

**Assignment 0 : Study/Review of SQL programming for data warehousing, Python for data mining / machine learning**

**Date : 07-09-2020**

**Submitted By : DWDM20G05**

**Objective :**

- Download / Install latest MySQL DB with workbench, study the data modeling for the warehouse.
- Download and Installation of Python, Demonstration of Python UI for data warehousing and data mining.
- Study of different machine learning / data mining libraries in Python.

**Introduction :**

MySQL, the most popular Open Source SQL database management system, is developed, distributed, and supported by Oracle Corporation. MySQL Workbench is a visual database design tool that integrates SQL development, administration, database design, creation and maintenance into a single integrated development environment for the MySQL database system.

Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. For decision making we have various libraries available in python for example pandas, Apache PySpark & Apache Airflow.

**Theory/Algorithms :**

Coding ETL processes in Python can take many forms, depending on technical requirements, business objectives, which libraries existing tools are compatible with, and how much developers feel they need to work from scratch. Python's strengths lie in working with indexed data structures and dictionaries, which are important in ETL operations.

Python is versatile enough that users can code almost any ETL process with native data structures. For example, filtering null values out of a list is easy with some help from the built-in Python math module:

```
import math data = [1.0, 3.0, 6.5, float('NaN'), 40.0, float('NaN')] filtered = [] for value in data: if not math.isnan(value): filtered.append(value)
```

Users can also take advantage of list comprehensions for the same purpose:

```
filtered = [value for value in data if not math.isnan(value)]
```

Coding the entire ETL process from scratch isn't particularly efficient, so most ETL code ends up being a mix of pure Python code and externally defined functions or objects, such as those from libraries mentioned above. For instance, users can employ pandas to filter an entire DataFrame of rows containing nulls:

```
filtered = data.dropna()
```

Python software development kits (SDK), application programming interfaces (API), and other utilities are available for many platforms, some of which may be useful in coding for ETL. For example, the Anaconda platform is a Python distribution of modules and libraries relevant for working with data. It includes its own package manager and cloud hosting for sharing code notebooks and Python environments.

Much of the advice relevant for generally coding in Python also applies to programming for ETL. For example, the code should be “Pythonic” — which means programmers should follow some language-specific guidelines that make scripts concise and legible and represent the programmer's intentions. Documentation is also important, as well as good package management and watching out for dependencies.

## Machine learning Algorithms / Libraries for Data Mining

1. **K-nearest neighbors** :The K-nearest neighbors (KNN) algorithm is a type of supervised machine learning algorithm. KNN is extremely easy to implement in its most basic form, and yet performs quite complex classification tasks. It is a lazy learning algorithm since it doesn't have a specialized training phase. Rather, it uses all of the data for training while classifying a new data point or instance. KNN is a non-parametric learning algorithm, which means that it doesn't assume anything about the underlying data. This is an extremely useful feature since most of the real world data doesn't really follow any theoretical assumption e.g. linear-separability, uniform distribution, etc.
2. **K means clustering** : To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.

- The defined number of iterations has been achieved.

### 3. ANN Algorithm

This is the type of computer architecture inspired by biological neural networks. They are used to approximate functions. That can depend on a large number of inputs and are generally unknown.

They are presented as systems of interconnected “neurons”. That can compute values from inputs. Also, they are capable of machine learning as well as pattern recognition. Due to their adaptive nature.

