Big Data Analytics

Dr Sirintra Vaiwsri | Email: sirintra.v@itm.kmutnb.ac.th

Databases

Relational Database Management System

(Holmes, 2017; Staragile, 2023)

- The major problem with using RDBMS for Big Data is scalability.
- RDBMS was designed to be used on one server which can be improved but still has its limits, called vertical scaling.
- RDBMS has **ACID** properties which are :
    - **A**tomicity - the incomplete transaction cannot update the database.
    - **C**onsistency - data is consistent before and after a transaction.
    - **I**solation - each transaction does not interfere with others.
    - **D**urability - a database must be updated to ensure data will not be lost if the system fails.

Not only SQL (NoSQL) (Holmes, 2017)

- Non-relational database
- NoSQL has **BASE** properties which are :
  - **B**asically **A**vailable - the system is always available even when a network failure occurs.
  - **S**oft state - means it has flexibility with consistent requirements.
  - **E**ventually consistent - the system eventually becomes consistent.
- NoSQL is horizontal scaling which allows more data to be inserted into the database.
- Working with NoSQL must concern the CAP theorem.

RDBMS VS NoSQL (Edward and Sabharwal, 2015)

	RDBMS	NoSQL
Schema flexibility	Inflexible, often ends up creating new tables	Column-oriented which allows adding more columns. It also supports semi-structured data.
Complex query	Often uses complex JOIN queries which are difficult to implement and maintain	It does not support relationships and foreign keys, thus, no complex query
Data update	If the system does not allow for updating multiple nodes at the same time, there is a risk of node failure	Synchronisation across nodes is challenging. However, NoSQL solutions offer synchronisation options.
Scalability	Low speed for large amounts of data	Provide great scalability

NoSQL

Key-Value Databases

(Bahga and Madisetti, 2019; DEV, 2023; KDnuggets, 2023; Janev et al., 2020)

- It is the simplest NoSQL database.
- It stores data in the form of key-value pairs.
- A key is unique for each data.
- Key is usually a string or an integer.
- A value contains data which can be in the form of attributes or collections.
- Value can be any type of data.

Key-Value Databases (DEV, 2023; KDnuggets, 2023)

- Advantages:
  - Scalability - it is horizontal scaling through partitioning and replication. It also has low overhead.
  - Mobility - it is easy to move from one to another system without changing in code/architecture required.
- Disadvantages:
  - All joins must be done in code.
  - No complex query filters.

Document Databases (Bahga and Madisetti, 2019)

- It is similar to key-value databases in that each document has a unique key (ID).
- Each document can store any type of data.
- Its query is JSON-like documents.
- Therefore, it requires a data format that a database can understand.

Document Databases (DEV, 2023)

- Advantages:
    - It collects data from RAM which is fast to access.
    - It is horizontal scaling.
- Disadvantages:
    - Selecting data from multiple collections requires multiple queries.
    - Data duplication can occur which makes it difficult to handle.

Column-Oriented Databases

(Bahga and Madisetti, 2019; DEV, 2023; KDnuggets, 2023; Janev et al., 2020)

- It stores data in the form of columns.
- Column - it contains name, value, and timestamp.
- Row - it contains one or more columns where different rows are not necessary to contain the same number of columns. It has row-key as a unique key (ID).
- Column family - it contains multiple rows where each row can contain multiple column families.
- Keyspace - it contains multiple column families.

Column-Oriented Databases (DEV, 2023)

- Advantages:
  - It is scalable and flexible.
  - Load and aggregation times are very fast.
- Disadvantages:
  - It is slow when deleting rows.
  - It can be slow when querying data using a join query.

Graph Databases (Bahga and Madisetti, 2019; DEV, 2023)

- It stores data that has a graph structure.
- It shows a relationship between data.
- Nodes have a set of attributes.
- Edges (links) also have a set of attributes.

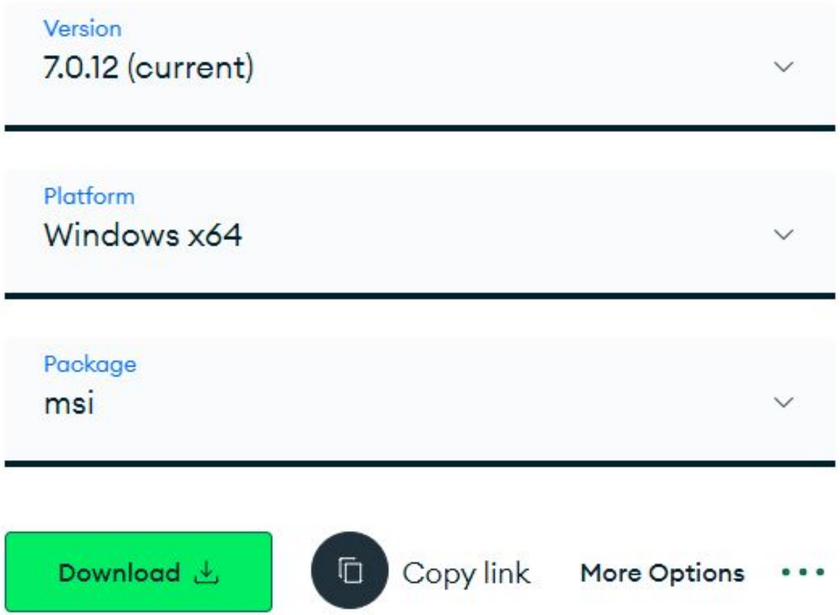
Graph Databases (DEV, 2023)

- Advantages:
  - It is easy to understand data and has descriptive queries.
  - It is flexible.
- Disadvantages:
  - It is difficult to scale.
  - It does not have a standard language.

MongoDB

Install MongoDB

Download MongoDB from here.

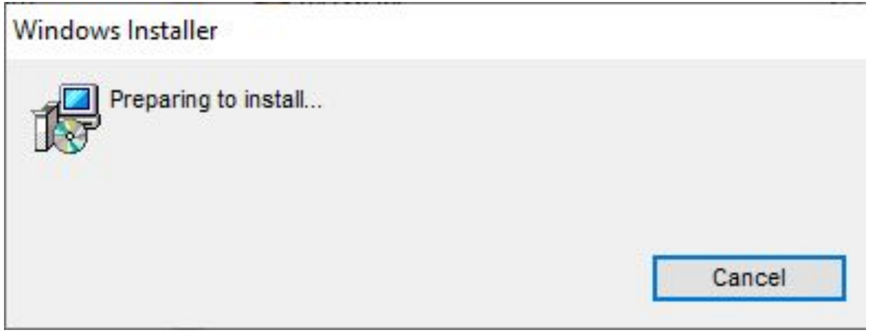


Version
7.0.12 (current)

Platform
Windows x64

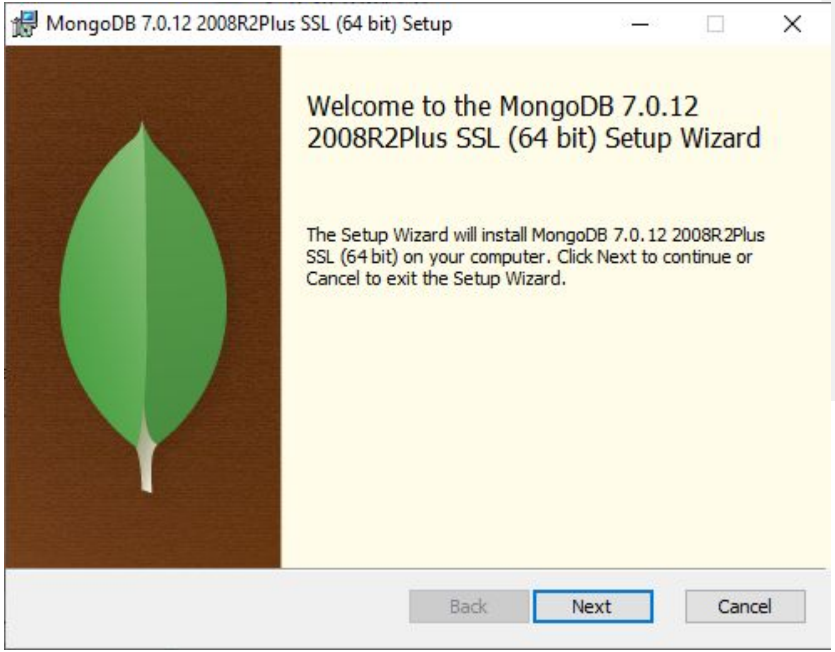
Package
msi

Install MongoDB



Windows Installer

Preparing to install...

