, 5 ] } },
                   { $inc: { inStock : -1 } },
                   { multi: true })
```

View the Oplog

The oplog is a capped collection in the `local` database of each replica set member:

```
use local
db.oplog.rs.find()
{ "ts" : Timestamp(1406944987, 1), "h" : NumberLong(0), "v" : 2, "op" : "n",
  "ns" : "", "o" : { "msg" : "initiating set" } }
...
{ "ts" : Timestamp(1406945076, 1), "h" : NumberLong("-9144645443320713428"),
  "v" : 2, "op" : "u", "ns" : "store.products", "o2" : { "_id" : 2 },
  "o" : { "$set" : { "inStock" : 19 } } }
{ "ts" : Timestamp(1406945076, 2), "h" : NumberLong("-787309683444l143322"),
  "v" : 2, "op" : "u", "ns" : "store.products", "o2" : { "_id" : 5 },
  "o" : { "$set" : { "inStock" : 49 } } }
```

Note:

- Note the last two entries in the oplog.
- These entries reflect the update command issued above.
- Note that there is one operation per document affected.
- More specifically, one operation for each of the documents with the _id values 2 and 5.

7.6 Write Concern

Learning Objectives

Upon completing this module students should understand:

- How and when rollback occurs in MongoDB.
- The tradeoffs between durability and performance.
- Write concern as a means of ensuring durability in MongoDB.
- The different levels of write concern.

What happens to the write?

- A write is sent to a primary.
- The primary acknowledges the write to the client.
- The primary then becomes unavailable before a secondary can replicate the write

Answer

- Another member might be elected primary.
- It will not have the last write that occurred before the previous primary became unavailable.
- When the previous primary becomes available again:
  - It will note it has writes that were not replicated.
  - It will put these writes into a `rollback file`.
  - A human will need to determine what to do with this data.
- This is default behavior in MongoDB and can be controlled using `write concern`.

Balancing Durability with Performance

- The previous scenario is a specific instance of a common distributed systems problem.
- For some applications it might be acceptable for writes to be rolled back.
- Other applications may have varying requirements with regard to durability.
- Tunable write concern:
  - Make critical operations persist to an entire MongoDB deployment.
  - Specify replication to fewer nodes for less important operations.

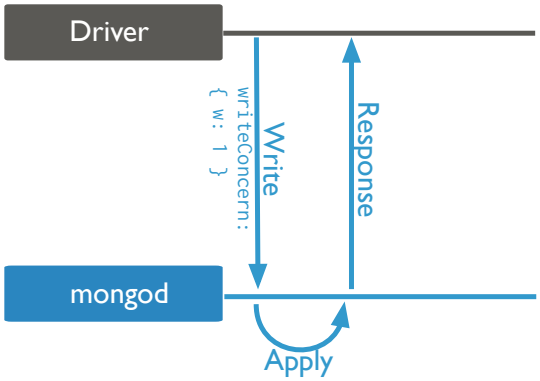
Note:

- MongoDB provides tunable write concern to better address the specific needs of applications.
- Clients may adjust write concern to ensure that the most important operations persist successfully to an entire MongoDB deployment.
- For other less critical operations, clients can adjust write concern to ensure faster performance.

Defining Write Concern

- Clients may define the write concern per write operation, if necessary.
- Standardize on specific levels of write concerns for different classes of writes.
- In the discussion that follows we will look at increasingly strict levels of write concern.

Write Concern: `{ w : 1 }`



Note:

- We refer to this write concern as "Acknowledged".
- This is the default.
- The primary sends an acknowledgement back to the client that it received the write operation (in RAM).
- Allows clients to catch network, duplicate key, and other write errors.

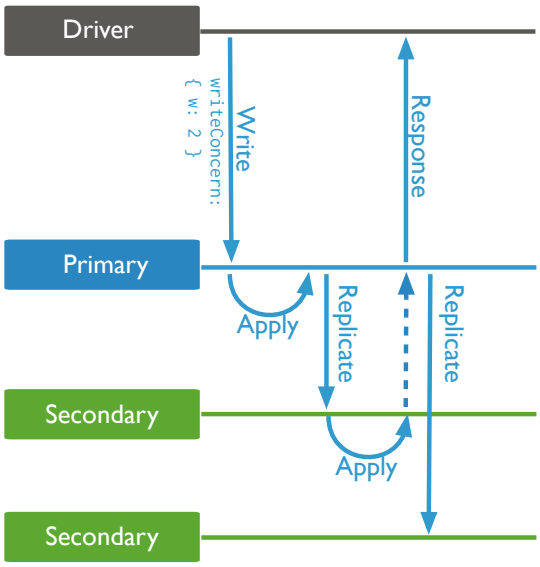
Example: `{ w : 1 }`

```
db.edges.insert( { from : "tom185", to : "mary_p" },
                 { writeConcern : { w : 1 } } )
```

Write Concern: `{ w : 2 }`

Note:

- Called "Replica Acknowledged"
- Ensures the primary completed the write.
- Ensures at least one secondary replicated the write.



Example: `{ w : 2 }`

```
db.customer.update( { user : "mary_p" },
                    { $push : { shoppingCart:
                      { _id : 335443, name : "Brew-a-cup",
                      price : 45.79 } } },
                    { writeConcern : { w : 2 } } )
```

Other Write Concerns

- You may specify any integer as the value of the `w` field for write concern.
- This guarantees that write operations have propagated to the specified number of members.
- E.g., `{ w : 3 }`, `{ w : 4}`, etc.

Write Concern: `{ w : "majority" }`

- Ensures the primary completed the write (in RAM).
- Ensures write operations have propagated to a majority of a replica set's **voting** members.
- Avoids hard coding assumptions about the size of your replica set into your application.
- Using majority trades off performance for durability.
- It is suitable for critical writes and to avoid rollbacks.

Example: `{ w : "majority" }`

```
db.products.update({ _id : 335443 },
                   { $inc : { inStock : -1 } },
                   { writeConcern : { w : "majority" }})
```

Quiz: Which write concern?

Suppose you have a replica set with 7 data nodes. Your application has critical inserts for which you do not want rollbacks to happen. Secondaries may be taken down from to time for maintenance, leaving you with a potential 4 server replica set. Which write concern is best suited for these critical inserts?

- { w : 1 }
- { w : 2 }
- { w : 3 }
- { w : 4 }
- { w : "majority" }

Note: Answer: { w : "majority"}. This is the same as 4 for a 7 member replica set.

Further Reading

See Write Concern Reference[26] for more details on write concern configurations, including setting timeouts and identifying specific replica set members that must acknowledge writes (i.e. tag sets[27]).

[26] http://docs.mongodb.org/manual/reference/write-concern
[27] http://docs.mongodb.org/manual/tutorial/configure-replica-set-tag-sets/#replica-set-configuration-tag-sets

7.7 Read Preference

What is Read Preference?

- Read preference allows you to specify the nodes in a replica set to read from.
- Clients only read from the primary by default.
- There are some situations in which a client may want to read from:
  - Any secondary
  - A specific secondary
  - A specific type of secondary
- Only read from a secondary if you can tolerate possibly stale data, as not all writes might have replicated.

Use Cases

- Running systems operations without affecting the front-end application.
- Providing local reads for geographically distributed applications.
- Maintaining availability during a failover.

Note:

- If you have application servers in multiple data centers, you may consider having a geographically distributed replica set[28] and using a read preference of `nearest`.
- This allows the client to read from the lowest-latency members.
- Use `primaryPreferred` if you want an application to read from the primary under normal circumstances, but to allow possibly stale reads from secondaries during failover.

Not for Scaling

- In general, do *not* read from secondaries to provide extra capacity for reads.
- Sharding[29] increases read and write capacity by distributing operations across a group of machines.
- Sharding is a better strategy for adding capacity.

[28] http://docs.mongodb.org/manual/core/replica-set-geographical-distribution
[29] http://docs.mongodb.org/manual/sharding

Read Preference Modes

MongoDB drivers support the following read preferences. Note that `hidden` nodes will never be read from when connected via the replica set.

- **primary**: Default. All operations read from the primary.
- **primaryPreferred**: Read from the primary but if it is unavailable, read from secondary members.
- **secondary**: All operations read from the secondary members of the replica set.
- **secondaryPreferred**: Read from secondary members but if no secondaries are available, read from the primary.
- **nearest**: Read from member of the replica set with the least network latency, regardless of the member's type.

Tag Sets

- There is also the option to used tag sets.
- You may tag nodes such that queries that contain the tag will be routed to one of the servers with that tag.
- This can be useful for running reports, say for a particular data center or nodes with different hardware (e.g. hard disks vs SSDs).

For example, in the mongo shell:

```
conf = rs.conf()
conf.members[0].tags = { dc : "east", use : "production" }
conf.members[1].tags = { dc : "east", use : "reporting" }
conf.members[2].tags = { use : "production" }
rs.reconfig(conf)
```

7.8 Lab: Setting up a Replica Set

Overview

- In this exercise we will setup a 3 data node replica set on a single machine.
- In production, each node should be run on a dedicated host:
  - To avoid any potential resource contention
  - To provide isolation against server failure

Create Data Directories

Since we will be running all nodes on a single machine, make sure each has its own data directory.

On Linux or Mac OS, run the following in the terminal to create the 3 directories `~/data/rs1`, `~/data/rs2`, and `~/data/rs3`:

```
mkdir -p ~/data/rs{1,2,3}
```

On Windows, run the following command instead in Command Prompt or PowerShell:

```
md c:\data\rs1 c:\data\rs2 c:\data\rs3
```

Launch Each Member

Now start 3 instances of `mongod` in the foreground so that it is easier to observe and shutdown.

On Linux or Mac OS, run *each* of the following commands in its *own terminal window*:

```
mongod --replSet myReplSet --dbpath ~/data/rs1 --port 27017 --oplogSize 200 --smallfiles
mongod --replSet myReplSet --dbpath ~/data/rs2 --port 27018 --oplogSize 200 --smallfiles
mongod --replSet myReplSet --dbpath ~/data/rs3 --port 27019 --oplogSize 200 --smallfiles
```

On Windows, run *each* of the following commands in its *own Command Prompt or PowerShell window*:

```
mongod --replSet myReplSet --dbpath c:\data\rs1 --port 27017 --oplogSize 200 --smallfiles
mongod --replSet myReplSet --dbpath c:\data\rs2 --port 27018 --oplogSize 200 --smallfiles
mongod --replSet myReplSet --dbpath c:\data\rs3 --port 27019 --oplogSize 200 --smallfiles
```

Status

- At this point, we have 3 `mongod` instances running.
- They were all launched with the same `replSet` parameter of "myReplSet".
- Despite this, the members are not aware of each other yet.
- This is fine for now.

Note:

- In production, each member would run on a different machine and use service scripts. For example on Linux, modify `/etc/mongod.conf` accordingly and run:

  ```
  sudo service mongod start
  ```

- To simplify this exercise, we run all members on a single machine.
- The same configuration process is used for this deployment as for one that is distributed across multiple machines.

Connect to a MongoDB Instance

- Connect to the one of the MongoDB instances with the mongo shell.
- To do so run the following command in the terminal, Command Prompt, or PowerShell:

```
mongo    // connect to the default port 27017
```

Configure the Replica Set

```
rs.initiate()
// wait a few seconds
rs.add   ('<HOSTNAME>:27018')
rs.addArb('<HOSTNAME>:27019')

// Keep running rs.status() until there's a primary and 2 secondaries
rs.status()
```

Note:

- `rs.initiate()` will use the FQDN. If we simply use `localhost` when adding the data node and arbiter, MongoDB will refuse to mix the two and return an error.

Problems That May Occur When Initializing the Replica Set

- bindIp parameter is incorrectly set
- Replica set configuration may need to be explicitly specified to use a different hostname:

```
> conf = {
    _id: "<REPLICA-SET-NAME>",
    members: [
        { _id : 0, host : "<HOSTNAME>:27017"},
        { _id : 1, host : "<HOSTNAME>:27018"},
        { _id : 2, host : "<HOSTNAME>:27019",
          "arbiterOnly" : true},
      ]
    }
> rs.initiate(conf)
```

Write to the Primary

While still connected to the primary (port 27017) with mongo shell, insert a simple test document:

```
db.testcol.insert({ a: 1 })
db.testcol.count()

exit    // Or Ctrl-d
```

Read from a Secondary

Connect to one of the secondaries. E.g.:

```
mongo --port 27018
```

Read from the secondary

```
rs.slaveOk()
db.testcol.find()
```

Review the Oplog

```
use local
db.oplog.rs.find()
```

Changing Replica Set Configuration

To change the replica set configuration, first connect to the primary via mongo shell:

```
mongo --port <PRIMARY_PORT>   # e.g. 27017
```

Let's raise the priority of one of the secondaries. Assuming it is the 3rd node (e.g. on port 27019):

```
cfg = rs.conf()
cfg["members"][2]["priority"] = 10
rs.reconfig(cfg)
```

Note:

- Note that `cfg["members"][2]["priority"] = 10` does not actually change the priority.
- `rs.reconfig(cfg)` does.

Verifying Configuration Change

You will see errors like the following, which are expected:

```
2014-10-07T17:01:34.610+0100 DBClientCursor::init call() failed
2014-10-07T17:01:34.613+0100 trying reconnect to 127.0.0.1:27017 (127.0.0.1) failed
2014-10-07T17:01:34.617+0100 reconnect 127.0.0.1:27017 (127.0.0.1) ok
reconnected to server after rs command (which is normal)
```

Verify that the replica set configuration is now as expected:

```
rs.conf()
```

The secondary will now become a primary. Check by running:

```
rs.status()
```

Further Reading

- Replica Configuration[30]
- Replica States[31]

[30] http://docs.mongodb.org/manual/reference/replica-configuration/
[31] http://docs.mongodb.org/manual/reference/replica-states/

8 Sharding

Introduction to Sharding **(page 132)** An introduction to sharding

Balancing Shards **(page 140)** Chunks, the balancer, and their role in a sharded cluster

Shard Tags **(page 143)** How tag-based sharding works

Lab: Setting Up a Sharded Cluster **(page 145)** Deploying a sharded cluster

8.1 Introduction to Sharding

Learning Objectives

Upon completing this module, students should understand:

- What problems sharding solves
- When sharding is appropriate
- The importance of the shard key and how to choose a good one
- Why sharding increases the need for redundancy

Contrast with Replication

- In an earlier module, we discussed Replication.
- This should never be confused with sharding.
- Replication is about high availability and durability.
  - Taking your data and constantly copying it
  - Being ready to have another machine step in to field requests.

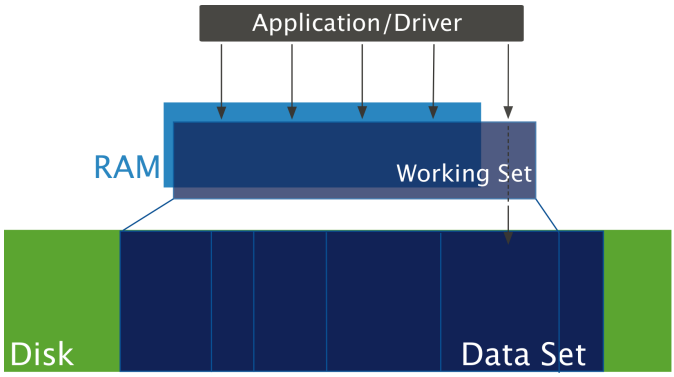
Sharding is Concerned with Scale

- What happens when a system is unable to handle the application load?
- It is time to consider scaling.
- There are 2 types of scaling we want to consider:
  - Vertical scaling
  - Horizontal scaling

Vertical Scaling

- Adding more RAM, faster disks, etc.
- When is this the solution?
- First, consider a concept called the `working set`.

The Working Set



Note:

- The working set for a MongoDB database is the portion of your data that clients access most often.
- Your working set should stay in memory, otherwise random disk operations will hurt performance.
- For the fastest processing, ensure that your indexes fit entirely in RAM so that the system can avoid reading the index from disk.
- In some cases, only recently indexed values must be in RAM.

Limitations of Vertical Scaling

- There is a limit to how much RAM one machine can support.
- There are other bottlenecks such as I/O, disk access and network.
- Cost may limit our ability to scale up.
- There may be requirements to have a large working set that no single machine could possible support.
- This is when it is time to scale horizontally.

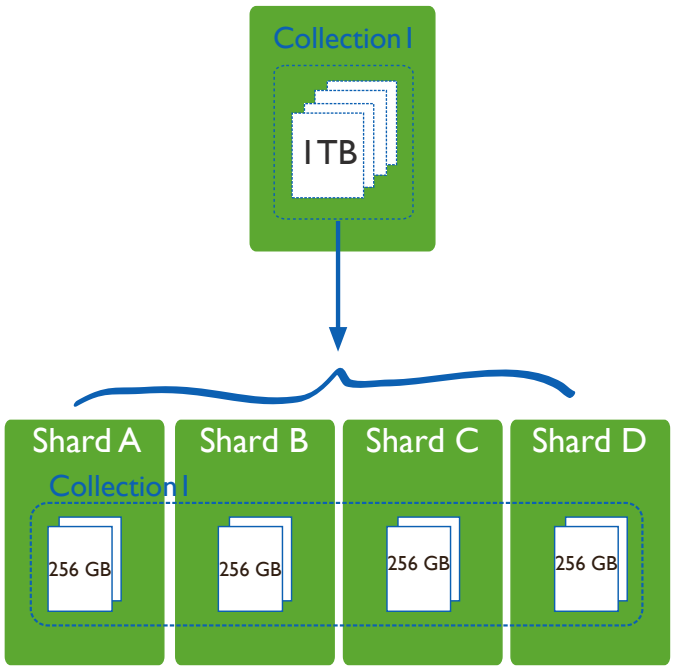
Sharding Overview

- MongoDB enables you to scale horizontally through sharding.
- Sharding is about adding more capacity to your system.
- MongoDB's sharding solution is designed to perform well on commodity hardware.
- The details of sharding are abstracted away from applications.
- Queries are performed the same way as if sending operations to a single server.
- Connections work the same by default.

When to Shard

- If you have more data than one machine can hold on its drives
- If your application is write heavy and you are experiencing too much latency.
- If your working set outgrows the memory you can allocate to a single machine.

Dividing Up Your Dataset



Note:

- When you shard a collection it is distributed across several servers.
- Each mongod manages a subset of the data.
- When you perform a read or write operation it will go to a router that will then direct the query to the appropriate server.
- Depending on the operation and how your cluster is configured you may need to touch only one server to complete the operation.

Sharding Concepts

To understanding how sharding works in MongoDB, we need to understand:

- Shard Keys
- Chunks

Shard Key

- You must define a shard key for a sharded collection.
- Based on one or more fields (like an index)
- Shard key defines a space of values
- Think of the key space like points on a line
- A key range is a segment of that line

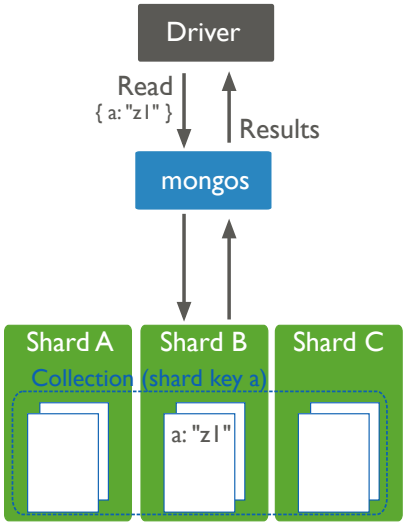
Shard Key Ranges

- A collection is partitioned based on shard key ranges.
- The shard key determines where documents are located in the cluster.
- It is used to route operations to the appropriate shard.
- For reads and writes
- Once a collection is sharded, you cannot change a shard key.

Note:

- For reads and updates, the shard key determines which shard holds the document and will handle the operation.
- When you insert a document, the shard key determines which server you will write to.

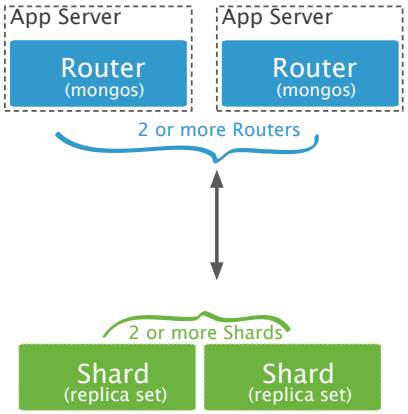
Targeted Query Using Shard Key



Chunks

- MongoDB partitions data into `chunks` based on shard key ranges.
- This is bookkeeping metadata.
- MongoDB attempts to keep the amount of data balanced across shards.
- This is achieved by migrating chunks from one shard to another as needed.
- There is nothing in a document that indicates its chunk.
- The document does not need to be updated if its assigned chunk changes.

Sharded Cluster Architecture



Note:

- This figure illustrates one possible architecture for a sharded cluster.
- Each shard is a self-contained replica set.
- Each replica set holds a partition of the data.
- As many new shards could be added to this sharded cluster as scale requires.
- At this point our cluster can handle a load approaching twice that of a single replica set using the same hardware.
- As mentioned, read/write operations go through a router.
- The server that routes requests is the mongos.

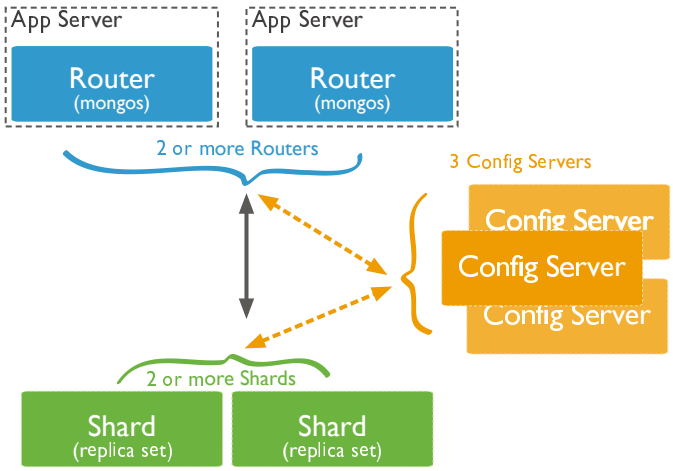
Mongos

- A mongos is responsible for accepting requests and returning results to an application driver.
- In a sharded cluster, nearly all operations go through a mongos.
- A sharded cluster can have as many mongos routers as required.
- It is typical for each application server to have one mongos.
- Always use more than one mongos to avoid a single point of failure.

Note:

- A mongos is typically deployed on an application server.
- There should be one mongos per app server.
- Scale with your app server.
- Very little latency between the application and the router.

Config Servers



Note:

- The previous diagram was incomplete; it was missing config servers.
- Use three config servers in production.
- These hold only metadata about the sharded collections.
  - Where your mongos servers are
  - Any hosts that are not currently available
  - What collections you have
  - How your collections are partitioned across the cluster
- Mongos processes use them to retrieve the state of the cluster.
- You can access cluster metadata from a mongos by looking at the `config` db.

Config Server Hardware Requirements

- Quality network interfaces
- A small amount of disk space (typically a few GB)
- A small amount of RAM (typically a few GB)
- The larger the sharded cluster, the greater the config server hardware requirements.

Possible Imbalance?

- Depending on how you configure sharding, data can become unbalanced on your sharded cluster.
  - Some shards might receive more inserts than others.
  - Some shards might have documents that grow more than those in other shards.
- This may result in too much load on a single shard.
  - Reads and writes
  - Disk activity
- This would defeat the purpose of sharding.

Balancing Shards

- If a chunk grows too large MongoDB will split it into two chunks.
- The MongoDB balancer keeps chunks distributed across shards in equal numbers.
- However, a balanced sharded cluster depends on a good shard key.

With a Good Shard Key

You might easily see that:

- Reads hit only 1 or 2 shards per query.
- Writes are distributed across all servers.
- Your disk usage is evenly distributed across shards.
- Things stay this way as you scale.

With a Bad Shard Key

You might see that:

- Your reads hit every shard.
- Your writes are concentrated on one shard.
- Most of your data is on just a few shards.
- Adding more shards to the cluster will not help.

Choosing a Shard Key

Generally, you want a shard key:

- That has high cardinality
- That is used in the majority of read queries
- For which the values read and write operations use are randomly distributed
- For which the majority of reads are routed to a particular server

More Specifically

- Your shard key should be consistent with your query patterns.
- If reads usually find only one document, you only need good cardinality.
- If reads retrieve many documents:
  - Your shard key supports locality
  - Matching documents will reside on the same shard

Cardinality

- A good shard key will have high cardinality.
- A relatively small number of documents should have the same shard key.
- Otherwise operations become isolated to the same server.
- Because documents with the same shard key reside on the same shard.
- Adding more servers will not help.
- Hashing will not help.

Non-Monotonic

- A good shard key will generate new values non-monotonically.
- Datetimes, counters, and ObjectIds make bad shard keys.
- Monotonic shard keys cause all inserts to happen on the same shard.
- Hashing will solve this problem.
- However, doing range queries with a hashed shard key will perform a scatter-gather query across the cluster.

Note:

- Documents will eventually move as chunks are balanced.
- But in the meantime one server gets hammered while others are idle.
- And moving chunks has its own performance costs.

Shards Should be Replica Sets

- As the number of shards increases, the number of servers in your deployment increases.
- This increases the probability that one server will fail on any given day.
- With redundancy built into each shard you can mitigate this risk.

8.2 Balancing Shards

Learning Objectives

Upon completing this module students should understand:

- Chunks and the balancer
- The status of chunks in a newly sharded collection
- How chunk splits automatically occur
- Advantages of pre-splitting chunks
- How the balancer works

Chunks and the Balancer

- Chunks are groups of documents.
- The shard key determines which chunk a document will be contained in.
- Chunks can be split when they grow too large.
- The balancer decides where chunks go.
- It handles migrations of chunks from one server to another.

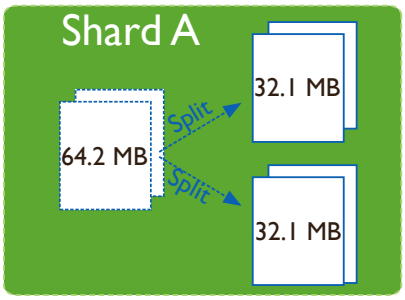
Chunks in a Newly Sharded Collection

- The range of a chunk is defined by the shard key values of the documents the chunk contains.
- When a collection is sharded it starts with just one chunk.
- The first chunk for a collection will have the range:

```
{ $minKey : 1 } to { $maxKey : 1 }
```

- All shard key values from the smallest possible to the largest fall in this chunk's range.

Chunk Splits



Note:

- When a chunk grows larger than the chunk size it will be split in half.
- The default chunk size is 64MB.
- A chunk can only be split between two values of a shard key.
- If every document on a chunk has the same shard key value, it cannot be split.
- This is why the shard key's cardinality is important
- Chunk splitting is just a bookkeeping entry in the metadata.
- No data bearing documents are altered.

Pre-Splitting Chunks

- You may pre-split data before loading data into a sharded cluster.
- Pre-splitting is useful if:
  - You plan to do a large data import early on
  - You expect a heavy initial server load and want to ensure writes are distributed

---

Note:

- A large data import will take time to split and balance without pre-splitting.

Start of a Balancing Round

- A balancing round may be initiated by any mongos in the cluster.
- This happens when the difference in the number of chunks between two shards becomes to large.
- Specifically, the difference between the shard with the most chunks and the shard with the fewest.
- A balancing round starts when the imbalance reaches:
  - 2 when the cluster has < 20 chunks
  - 4 when the cluster has 20-79 chunks
  - 8 when the cluster has 80+ chunks

Balancing is Resource Intensive

- Chunk migration requires copying all the data in the chunk from one shard to another.
- MonogDB can migrate only a single chunk at a time.
- MongoDB creates splits only after an insert operation.
- For these reasons, it is possible to define a balancing window to ensure the balancer will only run during scheduled times.

Chunk Migration Steps

1. The balancer process sends the `moveChunk` command to the source shard.
2. The source shard continues to process reads/writes for that chunk during the migration.
3. The destination shard requests documents in the chunk and begins receiving copies.
4. After receiving all documents, the destination shard receives any changes to the chunk.
5. Then the destination shard tells the config db that it has the chunk.
6. The destination shard will now handle all reads/writes.
7. The source shard deletes its copy of the chunk.

Concluding a Balancing Round

- Each chunk will move:
  - From the shard with the most chunks
  - To the shard with the fewest
- A balancing round ends when all shards differ by at most one chunk.

8.3 Shard Tags

Learning Objectives

Upon completing this module students should understand:

- The purpose for shard tags
- Advantages of using shard tags
- Potential drawbacks of shard tags

Tags - Overview

- Shard tags allow you to "tie" data to one or more shards.
- A shard tag describes a range of shard key values.
- If a chunk is in the shard tag range, it will live on a shard with that tag.

Example: DateTime

- Documents older than one year need to be kept, but are rarely used.
- You tag those ranges as "LTS" for Long Term Storage.
- Tag specific shards to hold LTS documents.
- These shards can be on cheaper, slower machines.
- Invest in high-performance servers for more frequently accessed data.

Example: Location

- You are required to keep certain data in its home country.
- You include the country in the shard tag.
- Maintain data centers within each country that house the appropriate shards.
- Meets the country requirement but allows all servers to be part of the same system.

Example: Premium Tier

- You have customers who want to pay for a "premium" tier.
- The shard key permits you to distinguish one customer's documents from all others.
- Tag the document ranges for each customer so that their documents will be located on shards of the appropriate tier.
- Shards tagged as premium tier run on high performance servers.
- Other shards run on commodity hardware.
- See Manage Shard Tags[32]

Note:

- As customers move from one tier to another it will be necessary to execute commands that either add a given customer's shard key range to the premium tag or remove that range from those tagged as "premium".
- During balancing rounds, if the balancer detects that any chunks are not on the correct shards per configured tags, the balancer migrates chunks in tagged ranges to shards associated with those tags.
- After re-configuring tags with a shard key range, and associating it with a shard or shards, the cluster may take some time to balance the data among the shards.

See: Tiered Storage Models in MongoDB: Optimizing Latency and Cost[33].

Tags - Caveats

- Because tagged chunks will only be on certain servers, if you tag more than those servers can handle, you'll have a problem.
  - You're not only worrying about your overall server load, you're worrying about server load for each of your tags.
- Your chunks will evenly distribute themselves across the available chunks. You cannot control things more fine grained than your tags.

[32] http://docs.mongodb.org/manual/tutorial/administer-shard-tags/
[33] http://blog.mongodb.org/post/85721044164/tiered-storage-models-in-mongodb-optimizing-latency

8.4 Lab: Setting Up a Sharded Cluster

Learning Objectives

Upon completing this module students should understand:

- How to set up a sharded cluster including:
  - Replica Sets as Shards
  - Config Servers
  - Mongos processes
- How to enable sharding for a database
- How to shard a collection
- How to determine where data will go

Our Sharded Cluster

- In this exercise, we will set up a cluster with 3 shards.
- Each shard will be a replica set with 3 members (including one arbiter).
- We will insert some data and see where it goes.

Sharded Cluster Configuration

- Three shards:
  1. A replica set on ports 27107, 27108, 27109
  2. A replica set on ports 27117, 27118, 27119
  3. A replica set on ports 27127, 27128, 27129
- Three config servers on ports 27217, 27218, 27219
- Two mongos servers at ports 27017 and 27018

Build Our Data Directories

On Linux or MacOS, run the following in the terminal to create the data directories we'll need.

