MongoDB and Python Integration

Applied Database Technologies DS-532
Ameya Dalvi
Prof. Olga Scrivner

MongoClient

1. The MongoClient() class for the PyMongo driver library for MongoDB creates client instances. Its main function is to enable you to connect to the MongoDB database efficiently and effortlessly. Whether you're changing data or retrieving it, every connection matters.

1. **from pymongo import MongoClient**

Workshop Setup

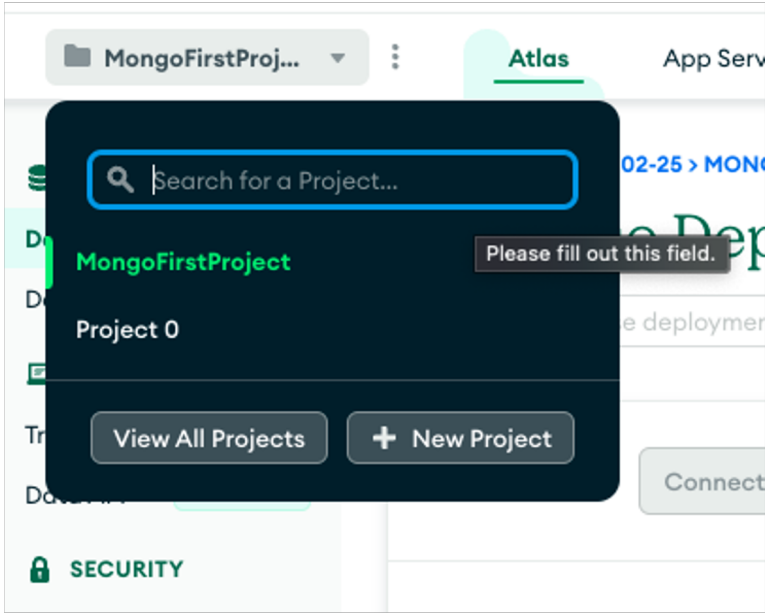
1. **Create an Atlas account:**

   https://www.mongodb.com/cloud/atlas

   Note - you can use the free tier for this workshop.

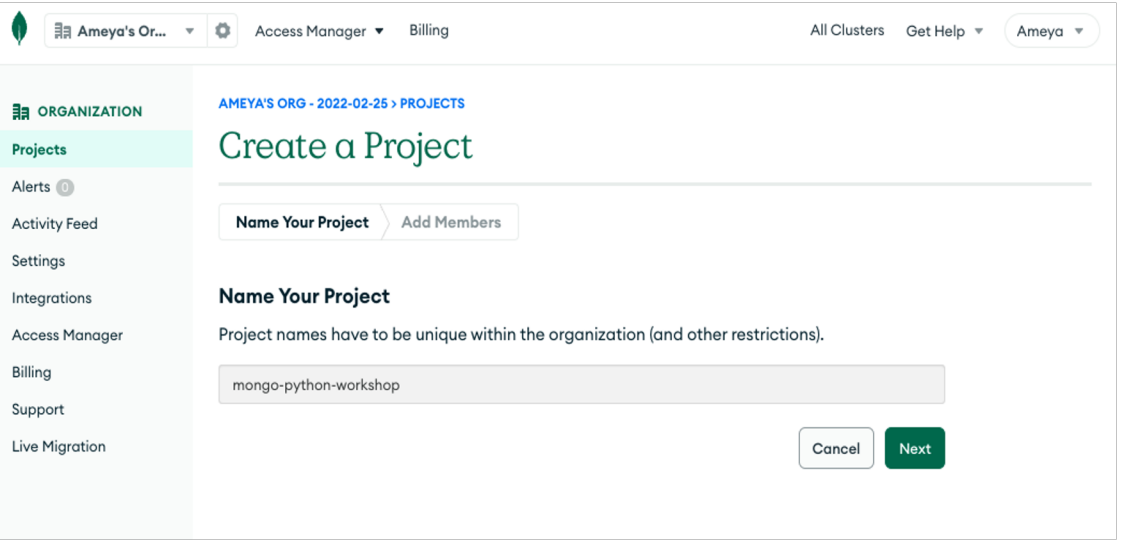
1. **Create a project:**

   Create a project, give it a name (no specific restrictions here, but choose something descriptive like "mongo-python-workshop")

