

Final Assignment: Data Engineering for a Consulting Firm



Estimated time needed: 45 mins

You are a data engineer at a data analytics consulting company. Your company prides itself in being able to efficiently handle data in any format on any database on any platform. Analysts in your office need to work with data on different databases, and data in different formats. While these analysts are good at analyzing data, they count on you to be able to move data from external sources into various databases, to be able to move data from one type of database to another, and be able to run basic queries on various databases.

Objectives

In this assignment you will:

- Import data into a MongoDB database.
- Query data in a MongoDB database.
- Export data from MongoDB.
- Import data into a Cassandra database.
- Query data in a Cassandra database.

About Skills Network Cloud IDE

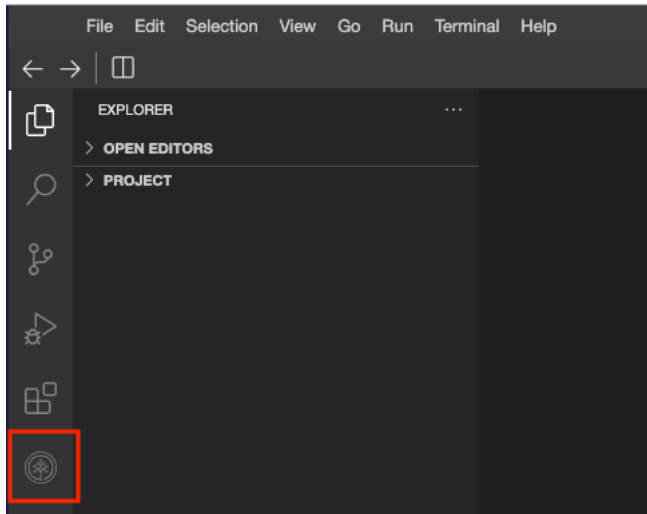
Skills Network Cloud IDE (based on Theia and Docker) provides an environment for hands on labs for course and project related labs. Theia is an open source IDE (Integrated Development Environment), that can be run on desktop or on the cloud. to complete this lab, we will be using the Cloud IDE based on Theia and MongoDB/Cassandra running in a Docker container.

Important Notice about this lab environment

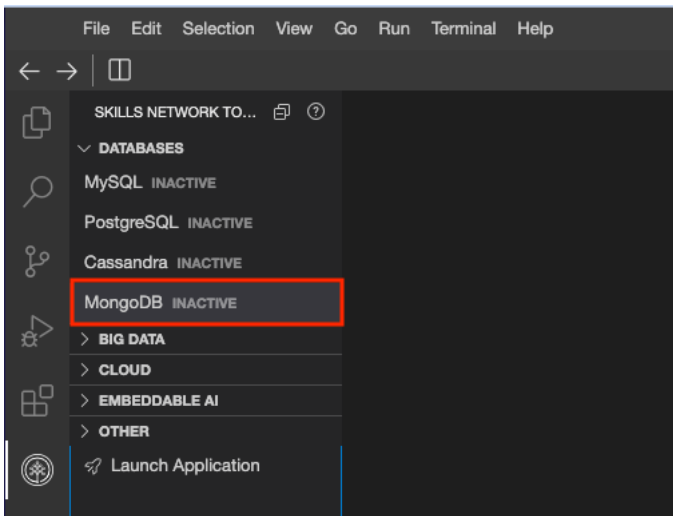
Please be aware that sessions for this lab environment are not persisted. Every time you connect to this lab, a new environment is created for you. Any data you may have saved in the earlier session would get lost. Plan to complete these labs in a single session, to avoid losing your data.

Set-up: Start MongoDB

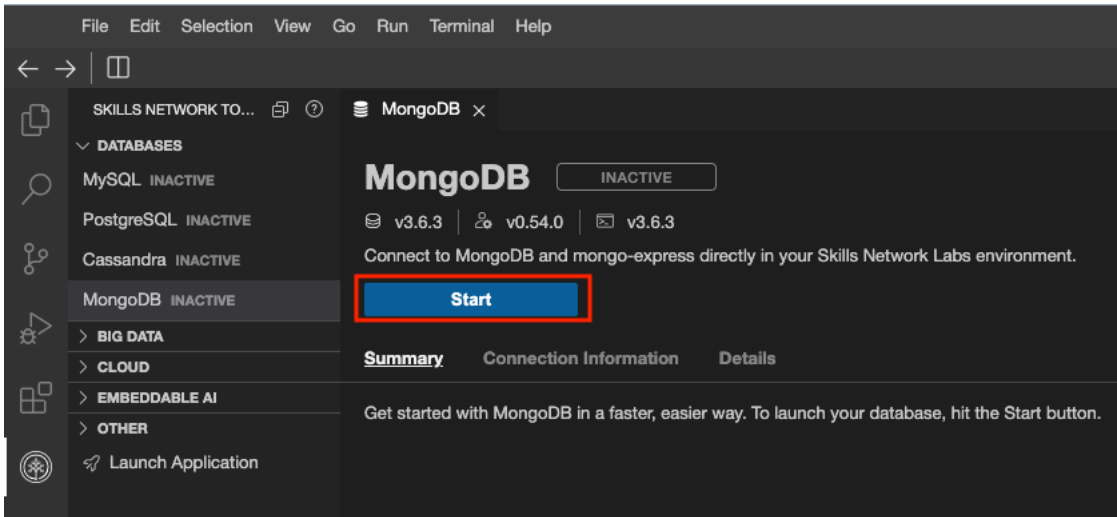
Navigate to Skills Network Toolbox.



