MongoDB Developer Training

MongoDB Developer Training

Release 3.0

MongoDB, Inc.

June 11, 2015

Contents

1 Introduction

1.1 Warm Up

Introductions

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

Notes:

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?

Notes:

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?

Notes:

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!"

Notes:

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.

Notes:

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

Notes:

MongoDB is a Document Database

Documents are associative arrays like:

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

Notes:

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?" }
  ]
}
```

Notes:

Vertical Scaling
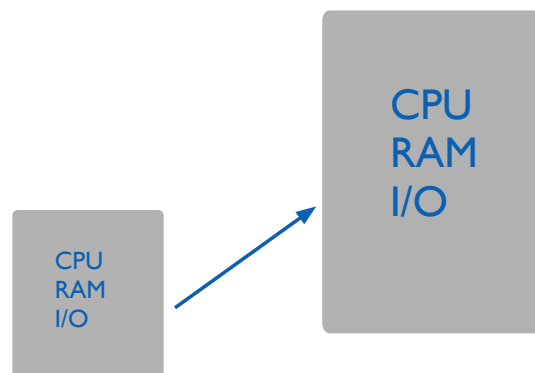
Notes:

Scaling with MongoDB

Notes:
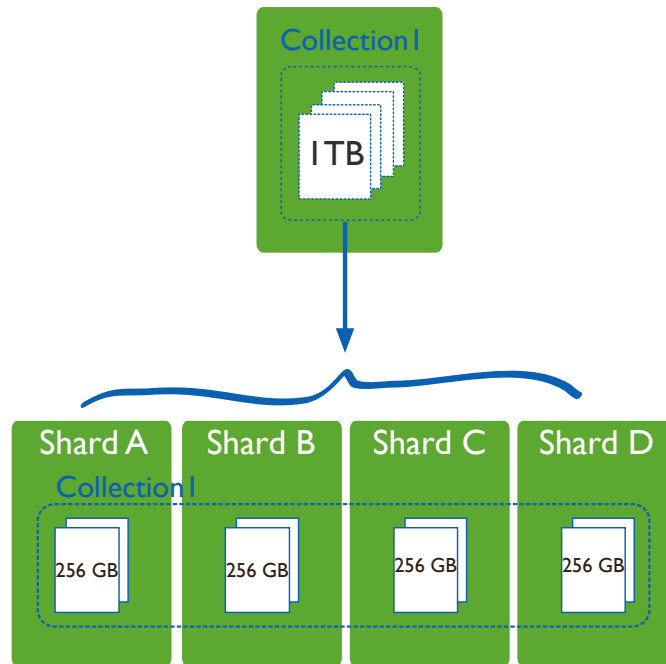
Database Landscape
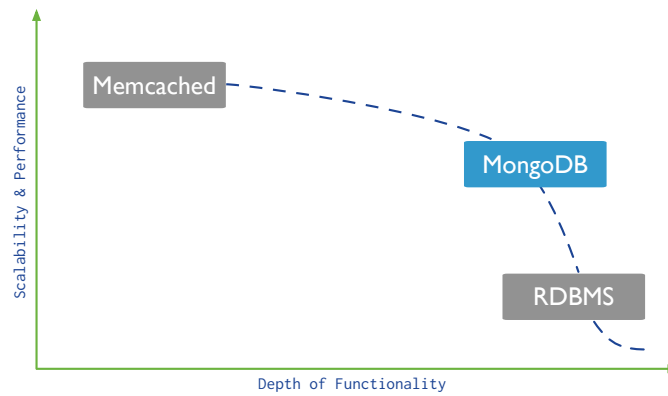
Notes:

MongoDB Deployment Models

Single

Replica

Sharded Cluster

Application Driver — Write Reads → mongod

Application Driver — Write Reads → mongod → mongod mongod

Application Driver — Write Reads → mongos ⇄ Config Server ⇄ Shards (replica sets)
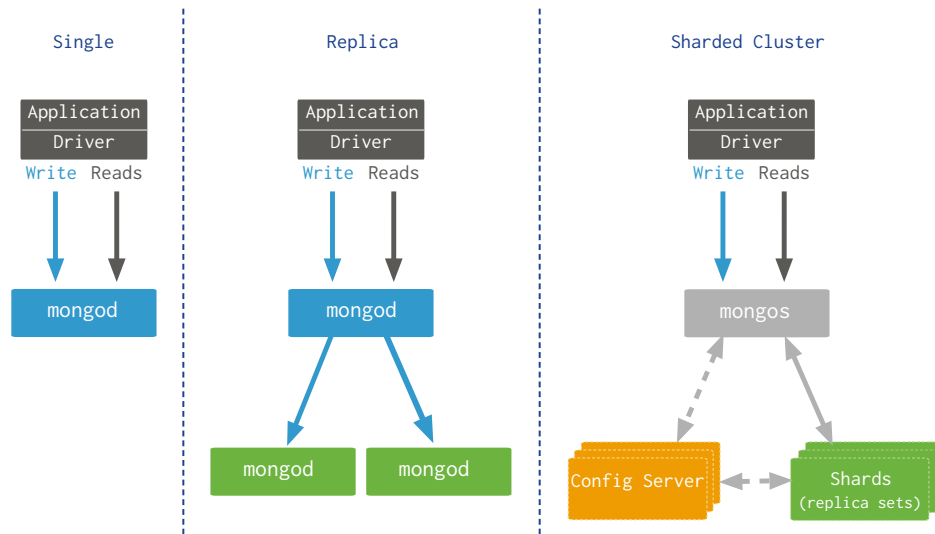
Notes:

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

Notes:

JSON

- JavaScript Object Notation

- Objects are associative arrays.

- They are composed of key-value pairs.

Notes:

A Simple JSON Object

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

Notes:

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].

Notes:

---

[1] http://json.org/

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" ]
  ]
}
```

Notes:

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].

Notes:

---

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

BSON Hello World

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

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

Notes:

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"
```

Notes:

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

Notes:

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).

Notes:

ObjectIds

```
                        Date        MAC address    PID       Counter
      ObjectId:  ___  ___  ___  ___ ___  ___  ___ ___  ___ ___  ___  ___
                        12 byte Hex String
```

Notes:

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

Notes:

What is a Database Storage Engine?

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.

