Cloud Computing for Science

Orienting

Dennis Gannon

Tutorial Goals

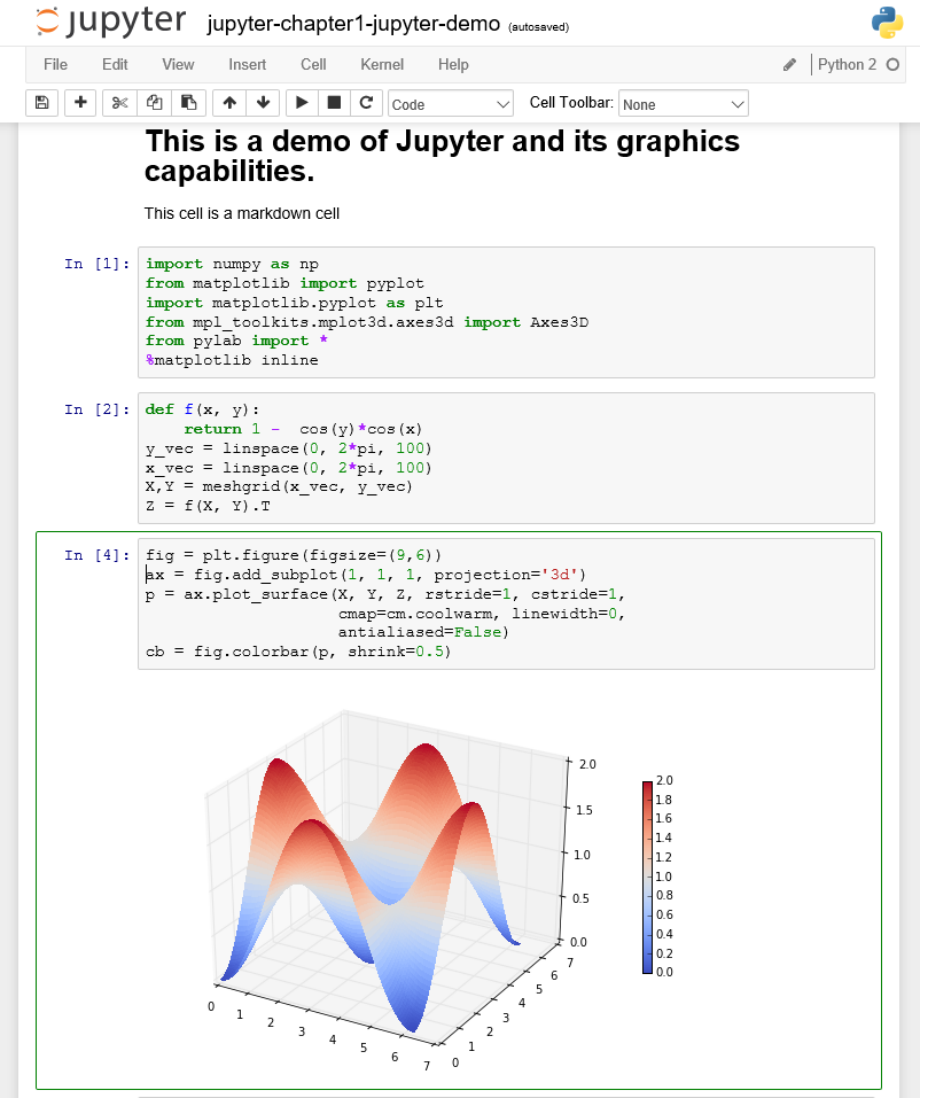
- An exploration of cloud computing for researchers
  - Scientists who need to go beyond their current resources
  - Computer science students who need to know what is possible
  - Data scientists who want to understand the potential of the cloud
- What will be covered
  - Cloud data services
  - VM and Container basics
  - Ways to scale: clusters, mapreduce, microservices
  - Data analytics in the cloud
  - Streaming data
  - Machine learning in the cloud

