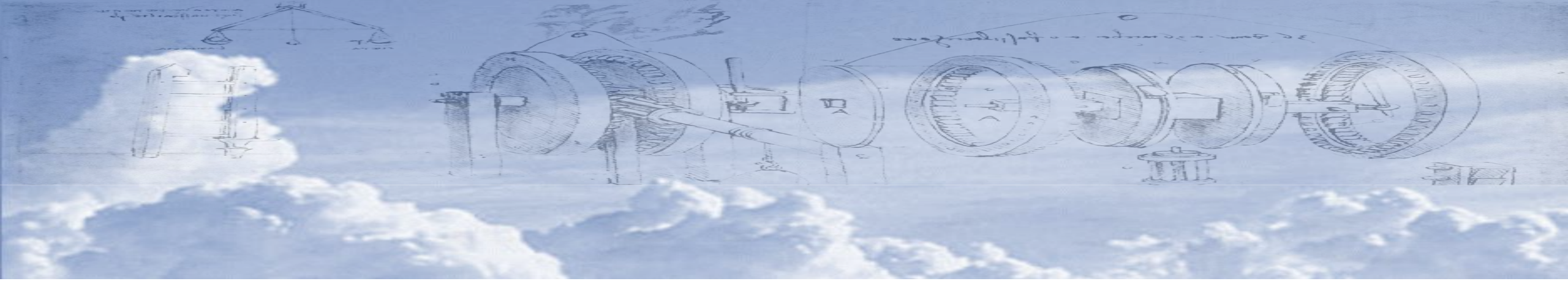


Cloud Computing for Science

Part 2 Virtual Machines and Containers

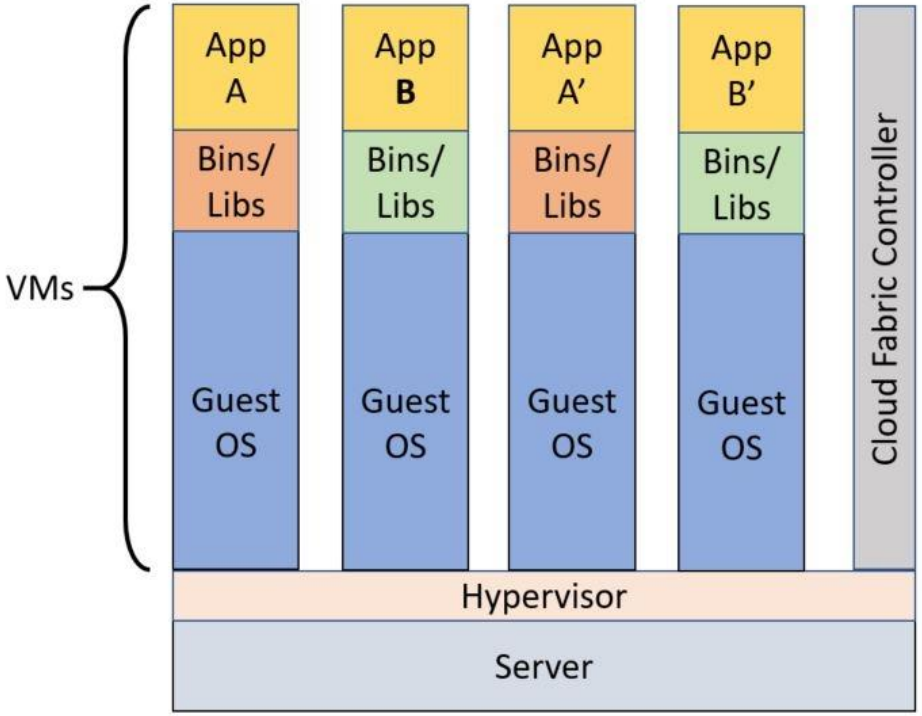
Dennis Gannon

What is a Virtual Machine?

- The foundation of Infrastructure as a Service (IaaS) Clouds
- Operating Systems manage multiple user processes by trapping "privileged" instructions they attempt to execute.
  - If safe to proceed the OS hands the process a "virtual" safe version of the instruction to execute.
- In the 1970 IBM and others figured out how to virtualize the entire computer.
  - A Hypervisor (or virtual machine monitor) is a manager of this virtualization that allows multiple distinct OSs to use the hardware simultaneously.
- This is the key to managing thousands of computer "VMs" for customers
  - A VM is a object that can be managed by a "fabric controller" and virtualized networks.