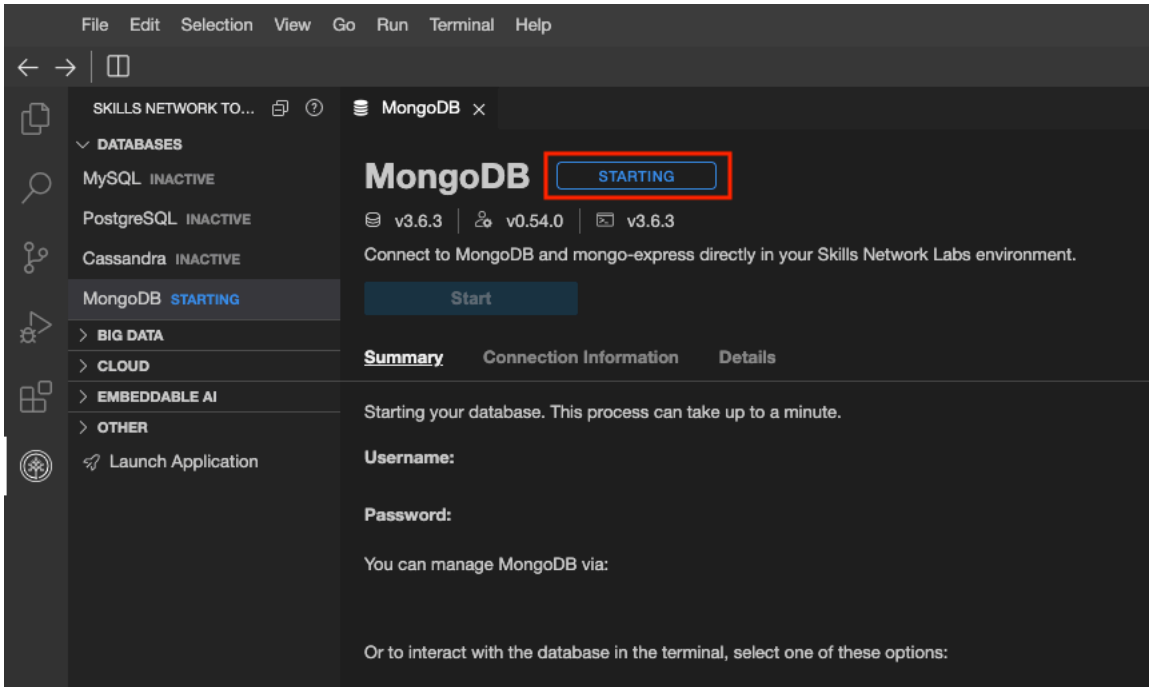
You will notice MongoDB listed there, but inactive, which means the database is not available to use.



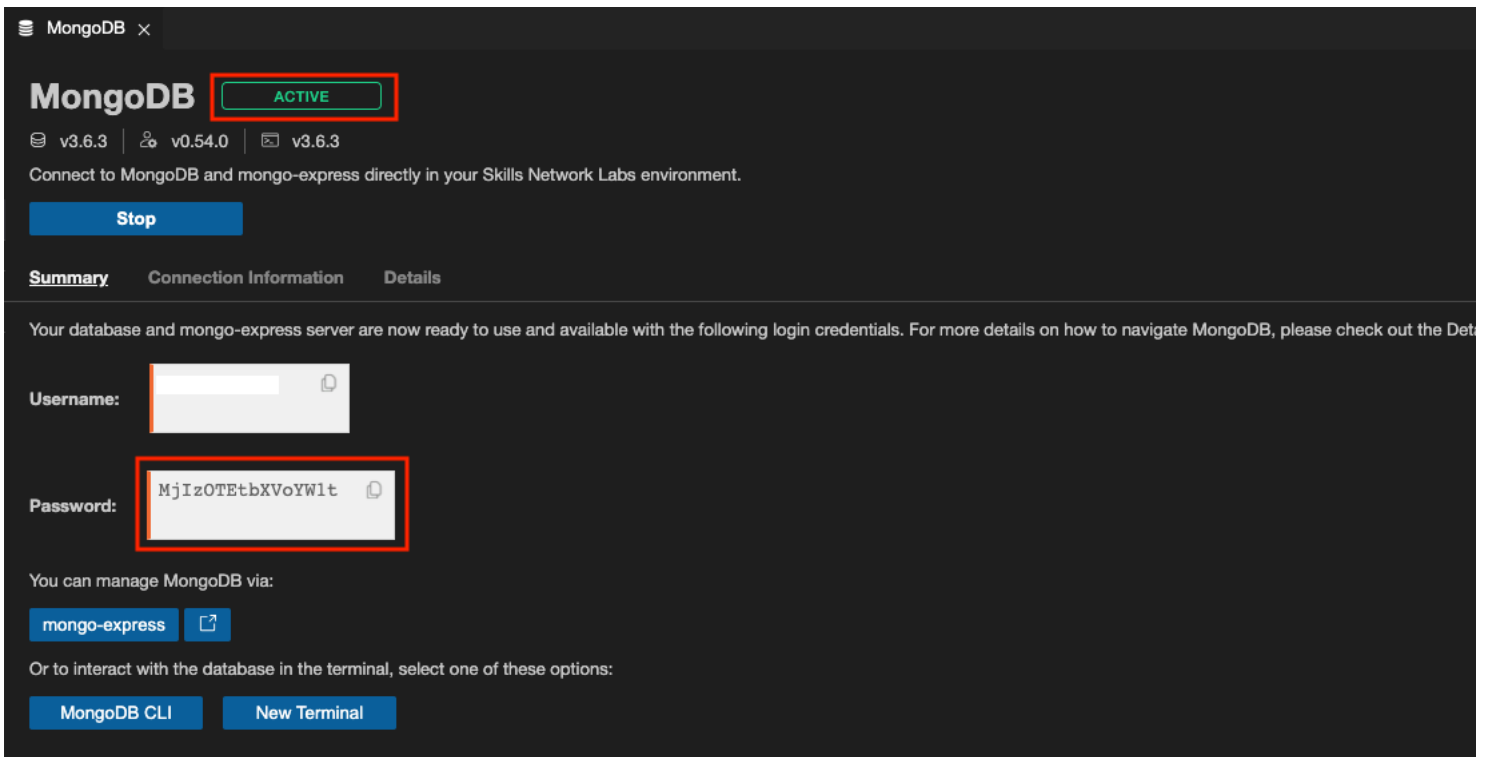
Once you click on the database, you will see more details and a button to start the database.



