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

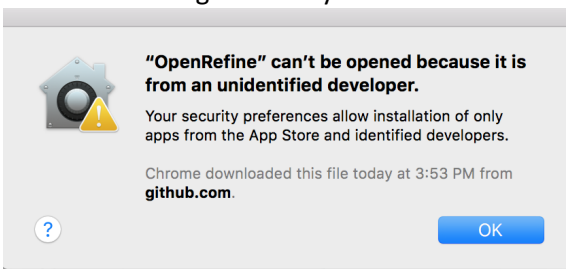
Java JRE is a prerequisite for OpenRefine

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

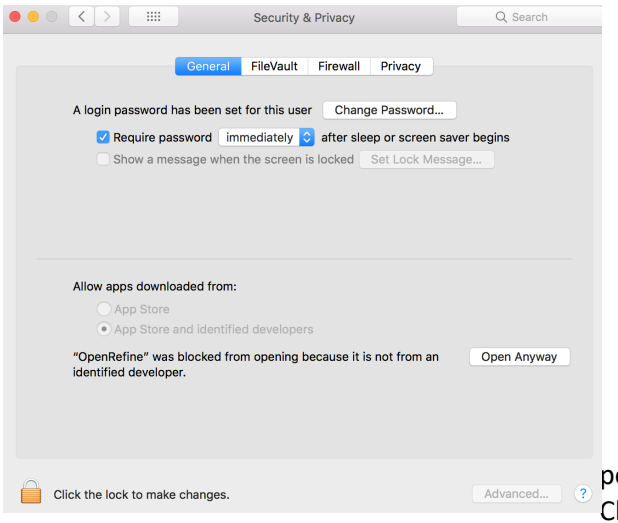
https://docs.oracle.com/javase/8/docs/technotes/guides/install/install_overview.html#CJAGAACB

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:

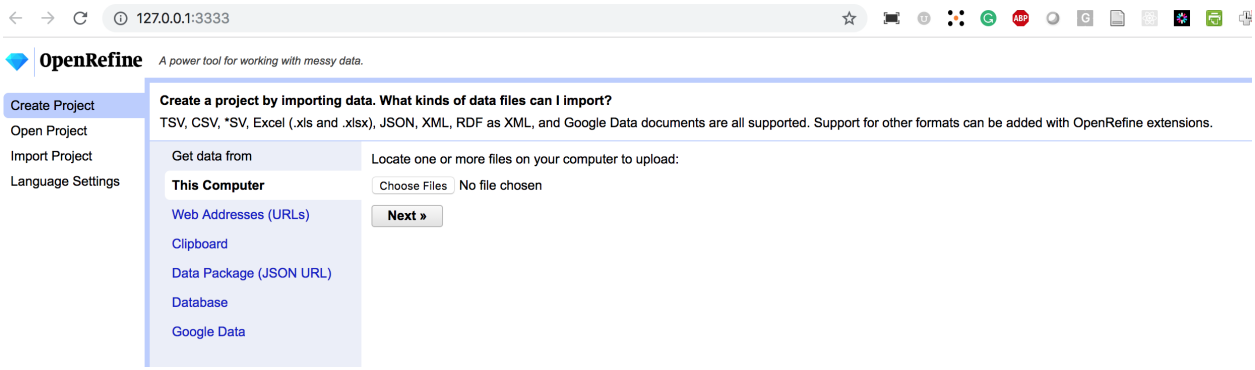


   b. If you receive the above error, go to your System Preferences and click on Security & Privacy.
   c. Click on the General tab and click on Open Anyway:

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

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

Installation & Setup

1. **Download FileZilla Client**
    a. Navigate to https://filezilla-project.org
    b. Click the grey "Download FileZilla Client" button (**not** FileZilla Server)
    c. Click the green "Download FileZilla Client" button on the next page to download the client for your given operating system
    d. If a pop-up is presented, click the green "Download" button for FileZilla (**not** FileZilla Pro)
2. **Install FileZilla Client**
    a. Once the program has downloaded, run the executable
    b. 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**
    Screenshots can be found here: https://2fa.rcc.uchicago.edu/getstarted.html > FileZilla tab
    a. Open FileZilla
    b. Add a new Site
        - 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 Midway
        - Input the following parameters:
            - Protocol: SFTP
            - Host: midway2.rcc.uchicago.edu
            - Port: 22
            - Logon Type: Interactive (you will need to use 2FA)
            - User: Your CNET ID
        - Select Transfer Settings tab and check the "Limit number…" and type 1
        - Click OK to save the settings
    c. Navigate to Edit > Settings
        - From the Settings menu, select Transfers from the left window pane
        - Input 1 in the "Maximum simultaneous transfers" box

d. Connect to your new saved connection
    - File > Site Manager
    - Select your saved connection on the left (Midway)
    - Click Connect
e. If prompted with a dialogue about an “Unknown host key”, check the box to “Always trust this host” and click OK.
f. Type your password in the first prompt
g. Select your 2FA method in the second prompt and complete the 2FA process
h. Confirm you see your file tree on the right side to indicate you successfully connected

Anaconda

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

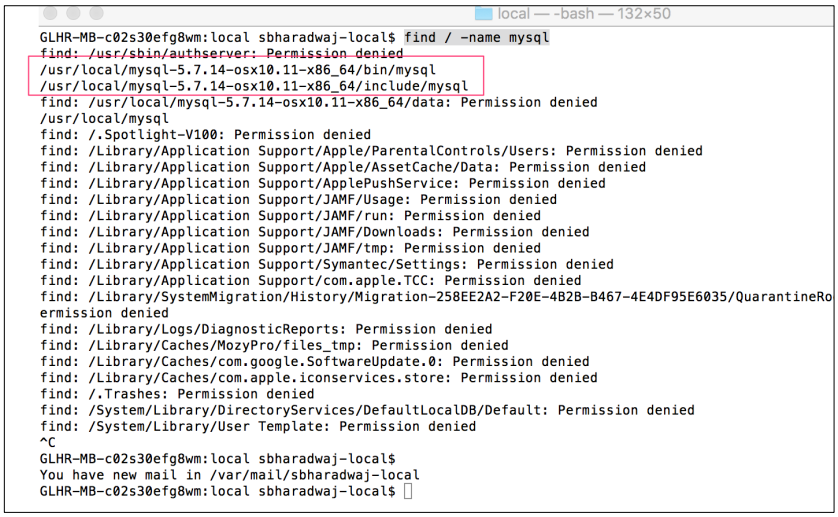
1. Download Anaconda for Python 3.7: https://www.anaconda.com/download/#macos (Anaconda 3.7 is the newest version - the screenshot below is old)

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.

MySQL Installation

Installation Instructions

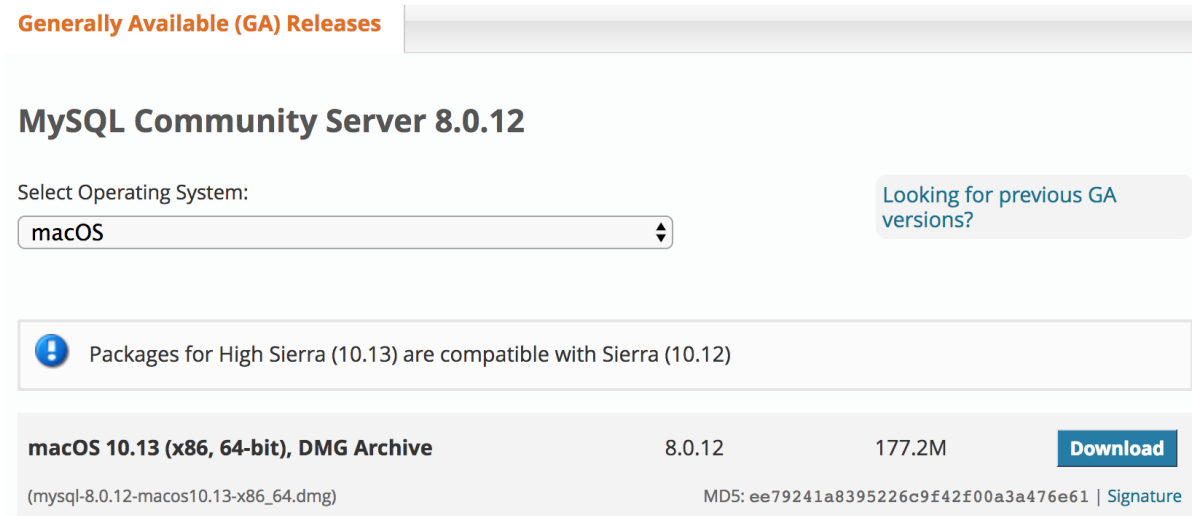
1. Before starting the MySQL installation make sure to uninstall any pre-existing installations of MySQL database and install the latest available version.
    a. Start Terminal and run the following commands
        - `find / -name mysql`