```
┌──────────┐ ┌──────────┐ ┌──────────┐ ┌──────────┐ ┌──────────┐
│ Content  │ │IoT Sensor│ │          │ │ Customer │ │          │
│  Repo    │ │ Backend  │ │Ad Service│ │Analytics │ │ Archive  │
└──────────┘ └──────────┘ └──────────┘ └──────────┘ └──────────┘
```

┌──┐ ┌───┐ ┌──┐
│ │ │   MongoDB Query Language + Native Drivers   │ │ │
│S │ └───┘ │M │
│e │ ┌───┐ │a │
│c │ │   MongoDB Document Data Model   │ │n │
│u │ └───┘ │a │
│r │ ┌─────────┐┌─────────┐┌─────────┐┌─────────┐┌──────┐ │g │
│i │ │ MMAPv1 ││WiredTiger││In-Memory││   ?   ││   ? │ │e │
│t │ └─────────┘└─────────┘└─────────┘└─────────┘└──────┘ │m │
│y │   │e │
└──┘   │n │
                                                            │t │
                                                            └──┘

Supported in MongoDB 3.0   Experimental   Future possible storage engines
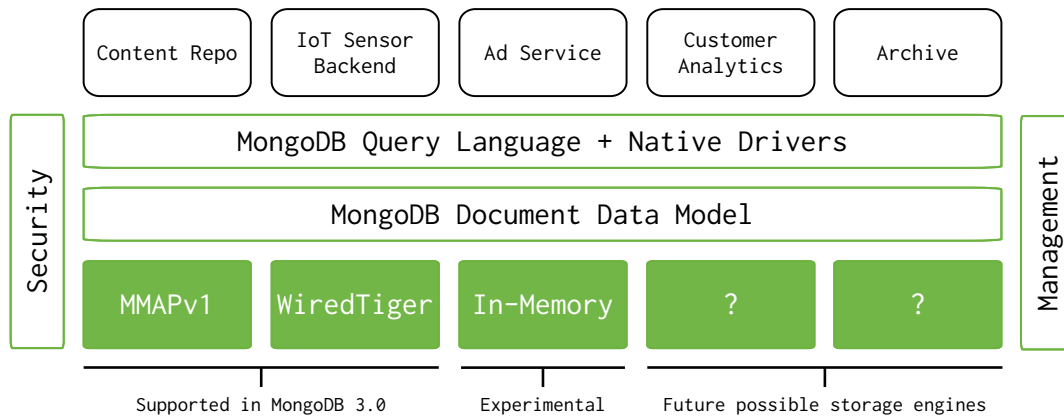
Notes:

How Storage Engines Affect Performance

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

Notes:

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

Notes:

MongoDB Storage Engines

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

- MMAPv1 (default)

- WiredTiger

Notes:

Specifying a MongoDB Storage Engine

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

```
mongod --storageEngine wiredTiger
```

Notes:

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

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

Notes:

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.

Notes:

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
```

Notes:

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

Notes:

Configuring Compression in WiredTiger

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

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

Notes:

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

Notes:

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

Notes:

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

Notes:

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

Notes:

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

Notes:

1.5 Exercise: Installing 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

Notes:

Production Releases

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

- Windows
- OSX
- Linux
- Solaris

Notes:

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 there.
- 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.

Notes:

Linux Setup

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

sudo mkdir -p /data/db

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

Notes:

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"

Notes:

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
    ```

Notes:

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
```

Notes:

The MongoDB Data Directory (MMAPv1)

```
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.

Notes:

The MongoDB Data Directory (WiredTiger)

```
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

Notes:

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.

Notes:

Launch a Mongo Shell

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

```
mongo
```

Display available commands.

```
help
```

Notes:

Explore Databases

Display available databases.

```
show dbs
```

To use a particular database we can type the following.

```
use <database_name>
```

```
db
```

Notes:

Exploring Collections

```
show collections

db.<COLLECTION>.help()

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

Notes:

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.

Notes:

2 CRUD

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

Notes:

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" } )
```

Notes:

Example: Inserting a Document

Experiment with the following commands.

```
use sample

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

db.movies.find()
```

Notes:

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.

Notes:

Example: Assigning _ids

Experiment with the following commands.

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

db.movies.find()
```

Notes:

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.

Notes:

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" } )
```

Notes:

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.

Notes:

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.

Notes:

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()
```

Notes:

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.

Notes:

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()
```

Notes:

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

Notes:

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()
```

Notes:

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.

Notes:

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`.

Notes:

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
```

Notes:

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.

Notes:

Example: Dropping a Collection

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

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

Notes:

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.

Notes:

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
```

Notes:

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()`

Notes:

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.

Notes:

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.

Notes:

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" } )
```

Notes:

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.

Notes:

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" ] } )
```

Notes:

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.

Notes:

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
```

Notes:

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" }
         ] },
    { type : "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
```

Notes:

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()`.

Notes:

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
   }
 })
```

Notes:

Projection: Example

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

Notes:

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.

Notes:

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 } )
```

Notes:

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.

Notes:

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
```

Notes:

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
```

Notes:

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.

Notes:

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(  )
```

Notes:

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 } )
```

Notes:

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.

Notes:

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.

Notes:

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)

Notes:

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" )

.. only:: instructor

   .. note::

      Returns

      .. code-block:: javascript

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

Notes:

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

Notes:

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

Notes:

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
  }
] )
```

Notes:

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" ] } } )
```

Notes:

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.

Notes:

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 } } } )
```

Notes:

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 } }
   ] }
 ] } )
```

Notes:

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.

Notes:

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 } } )
```

Notes:

---

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

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

Notes:

Example: Array Operators

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

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

Notes:

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"
      } } } )
```

Notes:

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?

Notes:

Exercise: Exams and Quizzes

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

Notes:

Exercise: Highest Quiz Score

Find the highest quiz score.

Notes:

Exercise: View Count > 1000

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

Notes:

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.

Notes:

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.)

Notes:

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

Notes:

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.

Notes:

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.

Notes:

`$set` and `$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`.

Notes:

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
  }
] )
```

Notes:

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 } } )
```

Notes:

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 } } )
```

Notes:

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` (already discussed)

- `$unset` (already discussed)

- `$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.

Notes:

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 } } )
```

Notes:

`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.

Notes:

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.

Notes:

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 } )
```

Notes:

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.

Notes:

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" } } )
```

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 } }
)
```

Notes:

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()
```

Notes:

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 } )
```

---

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

Notes:

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.

Notes:

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 } )
```

Notes:

save()

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

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

Notes:

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 })
```

Notes:

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.

Notes:

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".

Notes:

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?

Notes:

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.

Notes:

3 Indexes

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

Notes:

Why Indexes?
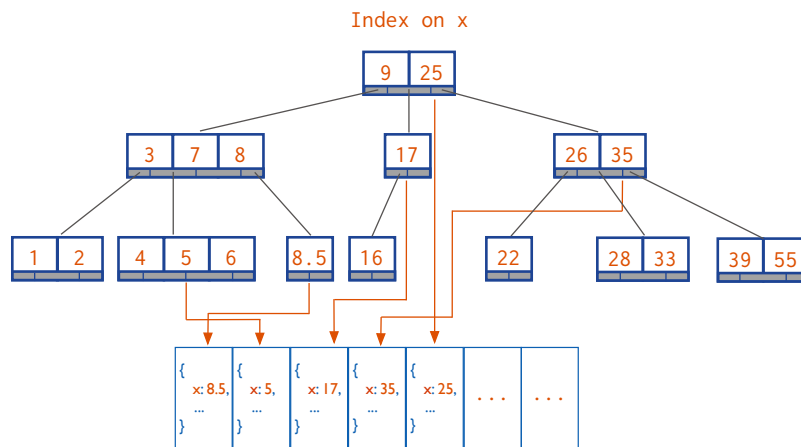
Notes:

Types of Indexes

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

Notes:

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()
```

Notes:

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
      }
    },
```

Notes:

Results of `explain()` - Continued

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

Notes:

`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

Notes:

`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
      }
    },
```

Notes:

`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
  }
  ...
}
```

Notes:

`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.

Notes:

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

Notes:

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()
```

Notes:

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 } } )
```

Notes:

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.

Notes:

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.

Notes:

Listing Indexes

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

List index keys:

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

Listing Indexes

List indexes for a collection:

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

List index keys:

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

Notes:

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

Notes:

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.

Notes:

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.

Notes:

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 } )
```

Additional Index Options

- Sparse
- Unique
- Background

Notes:

Sparse Indexes in MongoDB

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

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

Notes:

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 } )
  ```

Notes:

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 } )
  ```

Notes:

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.

