Data Engineering Platforms
MSCA 31012

Installation Documentation
for Mac

Contents:

Warning: Please read the documentation thoroughly. Some installations have special instructions for this course, including stipulations to wait until a certain week to install for demo versions.

NOTE: Always install the latest versions. This document may show a slightly older version number in the screenshots, but still use the latest version.

OpenRefine Installation & Setup Instructions

Purpose: OpenRefine is a data cleansing and transformation tool.

Source: https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions

OpenRefine is a desktop application in that you download it, install it, and run it on your own computer. However, unlike most other desktop applications, it runs as a small web server on your own computer and you point your web browser at that web server in order to use Refine. So, think of Refine as a personal and private web application.

Requirements

Java JRE is a prerequisite for OpenRefine

To check if you already have a JRE installed, open command terminal and type java. If not please follow JRE installation instructions below.

https://docs.oracle.com/javase/8/docs/technotes/guides/install/install_overview.html#CJAGAACB

Install Summary

1. OpenRefine requires you to have a working Java JRE, otherwise you will not be able to start OpenRefine.
2. Download OpenRefine here.
3. Install it as detailed below for your operating system (Windows, Mac OSX, Linux).
4. Go to your Applications, and click on OpenRefine.
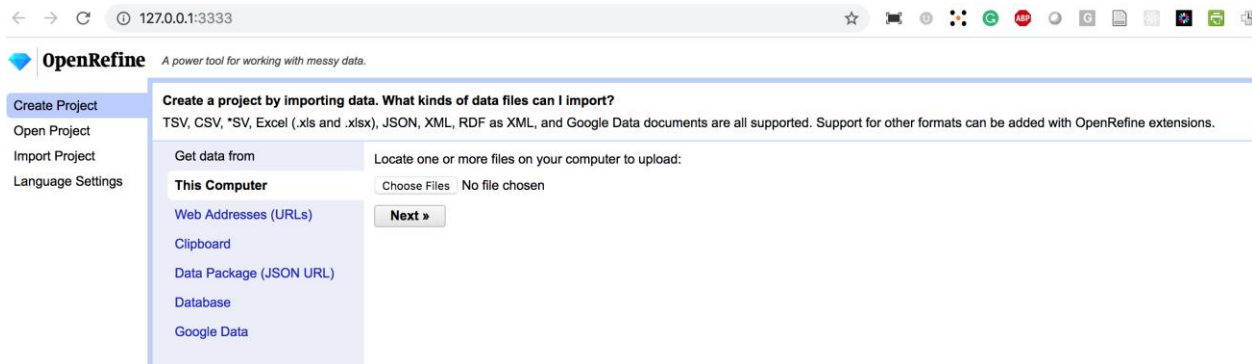   a. The following error may occur:
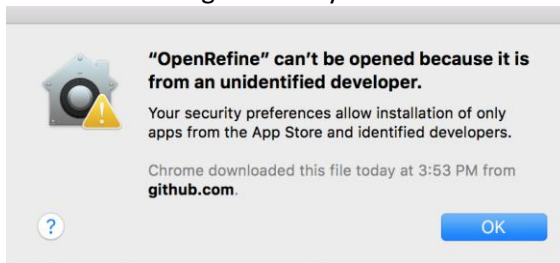
   b. If you receive the above error, go to your System Preferences and click on Security & Privacy.
   c. Click on the General tab and click on Open Anyway:
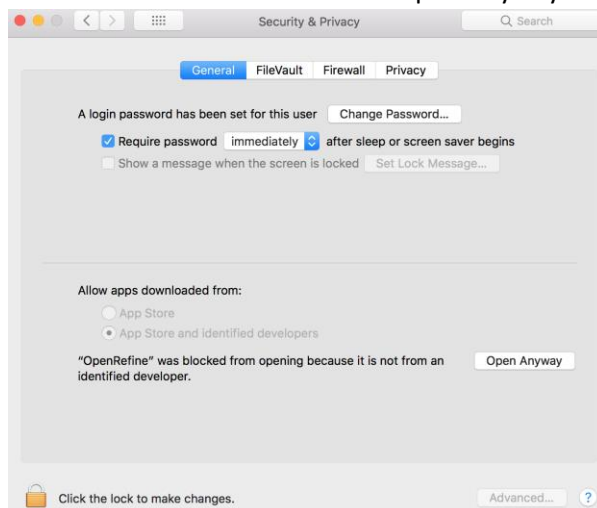
   d. A ___ pop will open asking if you're sure to open.

Click Open.

5. You can now go back to Applications and open OpenRefine. A new tab on your web browser should open.

6. As long as OpenRefine is running, you can point your browser at http://127.0.0.1:3333/ to use it, and you can even use it in several browser tabs and windows.
7. If you're running a proxy or get a BindException, you can change the IP configuration with -i and -p, see **Running & Configuration** below, or use refine -help for options.

By default (and for security reasons) Refine only listens to TCP requests coming from localhost (127.0.0.1 on port 3333). If you want to respond to TCP requests coming to any IP address the machine has, run refine like this from the command line:

./refine -i 0.0.0.0

On Mac OS X, you can add a specific entry to the Info.plist file located within the app bundle (/Applications/OpenRefine.app/Contents/Info.plist):

<string>-Drefine.host=0.0.0.0</string>

FileZilla Client Installation

Data Engineering Platforms for Analytics: MSCA 31012

Purpose: FileZilla will be used to connect to the RCC Midway server via SFTP

Installation & Setup

1. **Download FileZilla Client**
   - Navigate to https://filezilla-project.org
   - Click the grey "Download FileZilla Client" button (**not** FileZilla Server)
   - Click the green "Download FileZilla Client" button on the next page to download the client for your given operating system
   - If a pop-up is presented, click the green "Download" button for FileZilla (**not** FileZilla Pro)

2. **Install FileZilla Client**
   - Once the program has downloaded, run the executable
   - Use the default installation parameters EXCEPT for the following:
     - By default, FileZilla will check the boxes to opt you in to "offers" like McAfee or Yahoo. Uncheck these boxes during the installation process to avoid installing the unnecessary software.