VM advantages

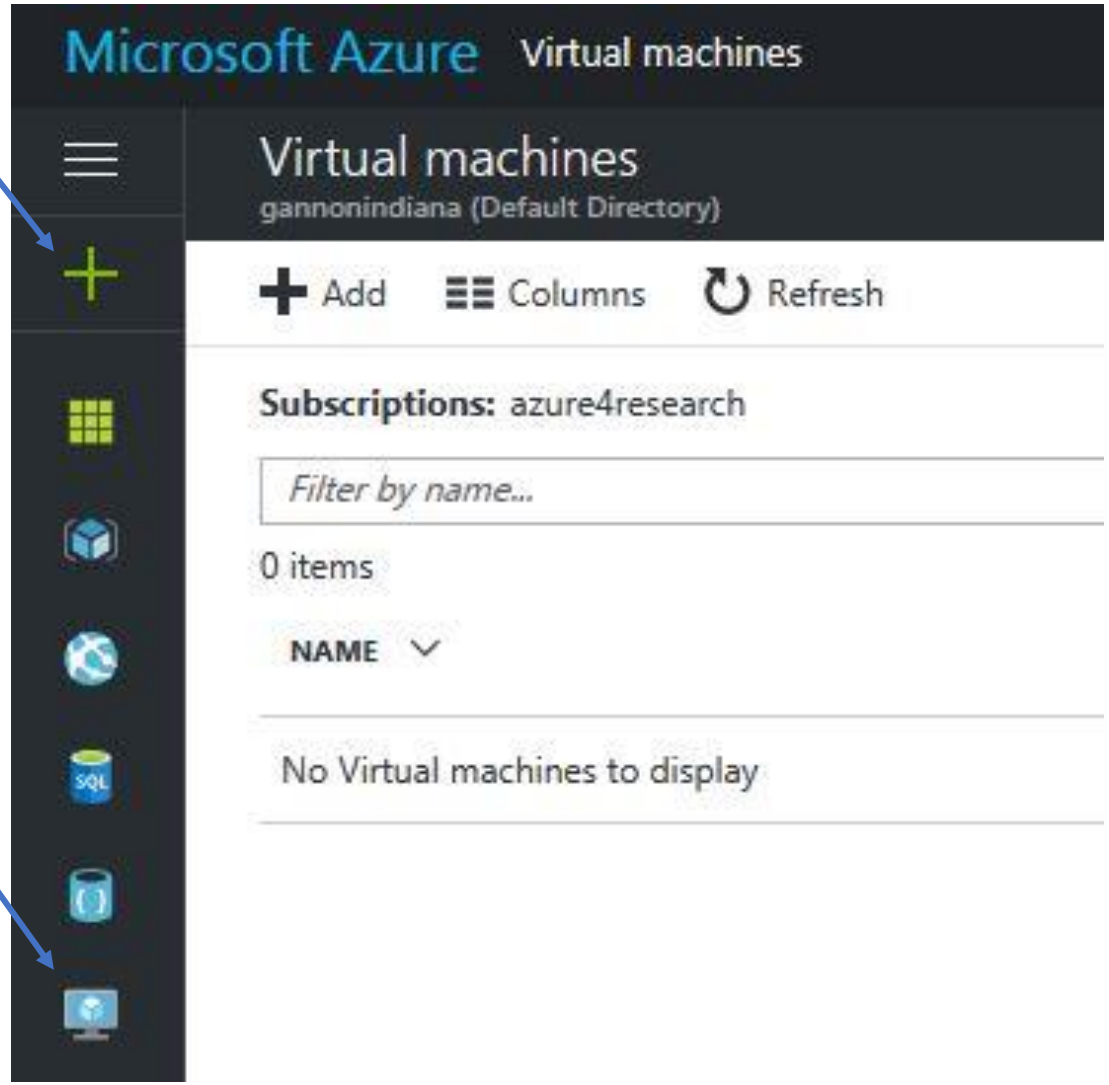
- A VM can completely wrap up an operating system, its file system and applications.
- VMs run on a software layer called a hypervisor which runs on the host server machine.
- You can run multiple VMs on a single host and the “virtual processors” share the host processors, memory and network.
- In the cloud the placement and management of VMs on servers is controlled by cloud software running on the host machines.



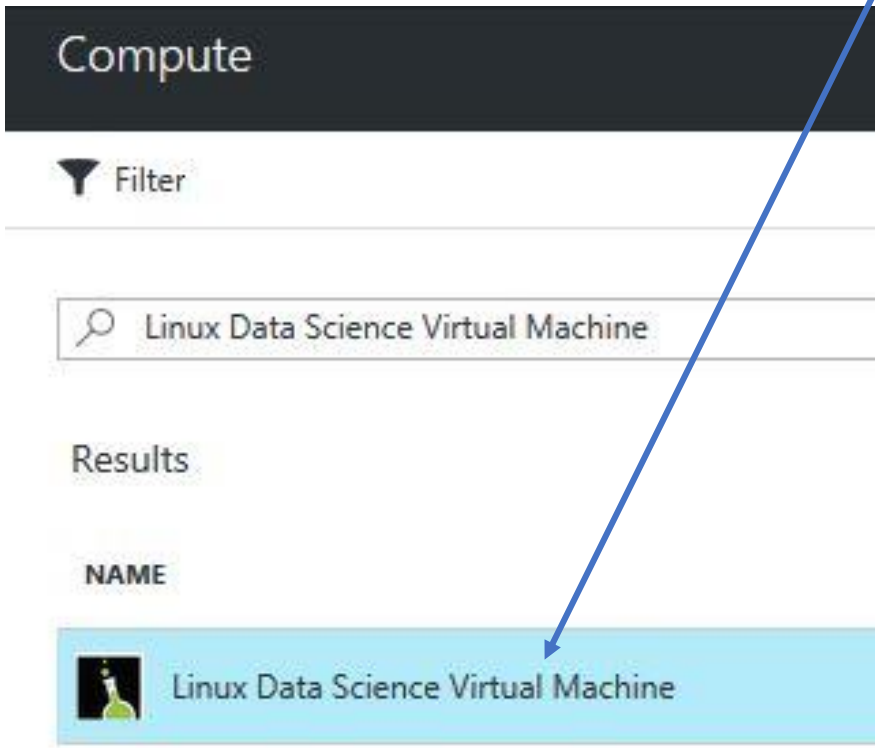
Using the Azure portal to create a VM

Click the plus
When you
Want to add
Something
new

Click this
To bring up
The VM page
Then click
Add



In the search box enter “data science” and you will see this choice. Pick this one and click here



The Linux Data Science VM

This Linux-based virtual machine contains popular tools for data science and development activities, including Microsoft R Open, Anaconda Python, Azure command line tools, and Jupyter notebooks for Python, R and Julia. It also has machine learning tools and algorithms like mxnet, CNTK, Vowpal Wabbit and xgboost.