Notes:

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.

Notes:

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 } )
  ```

Notes:

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.

Notes:

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.

Notes:

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)
```

Notes:

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.

Notes:

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.

Notes:

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?

Notes:

`totalKeysExamined > n`

timestamp	username
1	"anonymous"
2	"anonymous"
3	"sam"
4	"anonymous"
5	"martha"

Notes:

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.

Notes:

`totalKeysExamined == n`

username	timestamp
"anonymous"	1
"anonymous"	2
"anonymous"	4
"sam"	2
"martha"	5

Notes:

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.

Notes:

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.

Notes:

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.

Notes:

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.

Notes:

General Rules of Thumb

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

Notes:

Covered Queries
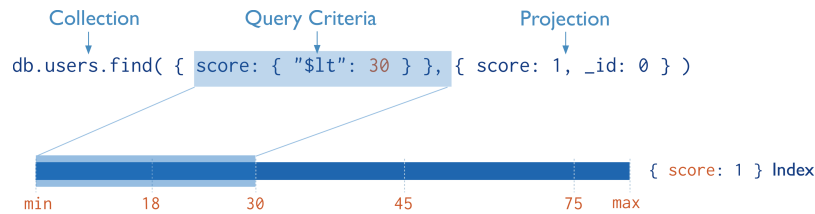
- When a query and projection include only the indexed fields, MongoDB will return results directly from the index.

- There is no need to scan any documents or bring documents into memory.

- These covered queries can be very efficient.

Notes:

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, "budget" : 1 }
                 ).explain("executionStats")
```

Notes:

3.3 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

Notes:

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.

Notes:

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 })
```

Notes:

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 } )
```

Notes:

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()
```

Notes:

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 ] } )
```

Notes:

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.

Notes:

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.

Notes:

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 } )
```

Notes:

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.

Notes:

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 ] } )
```

Notes:

3.4 Hashed Indexes

Learning Objectives

Upon completing this module, students should understand:

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

Notes:

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.

Notes:

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.

Notes:

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.

Notes:

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}.

Notes:

---

[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" } )
```

Notes:

4 Aggregation

Aggregation Tutorial **(page 89)** An introduction to the the aggregation framework, pipeline concept, and stages.

Optimizing Aggregation **(page 101)** Resource management in the aggregation pipeline.

Lab: Aggregating Zip Code Data **(page 104)** Aggregate zip code data.

4.1 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

Notes:

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.

Notes:

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 } )
```

Notes:

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)

Notes:

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.

Notes:

Exercise: The Match Stage

Select only the first two documents using a match stage in an aggregation pipeline.

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

// 2 docs are output from the aggregation pipeline
db.testcol.aggregate( [ { $match : { a : { $lte : 2 } } } ] )
```

Notes:

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.

Notes:

Exercise: Selecting fields with $project

Use the $project operator to pass specific fields in output documents.

```
db.testcol.drop()
for ( var i=1; i<=10; i++ ) {
    db.testcol.insert( { a : i, b : i*2, c : { d : i*4, e : i*8 } } ) }
db.testcol.find()

db.testcol.aggregate( [ { $project : { a : 1 } } ] )

db.testcol.aggregate( [ { $project : { _id : 0, a : 1 } } ] )

db.testcol.aggregate( [ { $project : { a : 1, "c.d": 1 } } ] )
```

Notes:

Exercise: Renaming fields with $project

Use the $project operator to rename a field

```
db.testcol.aggregate( [ { $project : { _id : 0,
                                       sequenceNumber : "$a",
                                       b : 1 } } ] )
```

Notes:

Exercise: Shaping documents with $project

Experiment with the following projections.

```
db.testcol.aggregate( [ { $project : { a : 1, b : 1, d : "$c.d" } } ] )

db.testcol.aggregate( [ { $project : { sequenceNumber : "$a",
                                       ratio : { $divide : [ "$c.d", "$c.e" ] } } } ] )
```

More about $divide[7] in another lesson.

Notes:

---

[7]http://docs.mongodb.org/manual/reference/operator/aggregation/divide/

A Twitter Dataset

- We now have a basic understanding of the aggregation framework.

- Let's look at some richer examples that illustrate the power of MongoDB aggregation.

- 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.

Notes:

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,
        ...
    },
}
```

Notes:

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

Notes:

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.

Notes:

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 } ] )
```

Notes:

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.

- Your result document should look something like the following:

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

Notes:

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 arithmetic or array operators.

Notes:

Group using $avg

```
db.testcol.aggregate( [ { $group : { _id : { a : "$a" },
                                      b_avg : { $avg : "$b" } } } ] )
```

Notes:

Group using $push

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" : 5 }
] )
```

Notes:

Group Aggregation Operators

The complete list of operators available in the group stage:

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

Notes:

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.
- We will use the aggregation framework to do this.

Notes:

Process

- Group together all tweets by a user for every user in our collection
- Count the tweets for each user
- Sort in decreasing order

Notes:

Exercise: Ranking Users by Number of Tweets

Try this aggregation pipeline for yourself.

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

Notes:

Exercise: 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.
- Write an aggregation pipeline that identifies the applications most frequently used to publish tweets.

Notes:

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 as within an embedded document for entities.

- This embedded document also lists any urls and hashtags used in the tweet.

Notes:

Example: User Mentions in a Tweet

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

Notes:

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 })
```

Notes:

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.

Notes:

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.

Notes:

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 }
] )
```

Notes:

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 } }])

Notes:

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 }, ...])

Notes:

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 }, ...])

Notes:

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" }

Notes:

4.2 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.

Notes:

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.

Notes:

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.

Notes:

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.

Notes:

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.

Notes:

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.

Notes:

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 Optimization[8]

Notes:

4.3 Lab: Aggregating Zip Code Data

For the following exercises, import the zips.json file from the USB stick:

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

Exercise: Average Population

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.

Notes:

---

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

Exercise: Cities Starting with a Digit

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] },
        }
    }
])
```

Notes:

5 Schema Design

Schema Design Core Concepts **(page 106)** An introduction to schema design in MongoDB.

Schema Evolution **(page 116)** Considerations for evolving a MongoDB schema design over an application's lifetime.

Common Schema Design Patterns **(page 120)** Common design patterns for representing 1-1, 1-M, and M-M relationships and tree structures in MongoDB.

Lab: Data Model for an E-Commerce Site **(page 127)** Schema design group exercise

5.1 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

Notes:

What is a schema?

- Maps concepts and relationships to data

- Sets expectations for the data

- Minimizes overhead of iterative modifications

- Ensures compatibility

Notes:

Example: Normalized Data Model

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

Notes:

Example: Denormalized Version

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

Notes:

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

Notes:

Three Considerations

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

Notes:

Case Study

- A Library Web Application
- Different schemas are possible.

Notes:

Author Schema

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

Notes:

User Schema

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

Notes:

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 } ]
}
```

Notes:

Example Documents: Author

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

Notes:

Example Documents: User

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

Notes:

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..." }
    ]
}
```

Notes:

Embedded Documents

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

Notes:

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..." }
```

```
    ]
}
```

Notes:

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

Notes:

Linked Documents

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

Notes:

Example: Linked Documents

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

Notes:

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?

Notes:

Arrays

- Array of scalars

- Array of documents

Notes:

Array of Scalars

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

Notes:

Array of Documents

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

Notes:

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)

Notes:

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")
}
```

Notes:

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")
      }
    ]
}
```

Notes:

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")
      }
    ]
}
```

Notes:

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

Notes:

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

Notes:

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.)