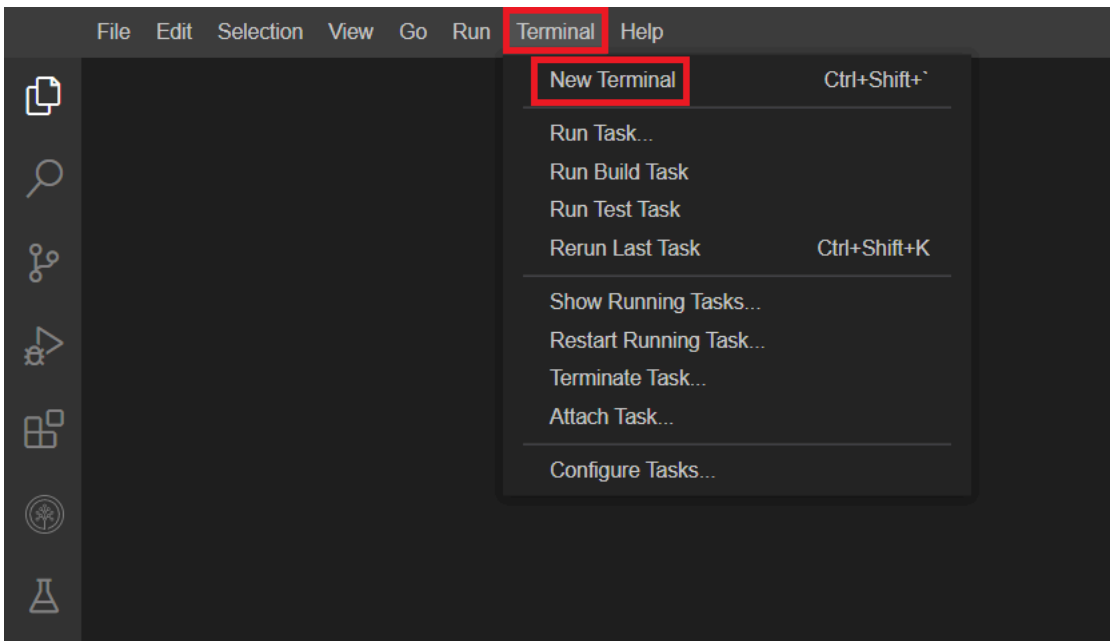
Clicking the start button runs a background process to configure and start your MongoDB server.



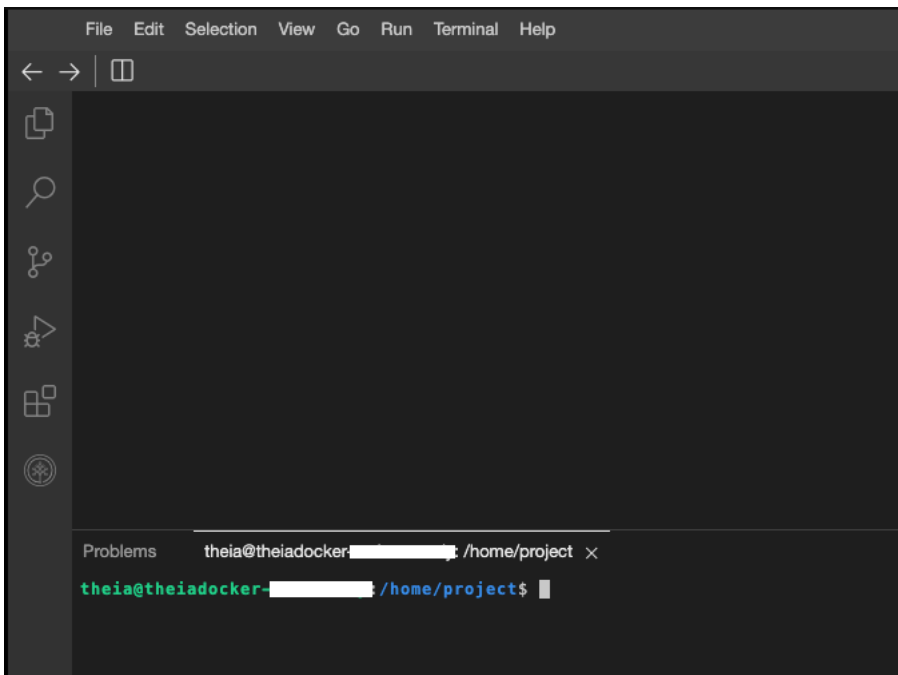
Soon, your server is ready for use. This deployment has access control enabled and MongoDB enforces authentication. So, take note of the password. You will need this password to login as root user.



You can now open a new terminal window.



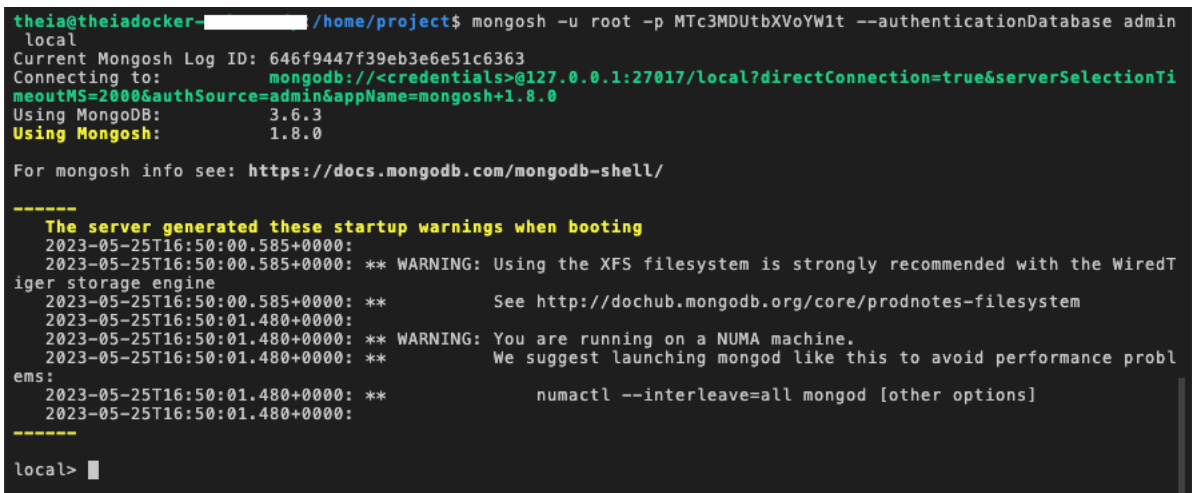
The new terminal window opens at the bottom of the screen as seen in the following image.



