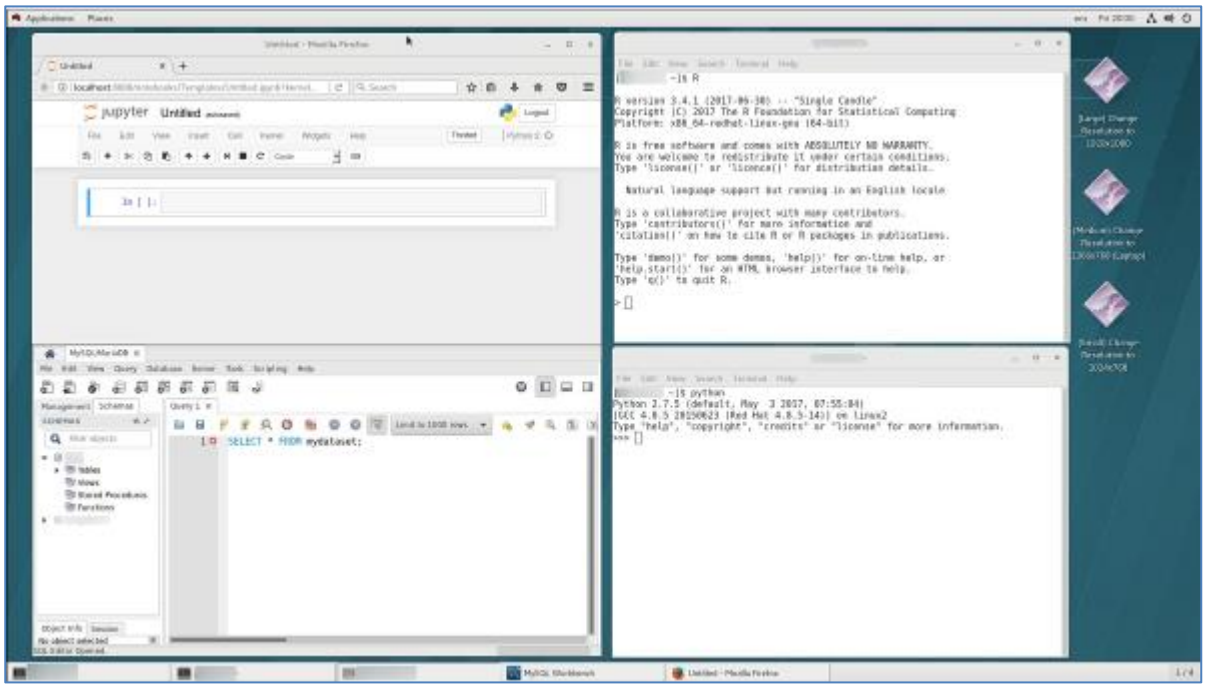
DATA ANALYTICS ENVIRONMENT TUTORIAL

school of continuing studies | YORK UNIVERSITÉ UNIVERSITY

1. DATA ANALYTICS ENVIRONMENT

1.1 OVERVIEW

The data analytics environment is designed to supplement your learning experience by providing you access to a server with the course software pre-installed. Each server is running Red Hat Enterprise Linux (RHEL) 7.4 (Linux), which will familiarize you with an operating system that is employed in many data analytics contexts (CentOS/RHEL). Linux in general provides advantages in terms of scalability, interoperability and reliability when working with big data applications. This environment will save you the work involved with installing these tools and let you dive directly into using them. The use of this environment is not mandatory; the software is available at no charge, some titles are available for Windows, and you have the option of installing the software on your own device or virtual server.



1.2 SOFTWARE

See below for a list of software that is available to users. The majority of the software pre-installed in the data analytics environment is free and open source and available to anyone. Many of the titles are also available for Windows, but are more commonly used in a Linux environment. In one or two cases, students may need to register with a commercial software vendor to obtain free student licenses (e.g. Denodo, RapidMiner) and further instructions will be provided in advance where this is necessary. Please note that some of the software titles, such as Cassandra and MongoDB, may be unavailable until they are introduced as part of the program.