3. **Setup FileZilla Client**
   - Open FileZilla
   - Navigate to File > Site Manager
   - Click "New Site" to begin configuring a new connection
   - Click "Rename" to give your connection a meaningful name such as RCCconnect
   - Input the following parameters:
     - Protocol: SFTP
     - Host: midway.rcc.uchicago.edu
     - Port: (leave blank - will default to 22)
     - Logon Type: Normal
     - User: Your CNET ID
     - Password: Your CNET password
   - Click Connect
   - Optional: Opt to save your CNET password
   - If prompted with a dialogue about an "Unknown host key", check the box to "Always trust this host" and click OK.
   - Confirm you see the file tree on the right side to indicate you successfully connected
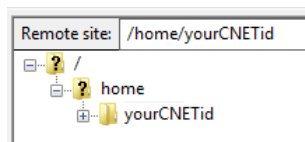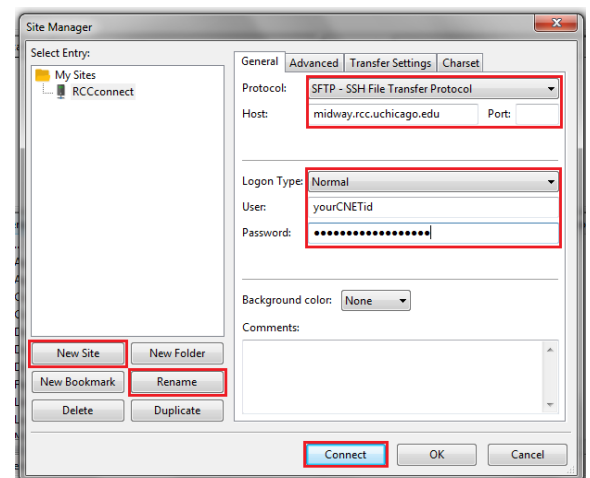
MySQL Installation

Installation Instructions

1. Before starting the MySQL installation make sure to uninstall any pre-existing installations of MySQL database and install the latest available version.
   a. Start Terminal and run the following commands
      - find / -name mysql

      - Delete all references and the mysql-<xxx> directory from /usr/local directory
      - Move MySQL workbench to trash

2. **Make a note of the below useful URLs that give additional insights into the installation process. If you run into any issues with the latest 5.7 version of MySQL please revert back to the 5.6 version.
   https://dev.mysql.com/doc/refman/5.6/en/osx-installation-pkg.html
   https://dev.mysql.com/doc/refman/5.6/en/osx-installation-launchd.html
   https://dev.mysql.com/downloads/workbench/

3. Download MySQL Installer Package from: https://dev.mysql.com/downloads/mysql/
   Choose the latest version of the DMG Archive for Mac download

4. The next page will have 2 buttons to login and signup. You can scroll to the bottom and bypass this by choosing No thanks, just start my download.

Login »
using my Oracle Web account

Sign Up »
for an Oracle Web account

MySQL.com is using Oracle SSO for authentication. If you already have an Oracle Web account, click the Login link. Otherwise, you can signup for a free account by clicking the Sign Up link and following the instructions.
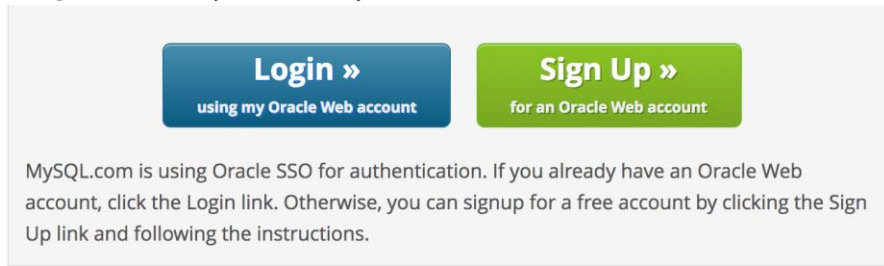
No thanks, just start my download.

5. Save the file to a folder on your machine. This will prevent you from having to download it again if something goes wrong with the install.
6. Double click on the pkg to open up the installer and follow the steps below.
   a. Click Continue on the 'The package will run a program to determine if the software can be installed.

   b. Click Continue on the Installation welcome prompt.