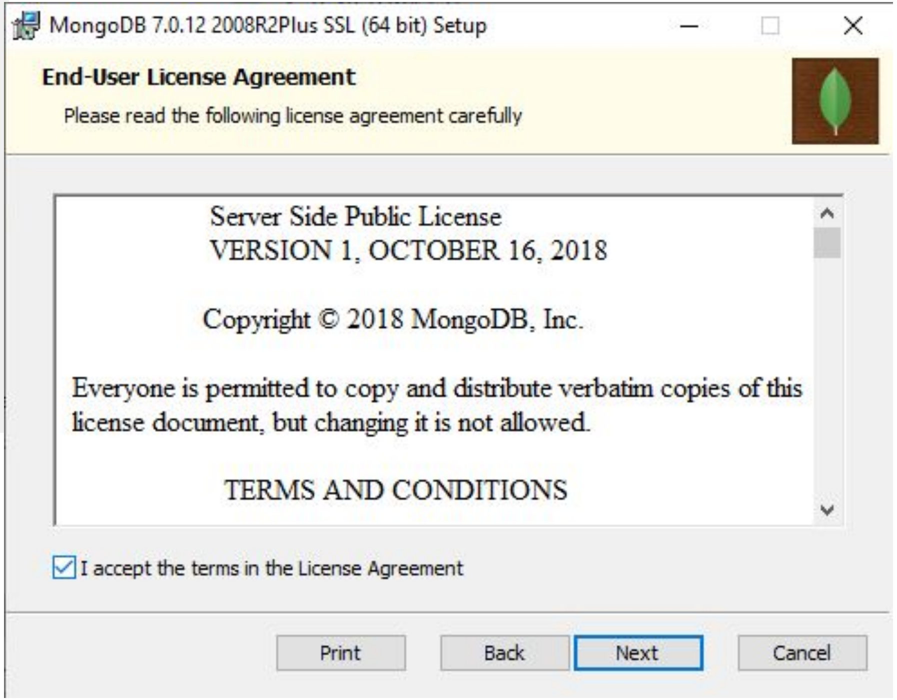


MongoDB 7.0.12 2008R2Plus SSL (64 bit) Setup

Welcome to the MongoDB 7.0.12 2008R2Plus SSL (64 bit) Setup Wizard

The Setup Wizard will install MongoDB 7.0.12 2008R2Plus SSL (64 bit) on your computer. Click Next to continue or Cancel to exit the Setup Wizard.

Install MongoDB



MongoDB 7.0.12 2008R2Plus SSL (64 bit) Setup

End-User License Agreement

Please read the following license agreement carefully

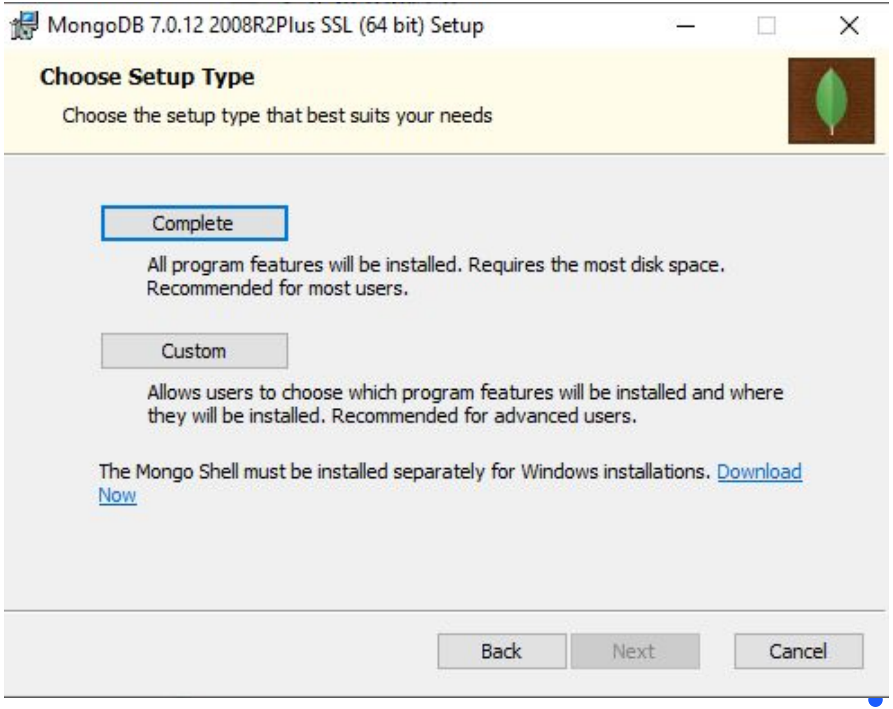
Server Side Public License
VERSION 1, OCTOBER 16, 2018

Copyright © 2018 MongoDB, Inc.

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

TERMS AND CONDITIONS

☑ I accept the terms in the License Agreement



MongoDB 7.0.12 2008R2Plus SSL (64 bit) Setup

Choose Setup Type

Choose the setup type that best suits your needs

Complete

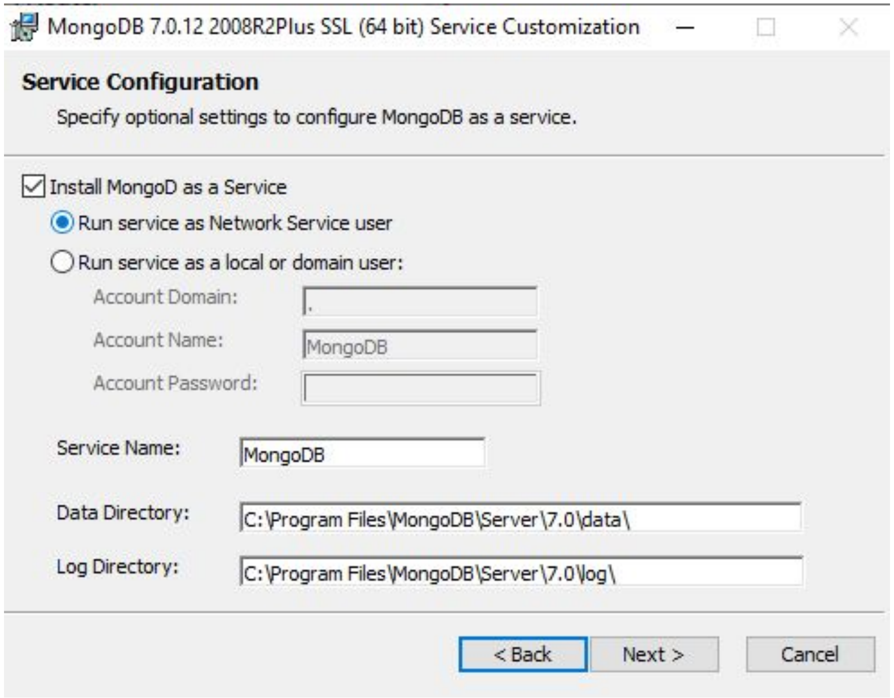
All program features will be installed. Requires the most disk space. Recommended for most users.

Custom

Allows users to choose which program features will be installed and where they will be installed. Recommended for advanced users.

The Mongo Shell must be installed separately for Windows installations. Download Now

Install MongoDB



MongoDB 7.0.12 2008R2Plus SSL (64 bit) Service Customization

Service Configuration

Specify optional settings to configure MongoDB as a service.

☑ Install MongoD as a Service

● Run service as Network Service user

○ Run service as a local or domain user:

Account Domain: .