Run the following command in the newly opened terminal window. (You can copy the code by clicking on the copy button on the bottom right of the following codeblock and then paste the code where needed.)

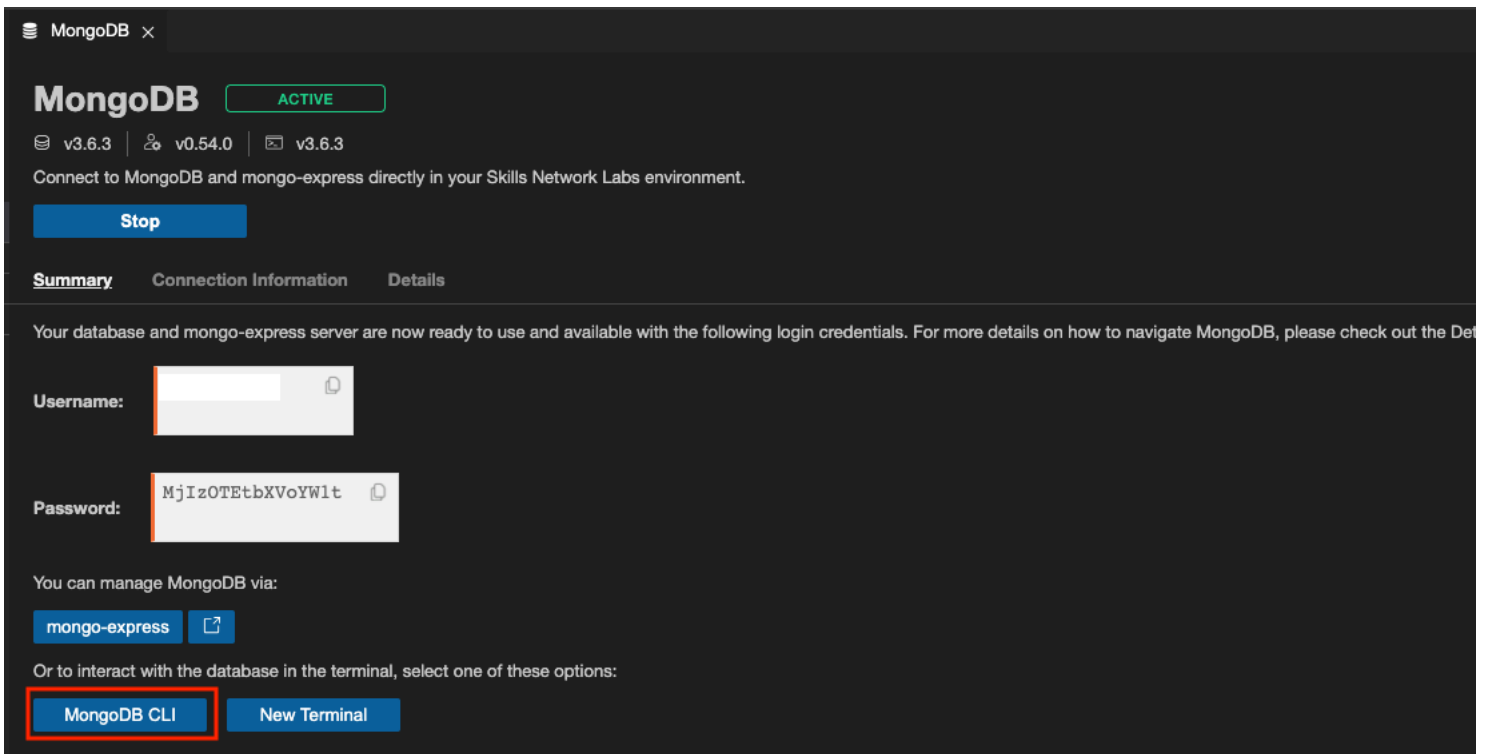
- 1
1. `mongosh -u root -p PASSWORD --authenticationDatabase admin local`

Copied! Executed!



The command contains the username and password to connect to mongodb server (the text after the -p option is the password). Your output would be different from the output shown in the prior image. Copy the command given to you, and keep this command handy. You will need this command in the next step.

Or, you can simply click on MongoDB CLI



In MongoDB CLI (the mongo shell), switch the context to the training database.

- 1
1. use training

Copied!

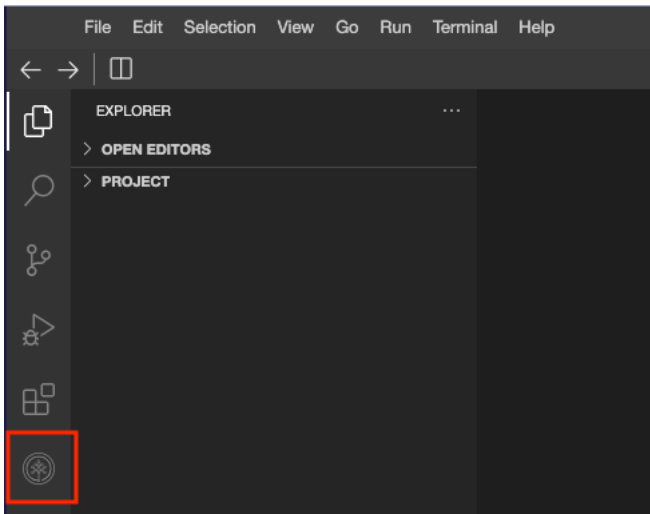
Create a collection called bigdata

- 1
1. db.createCollection("bigdata")

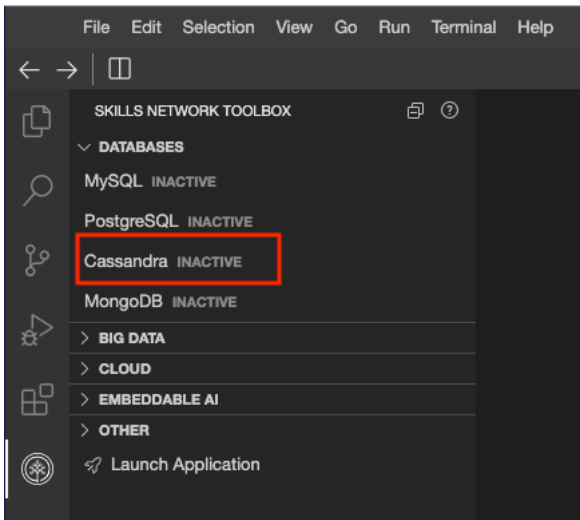
Copied!