   c. Click continue on the License Agreement

   d. Agree to the license agreement

   e. Click Install. The installation might need admin privileges so please enter admin credentials to install software.
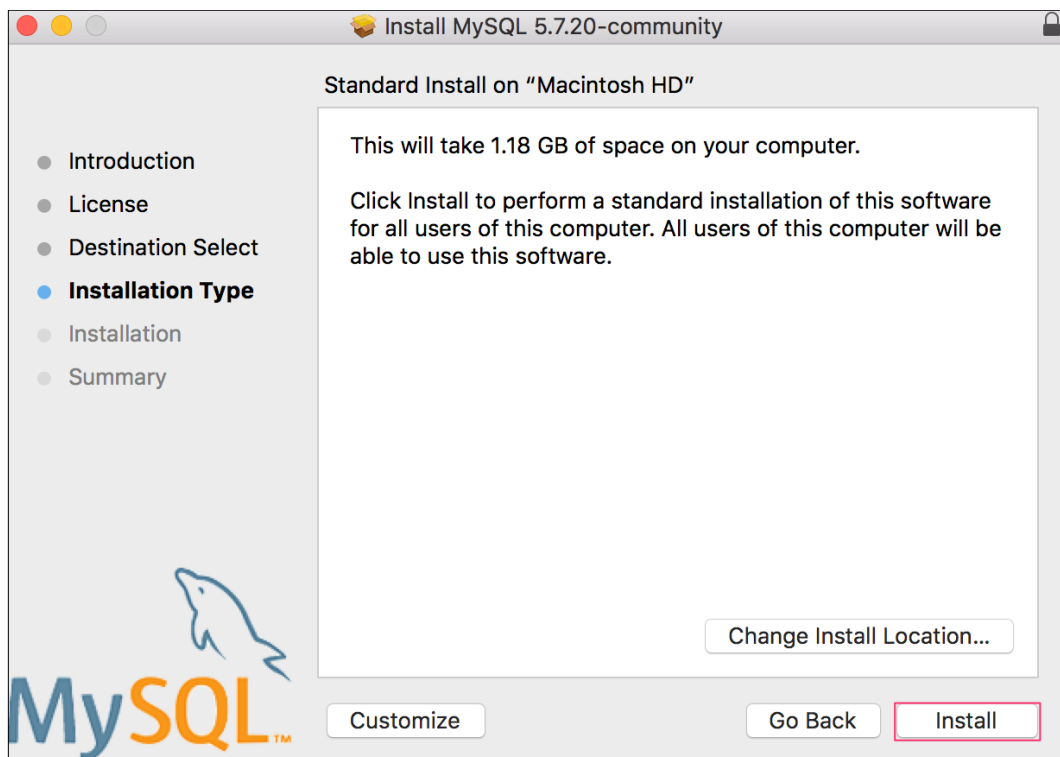
7. Once the Install is complete, the Configuration is next.
   a. On the Configuration screen, select Use Legacy Password Encryption
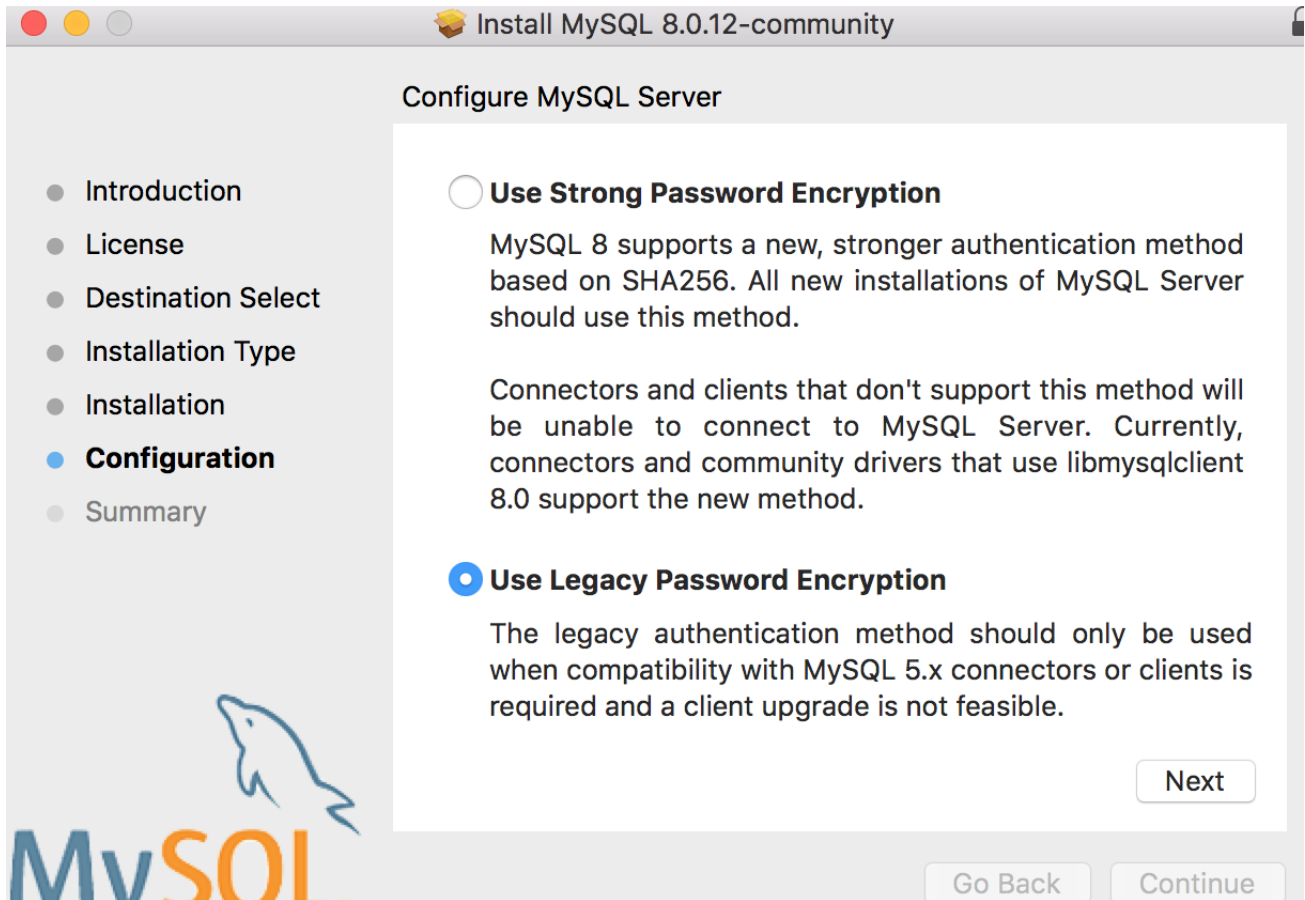
8. After clicking Next enter a password. PLEASE use 'root' as your password so the in-class scripts will run correctly. THen click Finish.

Install MySQL 8.0.12-community

Configure MySQL Server

- Introduction
- License
- Destination Select
- Installation Type
- Installation
- **Configuration**
- Summary

Please enter a password for the "root" user.

●●●●●●●●

A reasonable password consists of at least eight characters that are a mix of letters, numbers, and other characters.

☑ Start MySQL Server once the installation is complete.

