MIDS w205 - Fundamentals of Data Engineering

Course Description

Storing, managing, and processing datasets are foundational processes in data science. This course introduces the fundamental knowledge and skills of data engineering that are required to be effective as a data scientist. This course focuses on the basics of data pipelines, data pipeline flows and associated business use cases, and how organizations derive value from data and data engineering. As these fundamentals of data engineering are introduced, learners will interact with data and data processes at various stages in the pipeline, understand key data engineering tools and platforms, and use and connect critical technologies through which one can construct storage and processing architectures that underpin data science applications.

Course Format

The course is organized as an online inverted classroom. During each week, students first work through a set of asynchronous materials, including video lectures, readings, and other activities. Once a week, students meet for a 90-minute live session, in which they connect with an instructor and other students in an online classroom. A functioning webcam and an audio headset are required to participate in the live sessions. Students must complete all assigned asynchronous material before the scheduled live session begins.

Course Objectives

Tools and Technologies

Students will:

- Build practical experience building data pipelines.
- Build practical experience cleaning, anonymizing, and plumbing data.
- Learn tooling for queries and query management (e.g., BigQuery, SQL).
- Learn tooling for analytics (jupyter, python, py-based-libs).
- Get exposure to advanced tooling for analytics (spark, kafka, etc).
- Learn how to leverage revision control.
- Learn how to use docker to assemble common tools for analysis.
- Build practical experience leveraging cloud-based resources for data analytics.
- Build practical experience consuming data and services from APIs.
- Get exposure to events and event-log based analytics.

Concepts

Students will:

- Learn to keep their analysis grounded in business relevance.
- Get exposure to some basic distributed storage and compute concepts.
- Get exposure to some basic RDBMS concepts.
- Get exposure to RDB -vs- NoSQL tooling and approaches.
- Get exposure to some basic data warehousing concepts.
- Learn the basics of virtualization and containerization.
- Understand how analysis changes wrt scale / complexity of data.
- Learn about data partitioning.
- Learn about latency in data analysis.
- Get exposure to ETL -vs- NoETL.
- Learn the basic concepts of web-based applications.
- Understand how basic data privacy, security, and chain-of-custody works.

Course Tools

In this class we will be using cloud instances on Google Cloud Platform (GCP) for all class activities and projects. We will set these up in the first week of class, but if you'd like to get a head start, here are some notes.

We'll have you create an AI Platform Notebook instance for class. These are general instructions as well as a general video on AI Platform Notebooks that goes into a little more depth.

The only differences to note for class are:

- Sign up with your UCB ISchool email address and you should get a fresh $300 credit for GCP. This should more than cover the instance expense for the semester
- Please create your instance from the TensorFlow-2.0 image when prompted
- We recommend you customize the AI Platform Notebook Instance to be smaller (cheaper) than the default. A single cpu core with 4G RAM and no GPU should suffice for class. You can always increase/add these later if you want to use this for other projects/classes

Evaluation & Grading

Projects

There are three projects:

- Query Project
- Tracking User Activity
- Understanding User Behavior

that collectively form the core of this course. Your work on them is one of the best ways for you to learn, and they each count for a third of your grade in this course.

Due Dates for Projects

Please check with your section instructor for definitive due dates for your projects.

Readings

Most readings are available through a subscription to https://www.safaribooksonline.com/. Other readings are blog posts and links.

Prerequisites

- Previous experience with Python
- Basic knowledge of Unix/Linux commands and tools as well as concepts such as processes, file systems
- In addition we'll use Docker, Git, and SQL as well as other tools
- If you feel like you're not where you'd like to be with these technologies/tools, here are some resources to get up to speed. There are options, pick which one best suits your needs

SQL

```
SQL Tutorial
w3schools.com
https://www.w3schools.com/sql/default.asp

Learning SQL, 2nd Edition
by Alan Beaulieu
https://www.safaribooksonline.com/library/view/learning-sql-2nd/9780596801847/
```

The Command Line

```
Learning the Shell
by William E. Shotts, Jr.
http://linuxcommand.org/lc3_learning_the_shell.php
```