Set-up: Start Cassandra

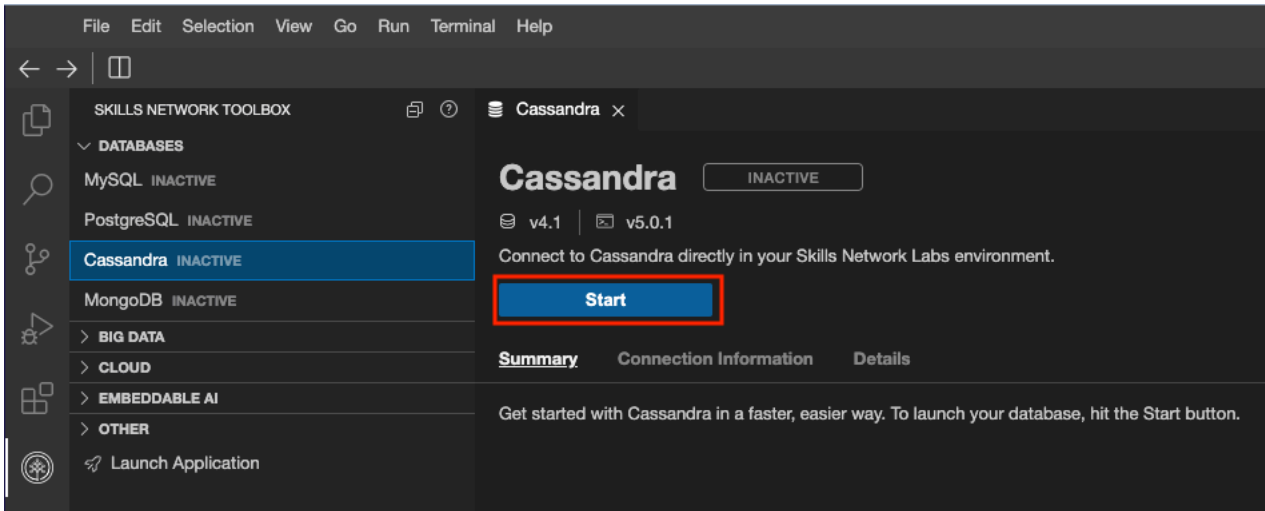
Navigate to Skills Network Toolbox.



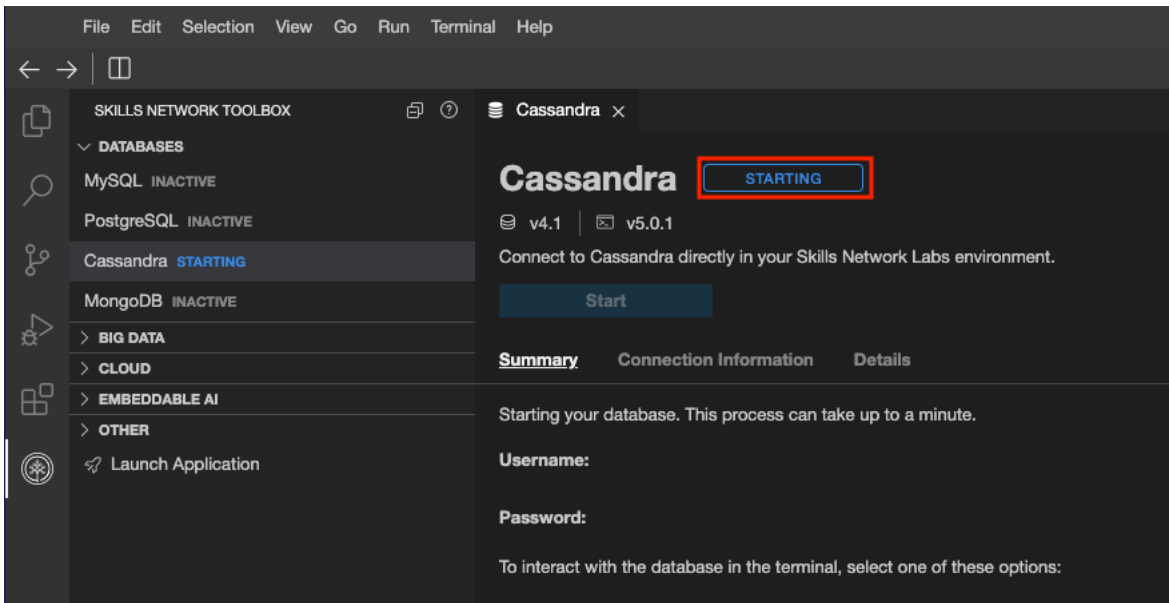
Cassandra is listed, but inactive, which means a database is not available to use.



Once you click on the database, you will see more details and the button to start the database.



Clicking the Start button runs a background process to configure and start your Cassandra server.



Shortly after that, your server is ready for use. This deployment has access control enabled and Cassandra enforces authentication. So, take note of the password as you will need the password to login as a cassandra user.

Cassandra

Cassandra

ACTIVE

v4.1 | v5.0.1

Connect to Cassandra directly in your Skills Network Labs environment.

Stop

Summary | Connection Information | Details

Your database server is now ready to use and available with the following login credentials. For more details on how to navigate Cassandra, please check out the Details section.