Preliminaries

- We will use Microsoft's Azure for Hands-on Exercises
  - Initializing your account
  - Download Azure Storage Explorer http://storageexplorer.com/
- Access tutorial tarball https://SciEngCloud.github.io/tutorial.tar.gz
  - On mac:
    - gunzip tutorial.tar.gz
    - tar –xf tutorial.tar
  - on PC use something like 7-zip or other decompression and extractor.
- Access to these slides and Jupyter Notebooks are also available
  - From the book "Cloud Computing for Science and Engineering" by Ian Foster an Dennis Gannon, to be published soon by MIT Press.
  - The link to the book is here
    - https://Cloud4SciEng.org

We will be using Python and Jupyter

- Jupyter in the cloud
  - Go to https://notebooks.azure.com
  - Signup – it's free.
    - If you are new to Jupyter do Welcome.ipynb
    - If you are new to Python do Python.ipynb
- Installing Jupyter on your laptop
  - Go to https://docs.continuum.io/anaconda/install
  - Do it. Then "Jupyter notebook" at the shell
- A Better solution: install Docker
  - https://docs.docker.com/engine/installation/
- and run Jupyter in a container
  - docker run –it -p 8888:8888 dbgannon/tutorial
  - Accept security exceptions and login with "tutorial"
  - Open notebook jupyter.ipynb to see the one on the right



Orienting in the Cloud

- We will discuss three different “public clouds” and a bit about a “science private clouds”
  - Public = anybody with a credit card has access. (aka commercial cloud)
  - Private = restricted to a special group of users. (aka Community Cloud or Academic Cloud)
  - (In Europe these terms are often reversed based on ownership.)
- They are:
  - Amazon Web Services (AWS) - 40% of all cloud resources on the planet.
  - Microsoft Azure – about 1/3 of AWS but growing
  - Google Cloud – third place
  - NSF JetStream – an OpenStack private cloud for US science researchers.
- There are *many* more clouds.
  - Public: IBM, DigitalOcean, Rackspace, 1&1, UpCloud, CityCloud, CloudSigma, CloudWatt, Aruba
  - Private Research Clouds: Aristotle, Bionimbus, Chameleon, Jetstream, RedCloud, indigo-datacloud, EU-Brazil Cloud, European Open Science Cloud
- What are the pros and cons of public vs private

Pros & Cons of Public vs Private Cloud

- Public cloud pros
  - Massive scale
  - Huge and growing list of services
  - Highly competitive on pricing due to economies of scale
  - Security is strong
  - Freedom from managing hardware
  - Hardware constantly upgraded
- Cons
  - Rules prohibit data moving to cloud
  - Funding models may make it hard to use
  - Fear of “vendor Lock-In”

- Private cloud pros
  - May be cheaper
  - You can keep it off the Internet so data can be very safe.
  - You can optimize your own hardware
  - You control everything
- Cons
  - You are responsible for everything
  - Not as many high level services
  - May not really be cheaper
  - You manage physical and system security

Two ways to access the cloud

- Portals and SDKs

- Web Portals
  - Dashboard that allow you to see and manage your cloud resources.

- Software Development Kits (SDKs)
  - Libraries that give you the tools to manage cloud resources from a program or script.
  - Based on REST web service calls