Git

```
Pro Git book
by Scott Chacon and Ben Straub
https://git-scm.com/book/en/v2
```

Python

```
Python for Data Analysis, 2nd Edition
by William Wesley McKinney
https://www.safaribooksonline.com/library/view/python-for-data/9781491957653/
```

Docker

```
Getting Started with Docker
https://docs.docker.com/get-started/

Using Docker
by Adrian Mouat
https://www.safaribooksonline.com/library/view/using-docker/9781491915752/
```

Course Outline

The course consists of 4 sections:

- Introduction
  - Week 01 - Introduction

A 3-week Introduction that covers the basics of storage and retrieval concepts and tools; a 5-week Basics section that provides a deeper exploration of working with data and data pipelines; a 4-week section that focuses on Streaming Data; and a concluding section, Putting it All Together, that integrates concepts and skills from the entire course into a cohesive model of the data pipeline.

In addition to the sequenced material covered, the course also includes Tutorial materials that focus on technical skills associated with data engineering technologies, tools, and platforms. These tutorials also provide a practical foundation for the discussions and activities that will take place in the live classroom for specific weeks in the term.

Part 1 - Introduction

Week 01 - Introduction

Themes: What is Data Engineering? Gathering event data. Get started with queries.

Readings

> Data Science vs. Data Engineering
> Insight Data Science blog https://blog.insightdatascience.com/data-science-vs-data-engineering-62da7678adaa
>
> Network Programmability and Automation by Jason Edelman, Matt Oswalt, Scott S. Lowe Chapter 8. Source Control with Git https://www.safaribooksonline.com/library/view/network-programmability-and/9781491931240/ch08.html#sourcecontrol
>
> Analytics For Hackers: How To Think About Event Data by Michelle Wetzler https://blog.keen.io/analytics-for-hackers-how-to-think-about-event-data/

Asynchronous Content

```
1.1 - Welcome
  1.1.1 What is Data Engineering? [lecture]
  1.1.2 What will I be able to do after this course? [lecture]

1.2 - This Course
  1.2.1 Organization and Approach [lecture]

1.3 - Data Pipelines
  1.3.1 What's a Data Pipeline? [lecture]
  1.3.2 Typical Pipeline Components [lecture]

1.4 - Events Everywhere
  1.4.1 Device Events [lecture]
  1.4.2 User Events [lecture]
  1.4.3 Event Streams [lecture]
```

Week 02 - Working with Data

Themes: Intro to data, metadata, and some basic tools for working with data.

Readings

> Network Programmability and Automation by Jason Edelman, Matt Oswalt, Scott S. Lowe Chapter 5. Data Formats and Data Models https://www.safaribooksonline.com/library/view/network-programmability-and/9781491931240/ch05.html#dataformats
>
> Python for Data Analysis, 2nd Edition by William Wesley McKinney Chapter 2.2 IPython Basics https://www.safaribooksonline.com/library/view/python-for-data/9781491957653/ch02.html#ipython_basics

Asynchronous Content

```
2.1 Working with Data
  2.1.1 Introduction [lecture]
  2.1.2 Formats [screencast]
  2.1.3 Schema/Types [screencast]


2.2 Tools for Working with Data
  2.2.1 Introduction to Tools  [screencast]
  2.2.2 Introduction to the Command Line [tutorial]
  2.2.3 Using the The Command Line [tutorial]
  2.2.4 Introduction to Jupyter Notebooks [tutorial]
  2.2.5 Introduction to Docker [tutorial]
```

Week 03 - Welcome to the Queryside

Themes: SQL, query tools

Readings

> Designing Data-Intensive Applications, 1st Edition Martin Kleppmann Chapter 2. Data Models and Query Languages https://www.safaribooksonline.com/library/view/designing-data-intensive-applications/9781491903063/ch02.html#ch_datamodels
>
> Learning SQL, 2nd Edition by Alan Beaulieu Chapter 3. Query Primer https://www.safaribooksonline.com/library/view/learning-sql-2nd/9780596801847/ch03.html
>
> The SQL tutorial for data analysis from basic to advanced Complete Basic, Intermediate and Advanced https://mode.com/sql-tutorial/
>
> SQL Zoo Learn SQL from basics to advanced step by step https://sqlzoo.net/wiki/SQL_Tutorial
>
> Introduction to Databases Stanford Online, by Jennifer Widom https://lagunita.stanford.edu/courses/Engineering/db/2014_1/about

Note: The above resources are useful to develop a stronger understanding of SQL and develop a sense of comfort when working with relational databases. However these same concepts can be used to join and query multiple tables in BigQuery by joining based on common fields or columns.