[Finish]

[Go Back]   [Continue]

9. The installer should now be complete. Click Close.

10. Verify the installation

```
[GLHR-MB-c02s30efg8wm:local sbharadwaj-local$ ls -ltr
 total 0
 drwxr-xr-x    3 root             wheel     96 Nov 26  2017 jamf
 drwxr-xr-x    3 root             wheel     96 Nov 26  2017 remotedesktop
 drwxr-xr-x    5 root             wheel    160 Nov 26  2017 man
 drwxr-xr-x    5 root             wheel    160 Nov 26  2017 share
 drwxr-xr-x   30 root             admin    960 Nov 26  2017 lib
 drwxr-xr-x   29 root             wheel    928 Nov 26  2017 include
 drwxr-xr-x   40 root             wheel   1280 Jan 31  2018 bin
 drwxr-xr-x    8 sbharadwaj-local staff    256 Sep 18 21:22 mongodb
 lrwxr-xr-x    1 root             wheel     30 Sep 20 20:48 mysql -> mysql-8.0.12-macos10.13-x86_64
 drwxr-xr-x   13 root             wheel    416 Sep 20 20:48 mysql-8.0.12-macos10.13-x86_64
 You have new mail in /var/mail/sbharadwaj-local
 GLHR-MB-c02s30efg8wm:local sbharadwaj-local$ █
```

11. Go to your System Preferences and click on MySQL.
12. **You may get an error when trying to open MySQL indicating MySQL preference pane is not working. If this occurs**:
    a. Go to Terminal to /usr/local/mysql/support-files. Find the mysql.server file:

```
/usr/local/mysql/support-files
[GLHR-MB-c02s30efg8wm:support-files sbharadwaj-local$ ls -ltr
total 48
-rw-r--r--  1 root  wheel    773 Jun 28 11:18 magic
-rwxr-xr-x  1 root  wheel   1061 Jun 28 12:53 mysqld_multi.server
-rwxr-xr-x  1 root  wheel   2048 Jun 28 12:53 mysql-log-rotate
-rwxr-xr-x  1 root  wheel  10622 Sep 20 21:37 mysql.server
GLHR-MB-c02s30efg8wm:support-files sbharadwaj-local$
```

    b.   Use VI to update the below fields:

edit /usr/local/mysql/support-files/mysql.server

change lines ~ 46 & 47

```
basedir=
datadir=
```

to

```
basedir=/usr/local/mysql
datadir=/usr/local/mysql/data
```

    c.   If you need to take this step, Restart your laptop.

13. After you double click on the MySQL icon under system preferences, the below screen will pop up. When the Active and Installed Instance is Green, the it is running correctly and successful, else please uninstall and reinstall MySQL Server
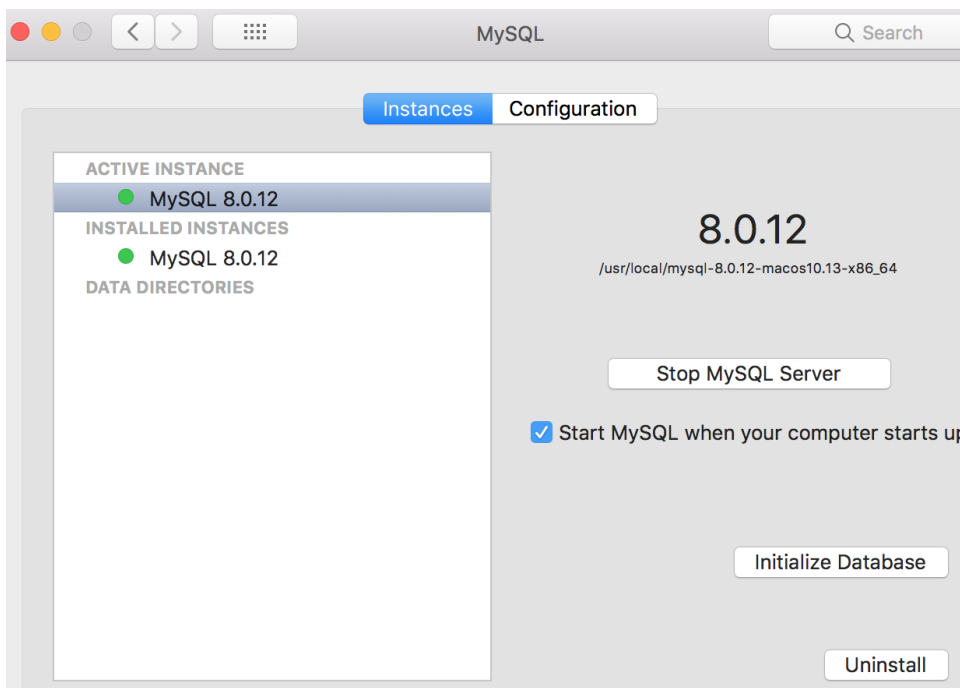
MySQL Workbench

14. Download the latest MySQL Workbench Package from:
https://dev.mysql.com/downloads/workbench/

15. Click "No thanks, just start my download" at the bottom

Note: The rest of the screen shots might show the older versions of installation. However, proceed with the latest version

16. Once downloaded, drag to install and authenticate.

17. Open MySQLWorkbench

18. Click on the Local instance card. You'll be prompted for a password. Please use the password from step 10 (should be **root**) to connect MySQL workbench to the MySQL Server. Click on the 'Save password to keychain' so you won't have to type in your password each time.

Welcome to MySQL Workb

MySQL Workbench is the official graphical user interface (GUI) tool for MySQL. It allows you create and browse your database schemas, work with database objects and insert data as design and run SQL queries to work with stored data. You can also migrate schemas and dat database vendors to your MySQL database.