- Let’s look at several Cloud Web Portals

Amazon AWS Portal

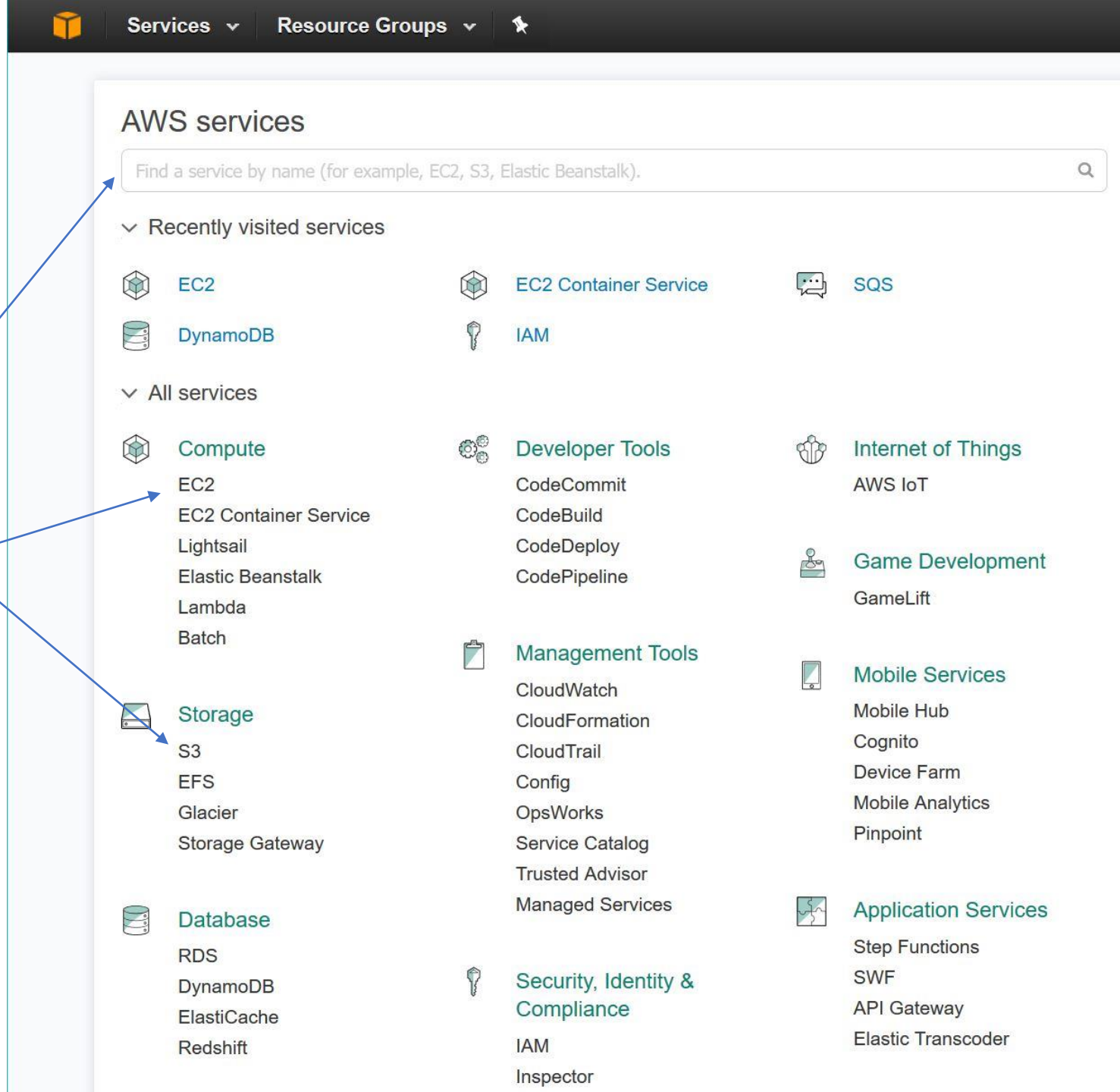
You see links to all the standard services

You also have a search bar to find others.