Notes:

5.2 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

Notes:

Development Phase

Support basic CRUD functionality:

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

Notes:

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 })
```

Notes:

Production Phase

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

Notes:

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 }});
```

Notes:

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 })
```

Notes:

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

Notes:

Exercise: Recent Reviews

- Display the 10 most recent reviews by a user.

- Make efficient use of memory and disk seeks.

Notes:

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")
        }
    ]
}
```

Notes:

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 }}
);
```

Notes:

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]);
    }
}
```

Notes:

Solution: Recent Reviews, Delete

Deleting a review

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

Notes:

5.3 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

Notes:

One-to-One Relationship

Let's pretend that authors only write one book.

Notes:

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,
}
```

Notes:

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...
}
```

Notes:

One-to-Many Relationship

In reality, authors may write multiple books.

Notes:

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]
}
```

Notes:

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...
}
```

Notes:

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...
}
```

Notes:

Many-to-Many Relationship

Some books may also have co-authors.

Notes:

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]
}
```

Notes:

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 });
```

Notes:

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"
}
```

Notes:

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 }});
```

Notes:

Tree Structures

E.g., modeling a subject hierarchy.

Notes:

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

Notes:

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"
}
```

Notes:

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"
}
```

Notes:

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.

Notes:

5.4 Lab: Data Model for an E-Commerce Site

Introduction

- In this group exercise, we're going to take what we've learned about MongoDB and develop a basic but reasonable data model for an e-commerce site.
- For users of RDBMSs, the most challenging part of the exercise will be figuring out how to construct a data model when joins aren't allowed.
- We're going to model for several entities and features.

Product Catalog

- **Products.** Products vary quite a bit. In addition to the standard production attributes, we will allow for variations of product type and custom attributes. E.g., users may search for blue jackets, 11-inch macbooks, or size 12 shoes. The product catalog will contain millions of products.
- **Product pricing.** Current prices as well a price histories.
- **Product categories.** Every e-commerce site includes a category hierarchy. We need to allow for both that hierarchy and the many-to-many relationship between products and categories.
- **Product reviews.** Every product has zero or more reviews and each review can receive votes and comments.

Notes:

Product Metrics

- **Product views and purchases.** Keep track of the number of times each product is viewed and when each product is purchased.

- **Top 10 lists.** Create queries for top 10 viewed products, top 10 purchased products.

- **Graph historical trends.** Create a query to graph how a product is viewed/purchased over the past.

- **30 days with 1 hour granularity.** This graph will appear on every product page, the query must be very fast.

Notes:

Deliverables

- Sample document and schema for each collection

- Queries the application will use

- Index definitions

Break into groups of two or three and work together to create these deliverables.

Notes:

6 Replica Sets

Introduction to Replica Sets **(page 129)** An introduction to replication and replica sets.

Write Concern **(page 133)** Balancing performance and durability of writes.

Read Preference **(page 138)** Configuring clients to read from specific members of a replica set.

6.1 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

Notes:

Use Cases for Replication

- High Availability
- Disaster Recovery
- Functional Segregation

Notes:

High Availability (HA)

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

Notes:

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

Notes:

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

Notes:

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)

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.

Notes:

Replica Sets
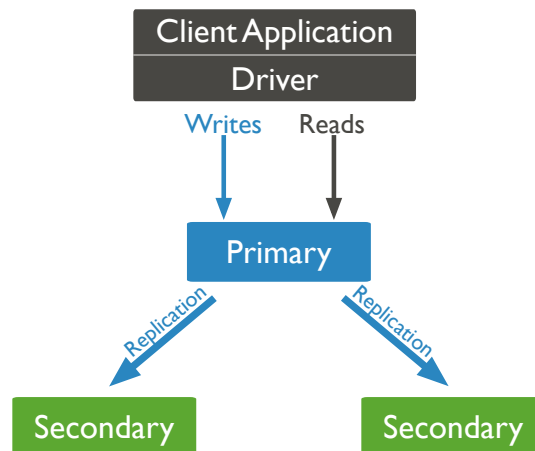
Notes:

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.

Notes:

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.

Notes:

Heartbeats
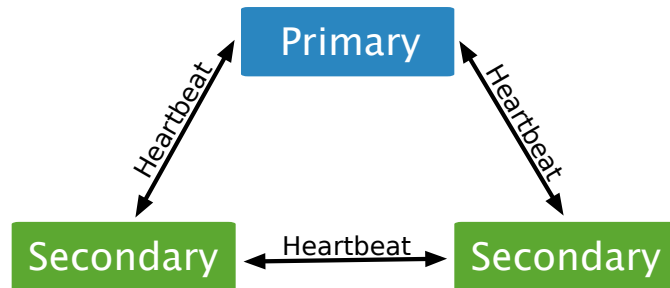
Notes:

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.

Notes:

6.2 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.

Notes:

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

Notes:

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`.

Notes:

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.

Notes:

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.

Notes:

Write Concern: `{ w :  1 }`
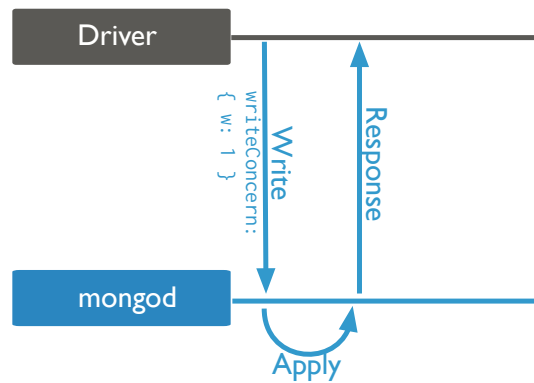
Notes:

Example: `{ w :  1 }`

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

Notes:

Write Concern: `{ w :  2 }`
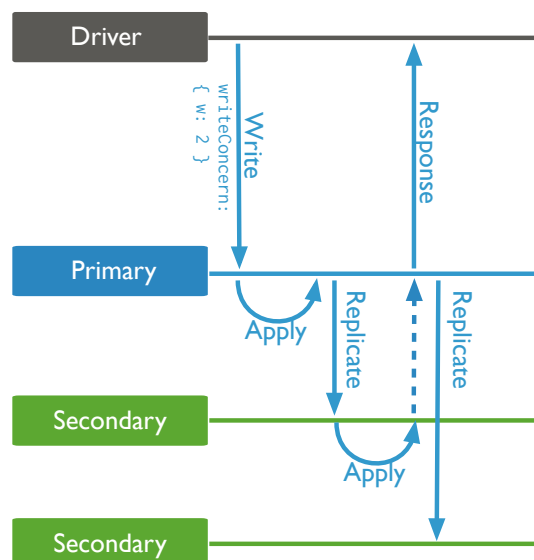
Example: `{ w :  2 }`

```
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" }})
```

Notes:

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" }

Notes:

Further Reading

See Write Concern Reference[9] for more details on write concern configurations, including setting timeouts and identifying specific replica set members that must acknowledge writes (i.e. tag sets[10]).

---

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

6.3 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.

Notes:

Use Cases

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

Notes:

Not for Scaling

- In general, do *not* read from secondaries to provide extra capacity for reads.
- Sharding[11] increases read and write capacity by distributing operations across a group of machines.
- Sharding is a better strategy for adding capacity.

Notes:

---

[11] 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.

Notes:

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)
```

Notes:

7 Sharding

Introduction to Sharding **(page 140)** An introduction to sharding.