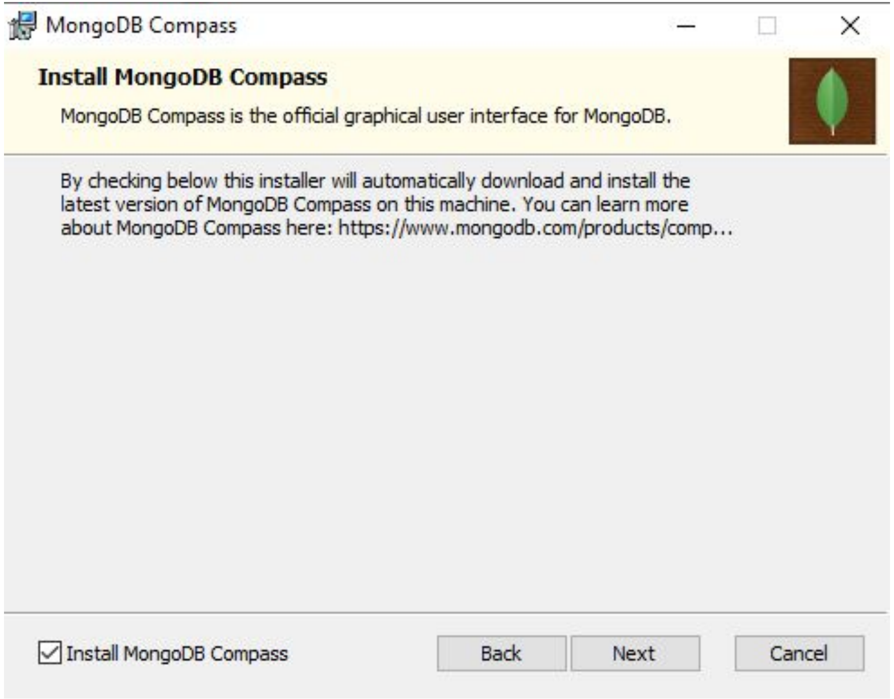
Account Name: MongoDB

Account Password:

Service Name: MongoDB

Data Directory: C:\Program Files\MongoDB\Server\7.0\data\

Log Directory: C:\Program Files\MongoDB\Server\7.0\log\



MongoDB Compass

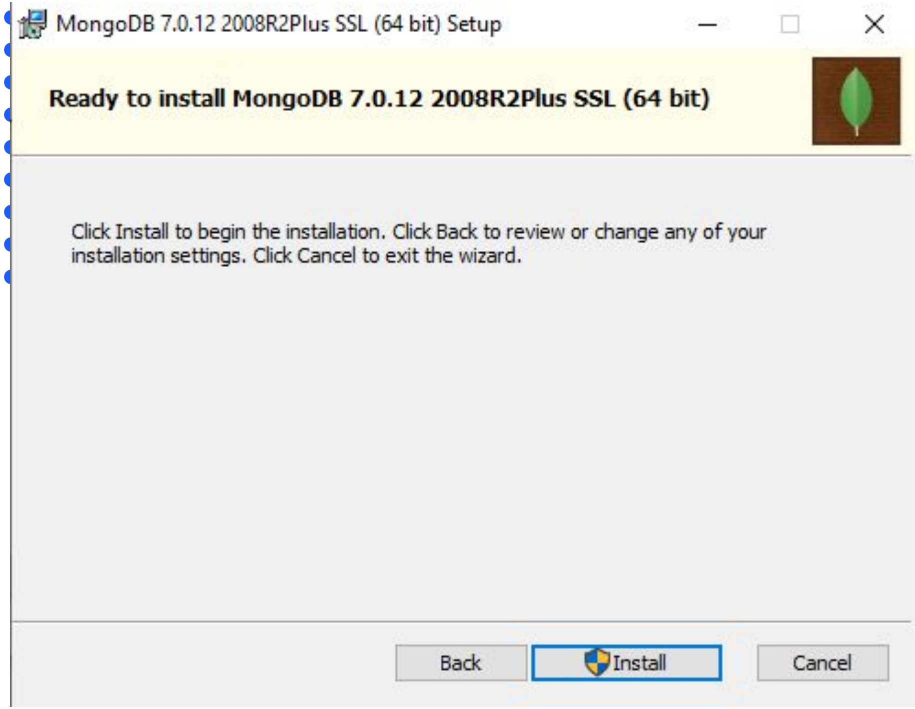
Install MongoDB Compass

MongoDB Compass is the official graphical user interface for MongoDB.

By checking below this installer will automatically download and install the latest version of MongoDB Compass on this machine. You can learn more about MongoDB Compass here: https://www.mongodb.com/products/comp...

☑ Install MongoDB Compass

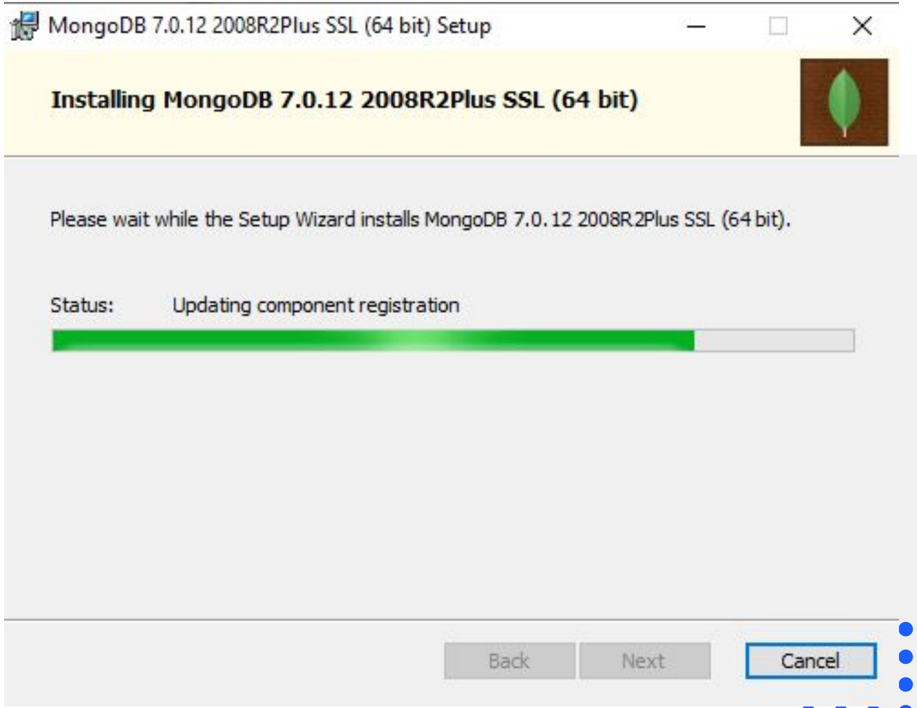
Install MongoDB



MongoDB 7.0.12 2008R2Plus SSL (64 bit) Setup

Ready to install MongoDB 7.0.12 2008R2Plus SSL (64 bit)

Click Install to begin the installation. Click Back to review or change any of your installation settings. Click Cancel to exit the wizard.



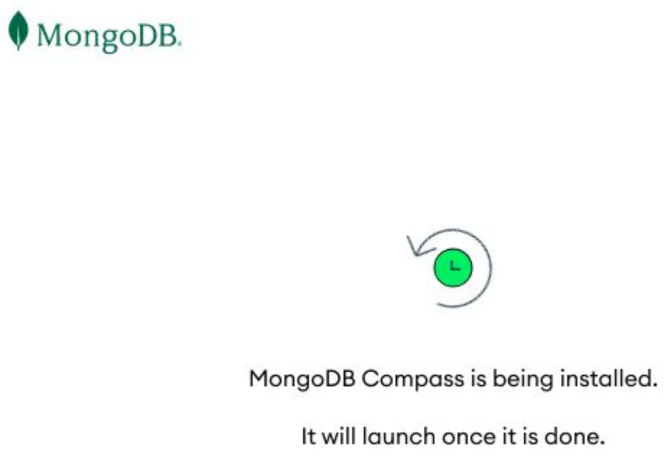
MongoDB 7.0.12 2008R2Plus SSL (64 bit) Setup

Installing MongoDB 7.0.12 2008R2Plus SSL (64 bit)

Please wait while the Setup Wizard installs MongoDB 7.0.12 2008R2Plus SSL (64 bit).