To create a storage account go to S3

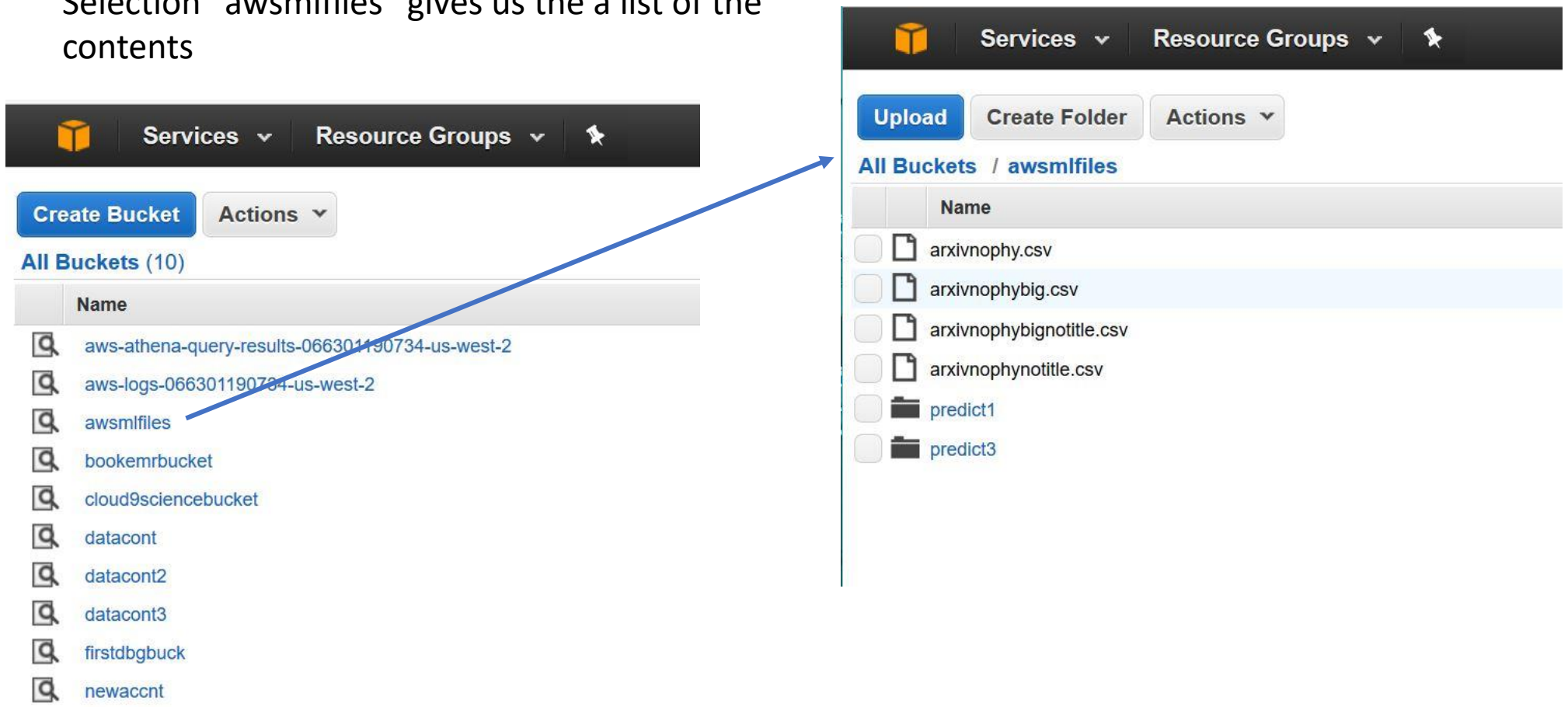
To launch a Virtual Machine go to EC2

Let's look at the S3 storage system