```
mkdir -p ~/data/cluster/config/{c0,c1,c2}
mkdir -p ~/data/cluster/shard0/{m0,m1,arb}
mkdir -p ~/data/cluster/shard1/{m0,m1,arb}
mkdir -p ~/data/cluster/shard2/{m0,m1,arb}
mkdir -p ~/data/cluster/{s0,s1}
```

On Windows, run the following commands instead:

```
md c:\data\cluster\config\c0 c:\data\cluster\config\c1 c:\data\cluster\config\c2
md c:\data\cluster\shard0\m0 c:\data\cluster\shard0\m1 c:\data\cluster\shard0\arb
md c:\data\cluster\shard1\m0 c:\data\cluster\shard1\m1 c:\data\cluster\shard1\arb
md c:\data\cluster\shard2\m0 c:\data\cluster\shard2\m1 c:\data\cluster\shard2\arb
md c:\data\cluster\s0 c:\data\cluster\s1
```

Initiate a Replica Set (Linux/MacOS)

```
mongod --replSet shard0 --smallfiles --nojournal --noprealloc \
       --dbpath ~/data/cluster/shard0/m0 \
       --logpath ~/data/cluster/shard0/m0/mongod.log \
       --fork --port 27107

mongod --replSet shard0 --smallfiles --nojournal --noprealloc \
       --dbpath ~/data/cluster/shard0/m1 \
       --logpath ~/data/cluster/shard0/m1/mongod.log \
       --fork --port 27108

mongod --replSet shard0 --smallfiles --nojournal --noprealloc \
       --dbpath ~/data/cluster/shard0/arb \
       --logpath ~/data/cluster/shard0/arb/mongod.log \
       --fork --port 27109

mongo --port 27107 --eval "\
    rs.initiate(); sleep(3000);\
    rs.add   ('<HOSTNAME>:27108');\
    rs.addArb('<HOSTNAME>:27109')"
```

Initiate a Replica Set (Windows)

```
mongod --replSet shard0 --smallfiles --nojournal --noprealloc \
       --dbpath c:\data\cluster\shard0\m0 \
       --logpath c:\data\cluster\shard0\m0\mongod.log \
       --port 27107 --oplogSize 10

mongod --replSet shard0 --smallfiles --nojournal --noprealloc \
       --dbpath c:\data\cluster\shard0\m1 \
       --logpath c:\data\cluster\shard0\m1\mongod.log \
       --port 27108 --oplogSize 10

mongod --replSet shard0 --smallfiles --nojournal --noprealloc \
       --dbpath c:\data\cluster\shard0\arb \
       --logpath c:\data\cluster\shard0\arb\mongod.log \
       --port 27109 --oplogSize 10

mongo --port 27107 --eval "\
    rs.initiate(); sleep(3000);\
    rs.add   ('<HOSTNAME>:27108');\
    rs.addArb('<HOSTNAME>:27109')"
```

Spin Up a Second Replica Set (Linux/MacOS)

```
mongod --replSet shard1 --smallfiles --nojournal --noprealloc \
       --dbpath ~/data/cluster/shard1/m0 \
       --logpath ~/data/cluster/shard1/m0/mongod.log \
       --fork --port 27117

mongod --replSet shard1 --smallfiles --nojournal --noprealloc \
       --dbpath ~/data/cluster/shard1/m1 \
       --logpath ~/data/cluster/shard1/m1/mongod.log \
       --fork --port 27118

mongod --replSet shard1 --smallfiles --nojournal --noprealloc \
       --dbpath ~/data/cluster/shard1/arb \
       --logpath ~/data/cluster/shard1/arb/mongod.log \
       --fork --port 27119

mongo --port 27117 --eval "\
    rs.initiate(); sleep(3000);\
    rs.add   ('<HOSTNAME>:27118');\
    rs.addArb('<HOSTNAME>:27119')"
```

Spin Up a Second Replica Set (Windows)

```
mongod --replSet shard1 --smallfiles --nojournal --noprealloc \
       --dbpath c:\data\cluster\shard1\m0 \
       --logpath c:\data\cluster\shard1\m0\mongod.log \
       --port 27117 --oplogSize 10

mongod --replSet shard1 --smallfiles --nojournal --noprealloc \
       --dbpath c:\data\cluster\shard1\m1 \
       --logpath c:\data\cluster\shard1\m1\mongod.log \
       --port 27118 --oplogSize 10

mongod --replSet shard1 --smallfiles --nojournal --noprealloc \
       --dbpath c:\data\cluster\shard1\arb \
       --logpath c:\data\cluster\shard1\arb\mongod.log \
       --port 27119 --oplogSize 10

mongo --port 27117 --eval "\
    rs.initiate(); sleep(3000);\
    rs.add   ('<HOSTNAME>:27118');\
    rs.addArb('<HOSTNAME>:27119')"
```

A Third Replica Set (Linux/MacOS)

```
mongod --replSet shard2 --smallfiles --nojournal --noprealloc \
       --dbpath ~/data/cluster/shard2/m0 \
       --logpath ~/data/cluster/shard2/m0/mongod.log \
       --fork --port 27127

mongod --replSet shard2 --smallfiles --nojournal --noprealloc \
       --dbpath ~/data/cluster/shard2/m1 \
       --logpath ~/data/cluster/shard2/m1/mongod.log \
       --fork --port 27128

mongod --replSet shard2 --smallfiles --nojournal --noprealloc \
       --dbpath ~/data/cluster/shard2/arb \
       --logpath ~/data/cluster/shard2/arb/mongod.log \
       --fork --port 27129

mongo --port 27127 --eval "\
    rs.initiate(); sleep(3000);\
    rs.add   ('<HOSTNAME>:27128');\
    rs.addArb('<HOSTNAME>:27129')"
```

A Third Replica Set (Windows)