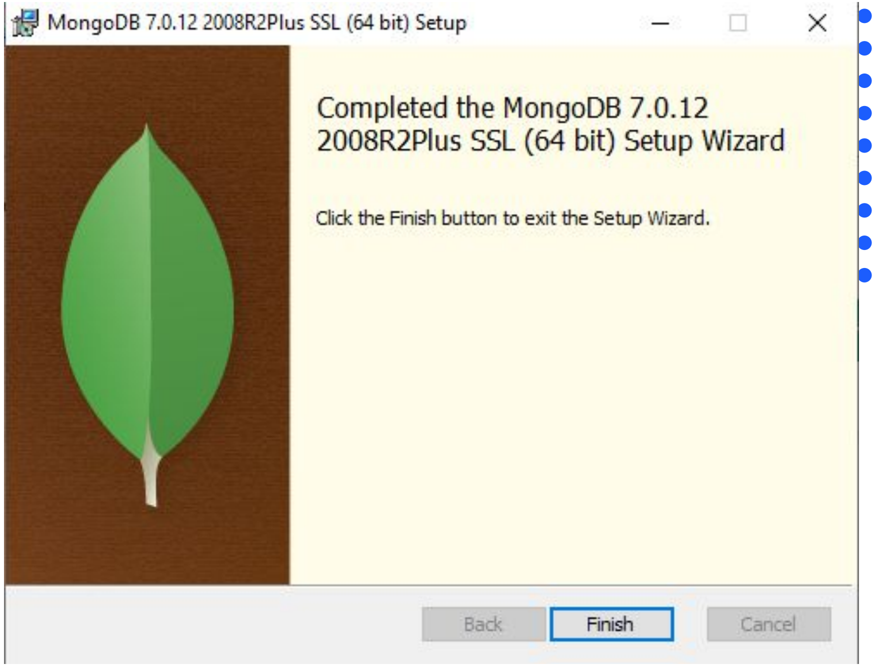
Status: Updating component registration

Install MongoDB



MongoDB Compass is being installed.

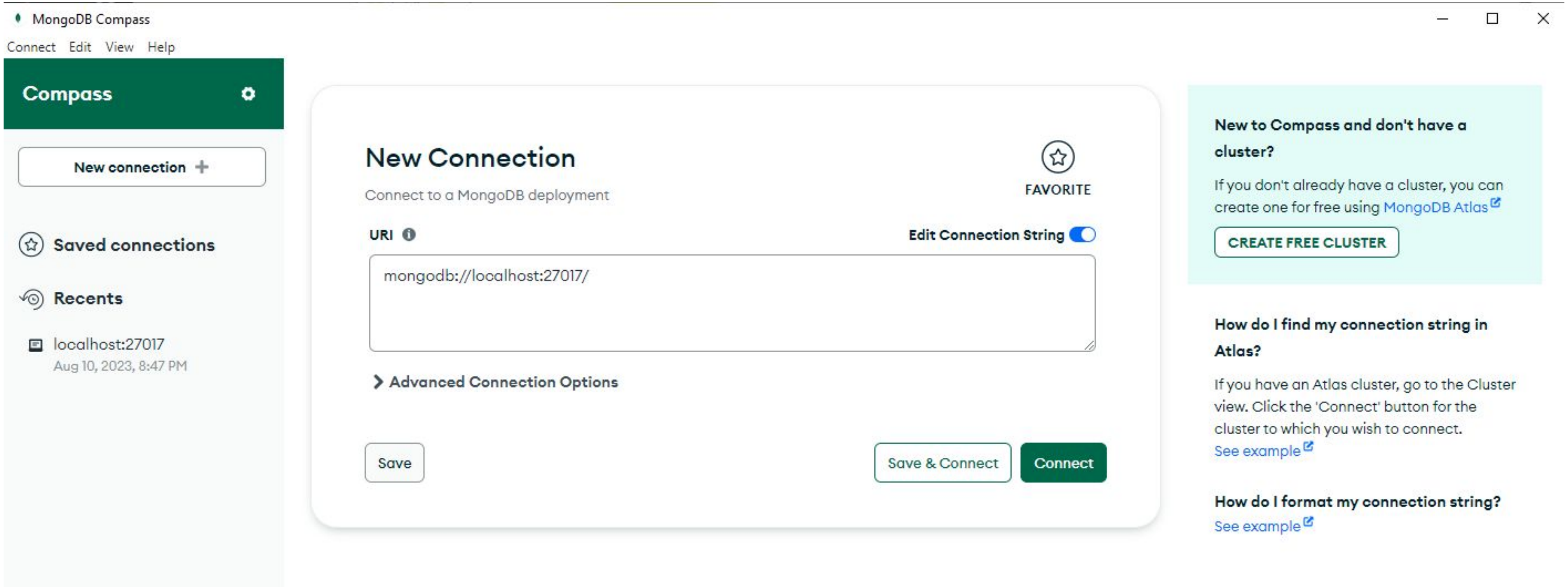
It will launch once it is done.



Completed the MongoDB 7.0.12 2008R2Plus SSL (64 bit) Setup Wizard

Click the Finish button to exit the Setup Wizard.

New Connection



localhost:27017
Aug 10, 2023, 8:47 PM

New Connection

Connect to a MongoDB deployment

URI

Edit Connection String

```
mongodb://localhost:27017/
```

New to Compass and don't have a cluster?

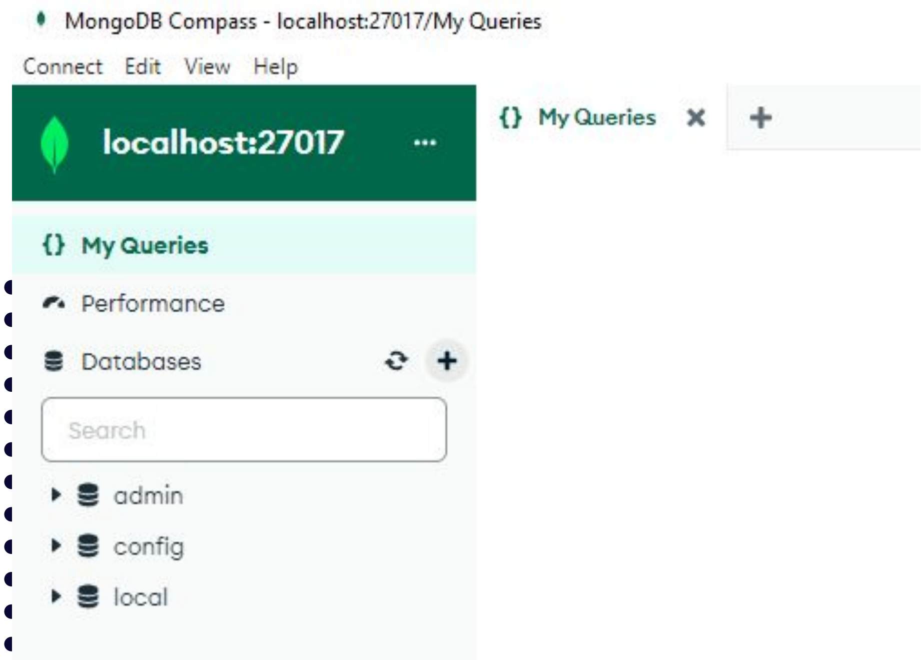
If you don't already have a cluster, you can create one for free using MongoDB Atlas

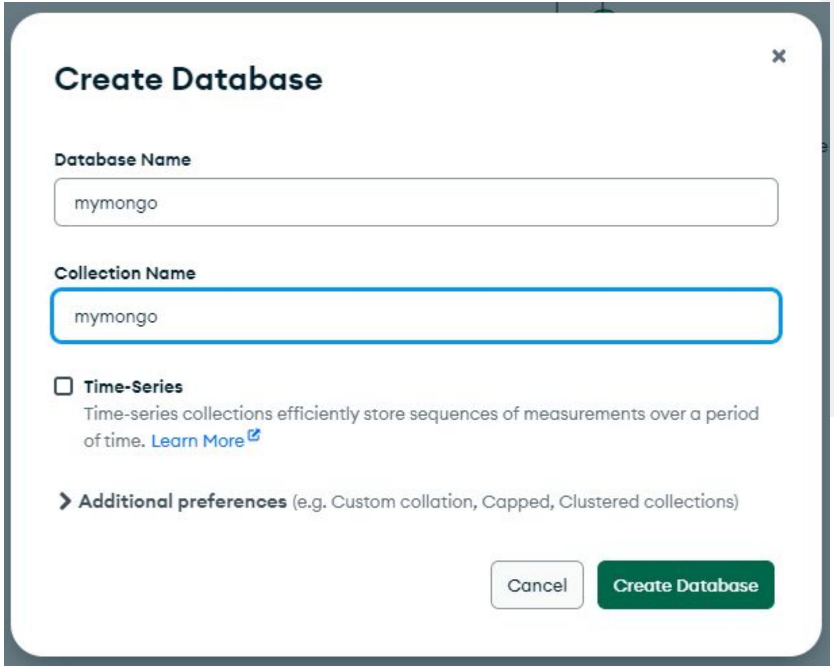
How do I find my connection string in Atlas?

