UVA | SCHOOL of DATA SCIENCE

Capstone Computation Resources

DS 6011: Capstone Part I

School of Data Science

University of Virginia

Last updated: March 21, 2024

Agenda

- Sensitive Data
- Sponsor Provided Resources
- Computing on Your Own Machine
- UVA Research Computing Resources
- Cloud Provider Services
- Example Scenarios

Compute and Storage

- To complete your capstone, you'll need compute and storage
- Compute – e.g., running Python scripts to train models, run inference
- Storage – e.g., large amounts of data files
- This deck will outline the options

Sensitive Data

- The presence of sensitive data can complicate matters

- Sensitive data includes:
  - Personally Identifiable Information (PII)
  - Protected Health Information (PHI)
  - Company/business data

- Secure environments are a must for handling this data

- Users cannot handle this data on their laptops

Sensitive Data, contd.

- Not all capstones work with sensitive data
- In some cases, sponsors can remove this data, or replace with meaningful data values that are not sensitive (de-identify the data)

Sponsor Provided Resources

- In some cases, the sponsor might give you access to their platform
- This is generally signaled early by the sponsor
- Will require login credentials and possibly training
- The platform might be a cloud environment (e.g., AWS)
- You would use the suggested tools (e.g., cloud storage, database)

Sponsor Provided Resources, contd.

- Users must NOT move the data off the platform
- Users are generally given access to environment that is not production

Computing on Your Own Machine

- In some cases, you might be able to compute on your machine

- May be feasible when:

- datasets are small
- the data is not sensitive (e.g., no PII)

Computing on Your Own Machine, contd.

- You will want to have a shared location for data

- In some cases, the sponsor might have tools or suggestions
- e.g., Google Drive

- This can work for non-sensitive data

UVA Research Computing

- UVA has on-premises computing resources
- Your faculty mentor can request an allocation, storage for the team
  https://www.rc.virginia.edu/

Options:

- **Rivanna** (to be renamed) : computing and storage
  For non-sensitive data

- **Ivy** : Compute for sensitive data
  High security VPN on machine (need training first)
  Need OPSWAT software on machine
  Can take 1-2 weeks onboard (watch for emails with instructions)

Cloud Provider Services

If the earlier options won't work, can contact Sue Haas, SDS IT Director (vsh@virginia.edu) to request setup with cloud account for the team

- **Microsoft Azure**
  Azure Virtual Labs – for non-sensitive data
  Azure Secure Enclave – for sensitive data.
  Same as Ivy,
  > High security VPN on machine (need training first)
  > Need OPSWAT on machine
  > Still being tested and may take extra time to onboard

- **Amazon Web Services (AWS)** :
  Currently supports non-sensitive data only
  User logs on and sees UVA AWS environment

Cloud Provider Services, contd.

- Cloud provider services come at additional cost to UVA SDS
- Resources: compute, GPU, storage
- Uses a virtual machine (VM)
- Students use browser, provide UVA credentials to log on
- Be sure to shut down tools when finished (e.g., stop server when done training)
- SDS IT is working to provide additional cloud providers

Example Scenarios

- If the sponsor can provide computational resources, this is a great option
- If the sponsor cannot provide resources:
- For non-sensitive data, options include **Rivanna, Azure Virtual Labs, AWS**
- For sensitive data, current recommendation is **Ivy**