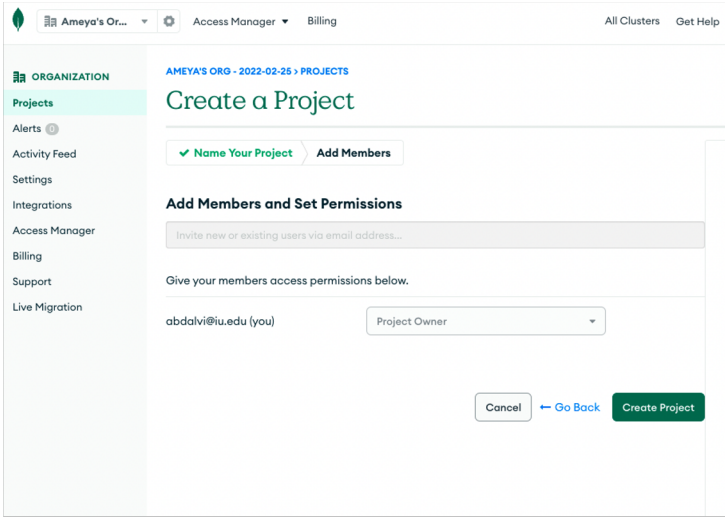


Follow along these steps:

Step1: Name your project



Step2: Add members, Set permissions

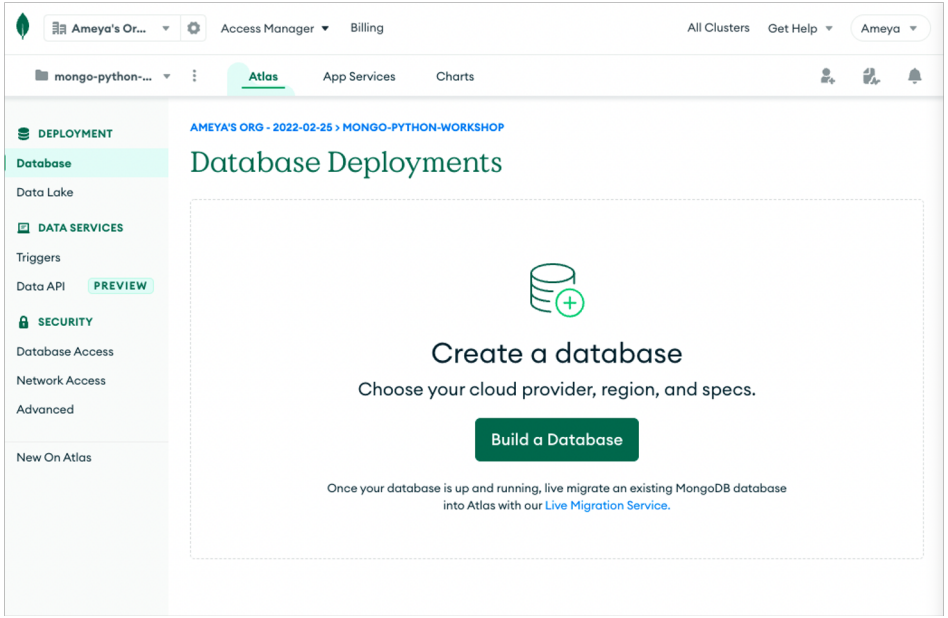


3. Set up a cluster

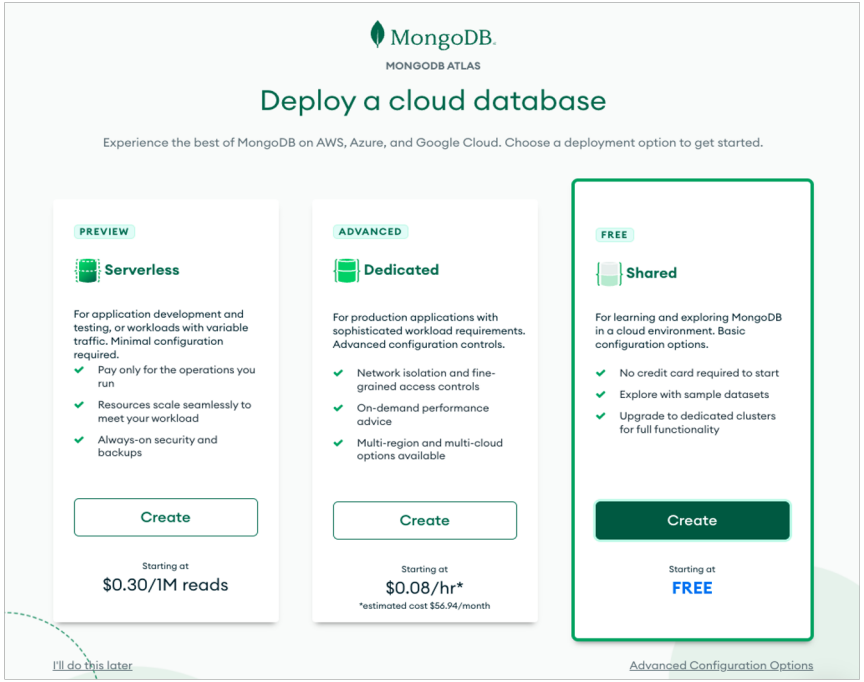