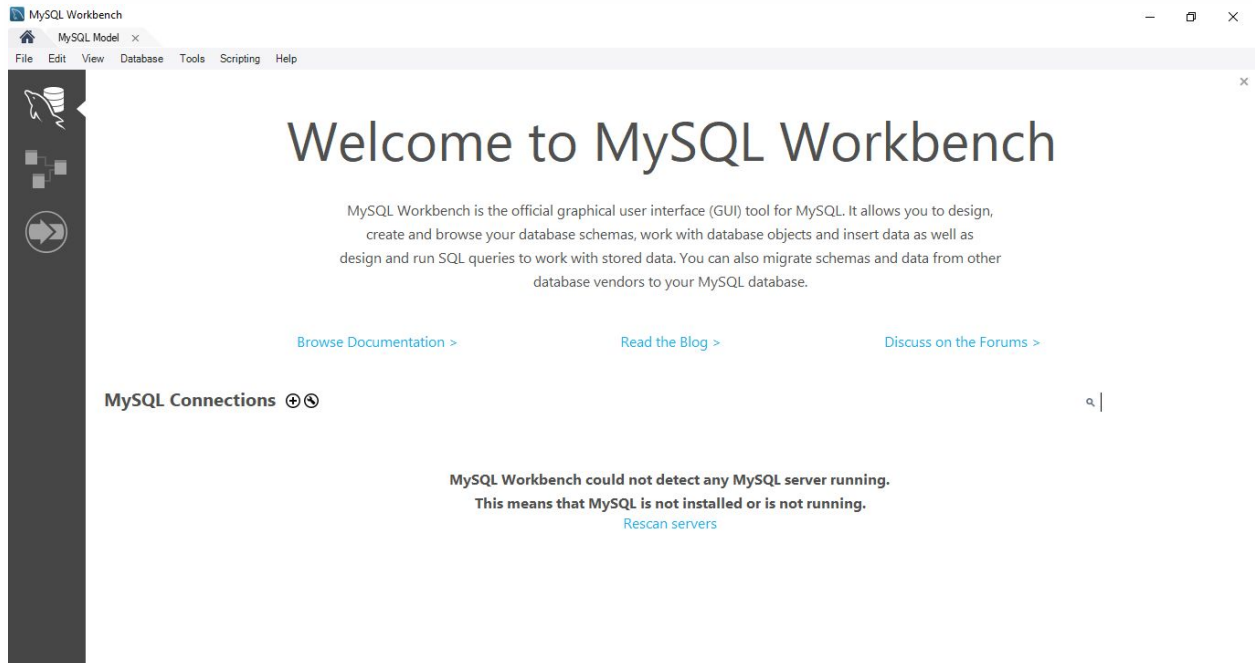
An artificial neural network operates by creating connections between many different processing elements. Each corresponding to a single neuron in a biological brain. These neurons may actually be constructed or simulated by a digital computer system. Each neuron takes many input signals. Then based on an internal weighting. That produces a single output signal that is sent as input to another neuron.

## Procedure for Installing Software Tools

- **Mysql WorkBench Community Version 8.0.21 Installation**

- 1) Visit the website <https://dev.mysql.com/downloads/workbench/> & select operating system where you want to install workbench.
- 2) Then download the .zip/.msi.
- 3) Double click the installer & click on the next button, please check the prerequisites.
- 4) Also it will alert you prerequisites are not installed, then click next in the next window you will get the download prerequisites button, download and install them.
- 5) Restart the setup.
- 6) Select installation directory, after successful installation start the workbench.