If you have an Atlas cluster, go to the Cluster view. Click the 'Connect' button for the cluster to which you wish to connect. See example

How do I format my connection string? See example

Create Database





Create Database

Database Name: mymongo

Collection Name: mymongo

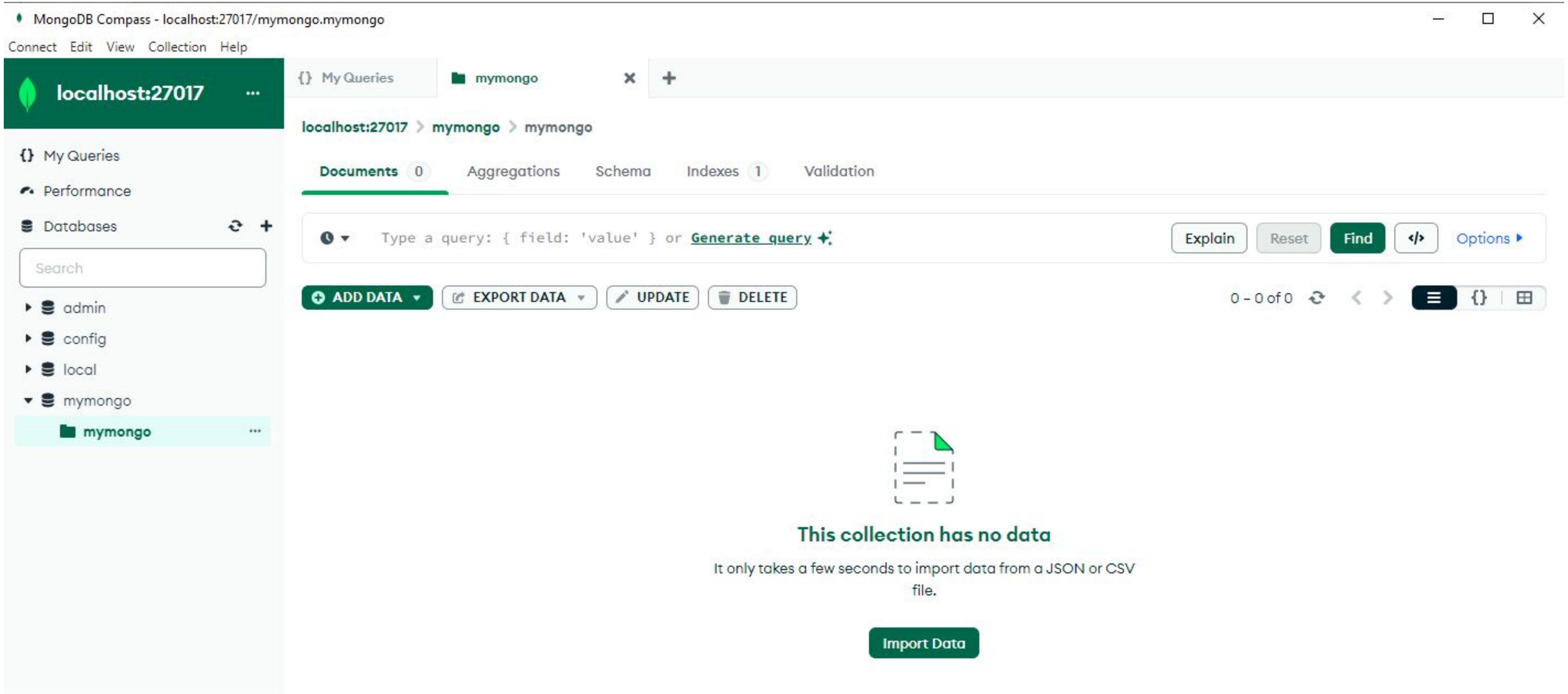
□ **Time-Series**

Time-series collections efficiently store sequences of measurements over a period of time. Learn More

> **Additional preferences** (e.g. Custom collation, Capped, Clustered collections)

Dr Sirintra Vaiwsri

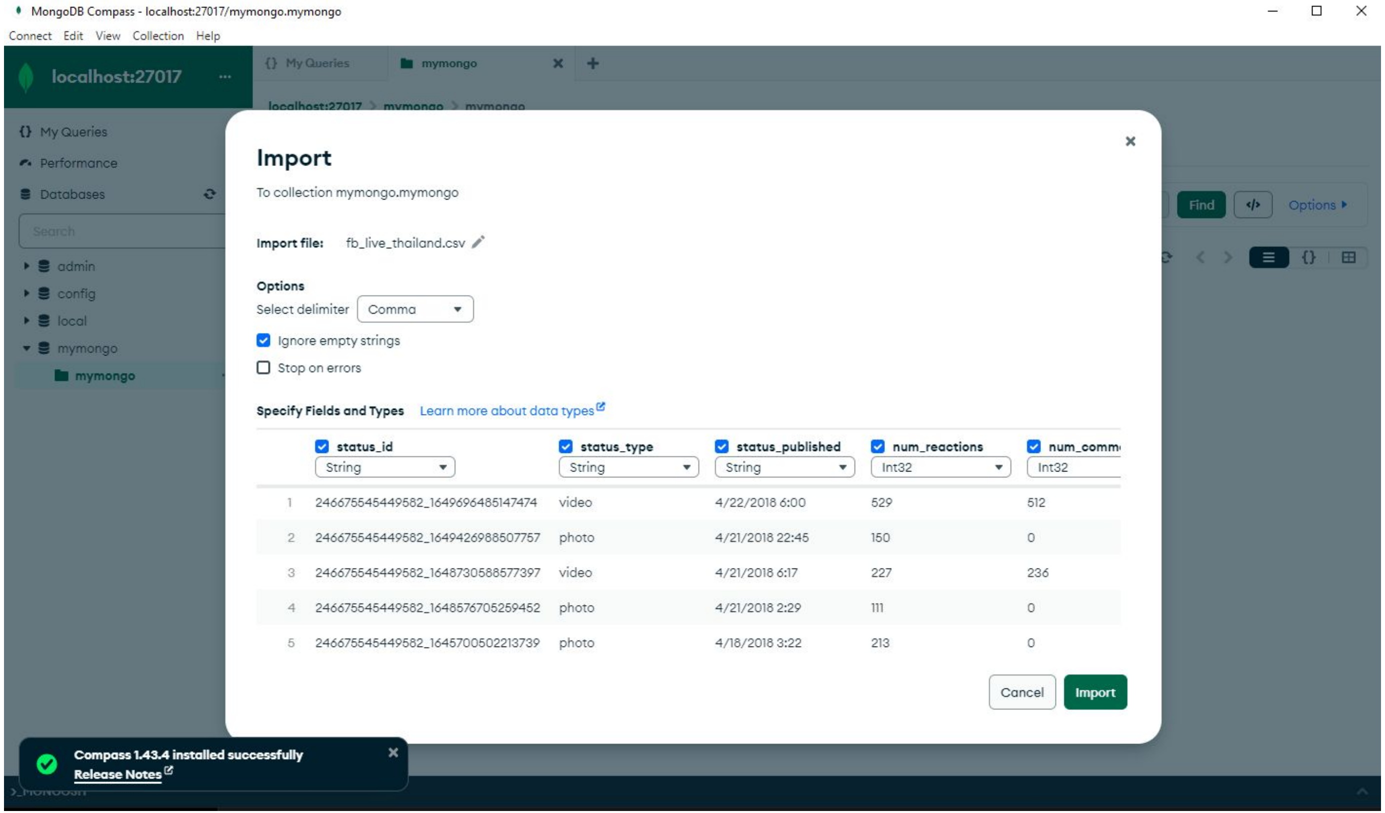
Create Database



This collection has no data

It only takes a few seconds to import data from a JSON or CSV file.