Asynchronous Content

```
3.1 What is the queryside?
  3.1.1 Revisiting Pipelines [lecture]
  3.1.2 Immutability [lecture]

3.2 Query Tools
  3.2.1 Queryside World  [screencast]
  3.2.2 Athena [screencast]
  3.2.3 BigQuery [screencast]

3.3 Sneak Peek
  3.3.1 Caching [screencast]
  3.3.2 Using Jupyter Notebooks [tutorial]
  3.3.3 Using Docker [tutorial]
  3.3.4 Using Containers to Run Services with Docker [tutorial]
```

Part 2 - The Basics

Week 04 - Storing Data

Themes: Relational and NoSQL datastores

Readings

Required

> Using Docker by Adrian Mouat Chapter 1. The What and Why of Containers https://www.safaribooksonline.com/library/view/using-docker/9781491915752/ch01.html#what_and_why

Optional

> Linux in a Nutshell by By Stephen Figgins, Arnold Robbins, Ellen Siever, Robert Love Chapter 1. Introduction https://www.oreilly.com/library/view/linux-in-a/9780596806088/ch01.html

Asynchronous Content

```
4.1 Introduction
  4.1.1 - Where are we in the Pipeline? [lecture]

4.2 Relational Data Stores
```

```
  4.2.1 - Relational Databases [lecture]
  4.2.2 - Relational Databases - Relations [lecture]
  4.2.3 - Relational Databases - Normalization [lecture]

4.3 NoSQL Data Stores
  4.3.1 - NoSQL Data Stores [lecture]
  4.3.2 - NoSQL - Relational Model [screencast]
  4.3.3 - NoSQL - Document Store [screencast]
  4.3.4 - NoSQL - Key-Value Store [screencast]
```

Week 05 - Storing Data II

Themes: Introduction to cloud concepts and Hadoop

Readings

> Hadoop: The Definitive Guide, 4th Edition by Tom White Chapters 1-3 https://www.safaribooksonline.com/library/view/hadoop-the-definitive/9781491901687/

Asynchronous Content

```
5.1 Hadoop
  5.1.1 - Hadoop [lecture]
  5.1.2 - Hadoop Walkthrough [screencast]
  5.1.3 - The Hadoop Ecosystem [screencast]
  5.1.4 - Using Hadoop [screencast]

5.2 Introduction to the Cloud
  5.2.1 - Virtualization and Containers [screencast]
  5.2.2 - Infrastructure Encapsulation and Isolation [screencast]
  5.2.3 - Composing Containers [screencast]
  5.2.4 - Failure in Distributed Systems [lecture]
```

Week 06 - Transforming Data

Themes: ETL and its discontents, more Hadoop, and container management

Readings

> Designing Data-Intensive Applications, 1st Edition Martin Kleppmann Chapter 10. Batch Processing https://www.safaribooksonline.com/library/view/designing-data-intensive-applications/9781491903063/ch10.html#ch_batch
>
> Designing Data-Intensive Applications, 1st Edition Martin Kleppmann Chapter 11. Stream Processing https://www.safaribooksonline.com/library/view/designing-data-intensive-applications/9781491903063/ch11.html#ch_stream
>
> Hadoop: The Definitive Guide, 4th Edition by Tom White Chapter 4. YARN https://www.safaribooksonline.com/library/view/hadoop-the-definitive/9781491901687/ch04.html