7.1 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

Notes:

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.

Notes:

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

Notes:

Vertical Scaling

- Adding more RAM, faster disks, etc.
- When is this the solution?
- First, consider a concept called the `working set`.

Notes:

The Working Set
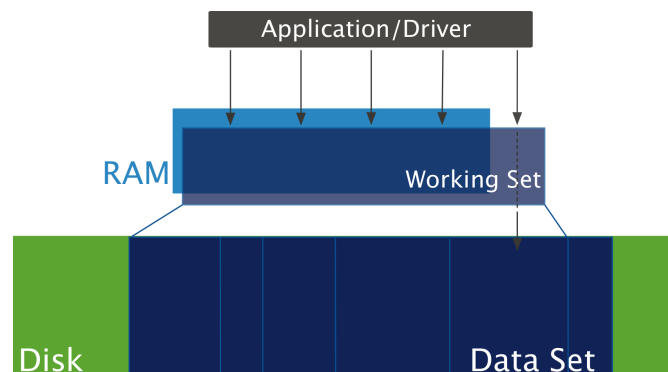
Notes:

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.

Notes:

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.

Notes:

A Model that Does Not Scale

Load   Latency
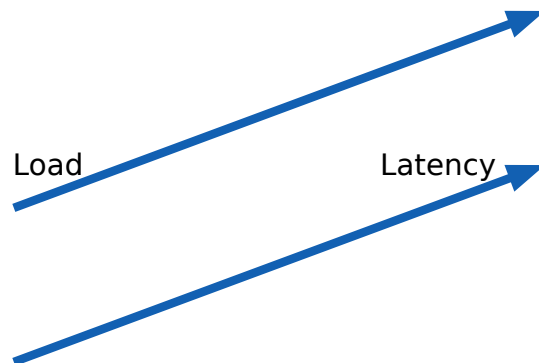
Notes:

A Scalable Model

Load   Latency
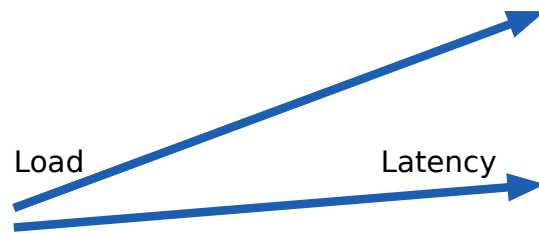
Notes:

Sharding Basics
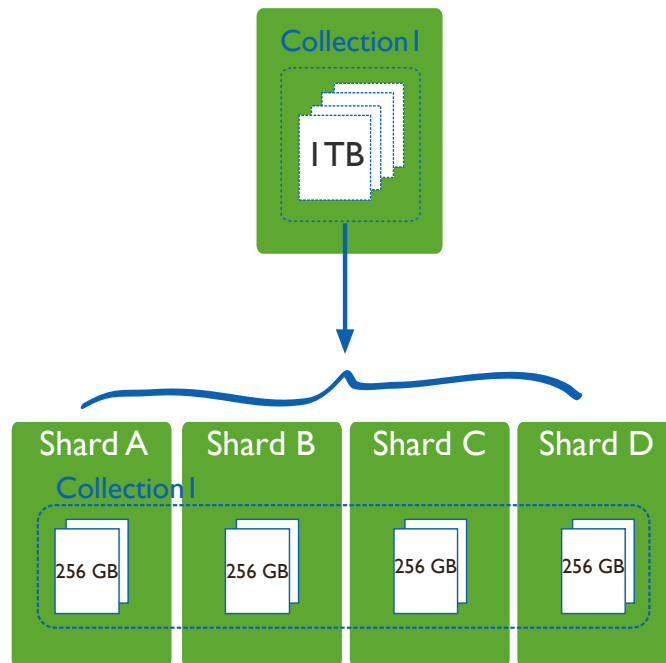
Notes:

Sharded Cluster Architecture
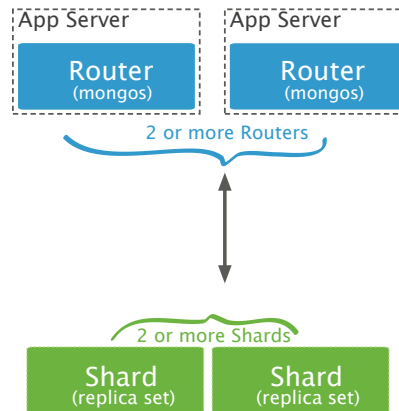
Notes:

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.

Notes:

Config Servers

Notes:
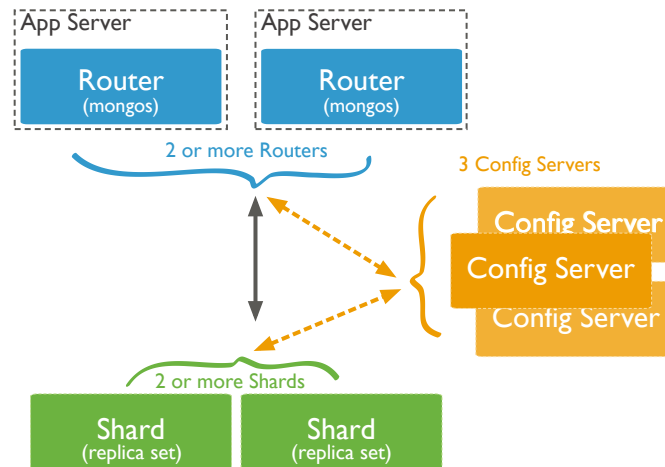
Config Server Hardware Requirements

- 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.

Notes:

When to Shard

- If you have more data than one machine can hold on its drives
- If your application is write heavy and you experiencing too much latency.
- If your working set outgrows the memory you can allocate to a single machine.

Notes:

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.

Notes:

Balancing Shards

- MongoDB divides data into `chunks`.
- This is bookkeeping metadata.
  - There is nothing in a document that indicates its chunk.
  - The document does not need to be updated if its assigned chunk changes.
- 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.

Notes:

What is a Shard Key?

- You must define a shard key for a sharded collection.
- Based on one or more fields that every document must contain.
- Is immutable.
- 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

Notes:

Targeted Query Using Shard Key
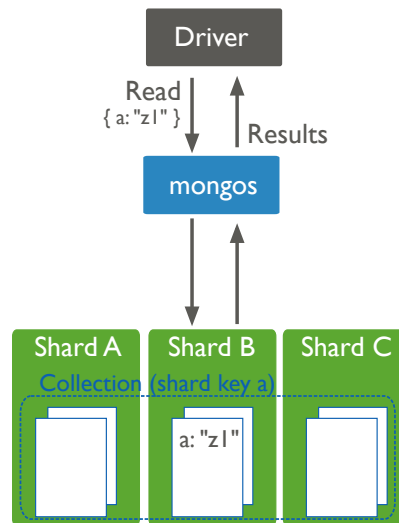
Notes:

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.

Notes:

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.