After creating a project, set up a **Cluster**

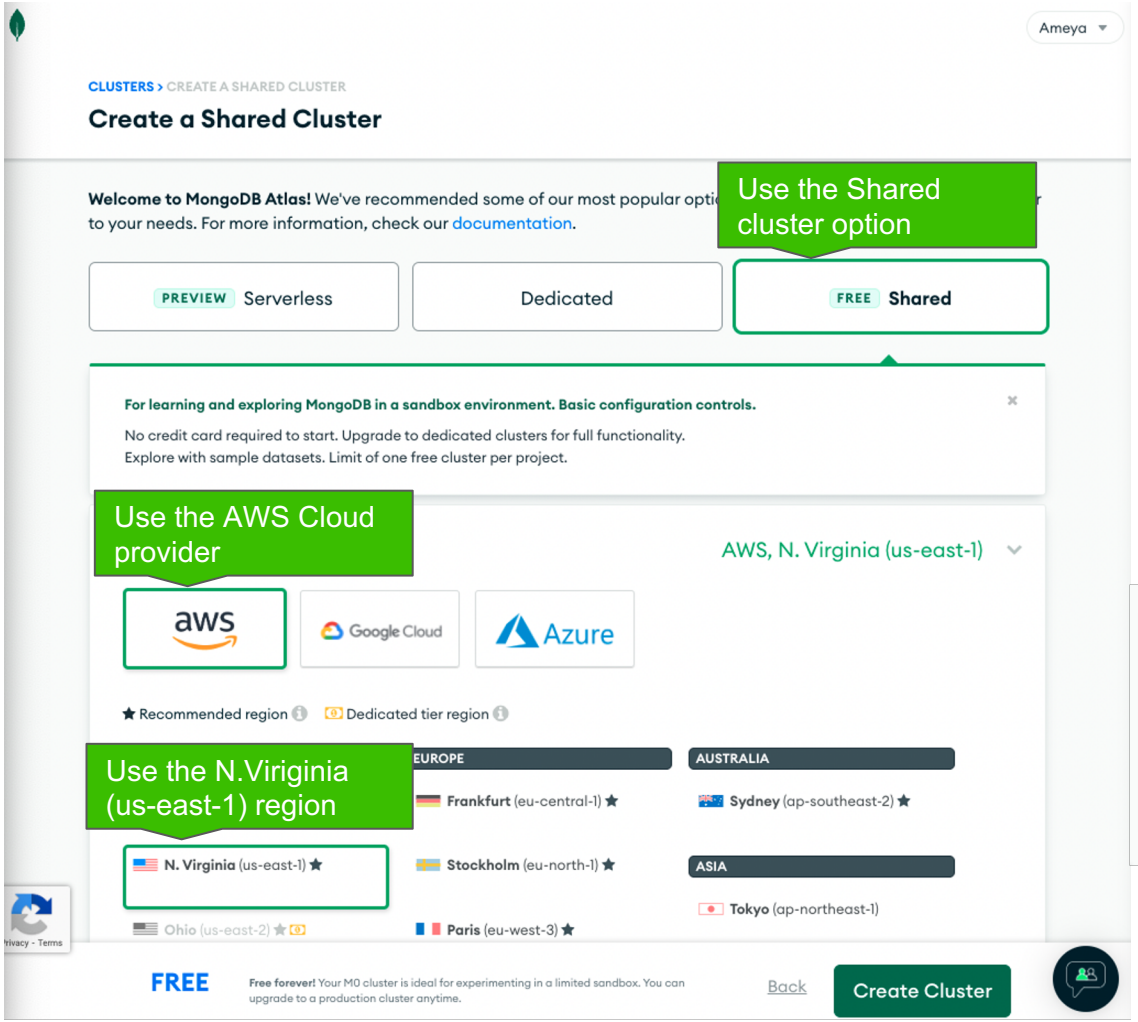
Make sure you name it! You may need to scroll to the bottom of the page. This is a one-time setup: once your cluster is created, you won't be able to change its name.