Asynchronous Content

```
6.1 Introduction
  6.1.1 - Transform Section of Pipeline [lecture]

6.2 ETL
  6.2.2 - You don't always (aka usually) get what you want [screencast]

6.3 Hadoop ETL
  6.3.1 - Running Hadoop Jobs [lecture]
  6.3.2 - Hadoop ETL [screencast]
  6.3.3 - Introduction to Spark [screencast]

6.4 Container Management
  6.4.1 - Distributed Execution Models [screencast]
  6.4.2 - Introduction to Schedulers [screencast]
  6.4.3 - Using Schedulers [screencast]
```

Week 07 - Sourcing Data

Themes: Data and its provenance, security and privacy

Readings

- Learning Apache Kafka By Nishant Garg Chapter 1. Introducing Kafka
- https://www.oreilly.com/library/view/learning-apache-kafka/9781784393090/ch01.html

Asynchronous Content

```
7.1 Pipeline Context
  7.1.1 - Pipeline Context [lecture]

7.2 Where did the data come from?
  7.2.1 - APIs [screencast]
  7.2.2 - Web pages [screencast]
  7.2.3 - Email attachments [screencast]
  7.2.4 - Databases [screencast]
  7.2.5 - Cloud Storage [screencast]
  7.2.6 - Application Pipelines [screencast]

7.3 What do you want to know about the data you get?
  7.3.1 - Lineage [screencast]

7.4 Security and Privacy
  7.4.1 - Security and Privacy [lecture]
  7.4.2 - Sensitive Data [screencast]
```

Week 08 - Querying Data

Themes: Querying with partition keys and query planning

Readings

> Designing Data-Intensive Applications, 1st Edition by Martin Kleppmann Chapter 6. Partitioning https://www.safaribooksonline.com/library/view/designing-data-intensive-applications/9781491903063/ch06.html#ch_partitioning

Asynchronous Content

```
8.1 Querying
  8.1.1 - Pipeline Context [lecture]
  8.1.2 - Base decisions on queries [screencast]
  8.1.3 - Query Walkthrough [screencast]

8.2 Partitions
  8.2.1 - Querying with Partition Keys [screencast]

8.3 Query Planning
  8.3.1 - Query Optimization [screencast]
  8.3.2 - Using the Tools [screencast]
```

Part 3 - Streaming

Week 09 - Ingesting Data

Themes: Ingesting streaming data, using Kafka, considerations of latency.

Readings

> Kafka: The Definitive Guide, 1st Edition by Gwen Shapira, Neha Narkhede, Todd Palino Chapter 1. Meet Kafka https://www.oreilly.com/library/view/kafka-the-definitive/9781491936153/ch01.html
>
> Flask Web Development, 2nd Edition by Miguel Grinberg Chapter 2. Basic Application Structure https://www.oreilly.com/library/view/flask-web-development/9781491991725/ch02.html

Asynchronous Content

```
9.1 Pipeline Context
  9.1.1 - Where are we in the Pipeline? [lecture]

9.2 Kafka
  9.2.1 - Distributed Messaging [lecture]
  9.2.2 - Kafka Walkthrough [screencast]
  9.2.3 - Where are the data coming from? [screencast]
  9.2.4 - Latency [lecture]
  9.2.5 - Batch vs. Real Time [screencast]
  9.2.6 - Driven by Queries [screencast]
```

Week 10 - Transforming Streaming Data

Themes: NoETL, batch vs streaming, in-memory computing

Readings

> High Performance Spark, 1st Edition by Holden Karau, Rachel Warren Chapter 2. How Spark Works https://www.safaribooksonline.com/library/view/high-performance-spark/9781491943199/ch02.html

Asynchronous Content

```
10.1 NoETL
  10.1.1 - Pipeline Context [lecture]

10.2 Modes of Execution
  10.2.1 - Batch -vs- Streaming [screencast]
  10.2.2 - Single Event Processing [screencast]
  10.2.3 - Microbatch [screencast]
  10.2.4 - Continuous Applications - Handling Batch and Streaming Data in the Same System [screencast]

10.3
  10.3.1 - In Memory Computing [screencast]

10.4 Take Action!
  10.4.1 * Take Action! [lecture]
```

