



UMD DATA605: Big Data Systems

Lesson 1.1: Introduction

Instructor: Dr. GP Saggese, gsaggese@umd.edu

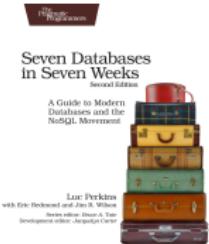
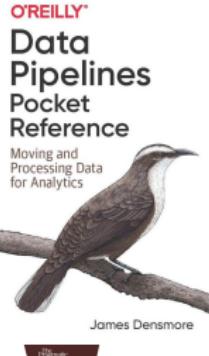
Invariants of a Class Lecture

- **Invariants of a class lecture**
 - Focus on intuition
 - Interactive Jupyter notebook tutorials
 - Tutorials done at home
 - Videos added over time
- **Class flow**
 - Alternate between slides, whiteboard, tutorials
- **Labs**
 - Review complete class project examples
 - Collaborate on class project

Books of the Class

- **Slides**

- Are extracted from books, technical articles, Internet
- Aim to be self-sufficient



Learning Outcomes

- Model and reason about data
- Process and manipulate data
 - E.g., Python, Pandas
- Introduce a variety of data models
 - E.g., relational, NoSQL, graph DBs
 - Decide appropriate data model for different applications
- Use data management systems
 - E.g., PostgreSQL, MongoDB, HBase
 - Decide appropriate system for scenarios
- Build data processing pipelines
 - E.g., Docker, Airflow
- Build a big-data system end-to-end
 - Class project
 - Contribute to an open-source project



Tools We Will Learn To Use

- **Programming languages**
 - Python
- **Development tools**
 - Bash/Linux OS
 - Git: data model, branching
 - GitHub: Pull Requests (PR), issues
 - Jupyter notebooks
 - Docker
- **Big data tools**
 - Extract-Transform-Load (ETL) pipelines
 - Relational DBs (PostgreSQL)
 - NoSQL DBs (HBase, MongoDB, Couchbase, Redis)
 - Graph DBs (Neo4j, GraphX, Giraph)
 - Computing framework (Hadoop, Spark, Dask)
 - Workflow manager (Airflow)
 - Cloud services (AWS)
- **Tutorials** for tools used in the class projects

Todos

- Study slides and materials
- DATA605 - ELMS/Canvas site
 - Enable notifications
 - Contact info for me and TAs
 - Always keep the TAs in cc
- Check [DATA605 Schedule](#)
- Check [DATA605 GitHub repo](#)
- Get these [slides](#)
- Check [DATA605 FAQs](#)
- Setup computing environment
 - Install Linux/VMware
 - Install Docker on laptop
 - Instructions in class repo
- Bring laptop to class
- Lessons recorded
 - Still attend class, when possible

Grading

- **Quizzes**
 - 40% of grade
 - Multi-choice on previous 2 lessons
 - 20 questions in 20 minutes
 - 4-5 quizzes to encourage study during semester
- **Final Project**
 - 60% of grade
 - Comprehensive application of course concepts
 - Big data project in Python from a list of topics
 - Individual or group

Class Projects

- **Project** is “*Build X with Y*”, where
 - X is a “use case”
 - Y is a “technology”
- **Activities**
 - Study and describe technology Y
 - Implement use case X using technology Y
 - Create Jupyter notebooks to demo your project
 - Commit code to GitHub, contribute to open-source repo
 - Write a blog entry
 - Present your project in a video
 - In-class labs + reviews
- You choose from list of X and Y, e.g.,
 - Data engineering
 - Emerging technologies (e.g., large language models)
 - ...
- **Each project:**
 - Individual or group ($n < 4$)
 - Varying difficulty levels

Soft Skills to Succeed in the Workplace

- **Goal:** model class project for workplace preparation
 - Work in a team
 - Design software architecture (OOP, Agile, Design Patterns)
 - Comment your code
 - Write external documentation (tutorials, manuals, how-tos)
 - Write understandable code (including for future-you)
 - Read others' code
 - Follow code conventions (PEP8, Google Code)
 - Communicate clearly (emails, Slack)
 - File a bug report
 - Reproduce a bug
 - Intuition of CS constants
 - Basic understanding of OS (virtual memory, processes)

Yours Truly

- **GP Saggese**
 - 2001-2006, PhD / Postdoc at the University of Illinois at Urbana-Champaign
 - [LinkedIn](#)
 - gsaggese@umd.edu
- **University of Maryland:**
 - 2023-, Lecturer for UMD DATA605: Big Data Systems
 - 2025-, Lecturer for UMD MSML610: Advanced Machine Learning
- **In the real-world**
 - Research scientist at NVIDIA, Synopsys, Teza, Engineers' Gate
 - 3x AI and fin-tech startup founder (ZeroSoft, June, Causify AI)
 - 20+ academic papers, 2 US patents