Click on Build a Database

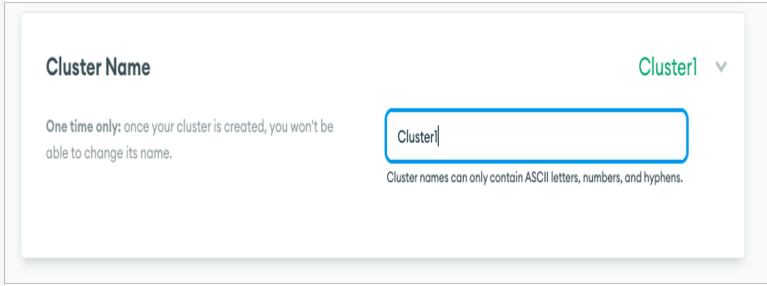


Click on the Create button for the Shared db





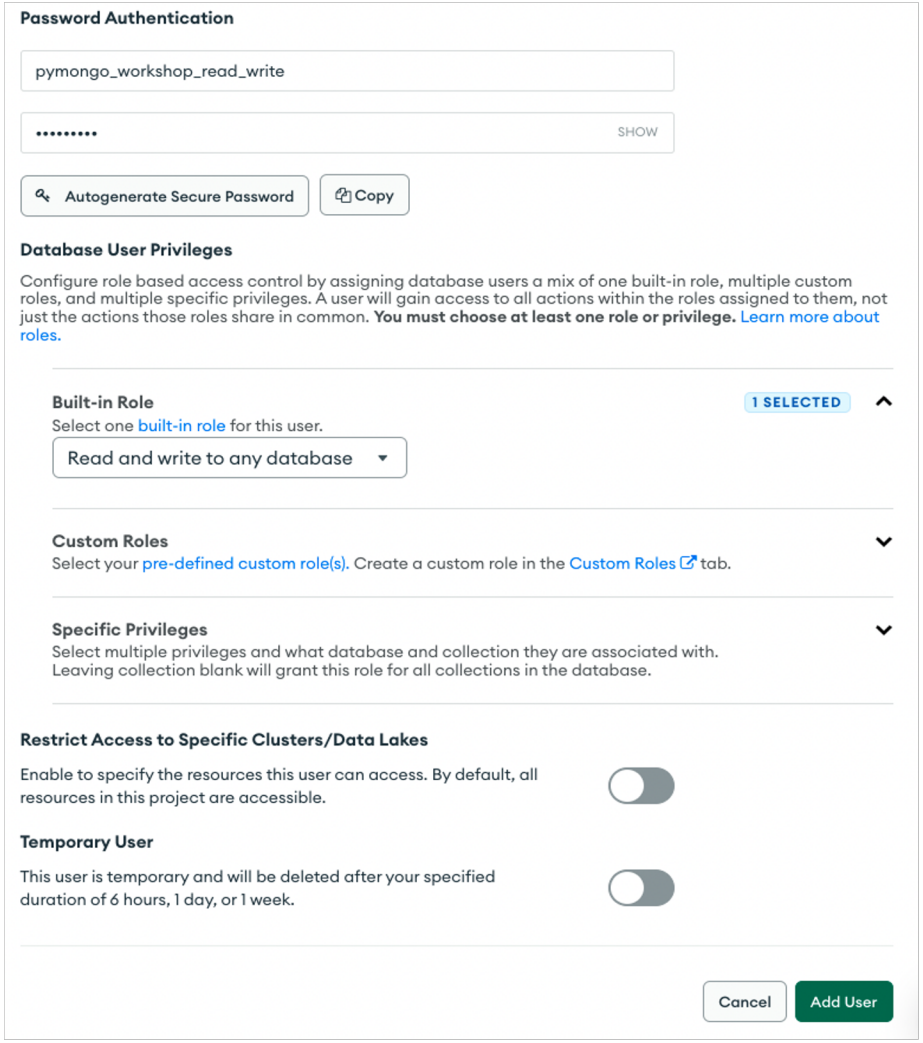
Create a cluster name:



4. Create use and provide r/w access

Create a new database user name:
pymongo_workshop_read_write

You can autogenerate a password or choose one (if you autogenerate it, save it somewhere) give this user write privileges, but not full admin

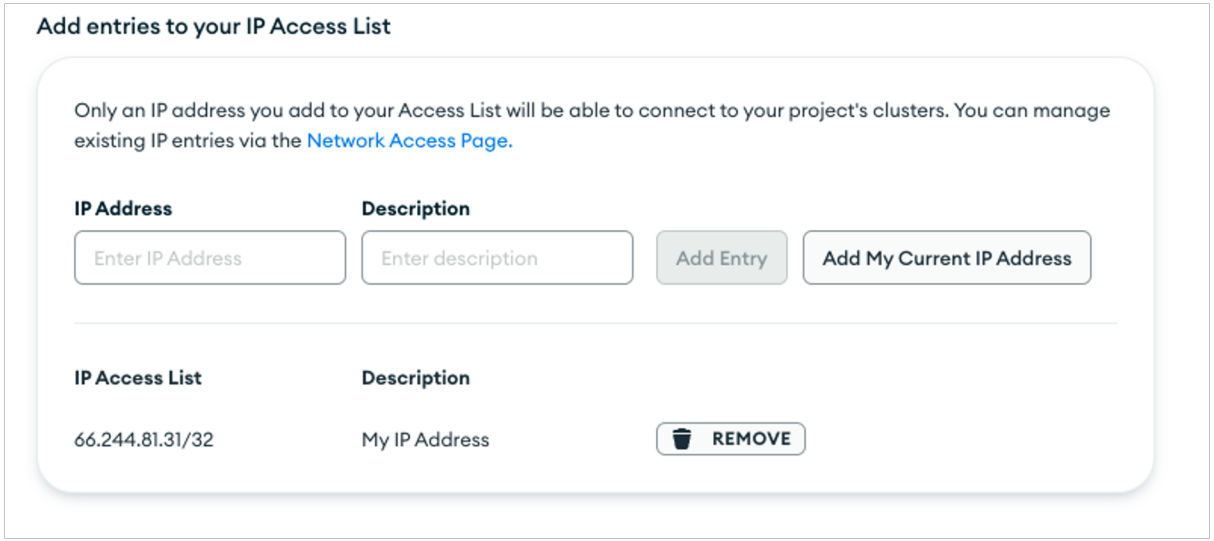


5. Provide network access

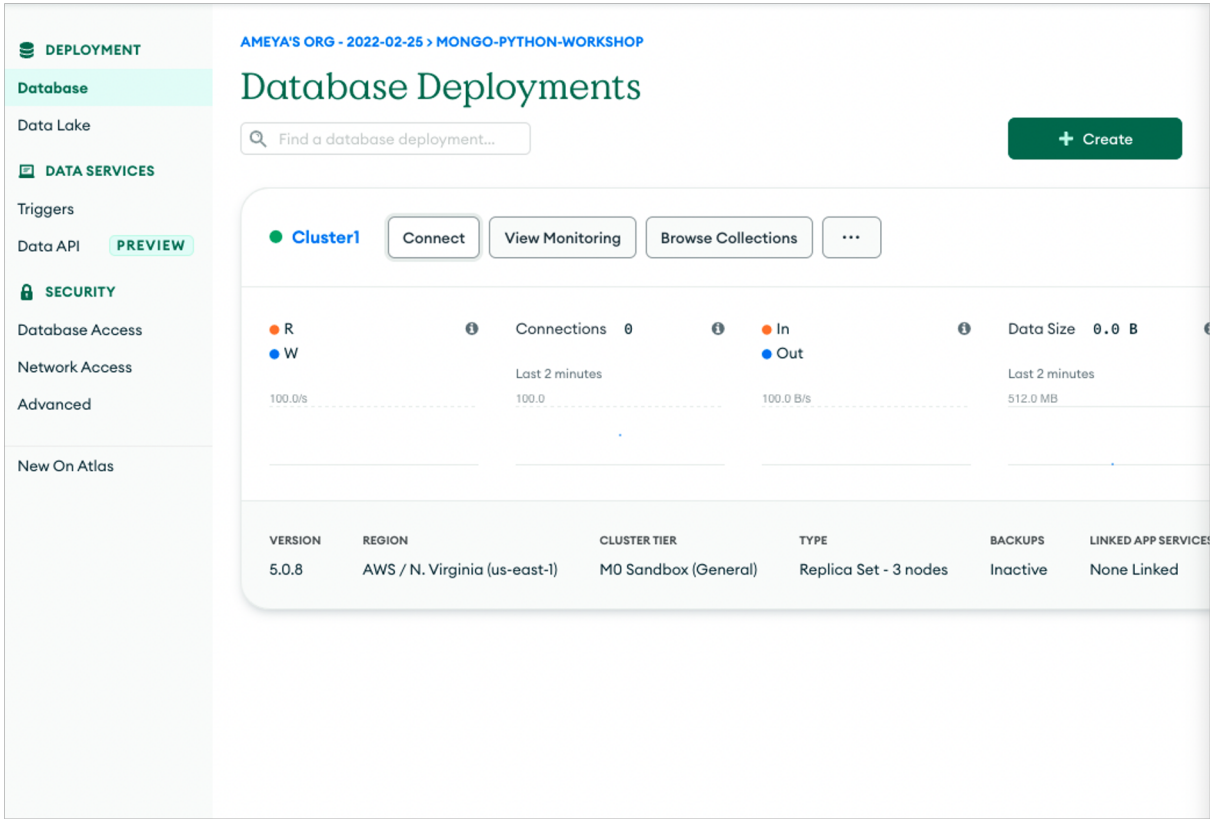
You can either make this world-readable or limit to specific IP address.

To limit to your current connection, use your IP address (note - if you don't have a persistent or permanent IP address, you won't have access next time you log in. If this is the case, you'll need to update your network access next time you try to read from this account outside your Atlas login).