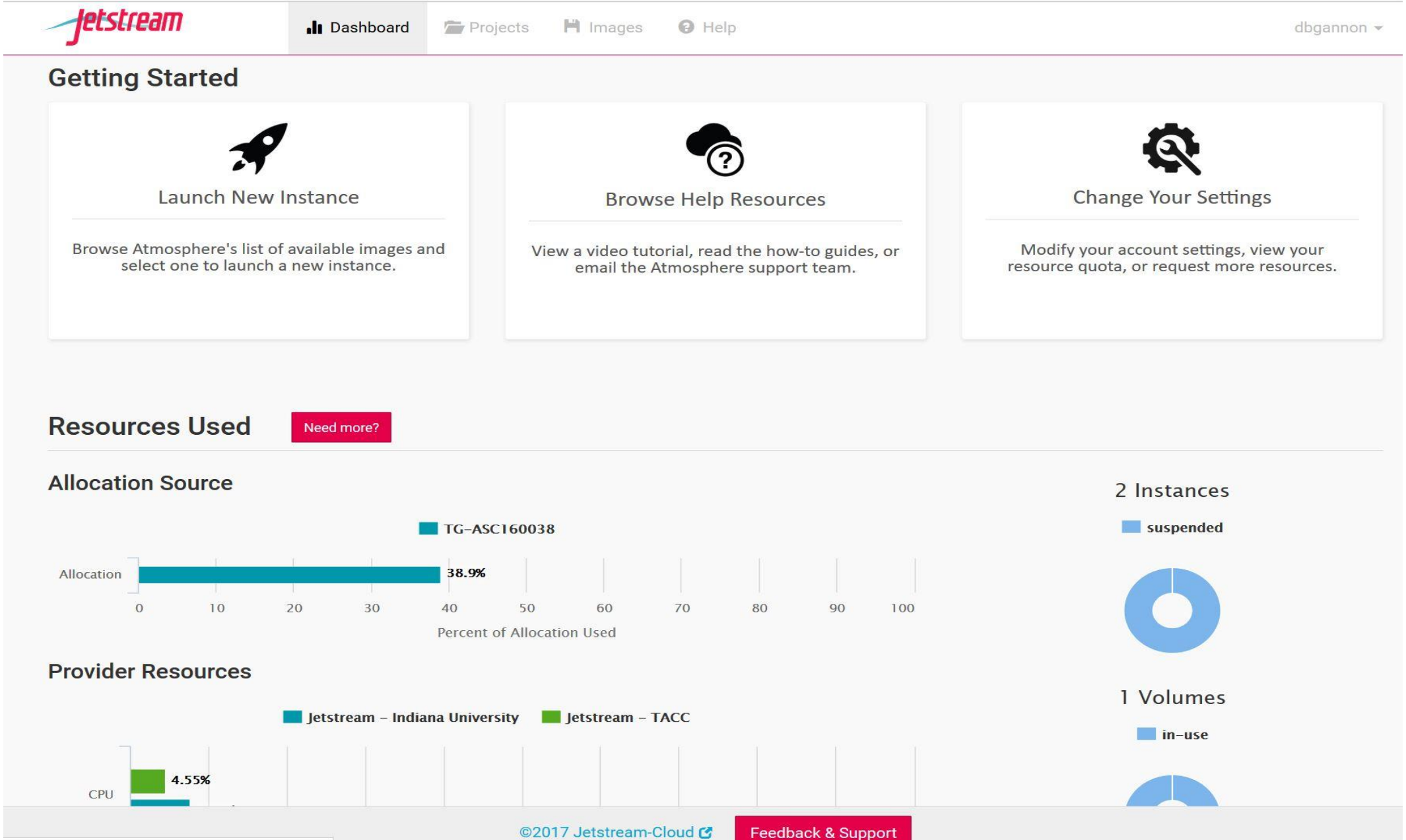
Selecting S3 we get the bucket list

Selection “awsmifiles” gives us the a list of the contents



Notice that it has regular objects AND folders.