Import Data


MongoDB Compass - localhost:27017/mymongo.mymongo
Connect Edit View Collection Help
localhost:27017
My Queries
mymongo
localhost:27017 > mymongo > mymongo
My Queries
Performance
Databases
Search
admin
config
local
mymongo
mymongo
Find
Options
Compass 1.43.4 installed successfully
Release Notes

Import

To collection mymongo.mymongo

Import file: fb_live_thailand.csv

Options

Select delimiter: Comma

☑ Ignore empty strings

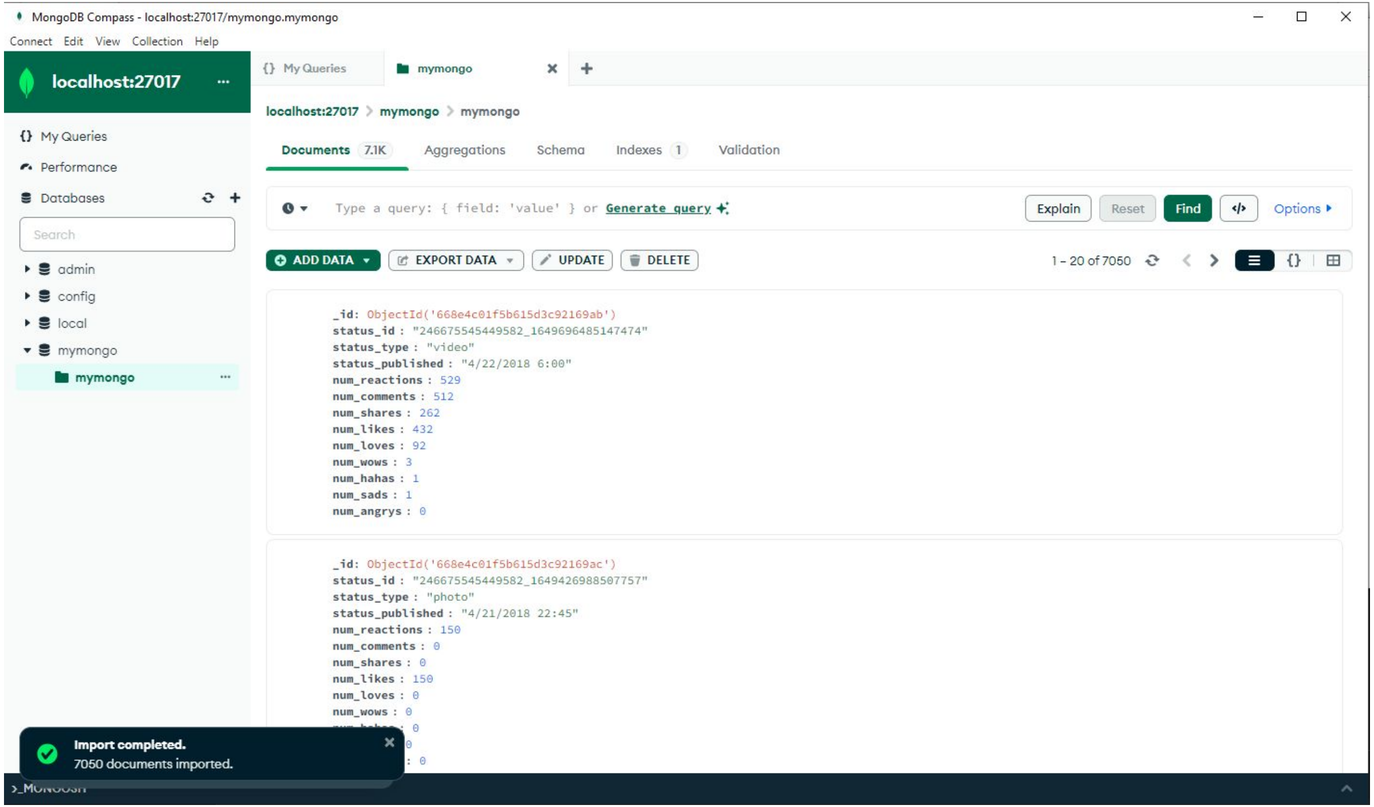
☐ Stop on errors

Specify Fields and Types Learn more about data types

	☑ status_id String	☑ status_type String	☑ status_published String	☑ num_reactions Int32	☑ num_comm Int32
1	246675545449582_1649696485147474	video	4/22/2018 6:00	529	512
2	246675545449582_1649426988507757	photo	4/21/2018 22:45	150	0
3	246675545449582_1648730588577397	video	4/21/2018 6:17	227	236
4	246675545449582_1648576705259452	photo	4/21/2018 2:29	111	0
5	246675545449582_1645700502213739	photo	4/18/2018 3:22	213	0

Cancel | Import

Import Data



```
_id: ObjectId('668e4c01f5b615d3c92169ab')
status_id : "246675545449582_1649696485147474"
status_type : "video"
status_published : "4/22/2018 6:00"
num_reactions : 529
num_comments : 512
num_shares : 262
num_likes : 432
num_loves : 92
num_wows : 3
num_hahas : 1
num_sads : 1
num_angrys : 0
```

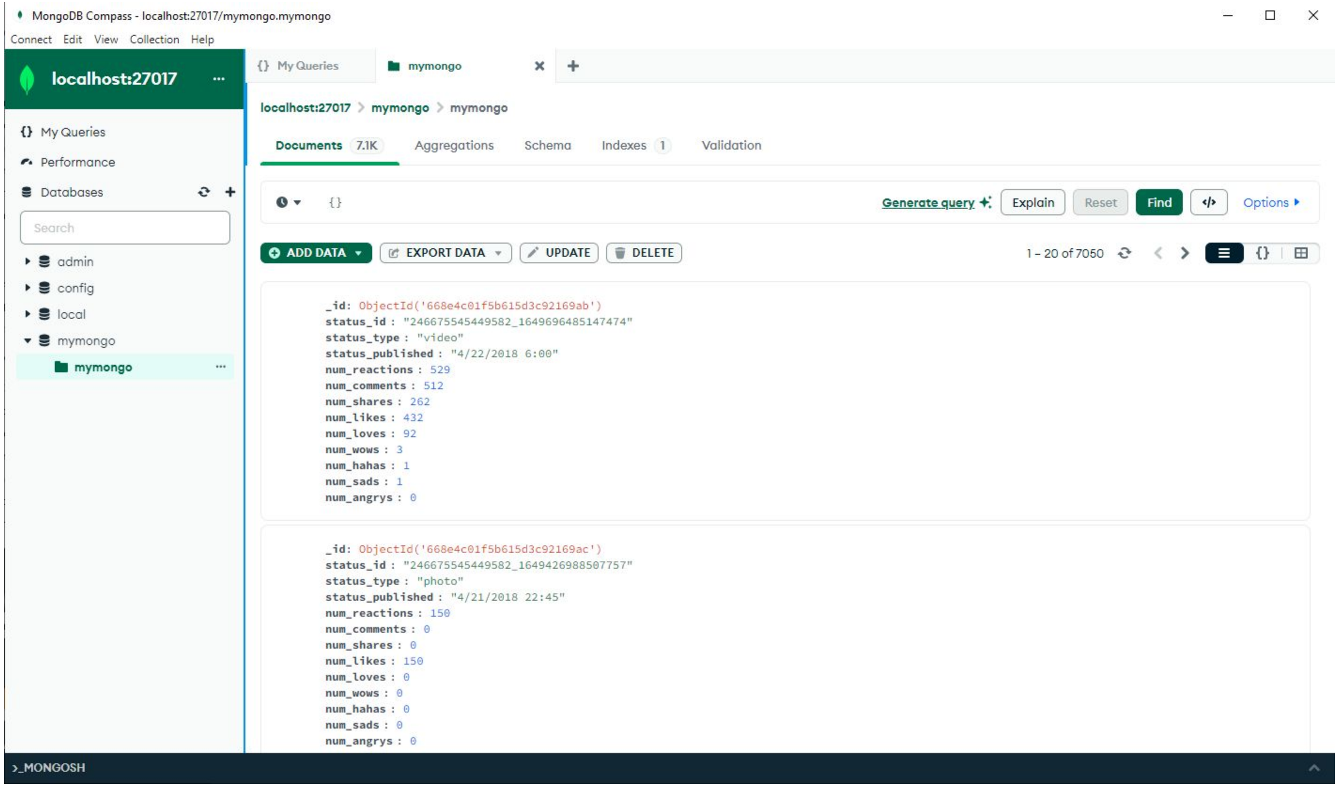
```
_id: ObjectId('668e4c01f5b615d3c92169ac')
status_id : "246675545449582_1649426988507757"
status_type : "photo"
status_published : "4/21/2018 22:45"
num_reactions : 150
num_comments : 0
num_shares : 0
num_likes : 150
num_loves : 0
num_wows : 0
num_hahas : 0
0
: 0
```