What's new

- The Linux data science virtual machine now includes Microsoft R Server 9.0, now with Microsoft R Open 3.3.2 and new options for operationalizing R models
- Weka for easy graphical exploration and machine learning
- Apache Drill for querying non-relational data using SQL
- Spark local 2.0.2 with a PySpark Jupyter kernel
- Single node local Hadoop (HDFS, Yarn)
- Visual Studio Code IDEs, IntelliJ IDEA, PyCharm, Atom
- Mxnet, tensorflow and CNTK for deep learning
- JuliaPro - a curated distribution of Julia Language and tools

Also Jupyter hub running on the ip address:8000

You need a public-private RSA key pair

Linux/Mac

>ssh-keygen

Will create a password protected private key and a public key

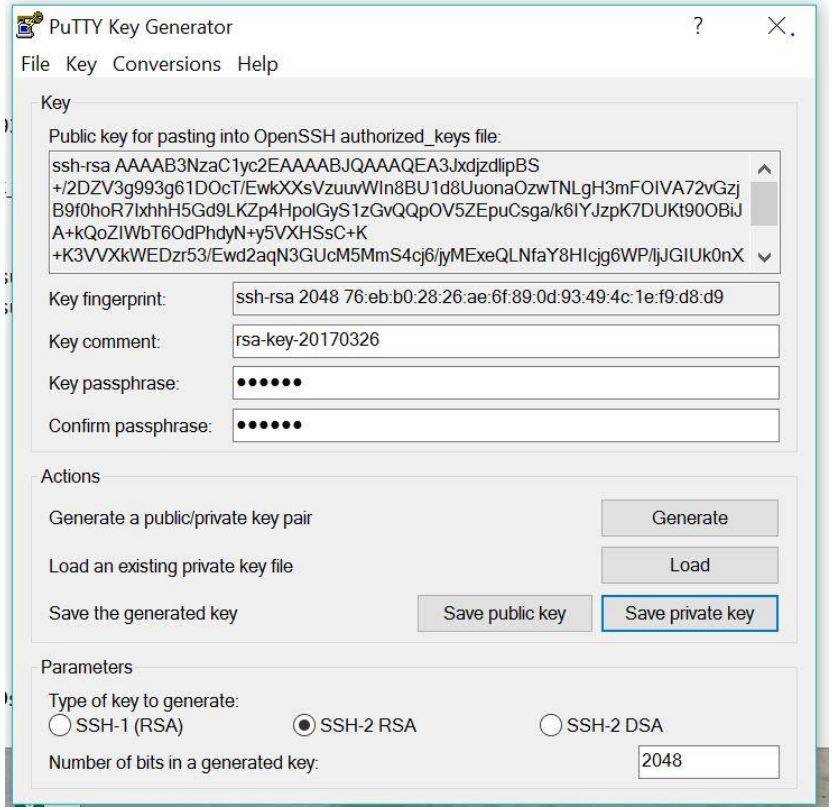
When machine is up, login with

>ssh –i privatekey userid@ipaddress

(it will ask for the password)

Windows

- On Windows10 you can use linux subsystem and follow linux method
- Or Install Putty
- Run PuTTYGen
- Runn Putty to log in.



After you log in you need to set your user password
> sudo passwd userid
>enter your password twice
We will need this later.

Configuring and launching

Create virtual machine

1 Basics
Configure basic settings

2 Size
Choose virtual machine size

3 Settings
Configure optional features

4 Summary
Linux Data Science Virtual Machine

5 Buy

Basics

* Name
myDataScienceVM

VM disk type
SSD

* User name
dbgannon

* Authentication type
SSH public key | Password

* SSH public key
---- BEGIN SSH2 PUBLIC KEY ----
Comment: "rsa-key-20170210"
AAAAB3NzaC1yc2EAAAABJQAAAQEAi+S
oqE+zhRcAt8wsF31YDgpwTQSnVMwQ5c

Subscription
azure4research

* Resource group
Create new | Use existing
bookRG

Location
South Central US

OK

2 Size
Choose virtual machine size

3 Settings
Configure optional features

4 Summary
Linux Data Science Virtual Machine

5 Buy

★ Recommended | View all

DS2_V2 Standard ★	DS3_V2 Standard ★	DS14_V2 Standard ★
2 Cores	4 Cores	16 Cores
7 GB	14 GB	112 GB
4 Data disks	8 Data disks	32 Data disks
6400 Max IOPS	12800 Max IOPS	50000 Max IOPS
14 GB Local SSD	28 GB Local SSD	224 GB Local SSD
Load balancing	Load balancing	Load balancing
Premium disk support	Premium disk support	Premium disk support
94.49 USD/MONTH (ESTIMATED)	189.72 USD/MONTH (ESTIMATED)	989.52 USD/MONTH (ESTIMATED)

Purchase

Offer details

Linux Data Science Virtual Machine — 0.0000 USD/hr
by Microsoft
Terms of use | privacy policy

Standard DS14 v2 — 1.3300 USD/hr
by Microsoft
Pricing for other VM sizes
Terms of use | privacy policy