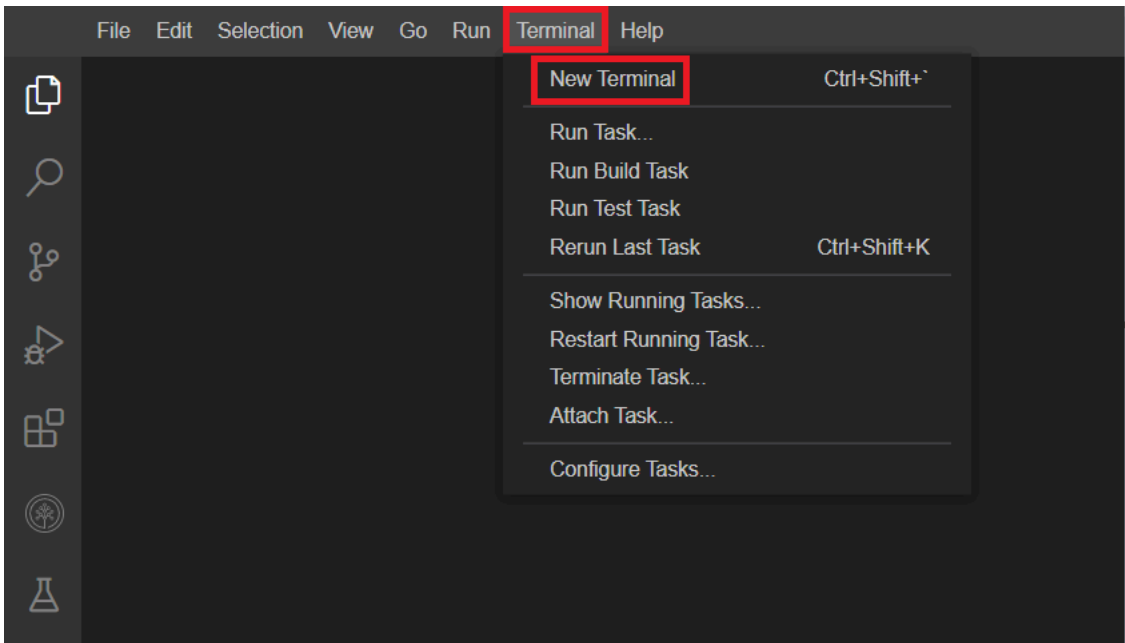
Username:

Password:

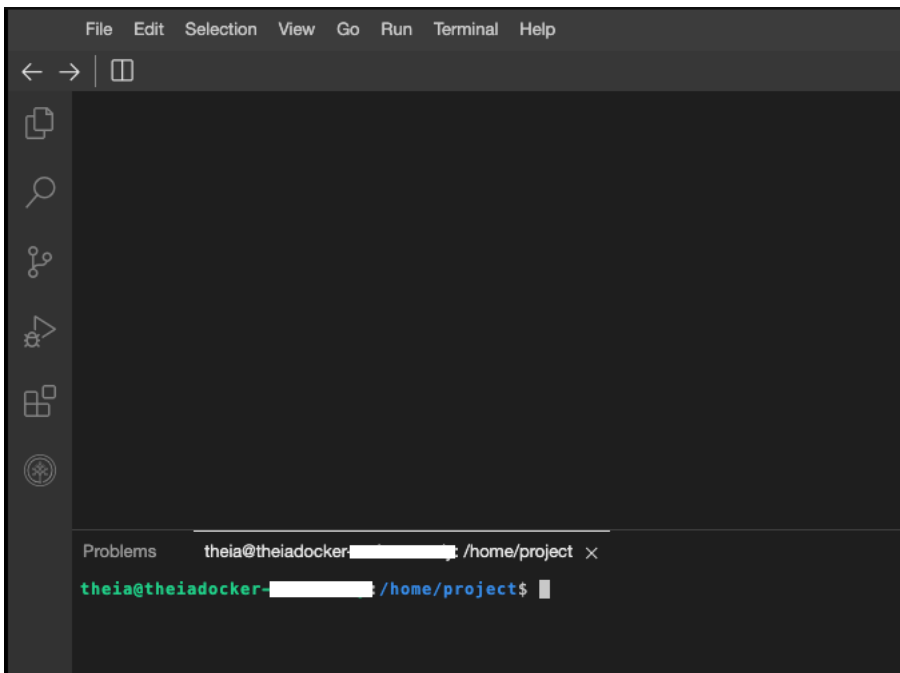
To interact with the database in the terminal, select one of these options:

Cassandra CLI | New Terminal

You can now type \ open terminal and enter the details.



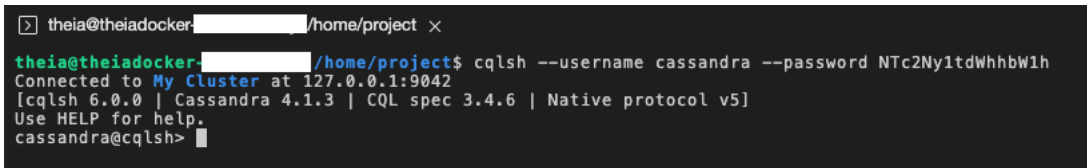
This action open a new terminal at the bottom of the screen as seen in the following image.



Run the following command in the newly opened terminal. (You can copy the code by clicking on the little copy button on the bottom right of the codeblock below and then paste the code where needed.)