Notes:

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 or reads are routed to a particular server

Notes:

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.

Notes:

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.

Notes:

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.

Notes:

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.

Notes:

8 MMS & Ops Manager

8.1 MongoDB Management Service (MMS) & Ops Manager

Learning Objectives

Upon completing this module students should understand:

- Features of the MongoDB Management Service (MMS) & Ops Manager
- Available deployment options
- The components of MMS & Ops Manager
- MMS demo

Notes:

MMS and Ops Manager

All services for managing a MongoDB cluster or group of clusters:

- Monitoring
- Automation
- Backups

Notes:

Deployment Options

- MMS: Hosted, http://mms.mongodb.com
- Ops Manager: On-premises

Notes:

Architecture

MongoDB Clusters

MMS / Ops Manager

Monitoring

Backup

Automation
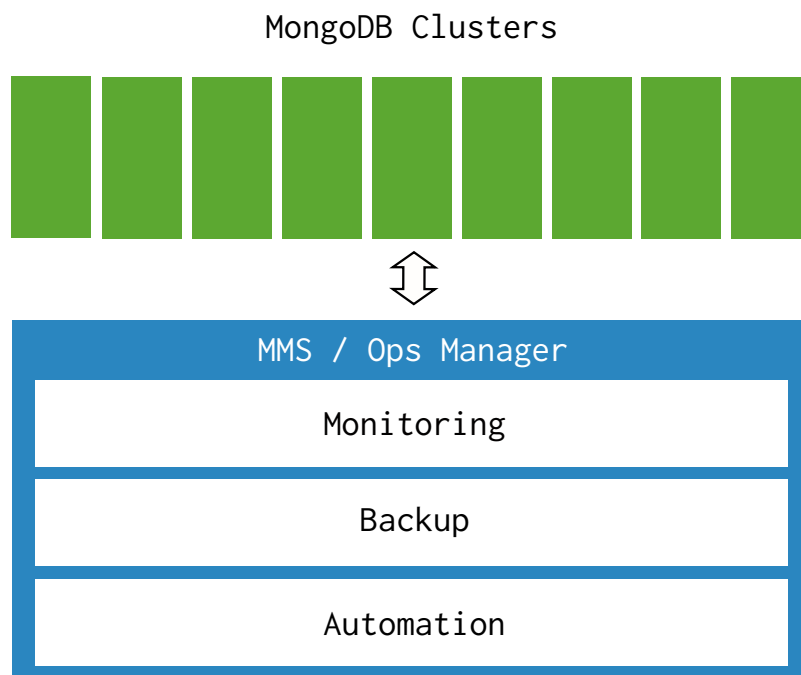
Notes:

MMS

- Manage MongoDB instances anywhere with a connection to MMS
- Option to provision servers via AWS integration

Ops Manager

On-premises MMS, with additional features for:

- Alerting (SNMP)
- Deployment configuration (e.g. backup redundancy across internal data centers)
- Global control of multiple MongoDB clusters

Notes:

MMS and Ops Manager Use Cases

- Manage a 1000 node cluster (monitoring, backups, automation)
- Manage a personal project (3 node replica set on AWS, using MMS)
- Manage 40 deployments (with each deployment having different requirements)

Notes:

Creating an MMS Account

Free account at mms.mongodb.com

Notes:

8.2 Automation

Learning Objectives

Upon completing this module students should understand:

- Use cases for MMS / Ops Manager Automation
- The MMS / Ops Manager Automation internal workflow

Notes:

What is Automation?

Fully managed MongoDB deployment on your own servers:

- Automated provisioning
- Dynamically add capacity (e.g. add more shards or replica set nodes)
- Upgrades
- Admin tasks (e.g. change the size of the oplog)

Notes:

How Does Automation Work?

- Automation agent installed on each server in cluster
- Administrator creates design goal for system (through MMS / Ops Manager interface)
- Automation agents periodically check with MMS / Ops Manager to get new design instructions
- Agents create and follow a plan for implementing cluster design
- Minutes later, cluster design is complete, cluster is in goal state

Notes:

Automation Agents

Machines in Data Center

Automation Agent	Automation Agent	Automation Agent

⇕ ⇕ ⇕

MMS / Ops Manager
Automation
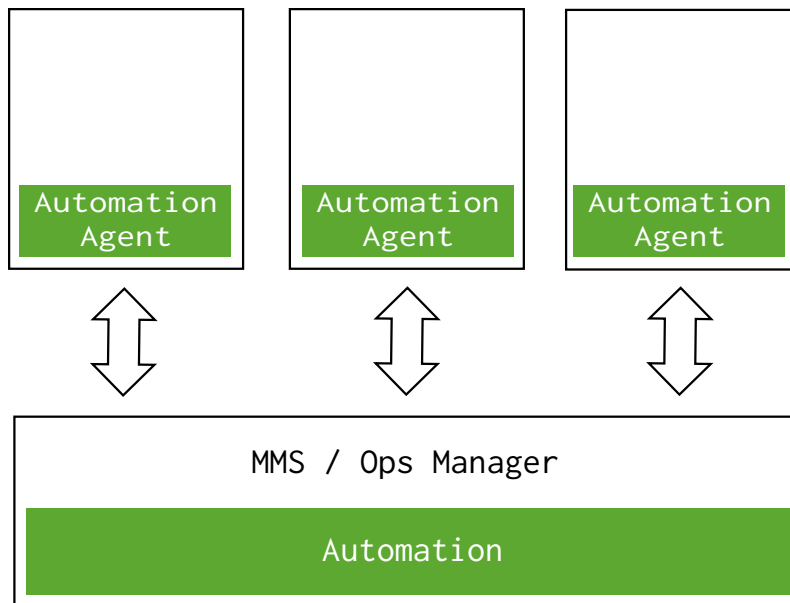
Notes:

Sample Use Case

Administrator wants to create a 100 shard cluster, with each shard comprised of a 3 node replica set:

- Administrator installs automation agent on 300 servers
- Cluster design is created in MMS / Ops Manager, then deployed to agents
- Agents execute instructions until 100 shard cluster is complete (usually several minutes)

Notes:

Upgrades Using Automation

- Upgrades without automation can be a manually intensive process (e.g. 300 servers)
- A lot of edge cases when scripting (e.g. 1 shard has problems, or one replica set is a mixed version)
- One click upgrade with MMS / Ops Manager Automation for the entire cluster

Notes:

Automation: Behind the Scenes

- Agents ping MMS / Ops Manager for new instructions
- Agents compare their local configuration file with the latest version from MMS / Ops Manager
- Configuration file in json
- All communications over SSL

```json
{
    "groupId": "55120365d3e4b0cac8d8a52a737",
    "state": "PUBLISHED",
    "version": 4,
    "cluster": { ...
```

Notes:

Configuration File

When version number of configuration file on MMS / Ops Manager is greater than local version, agent begins making a plan to implement changes:

```json
"replicaSets": [
{
    "_id": "shard_0",
    "members": [
        {
            "_id": 0,
            "host": "DemoCluster_shard_0_0",
            "priority": 1,
            "votes": 1,
            "slaveDelay": 0,
            "hidden": false,
            "arbiterOnly": false
        },
    ...
```

Notes:

Automation Goal State

Automation agent is considered to be in goal state after all cluster changes (related to the individual agent) have been implemented.