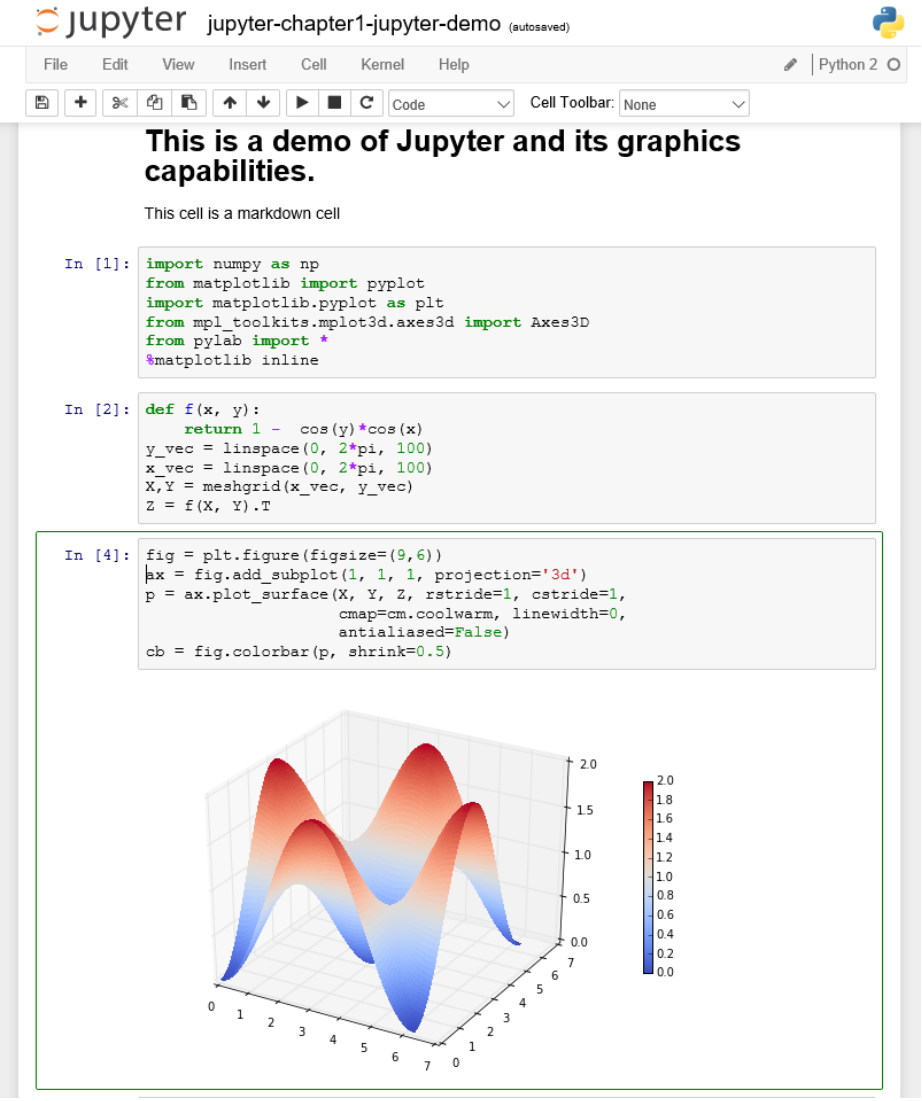
The highlighted Marketplace purchase(s) are not covered by your Azure credits, and will be billed separately.
You cannot use your Azure monetary commitment funds or subscription credits for these purchases. You will be billed separately for marketplace purchases.

Run Jupyter on your new VM

- Jupter Hub is already running.
- To start it you must have set your login password.
  - Login and do
  - > sudo passwd youruserID
- Then go to