- Jupyter Notebook
- MariaDB (MySQL) Relational Database
- Cassandra NoSQL Database
- MongoDB NoSQL Database
- Python Programming Language
- Rodeo (Python Development Environment)
- R Programming Language
- RStudio (R Development Environment)
- Apache Hadoop
- Apache Hive
- Apache Spark
- Neo4j
- RapidMiner
- ElasticSearch
- Logstash
- DenodoExpress
- Firefox / Chrome
- LibreOffice (open source equivalent to Microsoft Office)

In addition, the following Windows-only software is available in designated computer labs on campus (HNES B02, DB 2027, DB 2032):

- Tableau BI

2. HOW TO CONNECT

In order to connect to the data analytics environment, you will need to complete the following steps:

2.1 Download and install free remote control ("VNC client") and virtual private networking ("VPN client") software (first time only)

2.2 Connect to York University VPN (if connecting from off campus)

2.3 Connect to VNC server with server name, port and username/password

2.1 DOWNLOAD & INSTALL SOFTWARE

The remote control software used to access the server is based on the Virtual Network Computing (VNC) system. There are free VNC clients available for virtually every operating system and device. We recommend one of the following based on your operating system:

Windows 10 8.1 7	TightVNC Viewer	http://www.tightvnc.com/download.php Note: Use custom installation to only install TightVNC Viewer.
	MobaXTerm	http://mobaxterm.mobatek.net/download.html
Apple macOS 10.12 Sierra 10.11 El Capitan 10.10 Yosemite	VNC Viewer (RealVNC)	https://www.realvnc.com/en/connect/download/viewer/macos
	TigerVNC Viewer	https://bintray.com/tigervnc/stable/download_file?file_path=TigerVNC-1.8.0.dmg
Linux Ubuntu, Red Hat, CentOS, Fedora, Arch	TigerVNC	https://bintray.com/tigervnc/stable/tigervnc/1.8.0
	Remmina	https://github.com/FreeRDP/Remmina/wiki
Apple iOS **Android**	VNC Viewer (RealVNC)	https://www.realvnc.com/en/connect/download/viewer/ios/ https://www.realvnc.com/en/connect/download/viewer/android/

If you are connecting from a location off campus, you will also need to download and install the free Pulse Secure VPN application:

Operating System	Where to Get the Pulse Secure Client
Windows 7 and higher	Installation Instructions for Pulse on Windows 32 Bit / 64 Bit Direct Download: Windows 32 Bit | Windows 64 Bit
Mac OS 10.8x and higher	Installation Instructions for Pulse on Mac Direct Download: MacOS
Linux	Download the appropriate Linux clients for CentOS/RHEL or Ubuntu/Debian platforms and use the following instructions.
iOS (iPad / iPhone) 8.1 and higher	Installation Instructions for Pulse on iOS
Android Phone / Tablet 4.4.x and higher	Installation Instructions for Pulse on Android Devices

More information can be found here: http://student.computing.yorku.ca/how-to-connect-securely/.

2.2 CONNECT TO YORK UNIVERSITY VPN

Once the Pulse Secure VPN application is installed, you need to create a new connection profile for the York University VPN. Detailed instructions with screenshots can be found along with the installation instructions in section 2.1. The general steps are as follows:

Configure Profile (first time only)

1. Open the Pulse Secure application
2. Press the plus sign "+" to create a new connection
3. Enter the connection details:
   Name: YorkU
   Server URL: https://vpngateway.yorku.ca/vpnyork
4. Click Add

Connect

1. Click Connect next to the "YorkU" connection profile
2. Enter your Passport York username (e.g. yu308100) and password

Disconnect

1. Click Disconnect next to the "YorkU" connection profile

Important!