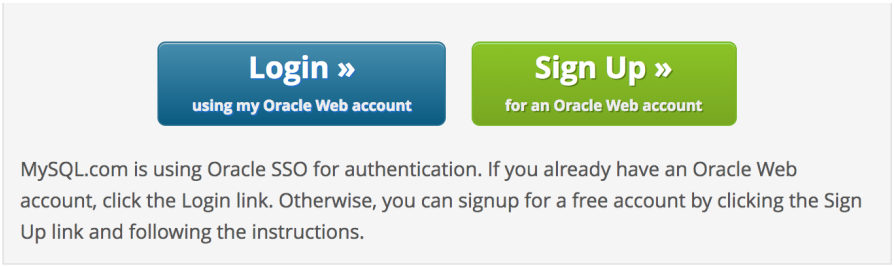
        - Delete all references and the mysql-<xxx> directory from /usr/local directory
        - Move MySQL workbench to trash

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.



No thanks, just start my download.

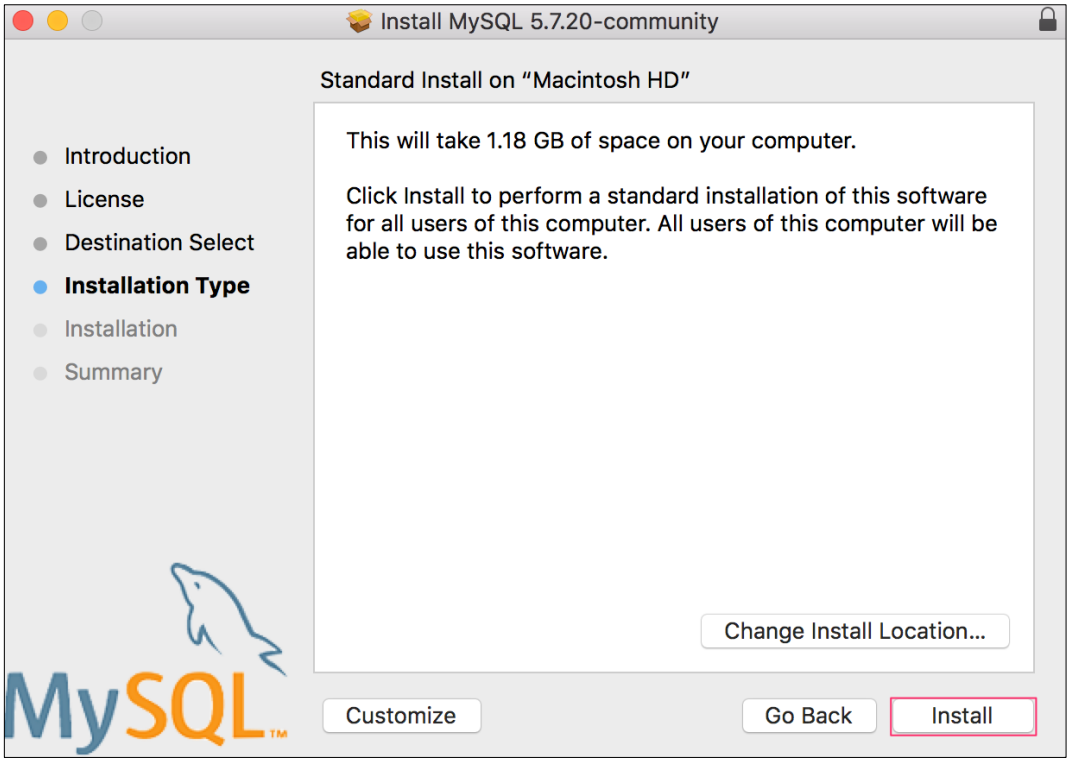
5. Save the file to a folder on your machine. This will prevent you from having to download it again if something goes wrong with the install.
6. Double click on the pkg to open up the installer and follow the steps below.
    a. Click Continue on the 'The package will run a program to determine if the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

