Prerequisite Knowledge Required for This Course

Prerequisite Areas

- Python programming
- SQL
- Linux CLI
- GitHub git CLI

Python Programming

- Python
- Object-oriented concepts: classes, objects, etc.
- Pandas
- Simple data visualizations using matplotlib, Seaborn, etc.
- Data structures
- Algorithms
- Jupyter Notebooks
- Must have taken or gone through the Python course plus bridge courses for data structures and algorithms

SQL

- Relational database concepts
  - Tables, primary keys, foreign keys, views, etc.
  - ERD (entity relationship diagrams) in 3NF (third normal form)—read, understand, write queries
  - Operational databases versus analytical databases
- SQL
  - Select, order by, where, aggregation, group by, having, pre- versus post-aggregation
  - Joins, type 1 subqueries, type 2 subqueries, set operations
  - Transactions

Linux CLI

- CLI (command line interface)
- Very different from GUI, but necessary for cloud and IoT (Internet of Things)
- Connecting, logging in, users, groups, permissions
- Creating, deleting, copying, moving, and setting permissions for files and directories
- Bash shell basic programming

Github Git CLI

- Not the GUI (graphical user interface)—cannot use in the cloud or in IoT
- Must use git CLI
- Clone repo, create branches, change branches, track changes, stage changes, commit changes, push a branch to GitHub, pull and sync changes from GitHub to git

The End

Data Engineering

Overview

Data Engineering

Support data scientists

- Data engineers design, create, and deploy data pipelines.
- Data scientists process the data using sophisticated mathematical/statistical models, artificial intelligence, machine learning, deep learning, etc.
- Data engineers design, create, and deploy data serving layers.

Data Science Skills

- Math, statistics
- Artificial intelligence, machine learning, deep learning, etc.
- Experiment
- Business/application
- Data engineering

Minimum Skills to Full Stack

- Minimum skills for a data scientist
  - Math, statistics
  - Artificial intelligence, machine learning, deep learning, etc.
  - Experiment
  - Business/application
- Data engineering skills
  - More is better
  - Full stack data scientist—can do their own data engineering

The End

How Data Scientists Work With Data Engineers

Degree of Separation of Roles

- Data engineers separate from data scientist
  - Typically found at larger companies
- Full stack data scientists
  - Typically found at smaller companies
  - Although becoming more popular at larger companies

Working Effectively

Data scientists must learn to work effectively with data engineers.

- Data engineers often feel they are doing the heavy lifting and data scientists are getting all the credit
- Data scientists need to be diplomatic, express gratitude, share credit
- Win battles but lose war

Data Engineers Are Hard to Keep

Hard to keep a good data engineer

- Typically want to transition to full stack data scientist
- Often have years of software engineering, advance math, etc.

Outsourcing Woes

Outsourcing data engineering

- Business deal to “save money”
- By nature, outsourcing companies maximize billing, minimize work output
- Trivial requests can come with high price tags, long delivery times
- No say in the hiring process

The End

Learning Skills for IT

Learning How to Learn

- Learning how to learn is the best IT skill to have.
- Technology and products come and go at fast pace.
- IT requires being able to constantly learn new skills.
- We must be able to learn on our own.
  - Videos
  - Tutorials
  - Books

Bleeding Edge

- New technologies
  - Less stable
  - Less documentation
  - Less training
  - No books
  - Change rate much higher than older technologies
- Data science often requires us to be on the bleeding edge

Corporate Training

- Very expensive
- Travel, flights, hotels, meals
- Time lost from job during training
- Managers have limited budgets for training
- People who can learn on their own save time and money

Frustration and Confusion in Learning

- Not necessarily a bad thing!
- Think of a time you learned something after initial frustration and confusion.
- Do you think you gained a deeper understanding than if you would have not struggled?