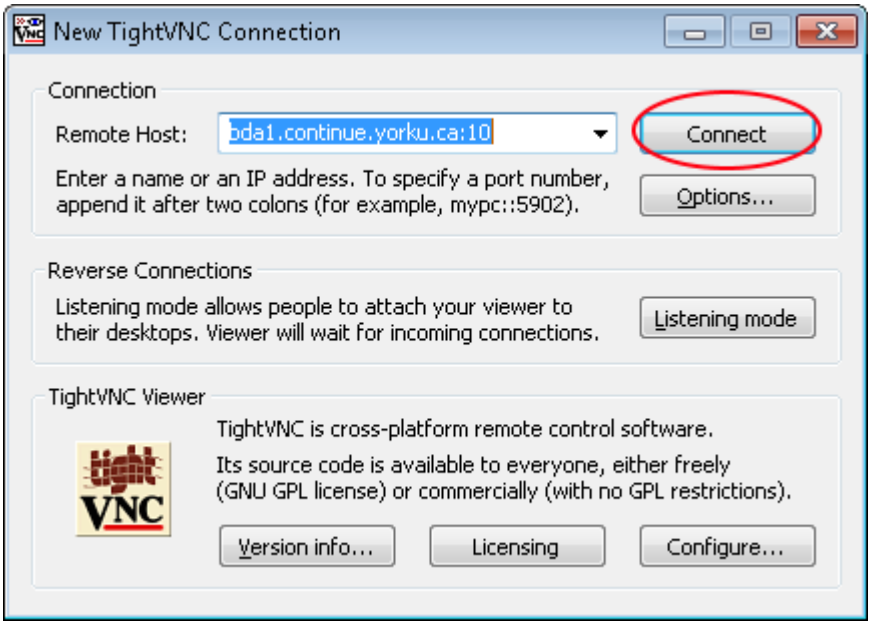
While you are connected to the York University VPN your remote device will be identified as part of the York University network and your internet traffic will be routed through this network. Be sure to disconnect from the VPN when you are not actively working in the data analytics environment.

2.3 CONNECT TO VNC SERVER

Each user has been assigned to a specific server (e.g. bdaX.continue.yorku.ca) and display number (e.g. 4). These two values combine to form the VNC server address. Here are a couple of examples:

Server Name	Display No.	Port No.	VNC Server Address
bda11.continue.yorku.ca	10	5910	bda11.continue.yorku.ca:10
bda12.continue.yorku.ca	4	5904	bda12.continue.yorku.ca:4
bda13.continue.yorku.ca	17	5917	bda13.continue.yorku.ca:17

1. Open the VNC client that you downloaded and installed (e.g. TightVNC Viewer)
2. Enter the VNC server address as the “Remote Host”
3. With some VNC clients, you may also need to enter the port number as above
4. Click Connect / OK / Open to connect to the VNC server
5. Enter your data analytics environment password



3. USING THE ENVIRONMENT

The VNC client will connect you to a persistent desktop environment that is specific to you. You can configure it according to your own preferences. Open windows and applications will remain even if you disconnect and reconnect later.

Some applications will have graphical user interfaces (GUI) and be accessible from the **Applications** menu at the top-left of the screen. Other applications will have command line (i.e. text only) interfaces and must be used through the **Terminal** application.

> You can access the Terminal by selecting Applications -> Favorites -> Terminal or by right-clicking on the desktop and selecting Terminal.

VNC Client Features

Most VNC client software will have the following usability features:

- Toggle between Window mode and Full screen mode (TightVNC: Ctrl + Shift + Alt + F)
- Copy and paste between your local computer and the remote server
- Zoom in and out

Desktop Resolution

You can change the resolution of the remote display to optimize it in relation to your own resolution by using the icons on the desktop, selecting Applications -> System Tools -> Settings -> Displays, or by entering the following command in the Terminal “xrandr -s 1920x1080”, which would change the resolution to 1920 by 1080 pixels

Below are instructions for some of the software applications you will use in the data analytics environment.