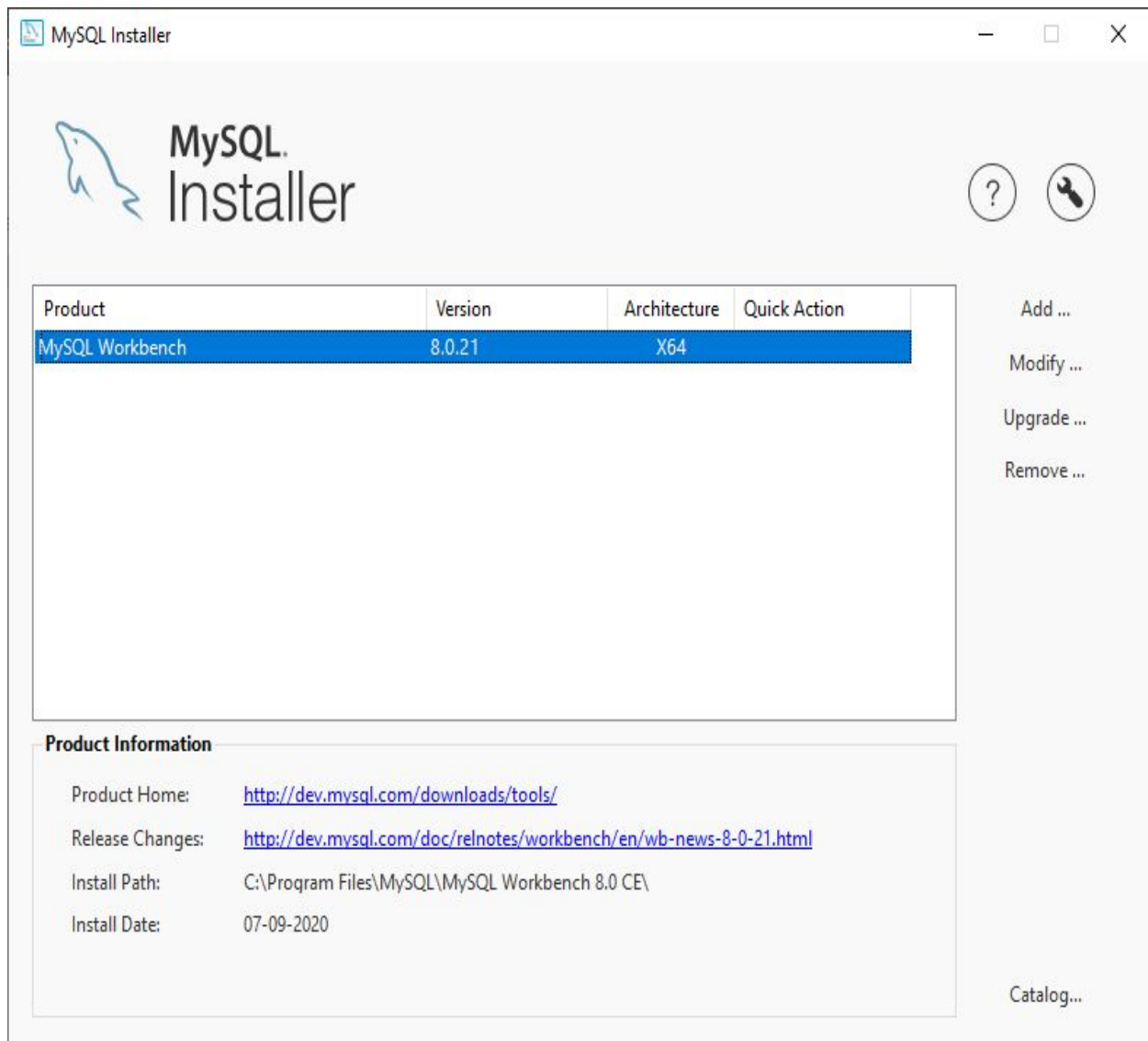
**Observations :**