    b. Click Continue on the Installation welcome prompt.

    c. Click continue on the License Agreement

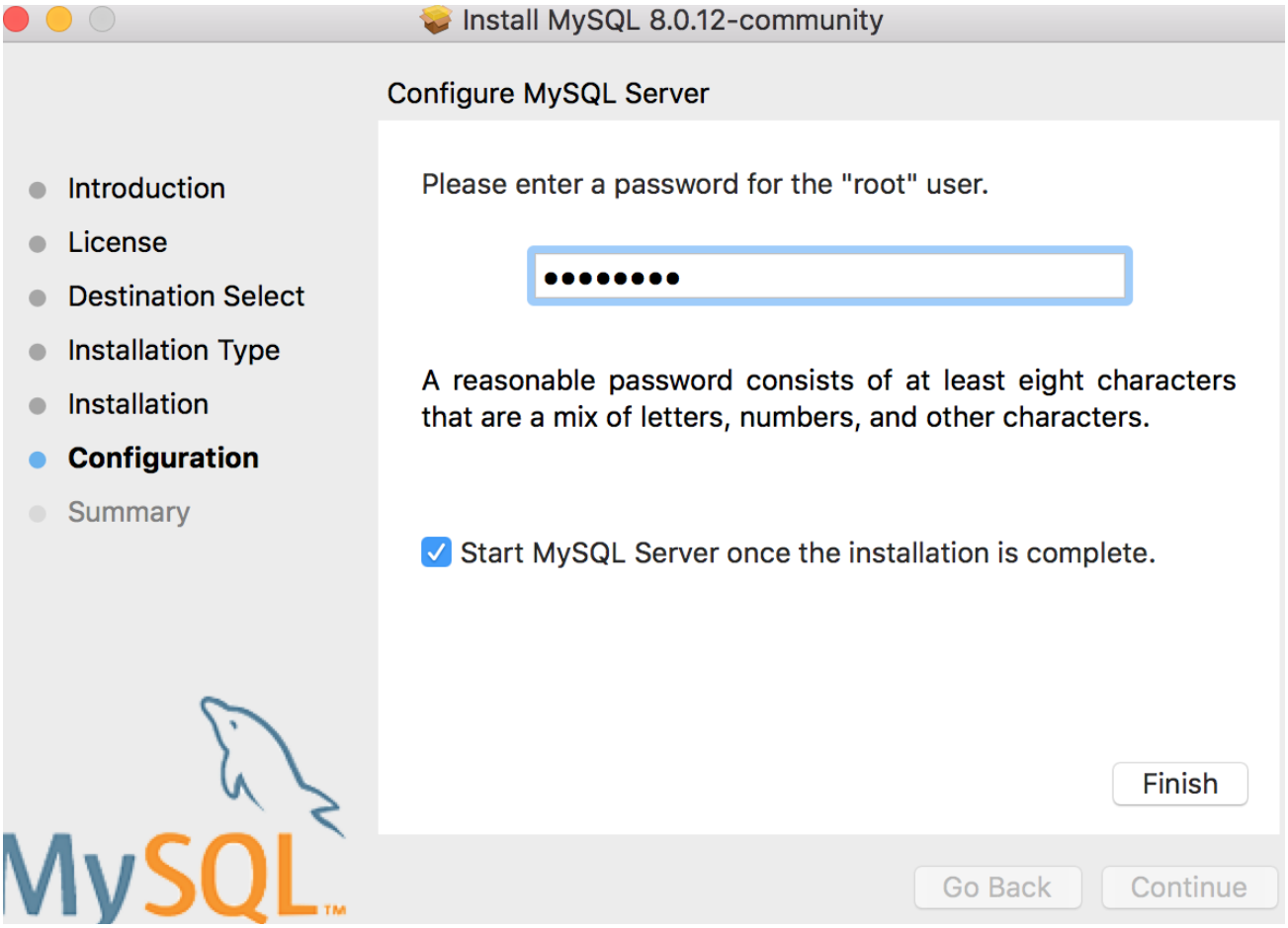
    d. Agree to the license agreement

    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 '**rootroot**' as your password. Using this common password will help with in-class scripts. Then click Finish.