Import completed.
7050 documents imported.

Select All

SQL: select * from <table>

MQL: { }

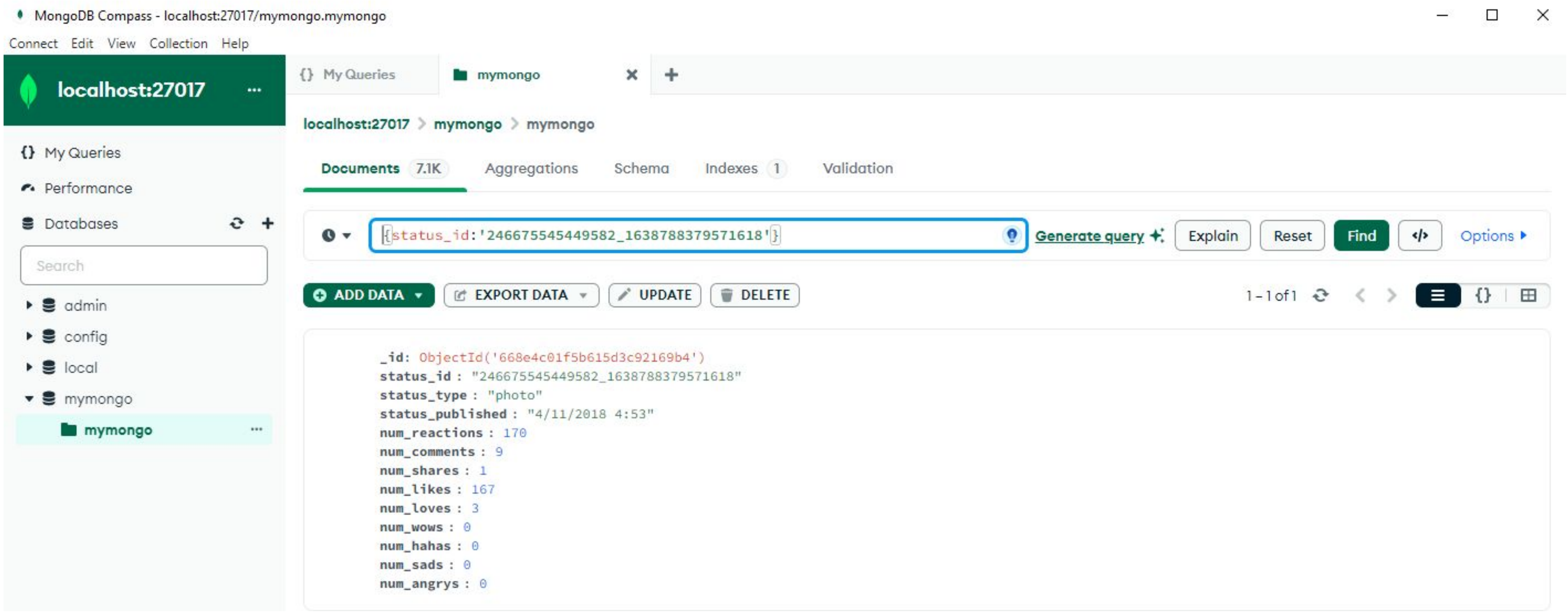


```
_id: ObjectId('668e4c01f5b615d3c92169ab')
status_id : "246675545449582_1649696485147474"
status_type : "video"
status_published : "4/22/2018 6:00"
num_reactions : 529
num_comments : 512
num_shares : 262
num_likes : 432
num_loves : 92
num_wows : 3
num_hahas : 1
num_sads : 1
num_angrys : 0
```

```
_id: ObjectId('668e4c01f5b615d3c92169ac')
status_id : "246675545449582_1649426988507757"
status_type : "photo"
status_published : "4/21/2018 22:45"
num_reactions : 150
num_comments : 0
num_shares : 0
num_likes : 150
num_loves : 0
num_wows : 0
num_hahas : 0
num_sads : 0
num_angrys : 0
```

Select ….. Where

SQL: select * from <table> where <column> = <value>   MQL: {<column>:<value>}



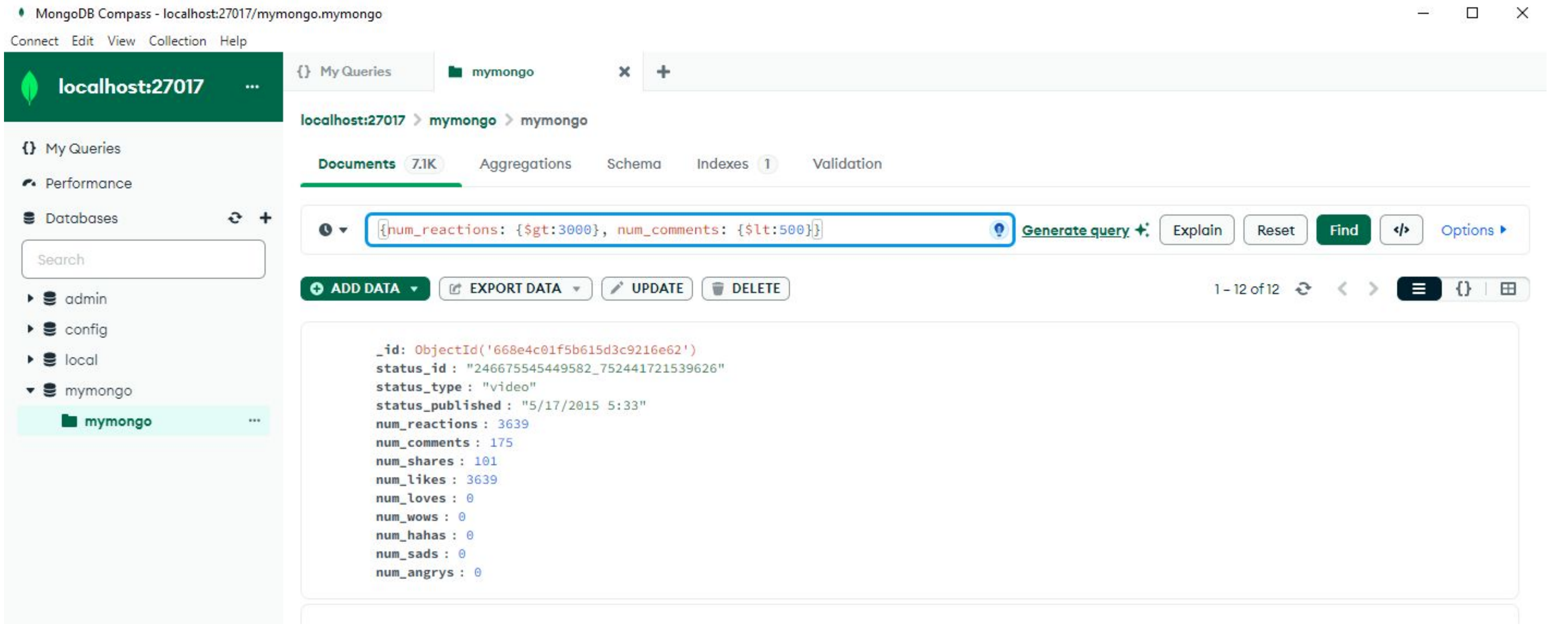
```
{status_id:'246675545449582_1638788379571618'}
```

```
_id: ObjectId('668e4c01f5b615d3c92169b4')
status_id : "246675545449582_1638788379571618"
status_type : "photo"
status_published : "4/11/2018 4:53"
num_reactions : 170
num_comments : 9
num_shares : 1
num_likes : 167
num_loves : 3
num_wows : 0
num_hahas : 0
num_sads : 0
num_angrys : 0
```