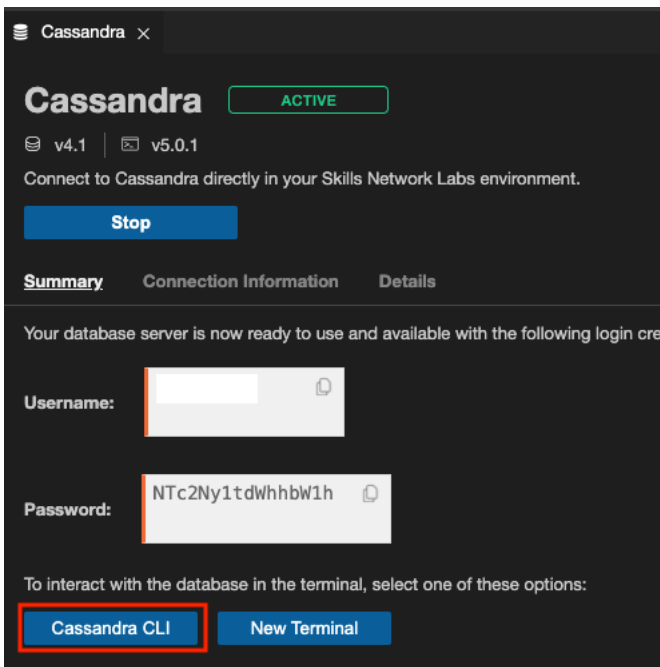
1. 1
1. `cqlsh --username cassandra --password PASSWORD`

Copied! Executed!



This command contains the username and password to connect to Cassanrda server. Your output could be different from the one shown above. Copy the command given to you, and keep the command handy. You will use this command in the next step.

Or you can simply click on Cassandra CLI which does that for you.



Exercise 1: Working with a MongoDB database

Download sample data file

Use the following command to download the data file in your Cloud IDE project directory.

1. 1

```
1. curl -O https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMSkillsNetwork-DB0151EN-edX/labs/FinalProject/movies.json
```

Copied! Executed!

A sample movie document

```
1. 1
2. 2
3. 3
4. 4
5. 5
6. 6
7. 7
8. 8
9. 9
10. 10
11. 11
12. 12
13. 13
14. 14

1. {
2.   _id: '9',
3.   title: 'The Lost City of Z',
4.   genre: 'Action,Adventure,Biography',
5.   Description: 'A true-life drama, centering on British explorer Col. Percival Fawcett, who disappeared while searching for a mysterious city in the
6.   Director: 'James Gray',
7.   Actors: 'Charlie Hunnam, Robert Pattinson, Sienna Miller, Tom Holland',
8.   year: 2016,
9.   'Runtime (Minutes)': 141,
10.  rating: 'unrated',
11.  Votes: 7188,
12.  'Revenue (Millions)': 8.01,
13.  Metascore: 78
14. }
```

Copied!

Task 1: Import movies.json into mongodb server into a database named entertainment and a collection named movies.

Now import the data in movies.json and take a screenshot of the command you used and the output.

```
theia@theiadocker-muhammady:/home/project$ mongoimport -u root -p MjUwMDMtYXV0YWlt --authenticationDatabase admin -d entertainment -c movies
2023-10-25T17:50:22.342-0400 connected to: mongodb://localhost/
2023-10-25T17:50:22.357-0400 100 document(s) imported successfully. 0 document(s) failed to import.
```

Assessment: Take a screen capture of the output and save the screen capture as 01-mongo-import.png (Save the images using either a .jpg or .png extension).

Task 2: Write a mongodb query to find the year in which most number of movies were released

► [Click here for Hint](#)

Take a screenshot of the command you used and the output showing 73 movies in year 2016.

```
entertainment> db.movies.aggregate([
...   {
...     "$group": {
...       "_id": "$year",
...       "moviecount": { "$sum": 1 }
...     }
...   },
...   {
...     $sort: { "moviecount": -1 }
...   },
...   {
...     $limit: 1
...   }
... ])
[ { _id: 2016, moviecount: 73 } ]
entertainment>
```

Assessment: Take a screen capture of the output and save the screen capture as 02-most-movies-year.png (You can save the image using either the .jpg or .png extension).

Task 3: Write a mongodb query to find the count of movies released after the year 1999

Take a screenshot of the command you used and the output showing 99 movies after year 1999.

```
entertainment> db.movies.countDocuments({ year: { $gt: 1999 } })
99
```

Assessment: Take a screen capture of the output and save the screen capture as 03-movies-count-1999.png (You can save the image using either the .jpg or .png extension).

Task 4: Write a query to find out the average votes for movies released in 2007

► [Click here for Hint](#)

Take a screenshot of the command you used and the output.

```
entertainment> db.movies.aggregate(
...
{
  $match: { year: 2007 }
},
{
  $group:
  {
    _id: "$year",
    averageVotes: { $avg: "$votes" }
  }
}
)
[ { _id: 2007, averageVotes: 192.5 } ]
```

Assessment: Take a screen capture of the output and save the screen capture as 04-average-votes.png You can save the image using either the .jpg or .png extension).

Task 5: Export the fields _id, title, year, rating and director from the movies collection into a file named partial_data.csv

Take a screenshot of the command you used and its output.



The screenshot shows a terminal window with a file explorer view of 'partial_data.csv'. The file contains a list of movies with their _id, title, year, rating, and director. Below the file list, the terminal shows the command used to export the data from the 'entertainment' database to the CSV file.

```
theia@theiadocker-muhammady: /home/project$ mongoexport -u root -p MjUwMDMtYXVl --authenticationDatabase admin -d entertainment -c movies --fields '_id,title,year,rating,director' --type=csv -o partial_data.csv
2023-10-25T18:09:33.467-0400 connected to: mongodb://localhost/
2023-10-25T18:09:33.472-0400 exported 100 records
```

Assessment: Take a screen capture of the output and save the screen capture as 05-mongo-export.png (images can be saved with either .jpg or .png extension).

Exercise 2 - Working with a Cassandra database

Download sample data file

If you haven't managed to successfully export data into partial_data.csv file, then download the data file already created for you with the following command.

- 1
1. `curl -O https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMSkillsNetwork-DB0151EN-edX/labs/FinalProject/partial_data.csv`

Copied! Executed!

Task 6 - Create a keyspace named entertainment

► [Click here for Hint](#)

Once created, use the describe command and use the output for assessment.

```
cassandra@cqlsh> CREATE KEYSPACE entertainment
... WITH replication = {'class': 'SimpleStrategy', 'replication_factor' : 3};

Warnings :
Your replication factor 3 for keyspace entertainment is higher than the number of nodes 1

cassandra@cqlsh> describe keyspaces

entertainment  system_auth      system_schema  system_views
system         system_distributed system_traces  system_virtual_schema
```

Assessment: Take a screen capture of the output and save the screen capture as 06-describe-keyspaces.png (You can save the image using either the .jpg or .png extension).