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

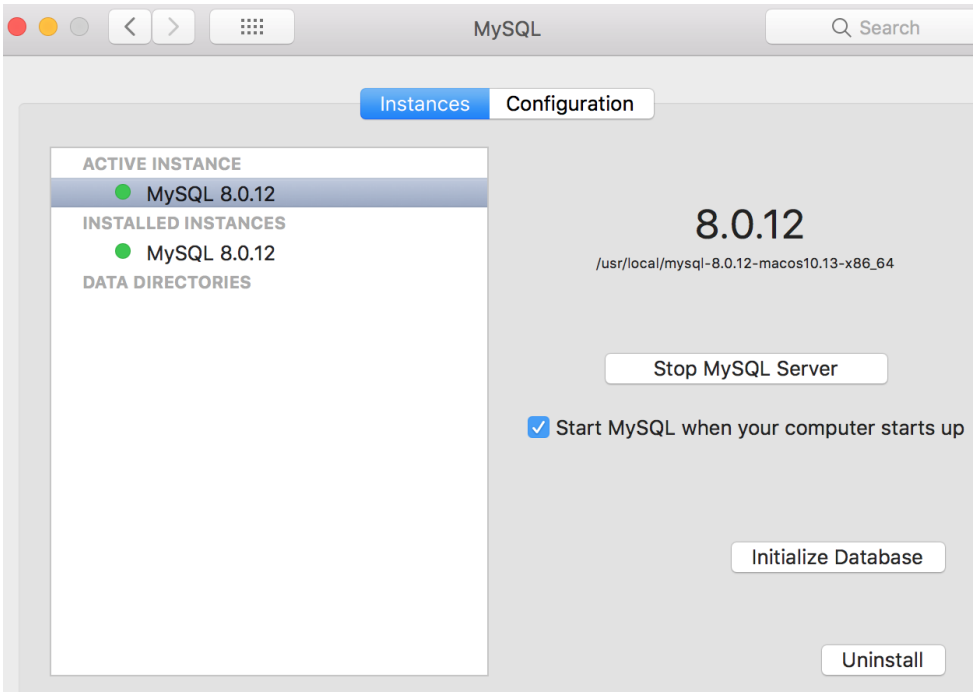
```
basedir=
datadir=
```

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/

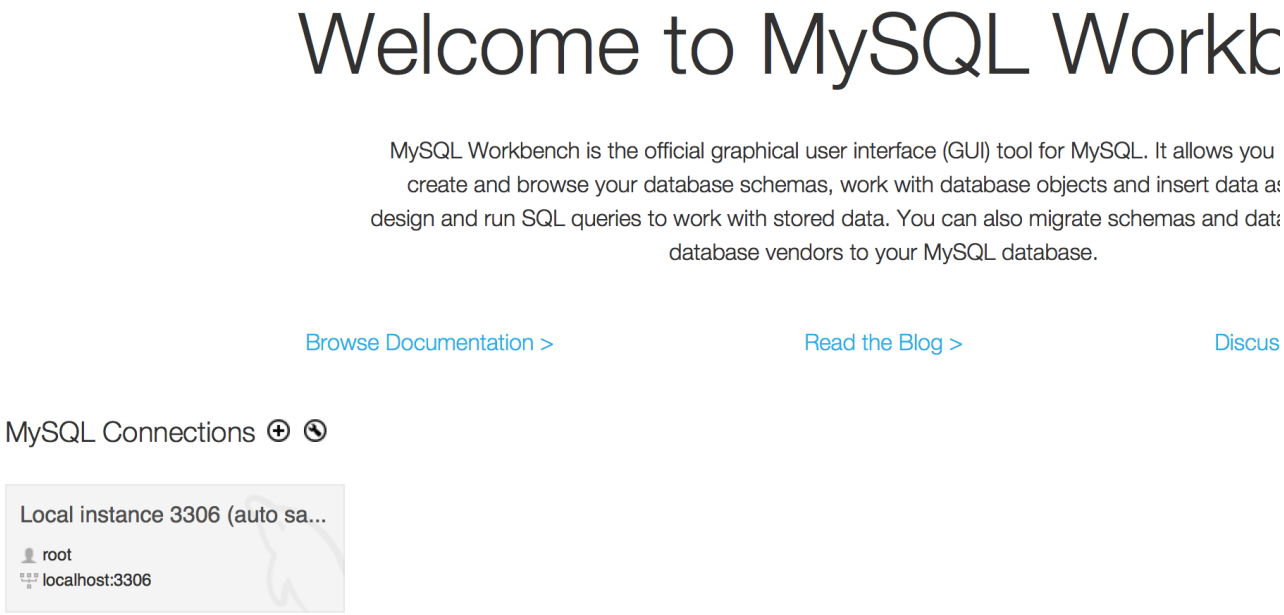