Browse Documentation >   Read the Blog >   Discus
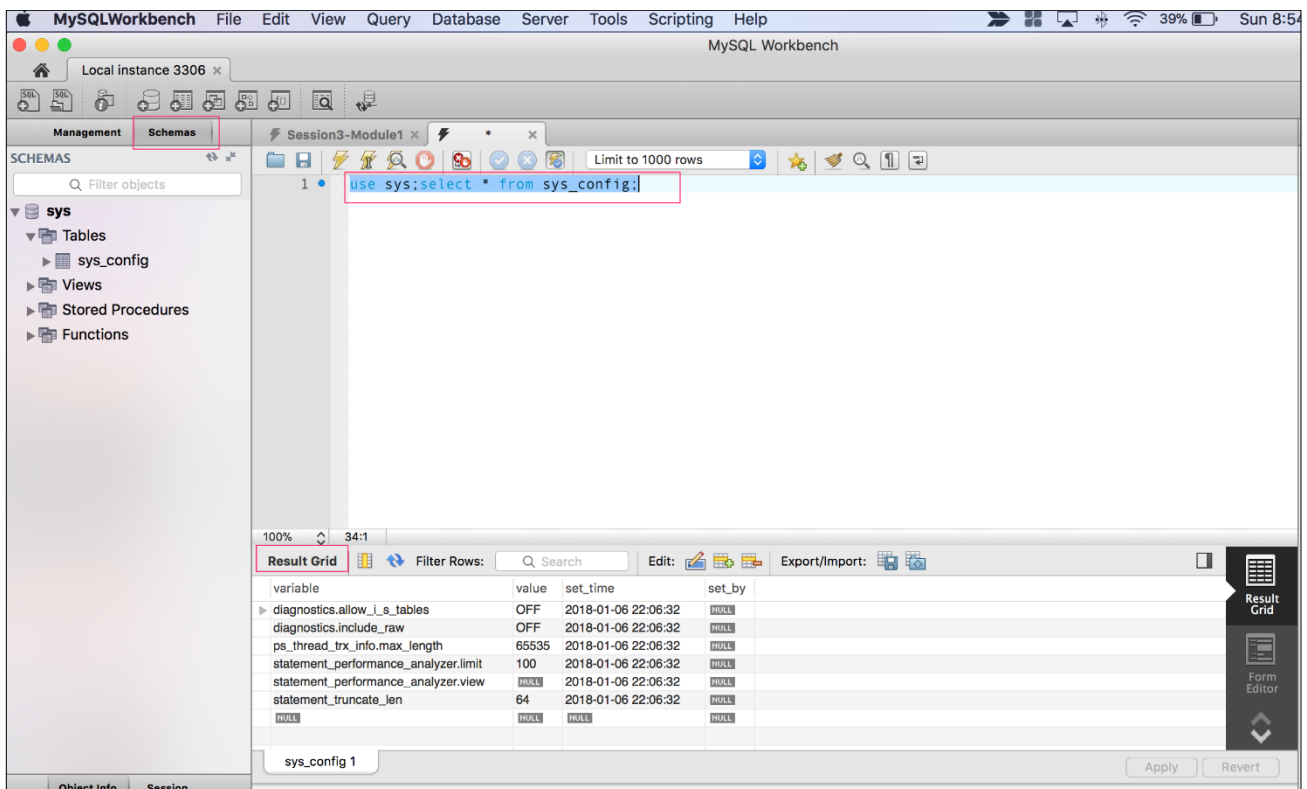
MySQL Connections ⊕ ⊙

Local instance 3306 (auto sa...
👤 root
localhost:3306

19. Type the following query and then run it (click lightning bolt or select and use CMD+ENTER to run)

20. In case of password related errors, please follow the steps below to reset the root password
    - Stop MySQL server
    - sudo /usr/local/mysql/support-files/mysql.server start --skip-grant-tables
    - /usr/local/mysql/bin/mysql
    - mysql> FLUSH PRIVILEGES;
    - mysql> ALTER USER 'root'@'localhost' IDENTIFIED BY 'root';
    - Ctrl + z
    - /usr/local/mysql/bin/mysql –u root –p
    - enter the new password i.e root

MongoDB Installation & Setup

Purpose: MongoDB is the NoSQL document store database used in this course.

Download MongoDB Community

1. Detailed installation installation documentation can be found here: https://docs.mongodb.com/manual/administration/install-community/

2. Download the Community Server installation package for your operating system

Install MongoDB Community

1. it.
2. Open the Terminal and follow these instructions

   Note: for the rest of the instructions, make sure to replace your download version number in the file name if it is newer than the instructions below

   a. 'cd' into the folder the .tgz is downloaded.

   b. Extract the files from the .tgz file by:

      i. Type in 'tar –zxvf mongodb-osx-ssl-x86_64-4.0.2.tgz' (use your file name/version)

```
GLHR-MB-c02s30efg8wm:~ sbharadwaj-local$ cd Downloads
GLHR-MB-c02s30efg8wm:Downloads sbharadwaj-local$ tar -zxvf mongodb-osx-ssl-x86_64-4.0.2.tgz
x mongodb-osx-x86_64-4.0.2/README
x mongodb-osx-x86_64-4.0.2/THIRD-PARTY-NOTICES
x mongodb-osx-x86_64-4.0.2/MPL-2
x mongodb-osx-x86_64-4.0.2/GNU-AGPL-3.0
x mongodb-osx-x86_64-4.0.2/LICENSE-Community.txt
x mongodb-osx-x86_64-4.0.2/bin/mongodump
x mongodb-osx-x86_64-4.0.2/bin/mongorestore
x mongodb-osx-x86_64-4.0.2/bin/mongoexport
x mongodb-osx-x86_64-4.0.2/bin/mongoimport
x mongodb-osx-x86_64-4.0.2/bin/mongostat
x mongodb-osx-x86_64-4.0.2/bin/mongotop
x mongodb-osx-x86_64-4.0.2/bin/bsondump
x mongodb-osx-x86_64-4.0.2/bin/mongofiles
x mongodb-osx-x86_64-4.0.2/bin/mongoreplay
x mongodb-osx-x86_64-4.0.2/bin/mongod
x mongodb-osx-x86_64-4.0.2/bin/mongos
x mongodb-osx-x86_64-4.0.2/bin/mongo
x mongodb-osx-x86_64-4.0.2/bin/install_compass
You have new mail in /var/mail/sbharadwaj-local
GLHR-MB-c02s30efg8wm:Downloads sbharadwaj-local$
```

3. Move the tarball to a new mongo directory, called '/usr/local/mongodb

   a. Type: sudo mv mongodb-osx-x86_64-4.0.2 /usr/local/mongodb

```
Downloads — -bash — 135×46
GLHR-MB-c02s30efg8wm:Downloads sbharadwaj-local$ sudo mv mongodb-osx-x86_64-4.0.2 /usr/local/mongodb
```

4. Create a MongoDB Data directory

   a. MongoDB stores the data into the /data/db directory by default, so you'll need to create this directory using the sudo mkdir –p /data/db command

   b. In addition, you'll need to assign proper permission using the 'chown' command. Type in 'whoami' to get your id. Then follow the sudo chown command below:

```
sbharadwaj-local — -bash — 135×46
GLHR-MB-c02s30efg8wm:~ sbharadwaj-local$ sudo mkdir -p /data/db
Password:
GLHR-MB-c02s30efg8wm:~ sbharadwaj-local$ whoami
sbharadwaj-local
GLHR-MB-c02s30efg8wm:~ sbharadwaj-local$ sudo chown sbharadwaj-local /data/db
GLHR-MB-c02s30efg8wm:~ sbharadwaj-local$
```