Jetstream –NSF Science Cloud (OpenStack)



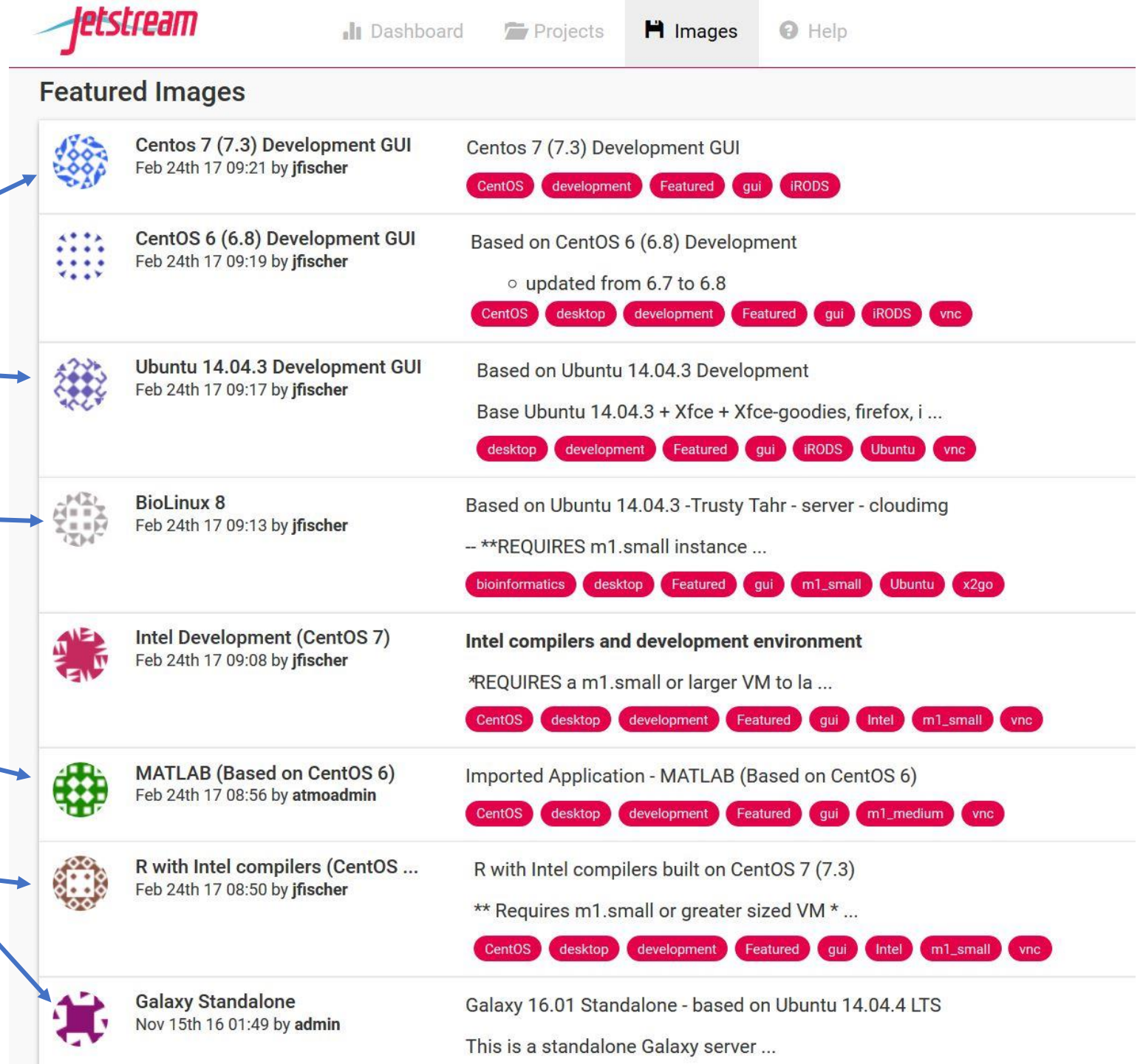
Jetstream is about virtual Machines for science