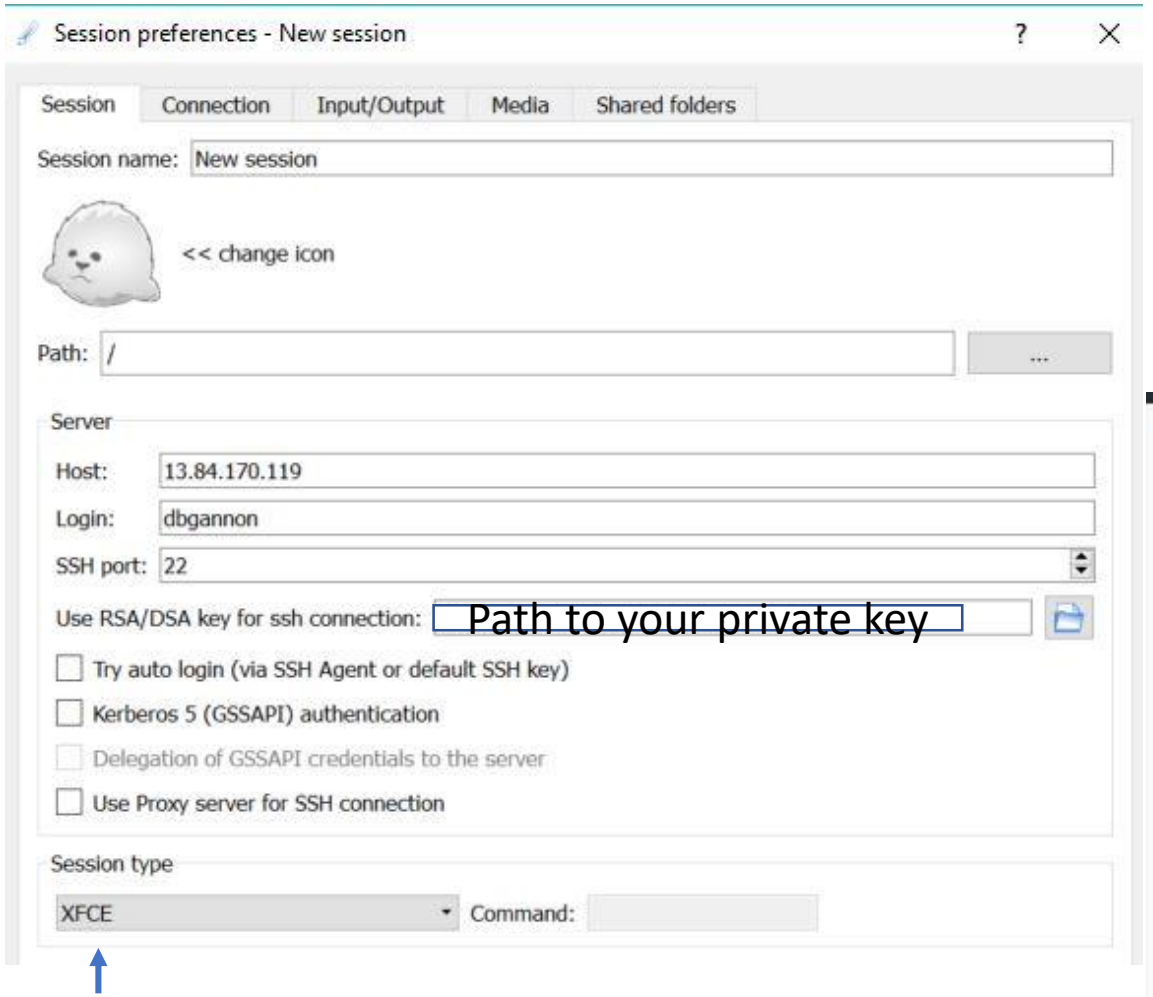
  https://yourvmIP:8000

Login with youruserID and passwd

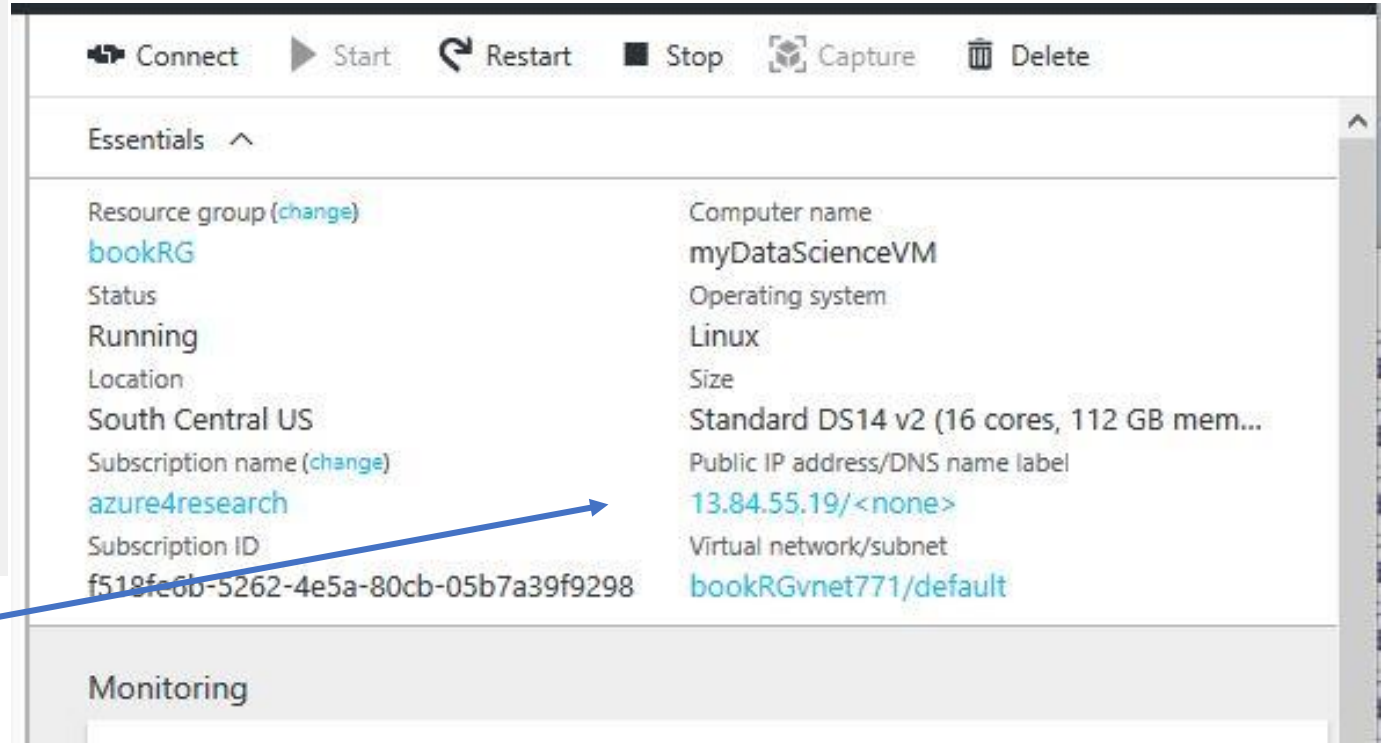


Install X2GO Client

- http://wiki.x2go.org
- A client to show the desktop of the Linux Data Science VM
- When VM up Start X2GO client