```
mongod --replSet shard2 --smallfiles --nojournal --noprealloc \
       --dbpath c:\data\cluster\shard2\m0 \
       --logpath c:\data\cluster\shard2\m0\mongod.log \
       --port 27127 --oplogSize 10

mongod --replSet shard2 --smallfiles --nojournal --noprealloc \
       --dbpath c:\data\cluster\shard2\m1 \
       --logpath c:\data\cluster\shard2\m1\mongod.log \
       --port 27128 --oplogSize 10

mongod --replSet shard2 --smallfiles --nojournal --noprealloc \
       --dbpath c:\data\cluster\shard2\arb \
       --logpath c:\data\cluster\shard2\arb\mongod.log \
       --port 27129 --oplogSize 10

mongo --port 27127 --eval "\
    rs.initiate(); sleep(3000);\
    rs.add   ('<HOSTNAME>:27128');\
    rs.addArb('<HOSTNAME>:27129')"
```

Status Check

- Now we have three replica sets running.
- We have one for each shard.
- They do not know about each other yet.
- To make them a sharded cluster we will:
  - Build our config databases
  - Launch our mongos processes
  - Add each shard to the cluster
- To benefit from this configuration we also need to:
  - Enable sharding for a database
  - Shard at least one collection within that database

Launch Config Servers (Linux/MacOS)

```
mongod --smallfiles --nojournal --noprealloc \
       --dbpath ~/data/cluster/config/c0 \
       --logpath ~/data/cluster/config/c0/mongod.log \
       --fork --port 27217 --configsvr

mongod --smallfiles --nojournal --noprealloc \
       --dbpath ~/data/cluster/config/c1 \
       --logpath ~/data/cluster/config/c1/mongod.log \
       --fork --port 27218 --configsvr

mongod --smallfiles --nojournal --noprealloc \
       --dbpath ~/data/cluster/config/c2 \
       --logpath ~/data/cluster/config/c2/mongod.log \
       --fork --port 27219 --configsvr
```

Launch Config Servers (Windows)

```
mongod --smallfiles --nojournal --noprealloc \
       --dbpath c:\data\cluster\config\c0 \
       --logpath c:\data\cluster\config\c0\mongod.log \
       --port 27217 --configsvr

mongod --smallfiles --nojournal --noprealloc \
       --dbpath c:\data\cluster\config\c1 \
       --logpath c:\data\cluster\config\c1\mongod.log \
       --port 27218 --configsvr

mongod --smallfiles --nojournal --noprealloc \
       --dbpath c:\data\cluster\config\c2 \
       --logpath c:\data\cluster\config\c2\mongod.log \
       --port 27219 --configsvr
```

Launch the Mongos Processes (Linux/MacOS)

Now our mongos's. We need to tell them about our config servers.

```
mongos --logpath ~/data/cluster/s0/mongos.log --fork --port 27017 \
       --configdb localhost:27217,localhost:27218,localhost:27219

mongos --logpath ~/data/cluster/s1/mongos.log --fork --port 27018 \
       --configdb localhost:27217,localhost:27218,localhost:27219
```

Launch the Mongos Processes (Windows)

Now our mongos's. We need to tell them about our config servers.

```
mongos --logpath c:\data\cluster\s0\mongos.log --port 27017 \
       --configdb localhost:27217,localhost:27218,localhost:27219

mongos --logpath c:\data\cluster\s1\mongos.log --port 27018 \
       --configdb localhost:27217,localhost:27218,localhost:27219
```

Add All Shards

```
echo 'sh.addShard( "shard0/localhost:27107" ); \
      sh.addShard("shard1/localhost:27117" ); \
      sh.addShard( "shard2/localhost:27127" ); sh.status()' | mongo
```

Note: Instead of doing this through a bash (or other) shell command, you may prefer to launch a mongo shell and issue each command individually.

Enable Sharding and Shard a Collection

Enable sharding for the test database, shard a collection, and insert some documents.

```
echo 'sh.enableSharding("test"); \
      sh.shardCollection("test.testcol", { a : 1, b : 1 } )' | mongo

echo 'for (i=0; i<10000; i++) { docArr = []; for (j=0; j<1000; j++) { \
     docArr.push( { a : i, b : j, c : "Filler String 000000000000000000000 \
     00000000000000000000000000000000000000000000000000000000000000000000 \
     000000000000000" } )  }; db.testcol.insert(docArr) }' | mongo
```

Observe What Happens

Connect to either mongos using a mongo shell and frequently issue:

```
sh.status()
```

Note:

- Point out to the students that you can see chunks get created and moved to different shards.
- Also useful to have students run a query or two.

  ```
  db.testcol.find( { a : { $lte : 100 } } ).explain()
  ```

---

9 Reporting Tools and Diagnostics

Performance Troubleshooting **(page 152)** An introduction to reporting and diagnostic tools for MongoDB

9.1 Performance Troubleshooting

Learning Objectives

Upon completing this module students should understand basic performance troubleshooting techniques and tools including:

- `mongostat`
- `mongotop`
- `db.setProfilingLevel()`
- `db.currentOp()`
- `db.<COLLECTION>.stats()`
- `db.serverStatus()`

`mongostat` and `mongotop`

- `mongostat` samples a server every second.
  - See current ops, pagefaults, network traffic, etc.
  - Does not give a view into historic performance; use MMS for that.
- `mongotop` looks at the time spent on reads/writes in each collection.

Exercise: `mongostat` (setup)

In one window, perform the following commands.

```
db.testcol.drop()
for (i=1; i<=10000; i++) {
    arr = [];
    for (j=1; j<=1000; j++) {
        doc = { _id: (1000 * (i-1) + j), a: i, b: j, c: (1000 * (i-1)+ j) };
        arr.push(doc)
    };
    db.testcol.insert(arr);
    var x = db.testcol.find( { b : 255 } );
    x.next();
    var x = db.testcol.find( { _id : 1000 * (i-1) + 255 } );
    x.next();
    var x = "asdf";
    db.testcol.update( { a : i, b : 255 }, { $set : { d : x.pad(1000) } });
    print(i)
}
```

Exercise: `mongostat` (run)

- In another window/tab, run mongostat.
- You will see:
  - Inserts
  - Queries
  - Updates

Exercise: `mongostat` (create index)

- In a third window, create an index when you see things slowing down:

```
db.testcol.createIndex( { a : 1, b : 1 } )
```

- Look at mongostat.
- Notice that things are going significantly faster.
- Then, let's drop that and build another index.

```
db.testcol.dropIndexes()
db.testcol.createIndex( { b : 1, a : 1 } )
```

Exercise: `mongotop`

Perform the following then, in another window, run `mongotop`.

```
db.testcol.drop()
for (i=1; i<=10000; i++) {
    arr = [];
    for (j=1; j<=1000; j++) {
        doc = {_id: (1000*(i-1)+j), a: i, b: j, c: (1000*(i-1)+j)};
        arr.push(doc)
    };
    db.testcol.insert(arr);
    var x = db.testcol.find( {b: 255} ); x.next();
    var x = db.testcol.find( {_id: 1000*(i-1)+255} ); x.next();
    var x = "asdf";
    db.testcol.update( {a: i, b: 255}, {$set: {d: x.pad(1000)}});
    print(i)
}
```

Note: Direct the students to the fact that you can see the activity on the server for reads/writes/total.

`db.currentOp()`

- currentOp is a tool that asks what the db is doing at the moment.
- currentOp is useful for finding long-running processes.
- Fields of interest:
  - microsecs_running
  - op
  - query
  - lock
  - waitingForLock

Exercise: `db.currentOp()`

Do the following then, connect with a separate shell, and repeatedly run `db.currentOp()`.