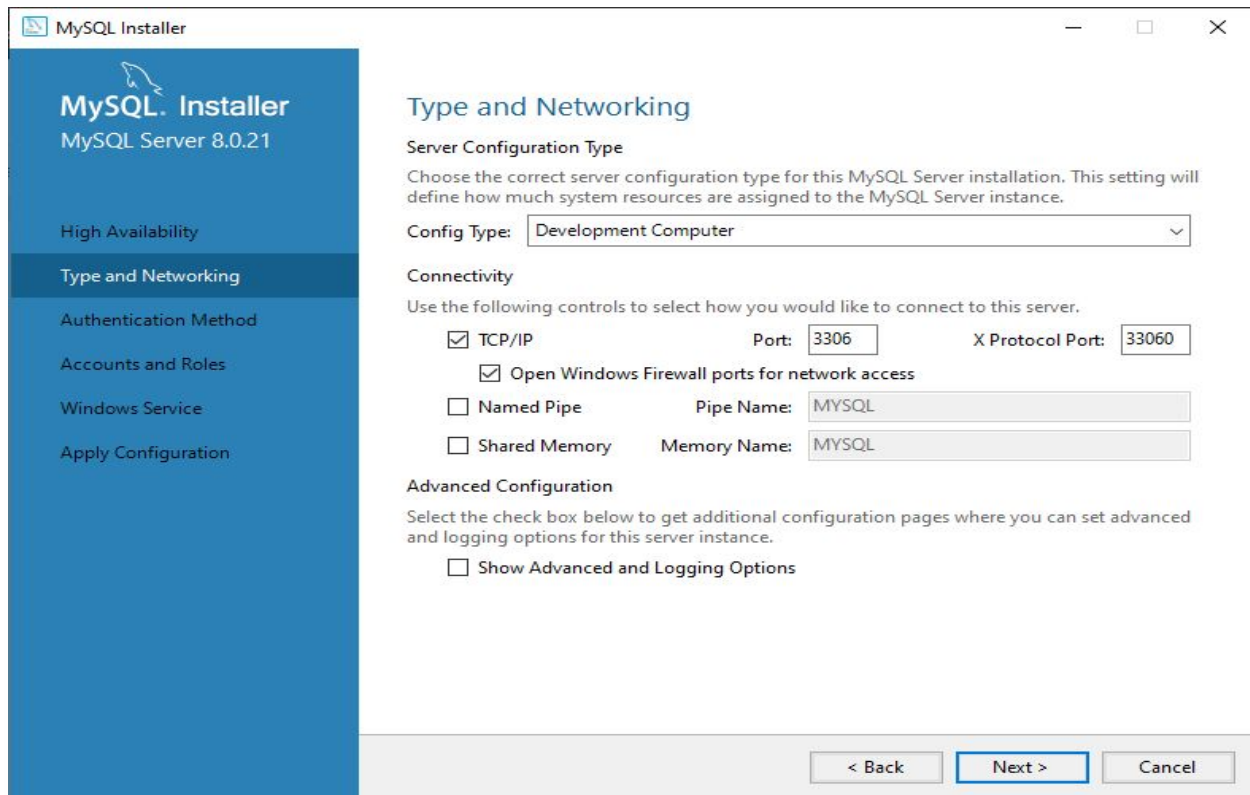
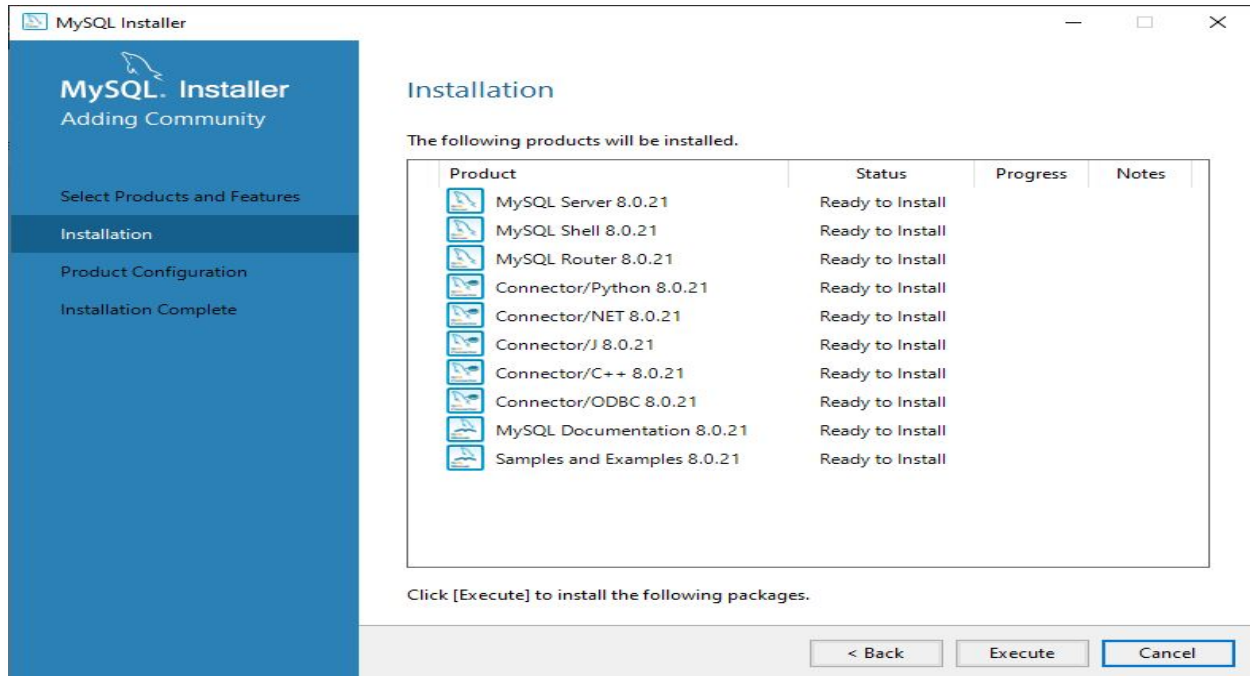