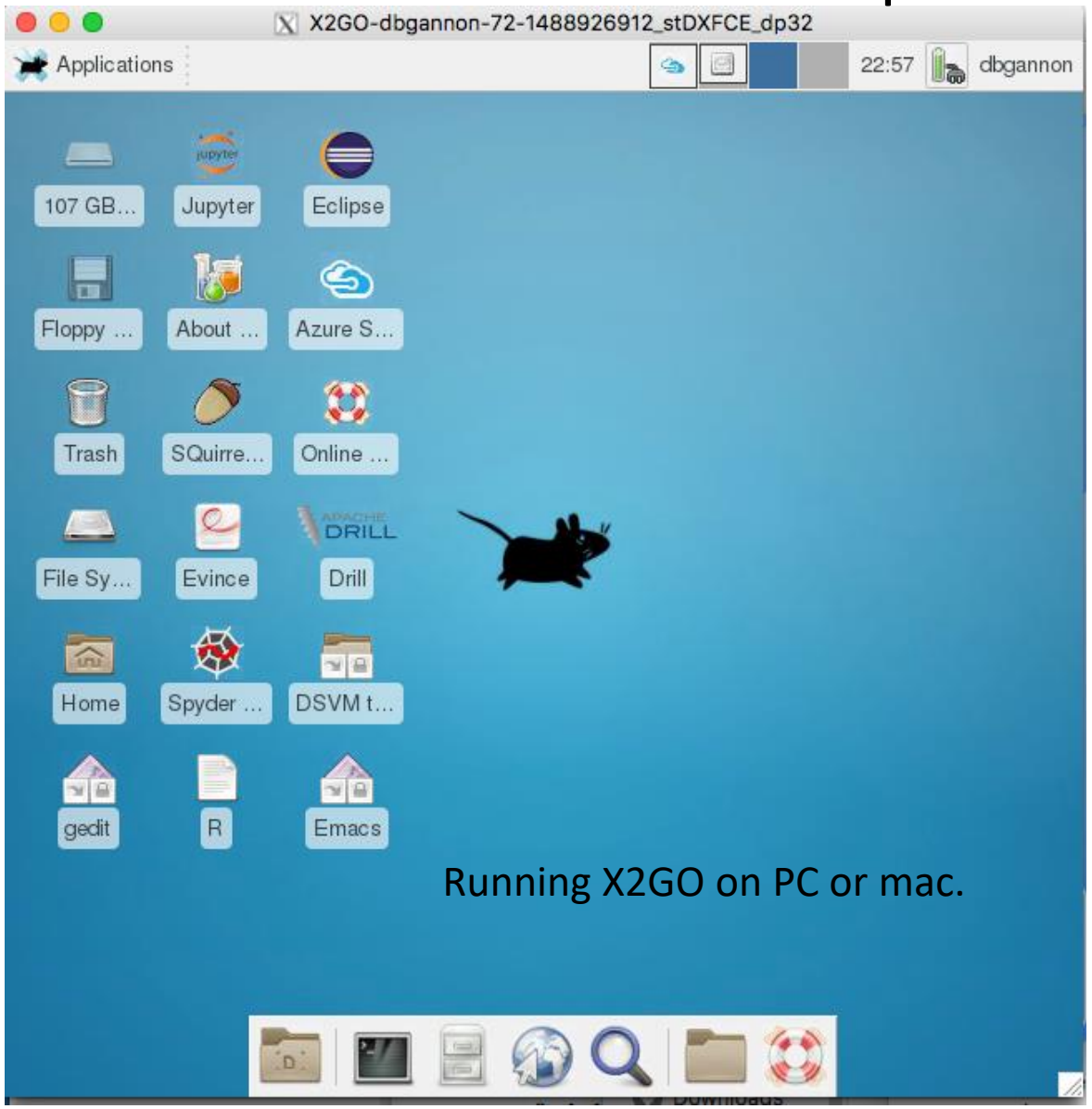


Set to this



Azure portal view and IP address

X2GO XFCE Desktop

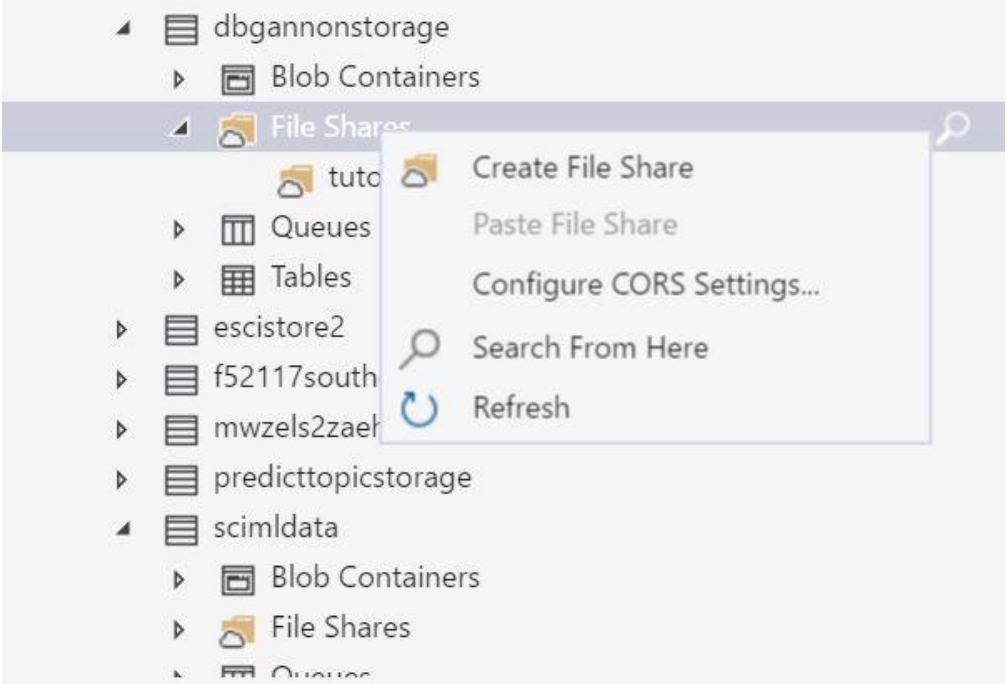


Running X2GO on PC or mac.

- To run jupyter use JupyterHub.
  - First you need to set you linux passwd
    - Open the shell tool
    - >sudo passwd yourID
      - Add the password twice
  - Go to https://yourDSVMip:8000
    - Enter you ID and password
  - Or launch a local Jupyter by clicking on the icon.

Adding a fileshare disk

- This is a disk in blob storage that you can see with AzureExplorer
  - Create a storage account in same location as your DSVM.
  - Create a file share in that account named XXX
  - Grab the storage account key
- Run these commands in your VM