Notes:

Demo

Notes:

8.3 Exercise: Cluster Automation

Learning Objectives

Upon completing this exercise students should understand:

- How to deploy, dynamically resize, and upgrade a cluster with MMS Automation

Notes:

MMS Automation Support

Windows machines are not supported at this time.

Notes:

Exercise #1

Using your personal computer, create a cluster using MMS automation with the following topology:

- 3 shards

- Each shard is a 3 node replica set (2 data bearing nodes, 1 arbiter)

- Version 2.6.8 of MongoDB

- **To conserve space on your local machine, set "smallfiles" = true and "oplogSize" = 10**

Exercise #2

Modify the cluster topology from Exercise #1 to the following:

- 4 shards (add one shard)

- Version 3.0.1 of MongoDB (upgrade from 2.6.8 -> 3.0.1)

Notes:

9 Supplementary Material (Time Permitting)

Geospatial Indexes **(page 158)** Geospatial indexes: both those on legacy coordinate pairs and those supporting queries that calculate geometries on an earth-like sphere.

TTL Indexes **(page 167)** Time-To-Live Indexes.

Text Indexes **(page 168)** Free text indexes on string fields.

9.1 Geospatial Indexes

Learning Objectives

Upon completing this module, students should understand:

- Use cases of geospatial indexes
- The two types of geospatial indexes
- How to create 2d geospatial indexes
- How to query for documents in a region
- How to create 2dsphere indexes
- Types of geoJSON objects
- How to query using 2dsphere indexes

Notes:

Introduction to Geospatial Indexes

We can use geospatial indexes to quickly determine geometric relationships:

- All points within a certain radius of another point
- Whether or not points fall within a polygon
- Whether or not two polygons intersect

Notes:

Easiest to Start with 2 Dimensions

- Initially, it is easiest to think about geospatial indexes in two dimensions.

- One type of geospatial index in MongoDB is a flat 2d index.

- With a geospatial index we can, for example, search for nearby items.

- This is the type of service that many phone apps provide when, say, searching for a nearby cafe.

- We might have a query location identified by an X in a 2d coordinate system.

Notes:

Location Field

- A geospatial index is based on a location field within documents in a collection.

- The structure of location values depends on the type of geospatial index.

- We will go into more detail on this in a few minutes.

- We can identify other documents in this collection with Xs in our 2d coordinate system.

Notes:

Find Nearby Documents

- A geospatial index enables us to efficiently query a collection based on geometric relationships between documents and the query.

- For example, we can quickly locate all documents within a certain radius of our query location.

- In this example, we've illustrated a `$near` query in a 2d geospatial index.

Notes:

Flat vs. Spherical Indexes

There are two types of geospatial indexes:

- Flat, made with a `2d` index
- Two-dimensional spherical, made with the `2dsphere` index
    - Takes into account the curvature of the earth
    - Joins any two points using a geodesic or "great circle arc"
    - Deviates from flat geometry as you get further from the equator, and as your points get further apart

Notes:

Flat Geospatial Index

- This is a Cartesian treatment of coordinate pairs.
- E.g., the index would not reflect the fact that the shortest path from Canada to Siberia is over the North Pole (if units are degrees).
- 2d indexes can be used to describe any flat surface.
- Recommended if:
    - You have legacy coordinate pairs (MongoDB 2.2 or earlier).
    - You do not plan to use geoJSON objects such as LineStrings or Polygons.
    - You are not going to use points far enough North or South to worry about the Earth's curvature.

Notes:

Spherical Geospatial Index

- Spherical indexes model the curvature of the Earth
- If you want to plot the shortest path from the Klondike to Siberia, this will know to go over the North Pole.
- Spherical indexes use geoJSON objects (Points, LineString, and Polygons)
- Coordinate pairs are converted into geoJSON Points.

Notes:

Creating a 2d Index

Creating a 2d index:

```
db.<COLLECTION>.createIndex(
 { field_name : "2d", <optional additional field> : <value> },
 { <optional options document> } )
```

Possible options key-value pairs:

- `min :  <lower bound>`
- `max :  <upper bound>`
- `bits :  <bits of precision for geohash>`

Notes:

Exercise: Creating a 2d Index

Create a 2d index on the collection `testcol` with:

- A min value of -20
- A max value of 20
- 10 bits of precision
- The field indexed should be `xy`.

Notes:

Inserting Documents with a 2d Index

There are two accepted formats:

- Legacy coordinate pairs
- Document with the following fields specified:
  - `lng` (longitude)
  - `lat` (latitude)

Notes:

Exercise: Inserting Documents with 2d Fields

- Insert 2 documents into the 'twoD' collection.

- Assign 2d coordinate values to the `xy` field of each document.

- Longitude values should be -3 and 3 respectively.

- Latitude values should be 0 and 0.4 respectively.

Notes:

Querying Documents Using a 2d Index

- Use `$near` to retrieve documents close to a given point.

- Use `$geoWithin` to find documents with a shape contained entirely within the query shape.

- Use the following operators to specify a query shape:

    - `$box`

    - `$polygon`

    - `$center` (circle)

Notes:

Example: Find Based on 2d Coords

Write a query to find all documents in the testcol collection that have an xy field value that falls entirely within the circle with center at [-2.5, -0.5] and a radius of 3.

```
db.testcol.find( { xy : { $geoWithin : { $center : [ [ −2.5, −0.5 ], 3 ] } } }
```

Notes:

Creating a 2dsphere Index

You can index one or more 2dsphere fields in an index.

```
db.<COLLECTION>.createIndex( { <location field> : "2dsphere" } )
```

Notes:

The geoJSON Specification

- The geoJSON format encodes location data on the earth.
- The spec is at http://geojson.org/geojson-spec.html
- This spec is incorporated in MongoDB 2dsphere indexes.
- It includes Point, LineString, Polygon, and combinations of these.

Notes:

geoJSON Considerations

- The coordinates of points are given in degrees (latitude then longitude).
- The LineString that joins two points will always be a geodesic.
- Short lines (around a few hundred kilometers or less) will go about where you would expect them to.
- Polygons are made of a closed set of LineStrings.

Notes:

Simple Types of 2dsphere Objects

Point: A single point on the globe

```
{ <field_name> : { type : "Point",
                   coordinates : [ <longitude>, <latitude> ] } }
```

Notes: **LineString**: A geodesic line that is defined by its two end Points

```
{ <field_name> : { type : "LineString",
                   coordinates : [ [ <longitude 1>, <latitude 1> ],
                                   [ <longitude 2>, <latitude 2> ],
                                   ...,
                                   [ <longitude n>, <latitude n> ] ] } }
```

Notes:

Polygons

Simple Polygon:

```
{ <field_name> : { type : "Polygon",
                   coordinates : [ [ [ <Point1 coordinate pair> ],
                                     [ <Point2 coordinate pair> ],
                                     ...
                                     [ <Point1 coordinate pair again> ] ]
               } }
```

Polygon with One Hole:

```
{ <field_name> : { type : "Polygon",
                   coordinates : [ [ <Points that define outer polygon> ],
                                   [ <Points that define inner polygon> ] ]
               } }
```

Notes:

Other Types of 2dsphere Objects

- **MultiPoint**: One or more Points in one document
- **MultiLine**: One or more LineStrings in one document
- **MultiPolygon**: One or more Polygons in one document
- **GeometryCollection**: One or more geoJSON objects in one document

Notes:

Exercise: Inserting geoJSON Objects (1)

Create a coordinate pair for each the following airports. Create one variable per airport.
- LaGuardia (New York): 40.7772° N, 73.8726° W
- JFK (New York): 40.6397° N, 73.7789° W
- Newark (New York): 40.6925° N, 74.1686° W
- Heathrow (London): 52.4775° N, 0.4614° W
- Gatwick (London): 51.1481° N, 0.1903° W
- Stansted (London): 51.8850° N, 0.2350° E
- Luton (London): 51.9000° N, 0.4333° W

Notes:

Exercise: Inserting geoJSON Objects (2)

- Now let's make arrays of these.
- Put all the New York area airports into an array called `nyPorts`.
- Put all the London area airports into an array called `londonPorts`.
- Create a third array for flight numbers: "AA4453", "VA3333", "UA2440".