- **MySQL Database Installation**

- 1) <https://dev.mysql.com/downloads/installer/> visit the website to download installer.
- 2) After downloading the installer, run it with admin privileges.
- 3) First window will show the packages required to install the mysql server, mysql shell & driver you needed.
- 4) After complete do the required configure like port no and protocol.
- 5) Configure account, roles and password.
- 6) Click on next and finish the installation.
- 7) Open mysql shell or workbench to connect with the database.

## Observations :





MySQL Installer

MySQL Server 8.0.21

High Availability

Type and Networking

Authentication Method

Accounts and Roles

Windows Service

Apply Configuration

Accounts and Roles

Root Account Password

Enter the password for the root account. Please remember to store this password in a secure place.

MySQL Root Password:

Repeat Password:

MySQL User Accounts

Create MySQL user accounts for your users and applications. Assign a role to the user that consists of a set of privileges.

MySQL User Name	Host	User Role
-----------------	------	-----------

Add User

Edit User

Delete

< Back

Next >

Cancel

Connect to Database

Stored Connection:

Select from saved connection settings

Connection Method:

Standard (TCP/IP)

Method to use to connect to the RDBMS

Parameters

SSL

Advanced

Hostname:

127.0.0.1

Port:

3306

Name or IP address of the server host - and TCP/IP port.

Username:

root

Name of the user to connect with.

Password:

Store in Vault ...