>sudo yum install samba-client samba-common cifs-utils

>sudo mkdir /mnt/tutorialshare

>sudo chmod 0777 /mnt/tutorialshare

>sudo mount -t cifs //yourstorageaccnt.file.core.windows.net/XXX /mnt/tutorialshare -o vers=3.0,user=yourstorageaccnt,password=you accntpasswrd ending in==,dir_mode=0777,file_mode=0777,serverino

>sudo vi /etc/fstab

Add this line at the end

//yourstorageaccnt.file.core.windows.net/XXX /mnt/tutorialshare vers=3.0,user=yourstorageaccnt,password=you accntpasswrd ending in==,dir_mode=0777,file_mode=0777,serverino

Containers

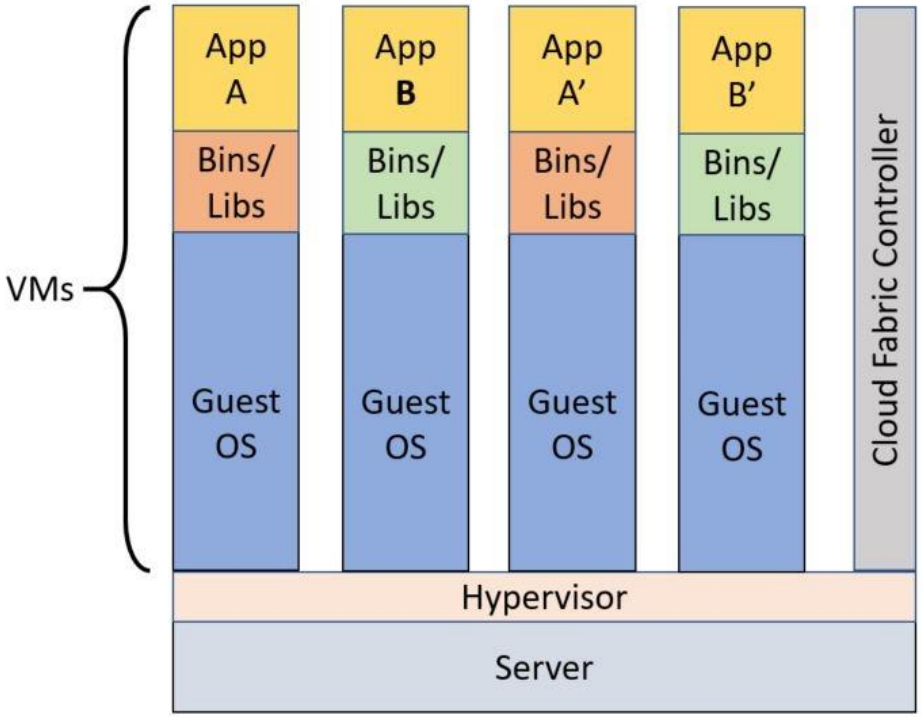
An alternative to virtual machines for encapsulating software

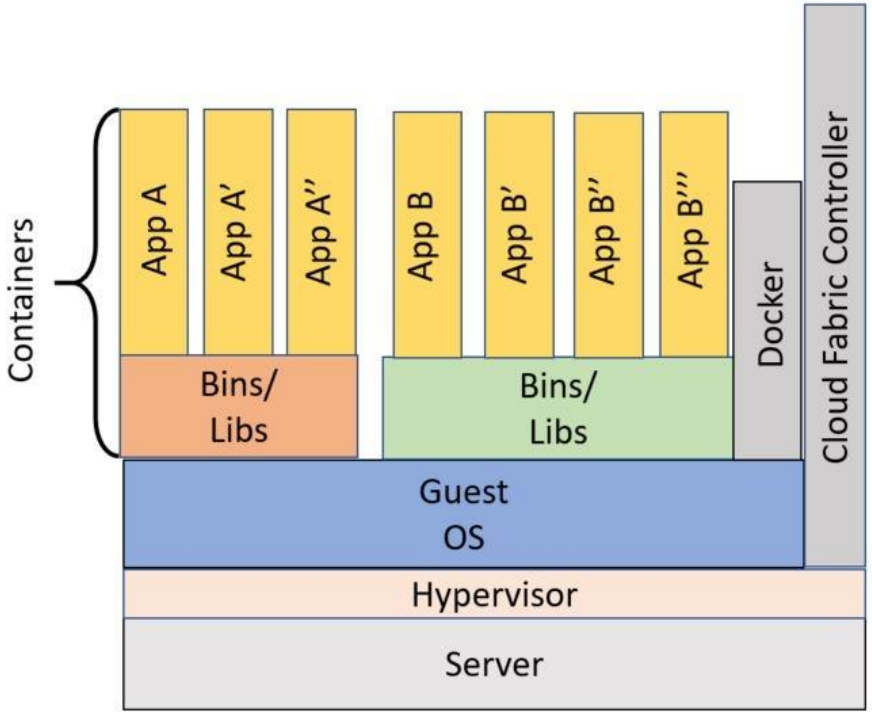
The challenge of software installation and deployment