5. Set mongodb/bin PATH to ~/.bash_profile.

a. You will need to set the environment variable for MongoDB. For that we need to add the mongodb/bin path to the bash_profile.
   i. 'cd' back to your local drive.
   ii. Type 'pwd' to ensure you're in the local directory.
   iii. Type 'touch .bash_profile'
   iv. Type open.bash_profile

```
● ● ●                                            🏠 sbharadwaj-local — -bash — 135×46
GLHR-MB-c02s30efg8wm:~ sbharadwaj-local$ cd
GLHR-MB-c02s30efg8wm:~ sbharadwaj-local$ pwd
/Users/sbharadwaj-local
GLHR-MB-c02s30efg8wm:~ sbharadwaj-local$ touch .bash_profile
GLHR-MB-c02s30efg8wm:~ sbharadwaj-local$ open .bash_profile
GLHR-MB-c02s30efg8wm:~ sbharadwaj-local$ ▯
```

b. A Text editor will open. Type the below and save:

```
export MONGO_PATH=/usr/local/mongodb
export PATH=$PATH:$MONGO_PATH/bin
```

6. You can check that it's installed and the version by typing: mongo –version

```
GLHR-MB-c02s30efg8wm:~ sbharadwaj-local$ mongo -version
MongoDB shell version v4.0.2
git version: fc1573ba18aee42f97a3bb13b67af7d837826b47
allocator: system
modules: none
build environment:
    distarch: x86_64
    target_arch: x86_64
You have new mail in /var/mail/sbharadwaj-local
GLHR-MB-c02s30efg8wm:~ sbharadwaj-local$ ;2D▮
```

7. Download and install the NoSQLBooster client here: https://nosqlbooster.com/

Mac OS X

Version 4.7.4

What's new in this version?

Change log

OS X 10.8 and above, Size: 59.7M

⬇ DOWNLOAD

Mirror:Japan

Previous Releases 3.5.7

8. Start MongoDB:
   a. To Start MongoDB, open your Terminal Shell. Type **mongod**
   b. Take note of the last line, it should say 'waiting for connections on port'

```
2018-09-19T20:17:17.299-0500 I CONTROL  [initandlisten]
2018-09-19T20:17:17.299-0500 I CONTROL  [initandlisten] ** WARNING: soft rlimits too low. Number of files is 256
2018-09-19T20:17:17.352-0500 I FTDC     [initandlisten] Initializing full-time diagnostic data capture with dire
2018-09-19T20:17:17.355-0500 I NETWORK  [initandlisten] waiting for connections on port 27017
```

9. Open up the NoSQLBooster application.
   a. The port should already be detected. Test the connection to ensure it works.

Anaconda Installation Instructions

Purpose: Anaconda is a Python distribution and package management tool. Python and Jupyter notebooks will be used for parts of this course.

Download & Install Anaconda

1. Download Anaconda for Python 3.6: https://www.anaconda.com/download/#macos
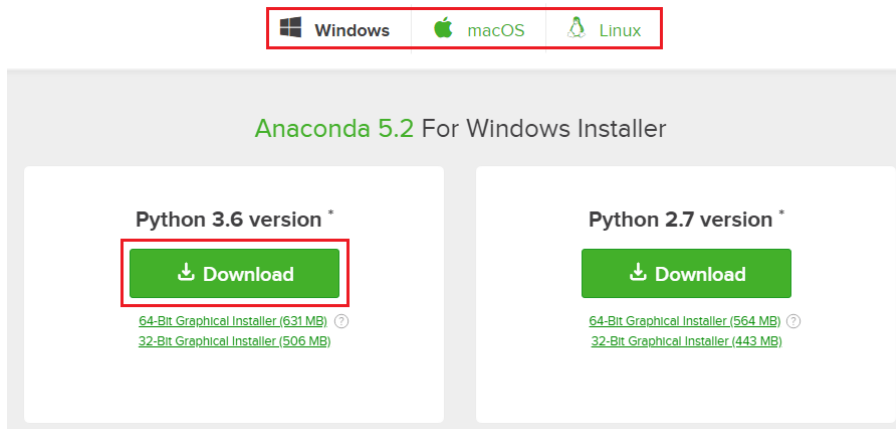
2. A pop up will open. Optional if you want a python cheatsheet or not. Otherwise, just skip.
3. Run the Installer
4. During the Install, you'll be asked if you want to download Microsoft VScode. This is optional. VSCode is a text editor.
5. Once it's finished installing, go to your terminal shell. 'cd' back to your local directory. Type 'jupyter notebook'. This will open a new tab in your default browser.
    a. If you ever want to open a jupyter notebook in a different directory than your local, cd to that directory first and then type jupyter notebook.

Neo4j Installation

Purpose: Neo4j is the graph database used for this course.

WARNING: This is a trial installation for 30 days. WAIT to install until just prior to the week we use Neo4j in class (week 8).

Download Neo4j

1. Download the free book on graph databases: https://neo4j.com/graph-databases-book/?ref=home

2. Download Neo4j by clicking the download button at:
   https://neo4j.com/download/

3. Register with your name and email then click "Download Desktop"

4. Run the installer.

5. Launch Neo4j

Setup Neo4j

1. Launch Neo4j Desktop

2. Accept the license terms

3. Confirm the location to store application data if prompted (default is fine)

4. Click to log in using a social account, then select Google.
   Log in using your uchicago.edu credentials.

5. Let the setup finish.

Tableau Installation

Purpose: Tableau is a data visualization tool.

License & Install:

1. Sign up for a 1 year educational license at https://www.tableau.com/academic/students

2. Click "Get Tableau for Free" and enter your information with uchicago.edu email

3. Verify your student status as prompted

4. Download and install Tableau as prompted

Google Cloud Platform (GCP) Cloud SQL Setup

Purpose: GCP will be used for your final project to collaborate with your team and get exposure to cloud computing.