To make it world-readable (no IP restrictions) 0.0.0.0/0



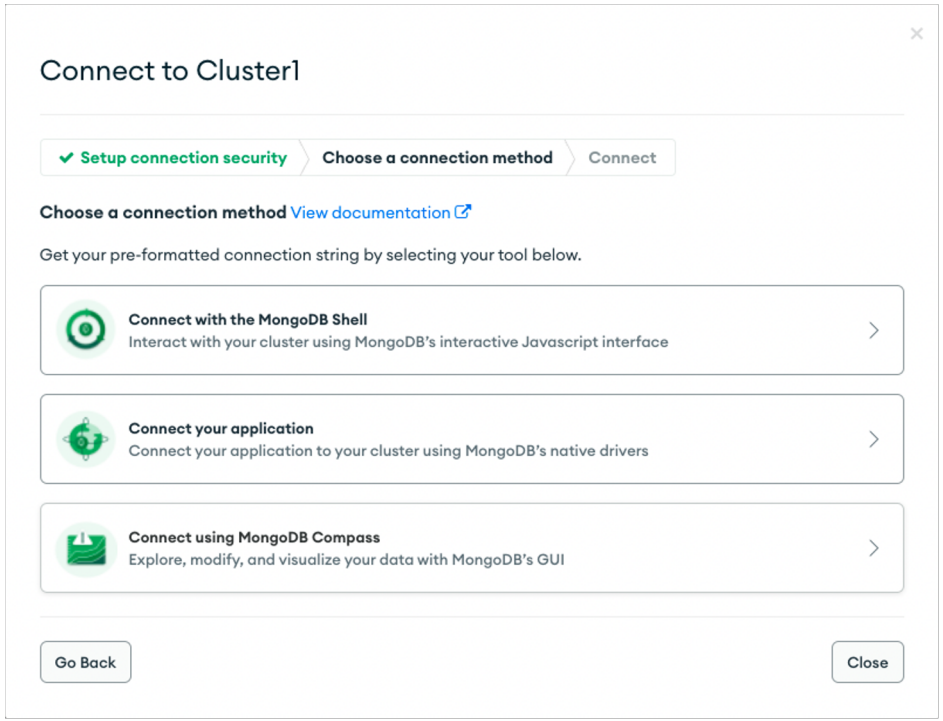
After configuring your cluster details, you will be redirected to this page :

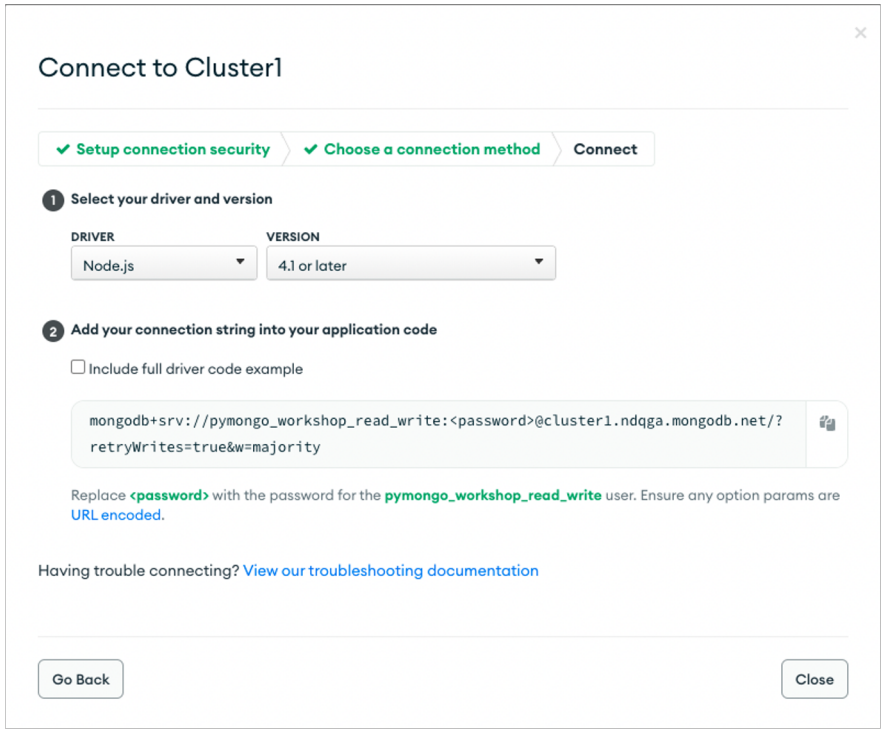


Click on Connect, to get the MongoDB connection string:In your cluster

Select "connect your application"
choose "python, 3.6 or later" and copy it
your connection string should look like this (the URI should be slightly different): **Store the connection string**

mongodb+srv://pymongo_workshop_read_write:<password>@cluster1.ndqga.mongodb.net/?retryWrites=true&w=majority