Notes:

Exercise: Inserting geoJSON Objects (3)

- Create documents for every possible New York to London flight.

- Include a `flightNumber` field for each flight.

Notes:

Exercise: Creating a 2dsphere Index

- Create two indexes on the collection `flights`.

- Make the first a compound index on the fields:

  – `origin`

  – `destination`

  – `flightNumber`

- Specify 2dsphere indexes on both `origin` and `destination`.

- Specify a simple index on `name`.

- Make the second index just a 2dsphere index on destination.

Notes:

Querying 2dsphere Objects

`$geoNear`: Finds all points, orders them by distance from a position.

```
{ <field name> : { $near : { $geometry : {
                               type : "Point",
                               coordinates : [ lng, lat ] },
                               $maxDistance : <meters> } } } }
```

`$near`: Just like `$geoNear`, except in very edge cases; check the docs.

`$geoWithin`: Only returns documents with a location completely contained within the query.

`$geoIntersects`: Returns documents with their indexed field intersecting any part of the shape in the query.

Notes:

9.2 TTL Indexes

Learning Objectives

Upon completing this module students should understand:

- How to create a TTL index
- When a TTL indexed document will get deleted
- Limitations of TTL indexes

Notes:

TTL Index Basics

- TTL is short for "Time To Live".
- TTL indexes must be based on a field of type `Date` (including `ISODate`) or `Timestamp`.
- Any `Date` field older than `expireAfterSeconds` will get deleted at some point.

Notes:

Creating a TTL Index

Create with:

```
db.<COLLECTION>.createIndex( { field_name : 1 },
                             { expireAfterSeconds : some_number } )
```

Notes:

Exercise: Creating a TTL Index

Let's create a TTL index on the `ttl` collection that will delete documents older than 30 seconds. Write a script that will insert documents at a rate of one per second.

```
db.sessions.drop()
db.sessions.createIndex( { "last_user_action" : 1 },
                         { "expireAfterSeconds" : 30 } )

i = 0
while (true) {
    i += 1;
    db.sessions.insert( { "last_user_action" : ISODate(), "b" : i } );
    sleep(1000);  // Sleep for 1 second
}
```

Notes:

Exercise: Check the Collection

Then, leaving that window open, open up a new terminal and connect to the database with the mongo shell. This will allow us to verify the TTL behavior.

```
// look at the output and wait. After a ramp-up of up to a minute or so,
// count() will be reset to 30 once/minute.
while (true) {
    print(db.sessions.count());
    sleep(100);
}
```

Notes:

9.3 Text Indexes

Learning Objectives

Upon completing this module, students should understand:

- The purpose of a text index

- How to create text indexes

- How to search using text indexes

- How to rank search results by relevance score

What is a Text Index?

- A text index is based on the tokens (words, etc.) used in string fields.

- MongoDB supports text search for a number of languages.

- Text indexes drop language-specific stop words (e.g. in English "the", "an", "a", "and", etc.).

- Text indexes use simple, language-specific suffix stemming (e.g., "running" to "run").

Notes:

Creating a Text Index

You create a text index a little bit differently than you create a standard index.

```
db.<COLLECTION>.createIndex( { <field name> : "text" } )
```

Notes:

Exercise: Creating a Text Index

Create a text index on the "dialog" field of the montyPython collection.

```
db.montyPython.createIndex( { dialog : "text" } )
```

Notes:

Creating a Text Index with Weighted Fields

- Default weight of 1 per indexed field.

- Weight is relative to other weights in text index.

```
db.<COLLECTION>.createIndex(
  { "title" : "text", "keywords": "text", "author" : "text" },
  { "weights" : {
    "title" : 10,
    "keywords" : 5
  }})
```

- Term match in "title" field has 10 times (i.e. 10:1) the impact as a term match in the "author" field.

Creating a Text Index with Weighted Fields

- The default weight is 1 for each indexed field.

- The weight is relative to other weights in a text index.

```
db.<COLLECTION>.createIndex(
    { "title" : "text", "keywords": "text", "author" : "text" },
    { "weights" : {
        "title" : 10,
        "keywords" : 5
    }})
```

- Term match in "title" field has 10 times (i.e. 10:1) the impact as a term match in the "author" field.

Notes:

Text Indexes are Similar to Multikey Indexes

- Continuing our example, you can treat the `dialog` field as a multikey index.
- A multikey index with each of the words in `dialog` as values.
- You can query the field using the `$text` operator.

Notes:

Exercise: Inserting Texts

Let's add some documents to our montyPython collection.

```
db.montyPython.insert( [
{ _id : 1,
  dialog : "What is the air-speed velocity of an unladen swallow?" },
{ _id : 2,
  dialog : "What do you mean? An African or a European swallow?" },
{ _id : 3,
  dialog : "Huh? I... I don't know that." },
{ _id : 45,
  dialog : "You're using coconuts!" },
{ _id : 55,
  dialog : "What? A swallow carrying a coconut?" } ] )
```

Notes:

Querying a Text Index

Next, let's query the collection. The syntax is:

```
db.<COLLECTION>.find( { $text : { $search : "query terms go here" } } )
```

Notes:

Exercise: Querying a Text Index

Using the text index, find all documents in the montyPython collection with the word "swallow" in it.

```
// Returns 3 documents.
db.montyPython.find( { $text : { $search : "swallow" } } )
```

Notes:

Exercise: Querying Using Two Words

- Find all documents in the montyPython collection with either the word 'coconut' or 'swallow'.

- By default MongoDB ORs query terms together.

- E.g., if you query on two words, results include documents using either word.

```
// Finds 4 documents, 3 of which contain only one of the two words.
db.montyPython.find( { $text : { $search : "coconut swallow" } } )
```

Notes:

Search for a Phrase

- To match an exact phrase, include search terms in quotes (escaped).

- The following query selects documents containing the phrase "European swallow":

```
db.montyPython.find( { $text: { $search: "\"European swallow\"" } } )
```

Notes:

Text Search Score

- The search algorithm assigns a relevance score to each search result.

- The score is generated by a vector ranking algorithm.

- The documents can be sorted by that score.

```
db.<COLLECTION>.find(
    { $text : { $search : "swallow coconut"} },
    { textScore: {$meta : "textScore" } }
).sort(
        { textScore: { $meta: "textScore" } }
) )
```

Notes:

Find out more
mongodb.com | mongodb.org
university.mongodb.com

Having trouble?
File a JIRA ticket:
jira.mongodb.org

Follow us on twitter
@MongoDBInc
@MongoDB