Obtain GCP Credits (Two Options - try to get both)

1. Sign up for $300 of free GCP credits at https://cloud.google.com/free
   NOTE: This promotion is only for google accounts that have not used GCP before. If you have previously used GCP and obtained these credits, create another google account. It is recommended to use your university Google account.

2. Obtain the $50 Education credits through the link on Canvas.
   NOTE: When accepting the $50 credits, make sure to use the same Google account as you used for the $300 credit above.

Setup GCP Project

1. After signing up for credits, your console should open.
   If not, navigate to https://console.cloud.google.com

2. If you are greeted with a welcome screen, click on "Google Cloud Platform" in the blue bar at the top left to go to your main dashboard

3. Click on "My First Project" in the top blue bar and then click "New Project"

   Give your project a name and click "Create"

Setup Cloud SQL

1. Type "Cloud SQL" in the search bar at the top of your project and select "Cloud SQL"
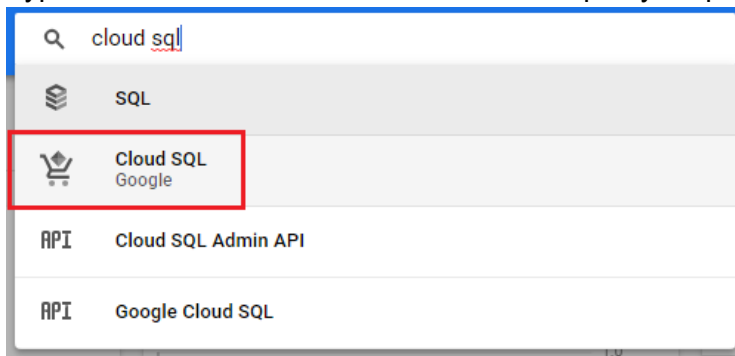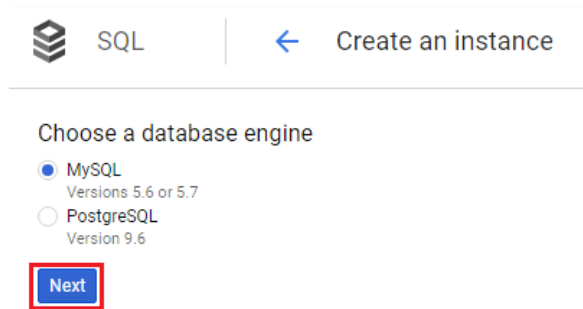

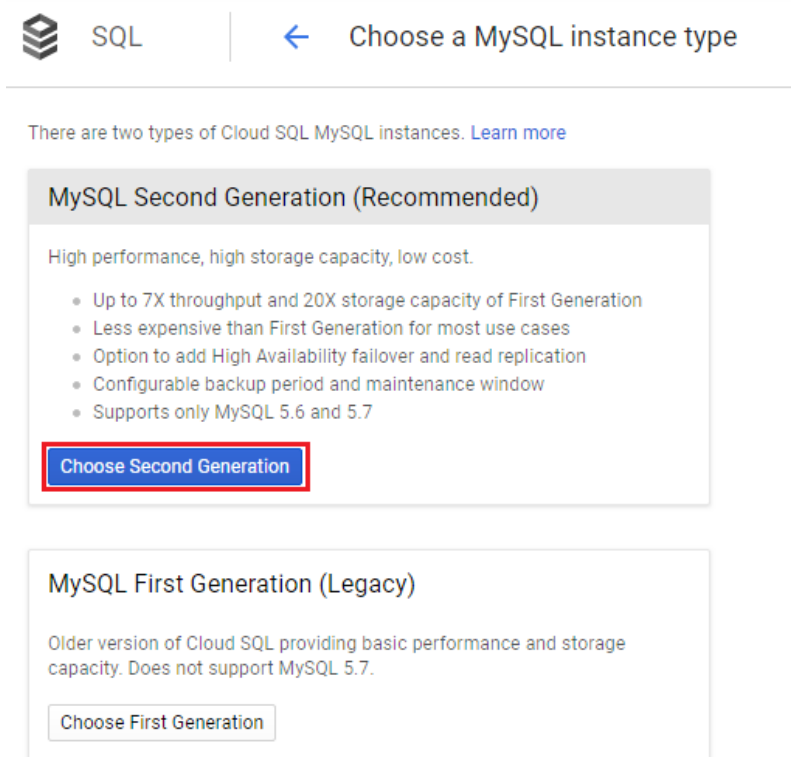2. Click "Go to Cloud SQL" on the next screen

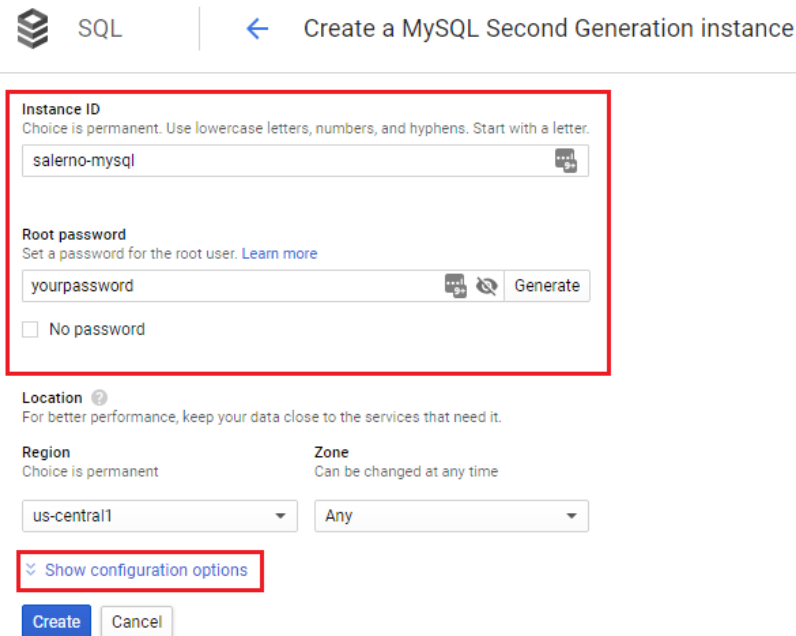3. Click "Create Instance" at the top of the next screen