Select ….. Where ….. And

SQL: select * from <table> where <column1> = <value1> and <column2> = <value2>

MQL: {<column1>:<value1>, <column2>:<value2>}



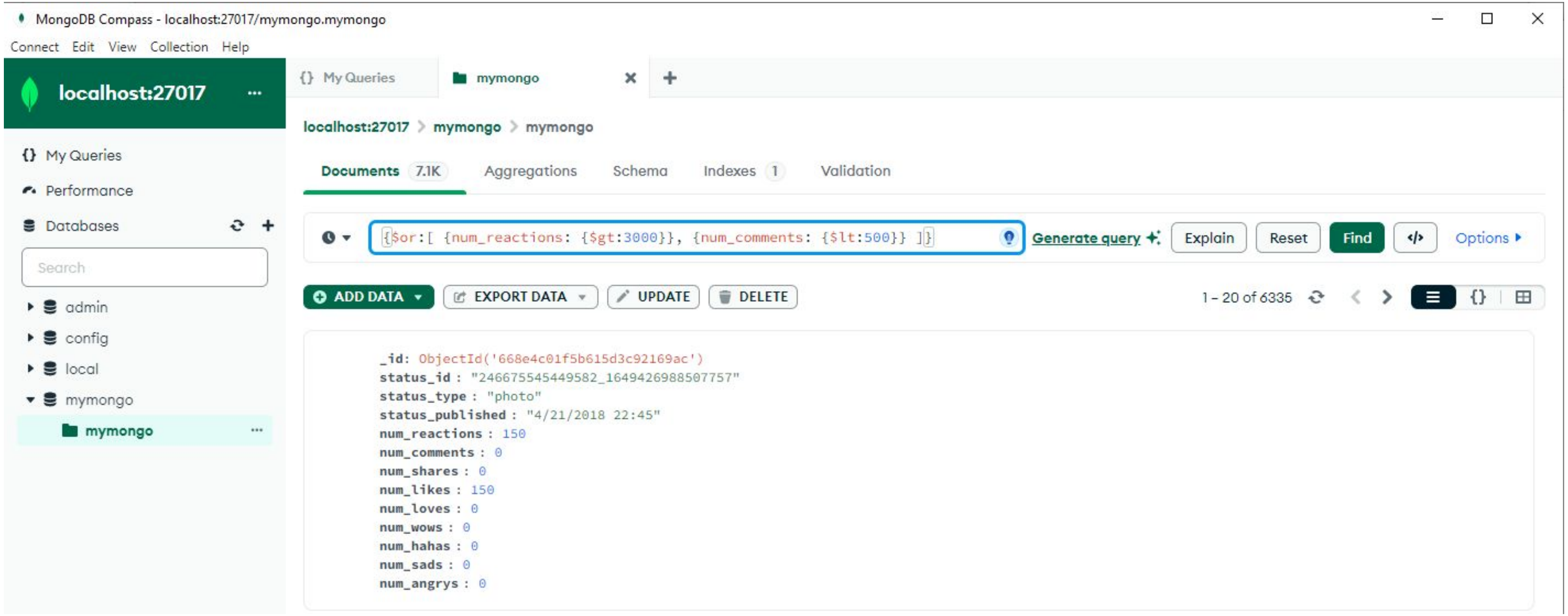
```
{num_reactions: {$gt:3000}, num_comments: {$lt:500}}
```

```
_id: ObjectId('668e4c01f5b615d3c9216e62')
status_id : "246675545449582_752441721539626"
status_type : "video"
status_published : "5/17/2015 5:33"
num_reactions : 3639
num_comments : 175
num_shares : 101
num_likes : 3639
num_loves : 0
num_wows : 0
num_hahas : 0
num_sads : 0
num_angrys : 0
```

Select ….. Where ….. Or

SQL: select * from <table> where <column1> = <value1> or <column2> = <value2>

MQL: {$or: [{<column1>:<value1>}, {<column2>:<value2>}] }



```
{$or:[ {num_reactions: {$gt:3000}}, {num_comments: {$lt:500}} ]}
```

```
_id: ObjectId('668e4c01f5b615d3c92169ac')
status_id : "246675545449582_1649426988507757"
status_type : "photo"
status_published : "4/21/2018 22:45"
num_reactions : 150
num_comments : 0
num_shares : 0
num_likes : 150
num_loves : 0
num_wows : 0
num_hahas : 0
num_sads : 0
num_angrys : 0
```

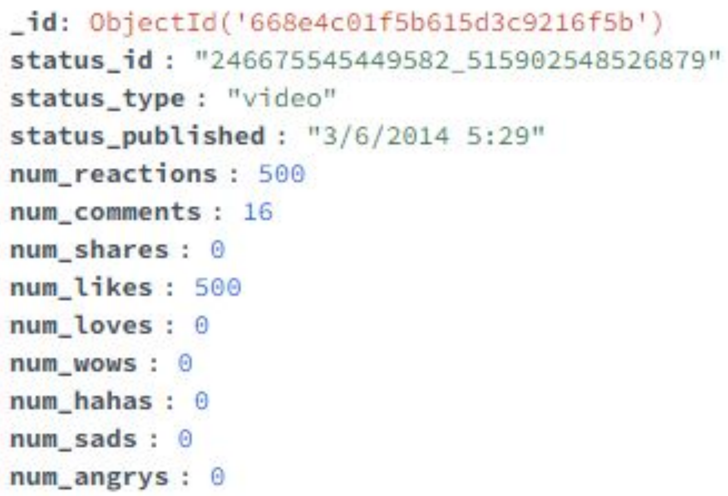
Select ..... Where ..... And ..... Or

SQL: select * from <table> where (<column1> = <value1>) and (<column2> = <value2> or <column3> = <value3>)

MQL: {{<column1>:<value1>}, $or: [{<column2>:<value2>},{<column3>:<value3>}] }

Assignment (1 point)

- Please select all records that have a number of likes equal to 500 and also have a number of reactions greater than 3,000 or a number of comments greater than 10



```
_id: ObjectId('668e4c01f5b615d3c9216f5b')
status_id : "246675545449582_515902548526879"
status_type : "video"
status_published : "3/6/2014 5:29"
num_reactions : 500
num_comments : 16
num_shares : 0
num_likes : 500
num_loves : 0
num_wows : 0
num_hahas : 0
num_sads : 0
num_angrys : 0
```

References

- Holmes, D. E. (2017). Big data: a very short introduction. Oxford University Press.
- Staragile. A Brief comparison between ACID vs BASE database model. https://staragile.com. Accessed: 2023-08-05.
- Edward, S. G., & Sabharwal, N. (2015). *Practical MongoDB: Architecting, Developing, and Administering MongoDB*. Apress.
- Bahga, A., & Madisetti, V. (2019). Big Data analytics: A hands-on approach.
- DEV. Intro to 4 types of NoSQL databases. https://dev.to. Accessed: 2023-08-14.
- Janev, V., Graux, D., Jabeen, H., & Sallinger, E. (2020). Knowledge graphs and big data processing (p. 209). Springer Nature.
- KDNuggets. https://www.kdnuggets.com/. Accessed: 2023-08-14.