Maximum Learning Potential

Frustration and confusion in learning demonstrates

- Stretching ourselves out of our comfort zone
- Learning at out maximum potential
- Not limiting ourselves to what is easy to learn

Setting Realistic Expectations

Expect and **make peace** with frustration and confusion in learning

- In this program
- In this course
- In your data science career

The End

Debugging Skills for IT

Debugging Skills

Working through a problem

- Calmly
- Systematically
- Logically
- Orderly

Optimized Debugging

- Orderly working through logical steps
- Each step cuts down search space as much as possible
- Enables us to come to a successful resolution as soon as possible

Working Independently

- Work on a problem and try everything reasonable to solve it before involving others.
- Don’t be the person who, at the first sign of trouble, throws up their hands and brings everyone into their issue.
- Companies and fellow workers value people who can debug most of their issues without involving others.

Setting Realistic Expectations

Expect and **make peace** with the idea.

- You will encounter problems on a daily or near daily basis
  - In this program
  - In this course
  - In your data science career

The End

Cloud

Concepts

Before the Cloud Era

Every company had to have its own:

- Data centers (primary and one or more secondary)
- Servers
- Storage
- Networks
- System software licenses
- Admins for servers, databases, network, storage

Technological Innovations

- VM (virtual machines)
- Containers
- Object store
- Edge servers
- Elastic computing

VM: Virtual Machines

- VM is a virtualization or emulation of a physical hardware computer.
- OS (operating system) runs on a VM instead of on a physical hardware computer.
- Multiple VMs can run on a single physical hardware computer.

VM Advantages

- Pool resources for computers that sit idle
- Migration of VMs to new hardware
  (typical hardware refresh is every three years)
- Configurations for software are often very specific and contradict configurations needed for other software
- VM images allow easy way to create new VMs

Containers

- Container is a virtualization or emulation of an OS.
- Some say it is a lightweight VM.
- Most modern computing is done in clusters of container.
- We will spend two weeks on containers.

Object Store

- Server-based storage independent of VMs, containers, clusters, etc. (“outlives”)
- Scales up with little or no effort on our part
- Easy to use
  - Folders
  - Files
  - Permissions

Edge Servers

- Servers located in all major cities all over the world
- Data from object store can be replicated to edge servers all over the world
- Fast, local access to data almost anywhere in the world
- Originally used for web servers, but many more uses

Elastic Computing

Quickly increase or decrease computing resources on an as needed basis

- Memory
- Storage
- CPUs
- VMs
- Containers
- Clusters

Cloud

- Leverages the main technological innovations we discussed: VMs, containers, object store, edge servers, elastic computing
- Data centers and edge servers around the world
- Instant access
- Pay-as-you-go—now cheaper than in-house

Managed Services

Cloud provides software ready to use with little effort on our part.

- Install, configuration, administration, elastic scale up and down, applies patches, security, troubleshooting, etc.
- Technical support
- Training
- Software license fees
- Pay-as-you-go
- Startups and small companies can use products that would be prohibitively expensive otherwise

The End

Cloud

Business Cases

Startup Company

- Don’t have to buy equipment, hire IT workers, etc.
- Low cost, pay-as-you-go
- Instantly have edge servers all around the world
- Quickly scale up clusters as needed
- Can add data centers all around the world as needed
- Managed services allow access to expensive products that would be out of reach budget-wise

Commonly Used Data

- Company has a lot of commonly used data that feeds into tens or hundreds of systems
- Solution:
  - Object store—design and organize folders, users, permission, etc.
  - All data stored in one central organized location
  - Edge servers all over the world—fast, easy access

Cluster Sharing

- User creates a very useful analytical cluster
- Everyone wants to use it
- Sharing is a hassle
- Solution:
  - Object store—holds data—cluster independent
  - If someone wants a cluster, they can create their own, load data, tear down the cluster when they are done

Point of Sale (POS) System

- POS system very busy at peak times
- POS not so busy at nonpeak times
- POS not used during closed hours
- Solution: elastic computing
  - Increase VMs for POS during peak times
  - Decrease VMs for POS during nonpeak times
  - Minimum VMs we never go below

Fortune 500 Company

- Fortune 500 company has had its own data centers since the 1960s—wants to move to cloud.
- Suggested strategy:
  - Use edge servers to boost web server static content
  - Move products to managed services
  - New systems start in the cloud going forward
  - Commonly used data to object store with edge servers
  - Migrate existing systems (slow and painful)

The End