```
db.testcol.drop()
for (i=1; i<=10000; i++) {
    arr = [];
    for (j=1; j<=1000; j++) {
        doc = {_id: (1000*(i-1)+j), a: i, b: j, c: (1000*(i-1)+j)};
        arr.push(doc)
    };
    db.testcol.insert(arr);
    var x = db.testcol.find( {b: 255} ); x.next();
    var x = db.testcol.find( {_id: 1000*(i-1)+255 }); x.next();
    var x = "asdf";
    db.testcol.update( {a: i, b: 255}, {$set: {d: x.pad(1000)}});
    print(i)
}
```

Note: Point out to students that the running time gets longer & longer, on average.

`db.<COLLECTION>.stats()`

- Used to view the current stats for a collection.
- Everything is in bytes; use the multiplier parameter to view in KB, MB, etc
- You can also use `db.stats()` to do this at scope of the entire database

Exercise: Using Collection Stats

Look at the output of the following:

```
db.testcol.drop()
db.testcol.insert( { a : 1 } )
db.testcol.stats()
var x = "asdf"
db.testcol2.insert( { a : x.pad(10000000) } )
db.testcol2.stats()
db.stats()
```

The Profiler

- Off by default.
- To reset, `db.setProfilerLevel(0)`
- At setting 1, it captures "slow" queries.
- You may define what "slow" is.
- Default is 100ms: `db.setProfilerLevel(1)`
- E.g., to capture 20 ms: `db.setProfilerLevel(1, 20)`

The Profiler (continued)

- If the profiler level is 2, it captures all queries.
  - This will severely impact performance.
  - Turns all reads into writes.
- Always turn the profiler off when done (set level to 0)
- Creates `db.system.profile` collection

Exercise: Exploring the Profiler

Perform the following, then look in your db.system.profile.

```
db.setProfilingLevel(0)
db.testcol.drop()
db.system.profile.drop()
db.setProfilingLevel(2)
db.testcol.insert( { a : 1 } )
db.testcol.find()
var x = "asdf"
db.testcol.insert( { a : x.pad(10000000) } )  // ~10 MB
db.setProfilingLevel(0)
db.system.profile.find().pretty()
```

Note:

- Mention to the students what the fields mean.
- Things to keep in mind:

  - op can be command, query, or update
  - ns is sometimes the db.<COLLECTION> namespace
    * but sometimes db.$cmd for commands
  - key updates refers to index keys
  - ts (timestamp) is useful for some queries if problems cluster.

`db.serverStatus()`

- Takes a snapshot of server status.
- By taking diffs, you can see system trends.
- Most of the data that MMS gets is from here.

Exercise: Using `db.serverStatus()`

- Open up two windows. In the first, type:

```
db.testcol.drop()
var x = "asdf"
for (i=0; i<=10000000; i++) {
    db.testcol.insert( { a : x.pad(100000) } )
}
```

- In the second window, type periodically:

```
var x = db.serverStatus(); x.metrics.document
```

Analyzing Profiler Data

- Enable the profiler at default settings.
- Run for 5 seconds.
- Slow operations are captured.
- The issue is there is not a proper index on the message field.
- Allow class to discover this as the data is examined.
- You will see how fast documents are getting inserted.
- It will be slow b/c the documents are big.

Performance Improvement Techniques

- Appropriate write concerns
- Bulk operations
- Good schema design
- Good Shard Key choice
- Good indexes

Performance Tips: Write Concern

- Increasing the write concern increases data safety.
- This will have an impact on performance, however.
- This is especially true when there are network issues.

Bulk Operations

- Using bulk operations can improve performance, especially when using write concern greater than 1.
- These enable the server to bulk write and bulk acknowledge.
- Can be done with both inserts and updates.

Exercise: Comparing bulk inserts with `mongostat`

Let's spin up a 3-member replica set:

```
mkdir -p /data/replset/{1,2,3}
mongod --logpath /data/replset/1/mongod.log \
       --dbpath /data/replset/1 --replSet mySet --port 27017 --fork
mongod --logpath /data/replset/2/mongod.log \
       --dbpath /data/replset/2 --replSet mySet --port 27018 --fork
mongod --logpath /data/replset/3/mongod.log \
       --dbpath /data/replset/3 --replSet mySet --port 27019 --fork

echo "conf = {_id: 'mySet', members: [{_id: 0, host: 'localhost:27017'}, \
     {_id: 1, host: 'localhost:27018'}, {_id: 2, host: 'localhost:27019'}]}; \
     rs.initiate(conf)" | mongo
```

`mongostat`, bulk inserts with `{w: 1}`

Perform the following, with writeConcern : 1 and no bulk inserts:

```
db.testcol.drop()
for (i=1; i<=10000; i++) {
    for (j=1; j<=1000; j++) {
        db.testcol.insert( { _id : (1000 * (i-1) + j),
                             a : i, b : j, c : (1000 * (i-1)+ j) },
                           { writeConcern : { w : 1 } } );
    };
    print(i);
}
```

Run `mongostat` and see how fast that happens.

Bulk inserts with `{w: 3}`

Increase the write concern to 3 (safer but slower):

```
db.testcol.drop()
for (i=1; i<=10000; i++) {
    for (j=1; j<=1000; j++) {
        db.testcol.insert(
            { _id: (1000 * (i-1) + j), a: i, b: j, c: (1000 * (i-1)+ j)},
            { writeConcern: { w: 3 } }
        );
    };
    print(i);
}
```

Again, run mongostat.

`mongostat`, bulk inserts with `{w: 3}`

- Finally, let's use bulk inserts to our advantage:
- Note that writeConcern is still { w: 3 }

```
db.testcol.drop()
for (i=1; i<=10000; i++) {
    arr = []
    for (j=1; j<=1000; j++) {
        arr.push(
            { _id: (1000 * (i-1) + j), a: i, b: j, c: (1000 * (i-1)+ j) }
        );
    };
    db.testcol.insert( arr, { writeConcern : { w : 3 } } );
    print(i);
}
```

Schema Design

- The structure of documents affects performance.
- Optimize for your application's read/write patterns.
- We want as few requests to the database as possible to perform a given application task.
- See the data modeling section for more information.

Shard Key Considerations

- Choose a shard key that distributes load across your cluster.
- Create a shard key such that only a small number of documents will have the same value.
- Create a shard key that has a high degree of randomness.
- Your shard key should enable a mongos to target a single shard for a given query.

Indexes and Performance

- Reads and writes that don't use an index will cripple performance.
- In compound indexes, order matters:
  - Sort on a field that comes before any range used in the index.
  - You can't skip fields; they must be used in order.
  - Revisit the indexing section for more detail.

10 Backup and Recovery

Backup and Recovery **(page 160)** An overview of backup options for MongoDB

10.1 Backup and Recovery

Disasters Do Happen

Human Disasters

Terminology: RPO vs. RTO

- **Recovery Point Objective (RPO)**: How much data can you afford to lose?
- **Recovery Time Objective (RTO)**: How long can you afford to be off-line?

Terminology: DR vs. HA

- **Disaster Recovery** (DR)
- **High Availability** (HA)
- Distinct business requirements
- Technical solutions may converge

Quiz

- Q: What's the hardest thing about backups?
- A: Restoring them!
- **Regularly test that restoration works!**

Backup Options

- Document Level
  - Logical
  - `mongodump`, `mongorestore`
- File system level
  - Physical
  - Copy files
  - Volume/disk snapshots

Document Level: `mongodump`

- Dumps collection to BSON files
- Mirrors your structure
- Can be run live or in offline mode
- Does not include indexes (rebuilt during restore)
- `--dbpath` for direct file access
- `--oplog` to record oplog while backing up
- `--query/filter` selective dump

`mongodump`

```
$ mongodump --help
Export MongoDB data to BSON files.