15. Click “No thanks, just start my download” at the bottom

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

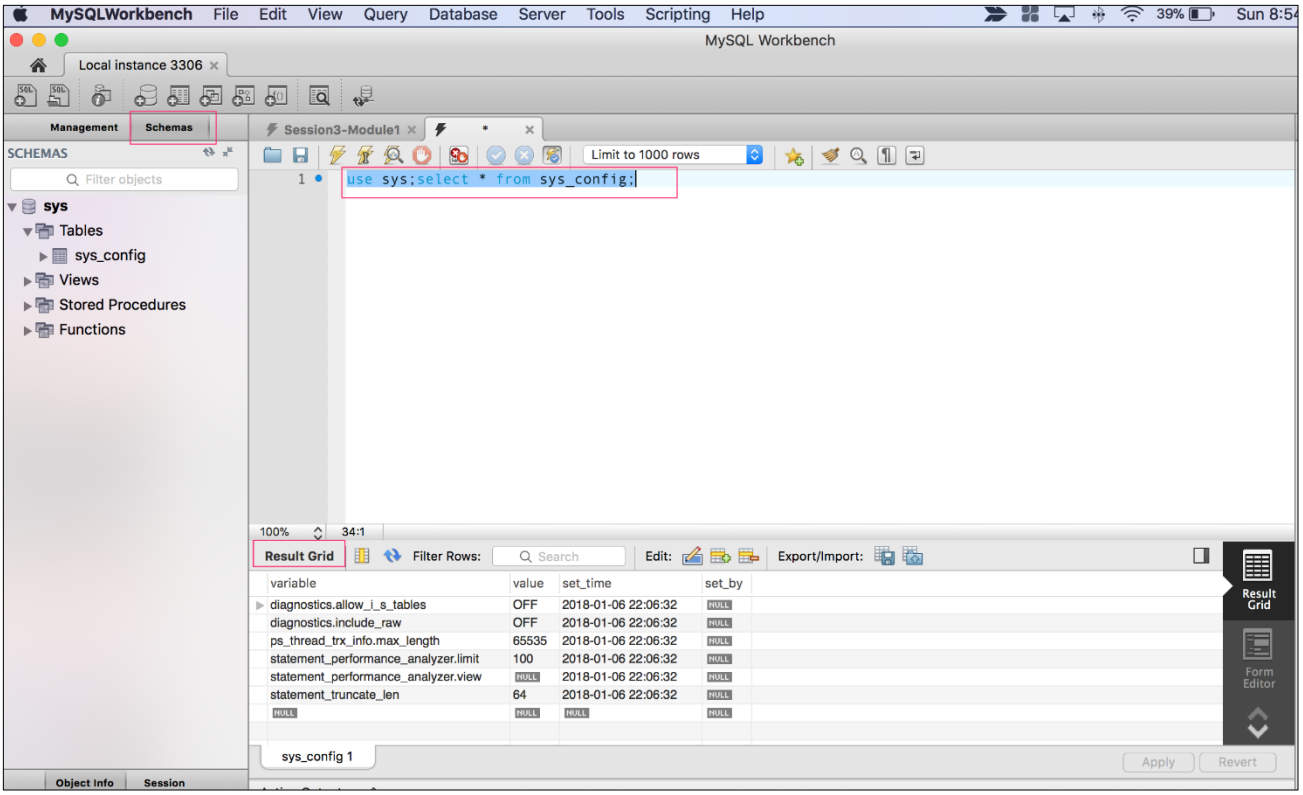
16. Once downloaded, drag to install and authenticate.

