The Data Science Virtual Machine for Linux (DSVM) is an Azure virtual machine image that is pre-installed with a collection of tools and libraries commonly used for doing data science and machine learning. It provides a way for data scientists to quickly create an environment for working on projects without having to first download, install, and configure programs. It also ensures that all team members use the same version of popular tools.

Connecting to your test drive

Using the test drive is easy. Once you start the test drive, a DSVM will be deployed for you, and the VM name, username, and password will be displayed on the page. Log in to your VM via SSH. You can use PuTTY for terminal access on Windows or X2Go for graphical access.

To use X2Go, download it from the homepage and install it. Run the X2Go client and select *New Session*. Enter the following configuration parameters:

- **Host** – the host name of your server
- **Login** – your username on the VM
- **Session Type** – select XFCE, which is the only supported desktop

Then click *OK*. Click on your new session to start it, enter your password, and click *Ok*.

Getting started

The DSVM includes many popular data science tools, including R, python, Jupyter and JupyterHub, Visual Studio Code, and others. Once you log in, you can start a terminal window and run *dsvm-more-info* to learn more about the installed tools.

Some of the key software components included are:

- Microsoft R Open
- Microsoft R Server Developer Edition
- Anaconda Python distribution (v 2.7 and v3.5), including popular data analysis libraries
- Jupyter Notebook (R, Python)
- Azure Storage Explorer
- Azure Command Line for managing Azure resources
- Azure SDK in Java, Python, node.js, Ruby, PHP
- Libraries in R and Python for use in Azure Machine Learning and other Azure services
- Development tools and editors (Visual Studio Code, Atom, emacs, gedit, vi)
- Spark local
- Julia
- And many more

It also includes several machine learning tools:

- The Microsoft Cognitive Toolkit (CNTK)
- TensorFlow
- mxnet
- Torch

- Theano
- Caffe
- Caffe2
- Vowpal Wabbit
- XGBoost
- Rattle (the R Analytical Tool To Learn Easily)

The article Provision the Data Science Virtual Machine for Linux (Ubuntu) also has more information on using some of the installed tools.

After your test drive is complete, any data you stored on the VM will be lost. Be sure to copy it to another location, like Azure storage, before your test drive ends.

Jupyter

The Linux data science virtual machine includes Jupyter, an online environment for creating notebooks that contain live code, text, and visualizations. The DSVM also includes JupyterHub, a multi-user server for Jupyter notebooks. Both R and python are supported in Jupyter.

Many sample notebooks are included on the DSVM. To access them, navigate to https://<your-test-drive-server>:8000/ in your browser, accept the security warning about the self-signed certificate, and log in with your Linux username and password. The sample notebooks will be visible after login. Consider starting with IntroToJupyterPython or IntroTutorialinR.

R

The Linux data science virtual machine (DSVM) includes Microsoft R Open and Microsoft R Server. Many popular R libraries are pre-installed, including ggplot2, rpart, randomForest, xgboost, and others. The AzureML package, to easily publish an R model to Azure Machine Learning, is also pre-installed.

The Azure Machine Learning Data Science repository on GitHub has an R sample to demo some of the DSVM's features. Rattle is also included in the DSVM. Rattle provides a graphical interface to quickly analyze datasets and build predictive models. To run it, start an R session, then type

```
require(rattle)

rattle()
```

The article Data science on the Linux Data Science Virtual Machine has some sample steps to get started with Rattle.

Microsoft R Server Developer edition is also installed. Microsoft R Server supports a variety of big data statistics, predictive modeling and machine learning capabilities. By using and extending open source R, Microsoft R Server is fully compatible with R scripts, functions, and CRAN packages. It also addresses the in-memory limitations of Open Source R by adding parallel and chunked processing of data in Microsoft R Server, enabling users to run analytics on data much bigger than what fits in main memory. Microsoft R Server Developer edition is for development and testing only. Production use requires a production virtual machine or license.

Python

The DSVM includes Anaconda Python, both 2.7 and 3.5. Many popular data science libraries are also installed, including pandas, scikit-learn, scipy, xgboost, and others. The azureml package, to easily publish a python model to Azure Machine Learning, is also pre-installed. The Azure Machine Learning Data Science repository on GitHub has a python sample to demo some of the DSVM's features.

Deep Learning

The Ubuntu DSVM includes many deep learning frameworks, including

- Caffe: A deep learning framework built for speed, expressivity, and modularity
- Caffe2: A cross-platform version of Caffe
- Computational Network Toolkit (CNTK): A deep learning software toolkit from Microsoft Research
- H2O: An open-source big data platform and graphical user interface
- Keras: A high-level neural network API in Python for Theano and TensorFlow
- MXNet: A flexible, efficient deep learning library with many language bindings
- NVIDIA DIGITS: A graphical system that simplifies common deep learning tasks
- TensorFlow: An open-source library for machine intelligence from Google
- Theano: A Python library for defining, optimizing, and efficiently evaluating mathematical expressions involving multi-dimensional arrays
- Torch: A scientific computing framework with wide support for machine learning algorithms

It also includes CUDA, cuDNN, and the NVIDIA driver for running on Azure NC-series GPU instances. Many sample Jupyter notebooks are included.