Week 11 - Storing Data III

Themes: Distributed in-memory storage and Spark

Readings

> Structured Streaming Programming Guide Apache Foundation https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html Through “Operations on streaming DataFrames/Datasets” section

Asynchronous Content

```
11.1 Pipeline Context
  11.1.1 - Pipeline Context [lecture]

11.2 In Memory
  11.2.1 - Streaming and Spark [screencast]
  11.2.2 - Structured Streaming [screencast]

11.3 Resource Selections
  11.3.1 - Dedicated Stream Processing [screencast]

11.4 Distributed In-memory Storage
  11.4.1 - Distributed In-memory Storage [screencast]

11.5 Activity
  11.5.1 - User Activity [lecture]
```

Week 12 - Querying Data II

Themes: Structured streaming, streaming queries, caching vs. stream queries

Readings

> Structured Streaming Programming Guide Apache Foundation https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html From “Streaming Queries” section to end

Asynchronous Content

```
12.1 Queries
  12.1.1 - Caching -vs- Stream Queries [screencast]
  12.1.2 - Caching and the Web [screencast]
  12.1.3 - A walk down memory lane - data mining [screencast]
  12.1.4 - Replay [screencast]
  12.1.5 - Caching queries [screencast]
```

Part 4 - Putting it All Together

Week 13 - Understanding Data

Themes: Sessionization and state and model validation and management

Readings

Scala:Applied Machine Learning by Alex Kozlov, Patrick R. Nicolas, Pascal Bugnion Sessionization (in Chapter 6. Working with Unstructured Data) https://www.safaribooksonline.com/library/view/scalaapplied-machine-learning/9781787126640/ch32s04.html

Advanced Analytics with Spark, 2nd Edition by Uri Laserson, Sandy Ryza, Sean Owen, Josh Wills Sessionization in Spark (in Chapter 8. Geospatial and Temporal Data Analysis on New York City Taxi Trip Data) https://www.safaribooksonline.com/library/view/Advanced+Analytics+with+Spark,+2nd+Edition/9781491972946/ch08.html#idm140398878264880

Asynchronous Content

```
13.1 Context
  13.1.1 - Pipeline Context [lecture]

13.2 Sessionization and State
  13.2.1 - Sessionization and state for action [lecture]

13.3 Using Public Clouds
  13.3.1 - AWS
  13.3.2 - GCP
```

Week 14 - Patterns for Data Pipelines

Themes: Conceptual DevOps and serverless architectures

Readings

Infrastructure as Code by Kief Morris Chapter 1. Challenges and Principles https://www.safaribooksonline.com/library/view/infrastructure-as-code/9781491924334/ch01.html#chapter-challenges

Network Programmability and Automation by Jason Edelman, Matt Oswalt, Scott S. Lowe Chapter 10. Continuous Integration https://www.safaribooksonline.com/library/view/network-programmability-and/9781491931240/ch10.html#cicd

Serverless Architectures on AWS by Peter Sbarski Chapter 1. Going serverless https://www.safaribooksonline.com/library/view/serverless-architectures-on/9781617293825/kindle_split_013.html

The Case for Learned Index Structures by Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, & Neoklis Polyzotis https://www.arxiv-vanity.com/papers/1712.01208v1/

Asynchronous Content

```
14.1 DevOps
  14.1.1 - Infrastructure as Code [screencast]

14.2 Cluster Examples
  14.2.1 - Docker Swarm
  14.2.2 - Kubernetes
```

Academic Integrity

Please read UC Berkeley’s policies around academic integrity: http://sa.berkeley.edu/conduct/integrity

Avoiding Plagiarism

Plagiarism is a serious academic offense, and students must take care not to copy code written by others. Beginning students sometimes have trouble identifying exactly when plagiarism takes place. Remember that it is generally fine to search for examples of code (for example, on forums like stackoverflow). This is a normal part of programming and can help you learn. However, it is important that you understand the code you find and use what you learn to write your own statements. If in doubt, simply document the place you found your example code and ask your instructors for further guidance.