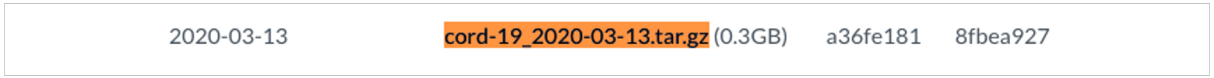




6. Load COVID-19 data into MongoDB on Atlas

We'll use the coronavirus research dataset from the Semantic Scholar team at the Allen Institute.
Go to:
https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/historical_releases.html

For the workshop, we'll only upload a subset of this data.
We are looking for something small, so we will download the original release, cord-19_2020-03-13.tar.gz (0.3 GB)

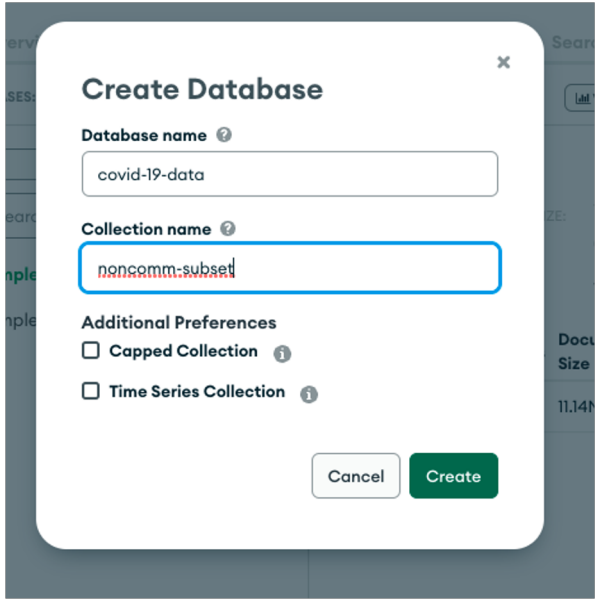


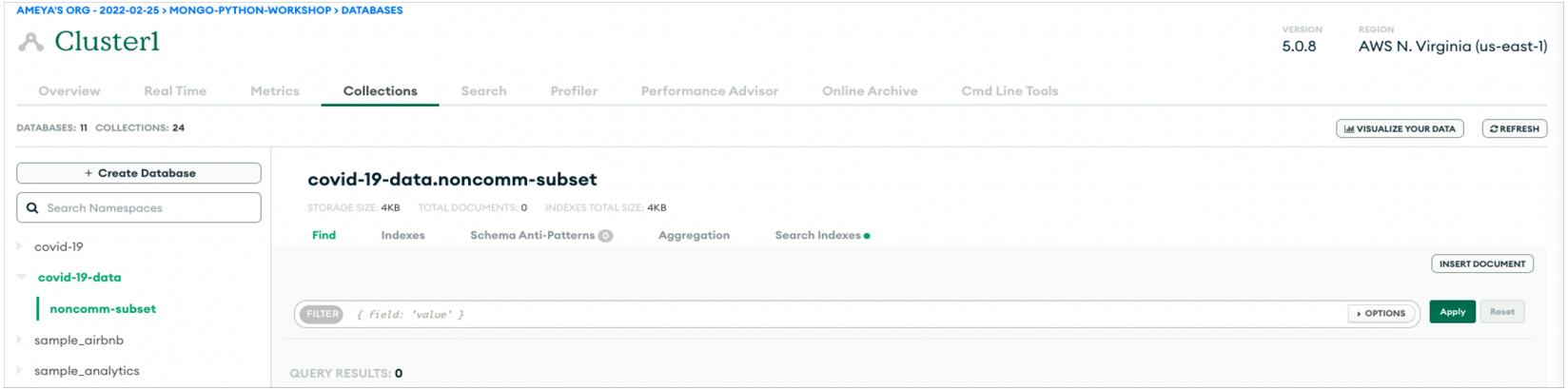
Untar it, you should get a folder named:
2020-13-13, within this folder, expand the **noncomm-use-subset**

7. Add the Data to a Collection

Back in your Atlas cluster:
Click on **Browse Collections**
Click on **Collections** tab
Click on the **+ Create Database** tab to create a database
db name: covid19-data collection name: noncomm-subset

This will basically be an empty database with 0 records currently. We will insert the downloaded CORD-19 data to it using the jupyter notebook





Switching to the notebook...