Task 7 - Import partial_data.csv into cassandra server into a keyspace named entertainment and a table named movies

While creating the table movies configure all of the columns as text columns including the id column.

- _id
- title
- year
- rating
- director

► Click here for Hint

```
cassandra@cqlsh:entertainment> CREATE TABLE movies(
...     id text PRIMARY KEY,
...     title text,
...     year text,
...     rating text,
...     director text
... );

cassandra@cqlsh:entertainment> describe movies

CREATE TABLE entertainment.movies (
  id text PRIMARY KEY,
  director text,
  rating text,
  title text,
  year text
) WITH additional_write_policy = '99p'
AND bloom_filter_fp_chance = 0.01
AND caching = {'keys': 'ALL', 'rows_per_partition': 'NONE'}
AND cdc = false
AND comment = ''
AND compaction = {'class': 'org.apache.cassandra.db.compaction.SizeTieredCompactionStrategy', 'max_threshold': '32', 'min_threshold': '4'}
AND compression = {'chunk_length_in_kb': '16', 'class': 'org.apache.cassandra.io.compress.LZ4Compressor'}
AND memtable = 'default'
AND crc_check_chance = 1.0
AND default_time_to_live = 0
AND extensions = {}
AND gc_grace_seconds = 864000
AND max_index_interval = 2048
AND memtable_flush_period_in_ms = 0
AND min_index_interval = 128
AND read_repair = 'BLOCKING'
AND speculative_retry = '99p';
```

And now import the data present in partial_data.csv

► Click here for Hint

```
cassandra@cqlsh:entertainment> COPY entertainment.movies(id,title,year,rating,director) FROM '/home/project/partial_data.csv' WITH DELIMITERS ',';
Using 15 child processes

Starting copy of entertainment.movies with columns [id, title, year, rating, director].
Processed 100 rows; Rate: 25 rows/s; Avg. rate: 47 rows/s
100 rows imported from 1 files in 0 day, 0 hour, 0 minute, and 2.116 seconds (0 skipped).
```

Assessment: Take a screen capture of the output and save the screen capture as 07-movies-imported.png (You can save the image using either the .jpg or .png extension).

Task 8 - Write a cql query to count the number of rows in the movies table

▼ Click here for Hint

use the SELECT COUNT(*) command

- 1
1. SELECT COUNT(*) FROM table_name;

Copied!

```
cassandra@cqlsh:entertainment> SELECT COUNT(*) FROM movies;

count
-----
288
(1 rows)

Warnings:
Aggregation query used without partition key
```

Assessment: Take a screen capture of the output and save the screen capture as 08-movies-count.png You can save the image using either the .jpg or .png extension).

Task 9 - Create an index for the rating column in the movies table using cql

► [Click here for Hint](#)

And then run the describe on movies that shows CREATE INDEX listed in the output.

Take a screenshot of the command you used and the output.

```
cassandra@cqlsh:entertainment> describe movies

CREATE TABLE entertainment.movies (
  id text PRIMARY KEY,
  director text,
  rating text,
  title text,
  year text
) WITH additional_write_policy = '99p'
AND bloom_filter_chance = 0.01
AND caching = {'keys': 'ALL', 'rows_per_partition': 'NONE'}
AND cdc = false
AND comment = ''
AND compaction = {'class': 'org.apache.cassandra.db.compaction.SizeTieredCompactionStrategy', 'max_threshold': '32', 'min_threshold': '4'}
AND compression = {'chunk_length_in_kb': '16', 'class': 'org.apache.cassandra.io.compress.LZ4Compressor'}
AND memtable_flush_period_in_ms = 0
AND gc_check_chance = 1.0
AND default_time_to_live = 0
AND extensions = {}
AND gc_grace_seconds = 864000
AND max_index_interval = 2048
AND memtable_flush_period_in_ms = 0
AND min_index_interval = 128
AND read_repair = 'BLOCKING'
AND speculative_retry = '99p';

CREATE INDEX rating_index ON entertainment.movies (rating);
```

Assessment: Take a screen capture of the output and save the screen capture as 09-movies-rating-index.png (You can save the image using either the .jpg or .png extension).

Task 10 - Write a cql query to count the number of movies that are rated 'G'.

▼ [Click here for Hint](#)

use the SELECT COUNT(*) FROM table_name WHERE criteria command

1. 1
1. SELECT COUNT(*) FROM table_name WHERE some_field='some value';

Copied!

```
cassandra@cqlsh:entertainment> SELECT COUNT(*) FROM movies WHERE rating='G';

count
-----
32
(1 rows)

Warnings:
Aggregation query used without partition key
```

Assessment: Take a screen capture of the output and save the screen capture as 10-g-rated-movies.png (You can save the image using either the .jpg or .png extension).

::page{title="Checklist for submission"}

1. **Assessment:** Take a screen capture of the output and save the screen capture as 01-mongo-import.png.
2. **Assessment:** Take a screen capture of the output and save the screen capture as 02-most-movies-year.png.
3. **Assessment:** Take a screen capture of the output and save the screen capture as 03-movies-count-1999.png.
4. **Assessment:** Take a screen capture of the output and save the screen capture as 04-average-votes.png.
5. **Assessment:** Take a screen capture of the output and save the screen capture as 05-mongo-export.png.
6. **Assessment:** Take a screen capture of the output and save the screen capture as 06-describe-keyspaces.png.
7. **Assessment:** Take a screen capture of the output and save the screen capture as 07-movies-imported.png.
8. **Assessment:** Take a screen capture of the output and save the screen capture as 08-movies-count.png.
9. **Assessment:** Take a screen capture of the output and save the screen capture as 09-movies-rating-index.png.
10. **Assessment:** Take a screen capture of the output and save the screen capture as 10-g-rated-movies.png.

Congratulations! That's a wrap!

Author: [Muhammad Yahya](#)

© IBM Corporation. All rights reserved.