Clear

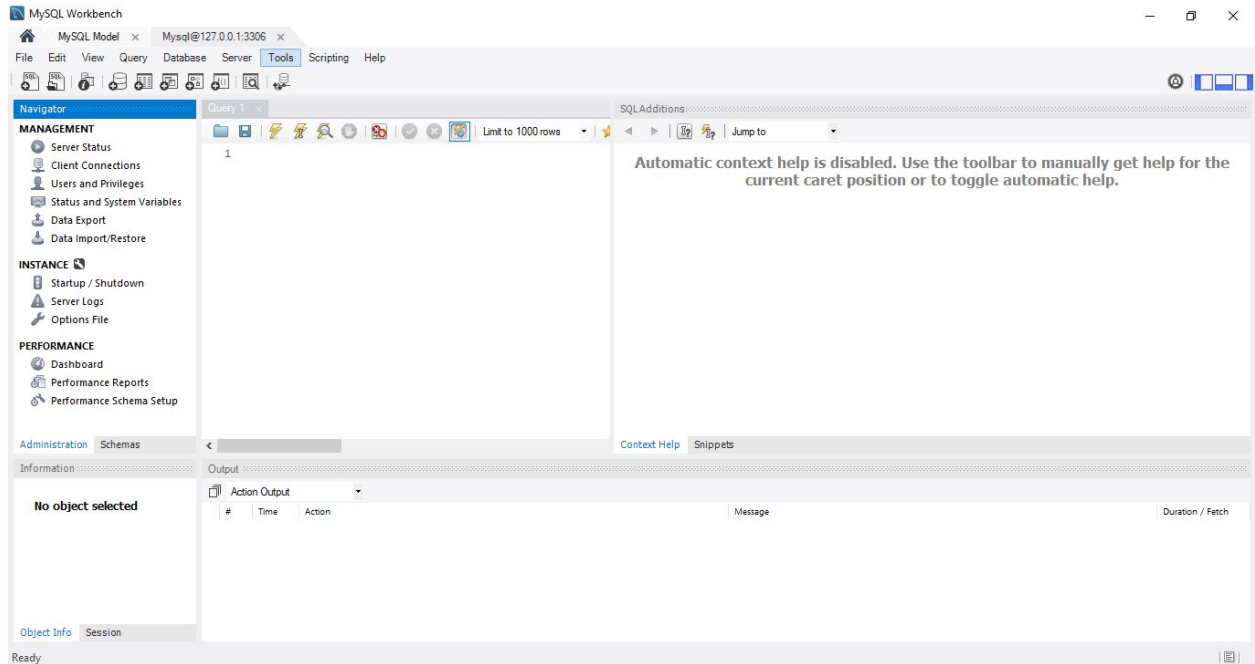
The user's password. Will be requested later if it's not set.

Default Schema:

The schema to use as default schema. Leave blank to select it later.

OK

Cancel



- **Installing python:**

- 1) Goto <https://www.python.org/downloads/>
- 2) Click the **Download for Windows** link and wait for completion of Download.
- 3) Double-click the icon labeling the file **python-3.7.6-amd64.exe**
- 4) Highlight the **Install Now** (or **Upgrade Now**) message, and then click it.

When run, a **User Account Control** pop-up window may appear on your screen. It asks, **Do you want to allow this app to make changes to your device.**
- 5) Click the **Yes** button.

A new **Python 3.7.6 (64-bit) Setup** pop-up window will appear with a **Setup Progress** message and a progress bar.
- 6) During installation, it will show the various components it is installing and move the progress bar towards completion. Soon, a new **Python 3.7.6 (64-bit) Setup** pop-up window will appear with a **Setup was successfully** message.
- 7) Close the window.
- 8) To verify open command prompt and type **python --version**, if successfully installed it will show python version.



## Observations:

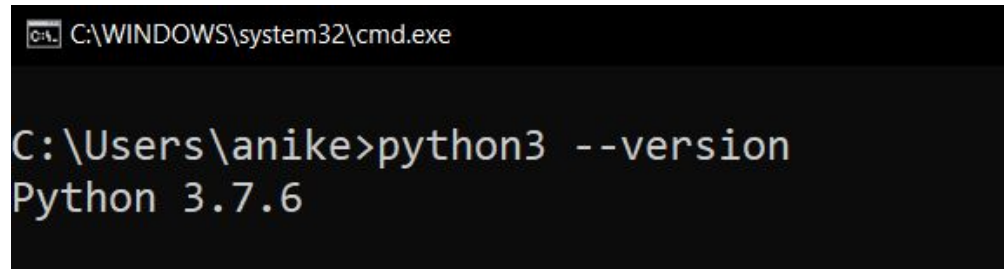
### 1) Download Page for python



### 2) Python installation wizard



### 3) Python version after installation

A screenshot of a Windows command prompt window. The title bar at the top reads "C:\WINDOWS\system32\cmd.exe". The command prompt shows the user's current directory as "C:\Users\anike" and the command "python3 --version" has been entered. The output of the command is "Python 3.7.6".

```
C:\WINDOWS\system32\cmd.exe  
  
C:\Users\anike>python3 --version  
Python 3.7.6
```

### Conclusion:

We have successfully installed & configured python 3.7.4 , mysql 8.0.21 and mysql workbench 8.0.21.