3.1 JUPYTER NOTEBOOK

1. Open Terminal (Applications -> Favorites -> Terminal)
2. Enter “jupyter notebook”
3. Wait for Jupyter Notebook to launch in the browser

3.2 MARIADB (MySQL) DATABASE

MariaDB is a community-developed fork of the MySQL relational database management system. It can be used as a “drop-in” replacement for MySQL since it uses the exact same commands. MariaDB (MySQL) can be accessed by command line, or using the graphical tool DBeaver or MySQL Workbench. Your default database is named after your student number (e.g. id301000). You can create and delete additional databases as long as they are prefixed with your username (e.g. id301000_bank, id301000_dataset).

Command Line

1. Open Terminal (Applications -> Favorites -> Terminal)
2. Enter “mysql” to connect to MariaDB
   Note: You will be connected automatically based on the connection profile at ~/.my.cnf.
3. Within the console, you can execute SQL statements (e.g. “SHOW DATABASES;”)
4. Enter “exit” when you are finished

MySQL Workbench

1. Open MySQL Workbench (Applications -> Programming -> MySQL Workbench)
2. Double-click on the “MySQL (MariaDB)” connection
3. Enter SQL statements (e.g. “SHOW DATABASES;”) within the SQL Editor pane
4. Close the application when you are finished

3.3 HADOOP

1. Open Terminal (Applications -> Favorites -> Terminal)
2. Enter “hdfs dfs <cmd>” to manipulate the hadoop file system
3. Please refer to the course material for more information about specific commands
4. Enter “exit” when you are finished

3.4 HIVE, SPARK, ZEPPELIN

- Hive (Applications -> Other -> Hive)
  Terminal Command: hive
- PySpark (Applications -> Other -> PySpark)
  Terminal Command: pyspark
- Spark Scala (Applications -> Other -> Spark Scala)
  Terminal Command: spark-shell
- Zeppelin (Applications -> Other -> Zeppelin)

3.5 PYTHON PROGRAMMING LANGUAGE

1. Open Terminal (Applications -> Favorites -> Terminal)
2. Enter “python”
3. Within the console, you can execute Python code and scripts
4. Enter “exit()” or press Ctrl+D when you are finished

3.6 R STUDIO

1. Open R Studio (Applications -> Other -> R Studio) OR browse to http://localhost:8787 from browser in the VNC (Firefox, Chrome)
2. Login with your username (e.g. id302000) and password
3. Logout and close the browser when you are finished

3.7 R PROGRAMMING LANGUAGE

1. Open Terminal (Applications -> Favorites -> Terminal)
2. Enter “R” (case-sensitive)
3. Within the console, you can execute R code and scripts
4. Enter “q()” or press Ctrl+D when you are finished

3.8 CASSANDRA NoSQL DATABASE

4. Open Terminal (Applications -> Favorites -> Terminal)
5. Enter “cqlsh”
6. Within the console, you can execute CQL statements
7. Enter “exit” when you are finished

3.9 MONGODB NoSQL DATABASE

1. Open Terminal (Applications -> Favorites -> Terminal)
2. Enter “mongo”
3. Within the console, you can execute MongoDB query statements
4. Enter “exit” when you are finished

3.10 LOGSTASH

1. Open Terminal (Applications -> Favorites -> Terminal)
2. Enter “logstash -f /path/to/config --path.data ~/Data/logstash/”
   Important: You must specify a --path.data option as above
3. Press Ctrl+C to close the logstash process if needed
4. Enter “exit” when you are finished

4. SUPPORT CONTACT

School of Continuing Studies IT

itscs@yorku.ca

We can assist with the following issues:

- Can't connect to server
- Software is not working
- Forgot username or password
- General inquiries regarding the data analytics environment