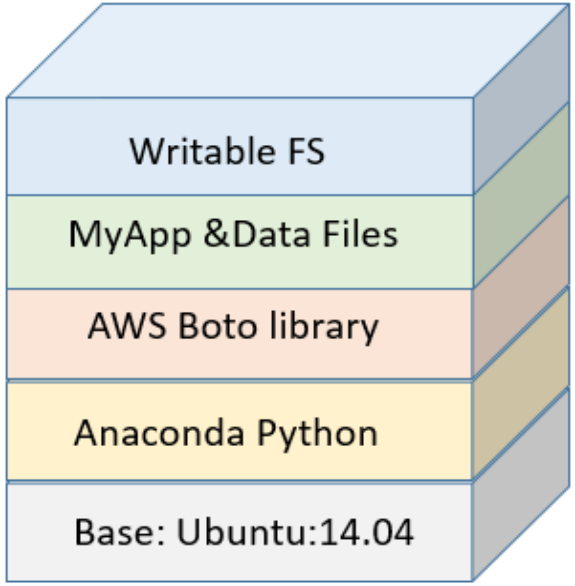
- Software systems are often complex and installation can be error prone
- When multiple packages need to interact conflicts arise
  - Different software libraries
  - Different operating system versions
- Containerization allows systems to be packaged with everything they need so they can run anywhere.
- Many containers can run on a single system

What are VMs? containers?

- A VM is an instance of a complete operating system and file system running on the virtualized hardware.
- Containers share components of the hosts OS and file system and are more similar to a process. (Uses special Linux features “control groups” and “name space isolation” to partition process space and layer new private file system components on the host filesystem.)







Advantages

Virtual machines	Containers
Heavyweight	Lightweight
Fully isolated and hence more secure	Process-level isolation; hence less secure
No automation for configuration	Script-driven configuration
Slow deployment	Rapid deployment
Easy port and IP address mapping	More abstract port and IP mappings
Custom images not portable across clouds	Completely portable

Using Docker to manage containers

- Download Docker for your pc or mac
  - https://docs.docker.com/engine/installation/
- Or login to your Azure data science linux vm
- Type

  sudo docker ps
- Type

  sudo docker run –it ubuntu
- You are now running Ubuntu linix in a container. -it give you the i/o for the shell. Type exit to exit.

Containers as a way to share science

- Lots of our sample jupyter notebooks
- For this tutorial we have a container

  >docker run –it –p 8888:8888 dbgannon/tutorial

  password = tutorial

  - contains lots of our sample jupyter notebooks
- Many other Science containers:
  - Bio - Galaxy and Hamburg genome toolkit
  - Geosciences – geoserver
  - Astronomy – LOFAR, PyImager, MegTree
  - Engineering – Ubercloud project (theubercloud.com)
  - Math – Matlab, Axiom

Creating a container from other containers

- A directory with
  - A “Dockerfile”
  - Things you want in the container
    - A script
    - A directory of data file: datadir
    - A directory of the notebook examples
    - An openssl certificate and key file
- A secure hash of a password “tutorial” as a ‘sha1:….” string
- To build the container I ran

docker build –t=“dbgannon/tutorial” .

Docker file for tutorial

```
# Version: 0.1.0
FROM jupyter/all-spark-notebook
MAINTAINER your name "dennis gannon"
RUN pip install azure-storage ==0.32.0
RUN pip install boto3
RUN easy_install pika
RUN easy_install bottle
COPY book-notebooks /tutorial_notebooks
COPY datadir /datadir
COPY script /
COPY mycert.pem /
COPY mycert.key /
CMD ["bash", "/script"]
```

Script file is

```
mkdir /home/jovyan/work/notebooks
cp /tutorial_notebooks/* /home/jovyan/work/notebooks
start-notebook.sh --certfile=/mycert.pem --keyfile=/mycert.key \
--NotebookApp.password='sha1:c02ed938ef17:0934044bb76008a364781d85db149a65fe9bb480'
```

Much more about containers

- You can mount your own directory in the container so data generated in the container can persist.
  - Docker run –v your_director:container_director -it –p …..
  - Containers can share mounted directories
- Docker compose – allows for container composition
- Singularity – a container system for supercomputer applications.

Section Summary

- Brief look at Virtual Machines
  - Installed Linux Data Science VM on Azure
- Containers
  - Some examples
  - An introduction to Container archeticture
  - Containers vs VMs
  - Building a container
- Later: Kubernetes = clusters of containers in the cloud