options:
  --help                  produce help message
  -v [ --verbose ]        be more verbose (include multiple times for
                          more verbosity e.g. -vvvvv)
  --version               print the program's version and exit
  -h [ --host ] arg       mongo host to connect to ( /s1,s2 for
  --port arg              server port. Can also use --host hostname
  -u [ --username ] arg   username
  -p [ --password ] arg   password
  --dbpath arg            directly access mongod database files in path
  -d [ --db ] arg         database to use
  -c [ --collection ] arg collection to use (some commands)
  -o [ --out ] arg (=dump)output directory or "-" for stdout
  -q [ --query ] arg      json query
  --oplog                 Use oplog for point-in-time snapshotting
```

File System Level

- **Must use journaling!**
- Copy `/data/db` files
- Or snapshot volume (e.g., LVM, SAN, EBS)
- *Seriously, always use journaling!*

Ensure Consistency

Flush RAM to disk and stop accepting writes:

- `db.fsyncLock()`
- Copy/Snapshot
- `db.fsyncUnlock()`

File System Backups: Pros and Cons

- Entire database
- Backup files will be large
- Fastest way to create a backup
- Fastest way to restore a backup

Document Level: `mongorestore`

- `mongorestore`
- `--oplogReplay` replay oplog to point-in-time

File System Restores

- All database files
- Selected databases or collections
- Replay Oplog

Backup Sharded Cluster

1. Stop Balancer (and wait) or no balancing window
2. Stop one config server (data R/O)
3. Backup Data (shards, config)
4. Restart config server
5. Resume Balancer

Restore Sharded Cluster

1. Dissimilar # shards to restore to
2. Different shard keys?
3. Selective restores
4. Consolidate shards
5. Changing addresses of config/shards

Tips and Tricks

- `mongodump/mongorestore`
  - `--oplog[Replay]`
  - `--objcheck/--repair`
  - `--dbpath`
  - `--query/--filter`
- `bsondump`
  - inspect data at console
- LVM snapshot time/space tradeoff
  - Multi-EBS (RAID) backup
  - clean up snapshots

11 Security

Security **(page 165)** An overview of security options for MongoDB

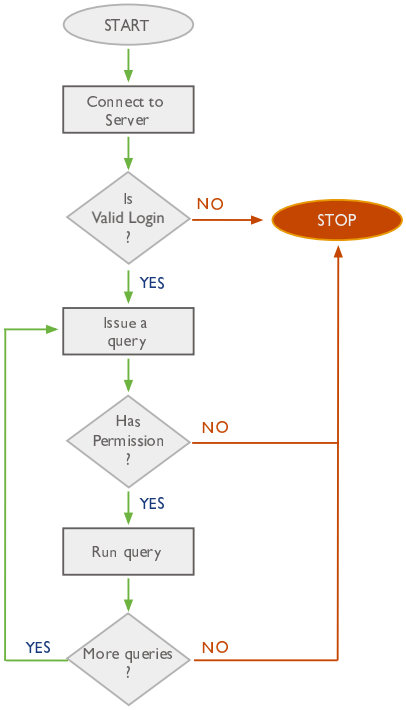
11.1 Security

Learning Objectives

Upon completing this module students should understand:

- Security options for MongoDB
- Basics of native auth for MongoDB
- User roles in MongoDB
- How to manage user roles in MongoDB

Overview



Note:

- You should only run MongoDB in a trusted environment.
- However, MongoDB offers security features from several angles:
  - Authentication: Is the client who they say they are?
  - Authorization: Is the client allowed to do this?
  - Network Exposure: Is this query or login coming from the place we expect?
- You are welcome to use any features you desire, or none.

- All security is off by default.

Authentication Options

- Challenge/response authentication using SCRAM-SHA-1 (username & password)
- x.509 Authentication (using x.509 Certificates)
- Kerberos (through an Enterprise subscription)
- LDAP

Note:

- Although there is a SCRAM-SHA-2 algorithm that addressed some vulnerabilities in SCRAM-SHA-1, it would not benefit MongoDB.

Authorization via MongoDB

- Each user has a set of potential roles
  - read, readWrite, dbAdmin, etc.
- Each role applies to *one* database
  - A single user can have roles on each database.
  - Some roles apply to all databases.
  - You can also create custom roles.

Network Exposure Options

- bindIp limits the ip addresses the server listens on.
- Using a non-standard port can provide a layer of obscurity.
- MongoDB should still be run only in a trusted environment.

Encryption (SSL)

- MongoDB can be configured at build time to run with SSL.
- To get it, build from the source code with –ssl.
- Alternatively, use MongoDB Enterprise.
- Allows you to use public key encryption.
- You can also validate with x.509 certificates.

Native MongoDB Auth

- Uses SCRAM-SHA-1 for challenge/response
- Sometimes called MongoDB-CR
- Start a mongod instance with `--auth` to enable this feature
- You can initially login using localhost
  - Called the "localhost exception".
  - Stops working when you create a user.

Note:

- Be careful to create a user who can create other users.
- Otherwise you'll be stuck with the users you initially create.
- When upgrading a node to 3.0, it will still use the legacy challenge response mechanism until authSchemaUpgrade is run
- When creating a new user on 3.0, you will not be able to connect to the node via a 2.6 shell (the 2.6 shell doesn't use SCRAM-SHA-1)

11.2 Lab: Creating an Admin User

Exercise: Create an Admin User, Part 1

- Launch a mongo shell.
- Create a user with the role, `userAdminAnyDatabase`
- Use name "roland" and password "12345".
- Enable this user to login on the admin database.

Note:

```
use admin
var role = "userAdminAnyDatabase"
var name = "roland"
var password = "12345"
newUserDoc = { user : name,
               pwd : password,
               roles : [ { role : role,
                           db : "admin" } ] }
db.createUser( newUserDoc )
exit
```

Exercise: Create an Admin User, Part 2

- Launch a mongo shell without logging in.
- Attempt to create a user.
- Exit the shell.
- Log in again as roland.
- Ensure that you can create a user.

Note:

```
mongo -u roland admin -p
```

Remember:

- Once a user is created, the localhost exception no longer applies.
- If that first user can't create other users, you will be extremely limited in your ability

---

Using MongoDB Roles

- Each user logs in on *one* database.
- The user inputs their password on login.
  - Use the -u flag for username.
  - Use the -p flag to enter the password.
- userAdmins may create other users
- But they cannot read/write without other roles.

Note:

- Users must specify the db when logging in.
- Trying to log into another db won't work
- Users may follow `-p` with the password
- They may also enter the password after hitting enter.
- The `mongo` command will prompt for the password before launching the shell.

11.3 Lab: Creating a `readWrite` User

Exercise: Creating a `readWrite` User, Part 1

- Create a user named *vespa*.
- Give *vespa* readWrite access on the *test* and *druidia* databases.
- Create this user so that the login database is *druidia*.

Note:

```
mongo -u roland admin -p

use druidia
var name = "vespa"
var password = "12345"
vespa = { user : name,
          pwd : password,
          roles : [ { role : "readWrite", db : "druidia" },
                    { role : "readWrite", db : "test" } ] }
db.createUser(vespa)
exit
```

Exercise: Creating a `readWrite` User, Part 2

Log in with the user you just created.

Note:

```
mongo -u vespa test -p  # won't work.
mongo -u vespa druidia -p  # will work.

show collections // should be empty
db.foo.insert ( { a : 1 } )
db.foo.find()  // see the doc
use test
db.foo.insert( { a : 1 } )
db.foo.find()
use test2
show collections  // can't
db.foo.find()  // can't
```

MongoDB Custom User Roles

- You can create custom user roles in MongoDB.
- You do this by modifying the system.roles collection.
- You can also inherit privileges from other roles into a given role.
- You won't remember how to do this, so if you need it, consult the docs[34].

[34] http://docs.mongodb.org/manual/core/security-introduction/