Generic Linux

Bioscience

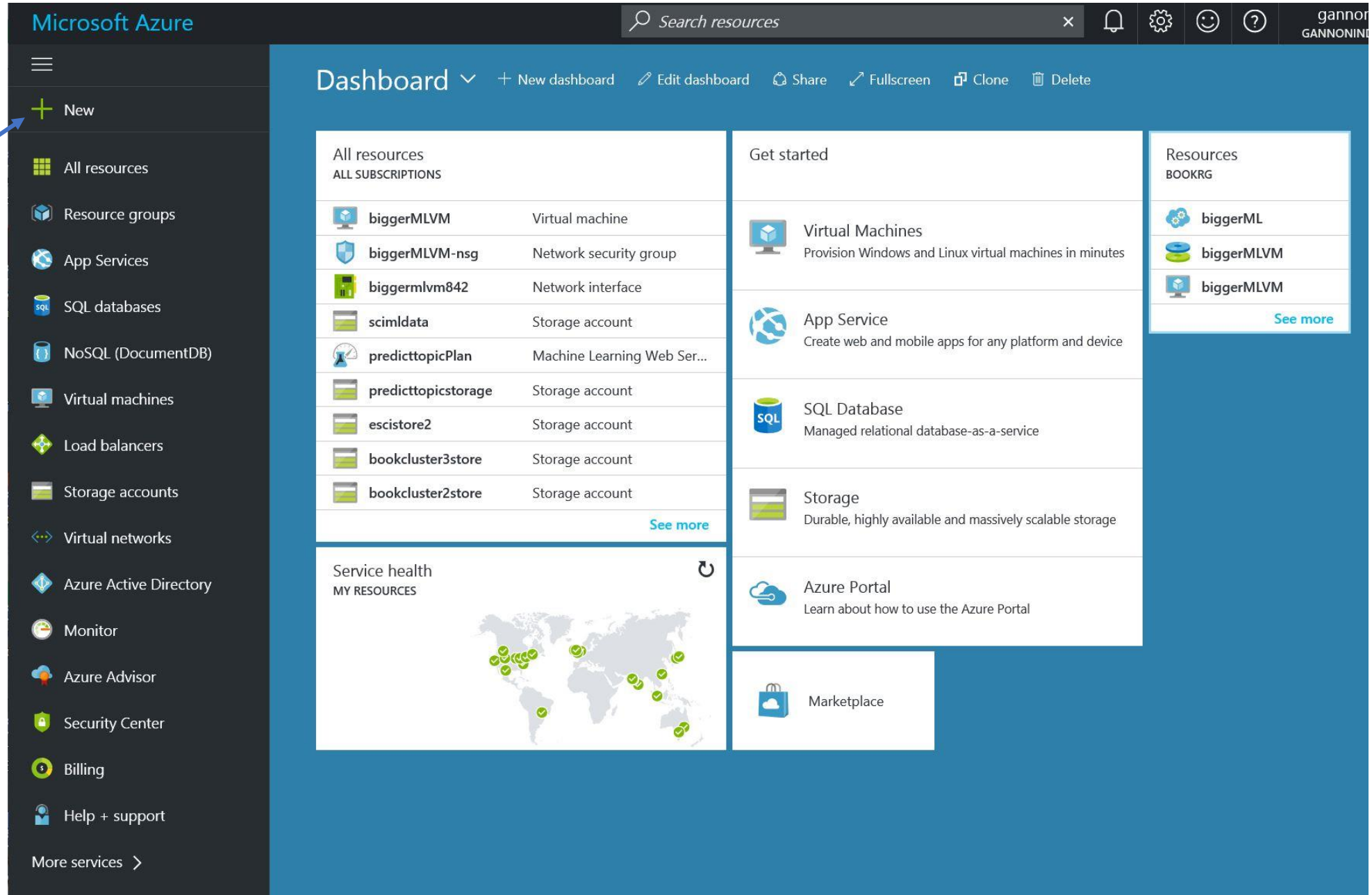
MATLAB

R



The Azure Portal

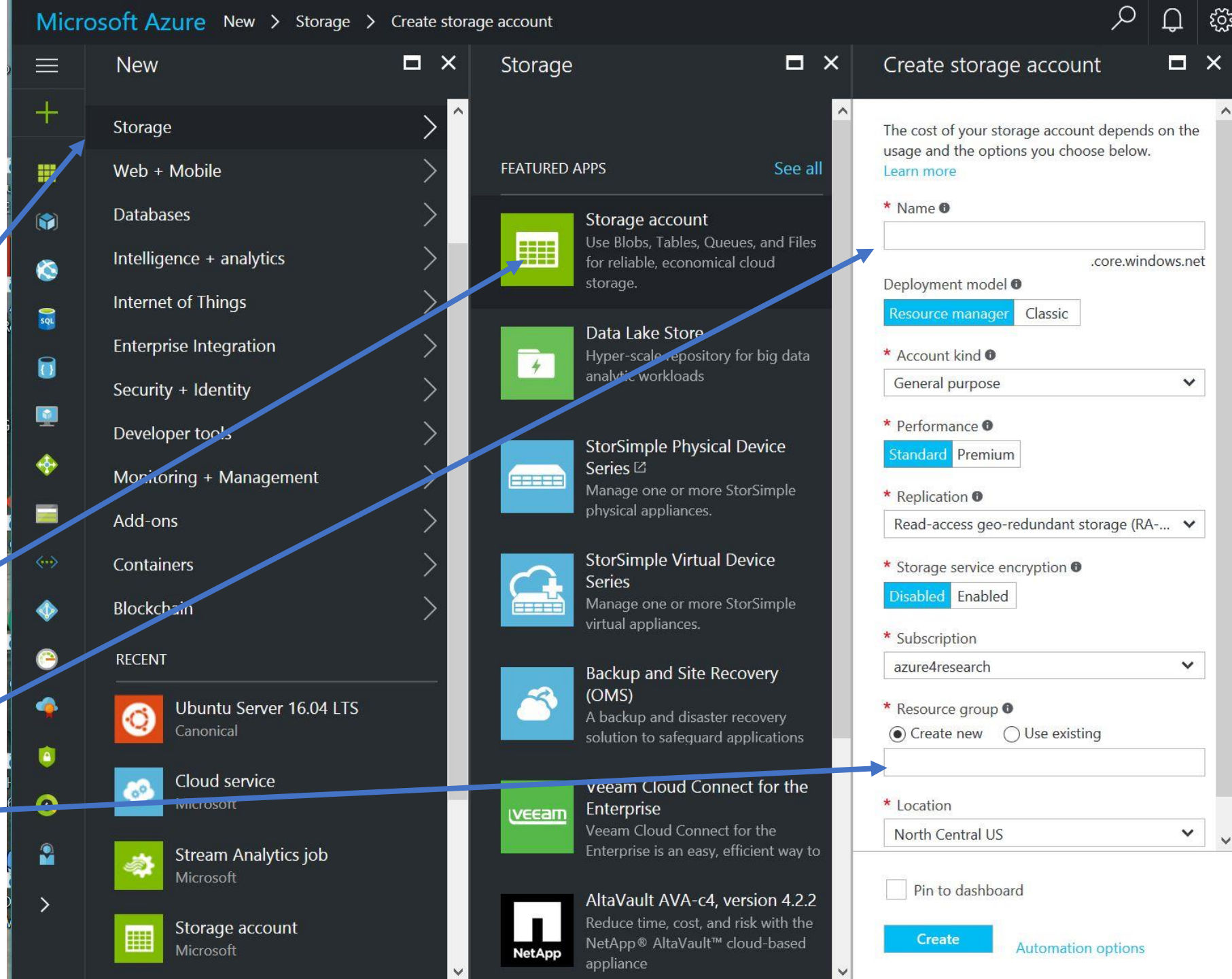
Select “+ New”



Selecting “+” gives this list of options.

Selecting “Storage” gives the secondary menu of types of storage apps.

To create a storage account select the top one.
- give it a name, research group and location.



Next

- A deeper look at storage
- Virtual Machines and Containers
- Scaling deployments and Microservices demo
- Analytics
- Machine Learning in the cloud