4. Choose MySQL and click Next

5. Choose Second Generation

6. Name your instance and specify a root password (pick something secure!)
THEN expand the "Show configuration options"

SQL | ← Create a MySQL Second Generation instance

Instance ID
Choice is permanent. Use lowercase letters, numbers, and hyphens. Start with a letter.

salerno-mysql

Root password
Set a password for the root user. Learn more

yourpassword   Generate

☐ No password

Location ⓘ
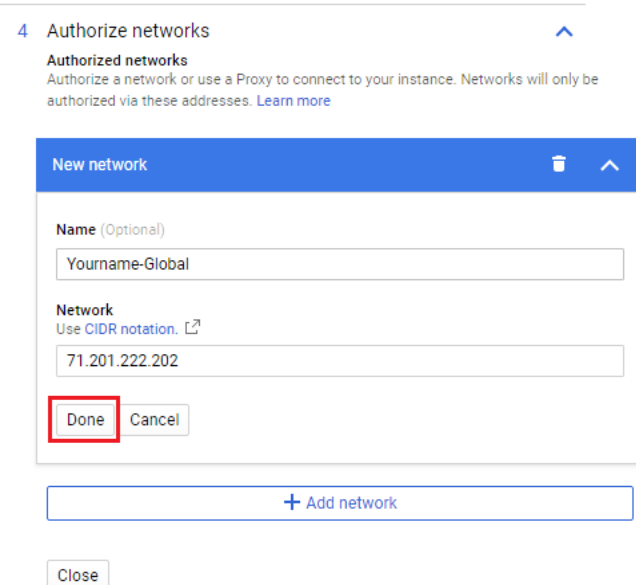For better performance, keep your data close to the services that need it.

Region
Choice is permanent

Zone
Can be changed at any time

us-central1 ▾   Any ▾

⌄ Show configuration options

Create   Cancel

7. Expand the "Authorize Networks" option and click "+ Add Network"
Add your public IP address (can be found by googling "what's my IP address") and give the connection a name. Note, this will have to be changed if you move locations.
Click Done

4   Authorize networks   ⌃
Authorized networks
Authorize a network or use a Proxy to connect to your instance. Networks will only be
authorized via these addresses. Learn more

New network   🗑 ⌃

Name (Optional)

Yourname-Global

Network
Use CIDR notation. ↗

71.201.222.202

Done   Cancel

+ Add network

Close

8. Click the blue Create button at the bottom below the options to create the instance.

9. You will see a spinning wheel next to your newly configured Cloud SQL instance. The setup will take a few/multiple minutes. Be patient.
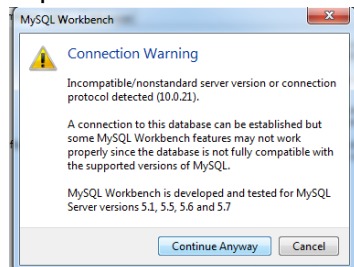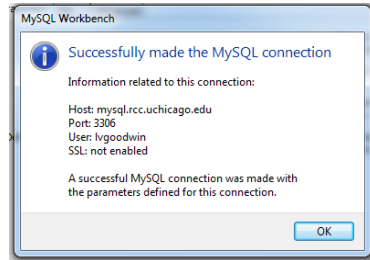
Connect MySQL to RCC

Setup

1. RCC - https://rcc.uchicago.edu/docs/tutorials/msca/ (tutorials about RCC)
2. VPN - Client – Cisco AnyConnect
   https://uchicago.service-now.com/it?id=kb_article&kb=KB00015292
   a. When connecting to RCC outside UC network, User needs to connect via VPN.
   b. cvpn.uchicago.edu (connection)
   c. use your CNET id to connect.
3. Open MySQL workbench and Choose Database > Manage Connections
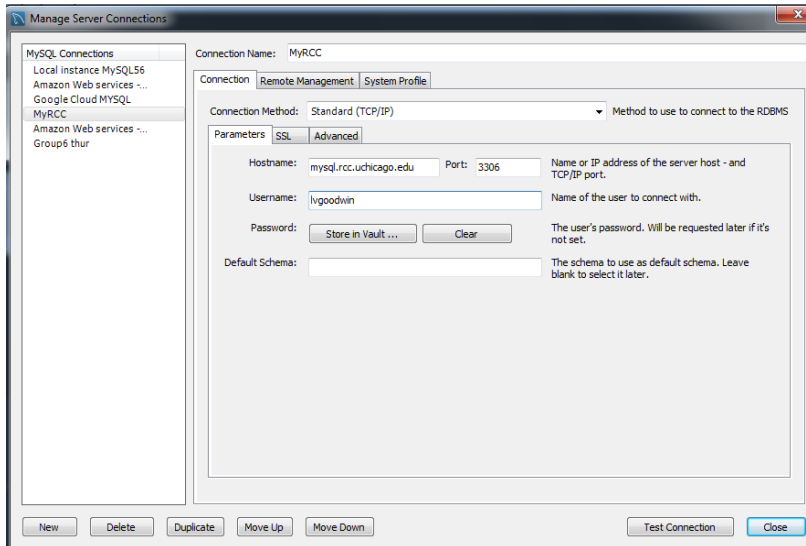
   a. Enter a name for your connection; (I suggest RCC)
   b. For Hostname enter: mysql.rcc.uchicago.edu
   c. Port: 3306
   d. **Username /Password (Credentials) will be provided in class. If you haven't received these yet, stop here for now.**
   e. Password: Click Store in Vault and then enter your password there to be saved
   f. Click Test Connection button
   g. When you Test your connection, you will get this warning error message. This is expected. Click Continue Anyway.

h. You should still get a successful message afterwards.

MySQL Workbench

Successfully made the MySQL connection

Information related to this connection:

Host: mysql.rcc.uchicago.edu
Port: 3306
User: lvgoodwin
SSL: not enabled

A successful MySQL connection was made with
the parameters defined for this connection.

OK

i. Click Close
j. Exit MySQL Workbench and then open it again.
k. You will now see your connection on the main page. Click it to open and you will be inside the tool.

You are now connected.