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.



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.

Install MongoDB using Homebrew

1. Follow the instructions for installing via Homebrew found here:
   https://docs.mongodb.com/manual/tutorial/install-mongodb-on-os-x/
   NOTE: Make sure your homebrew is up to date.

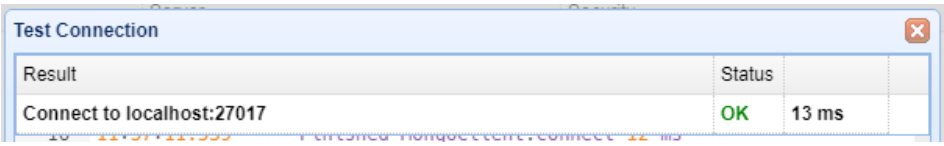
Install NoSQLBooster Client & Connect

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

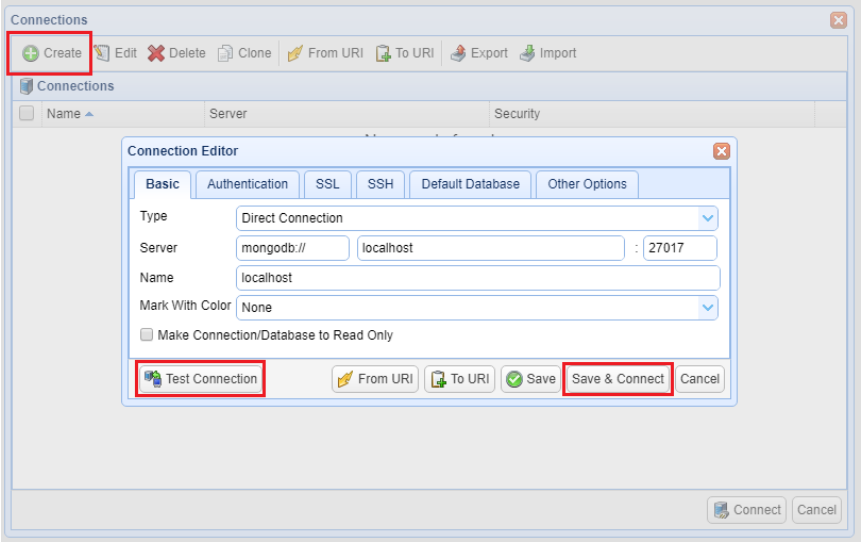
2. Make sure MongoDB is running as a MacOS service (if you haven't completed this step already from above), using the homebrew command found here:
   https://docs.mongodb.com/manual/tutorial/install-mongodb-on-os-x/#run-mongodb

3. Open up the NoSQLBooster application.

4. Connect to your MongoDB Server
   a. Click "Create"
   b. The server parameters should auto-fill
   c. Click Test Connection. Verify the Status is OK, then click "Close"



   NOTE: You may see some warnings. Those are OK, as long as the Status says OK in green.
   d. Click "Save & Connect"



   e. You are now connected to MongoDB in NoSQLBooster

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.