Data Engineering Project

Module 4
Project Description

September 27

Nektaria Tryfona, PhD

Electrical and Computer Engineering
Virginia Tech

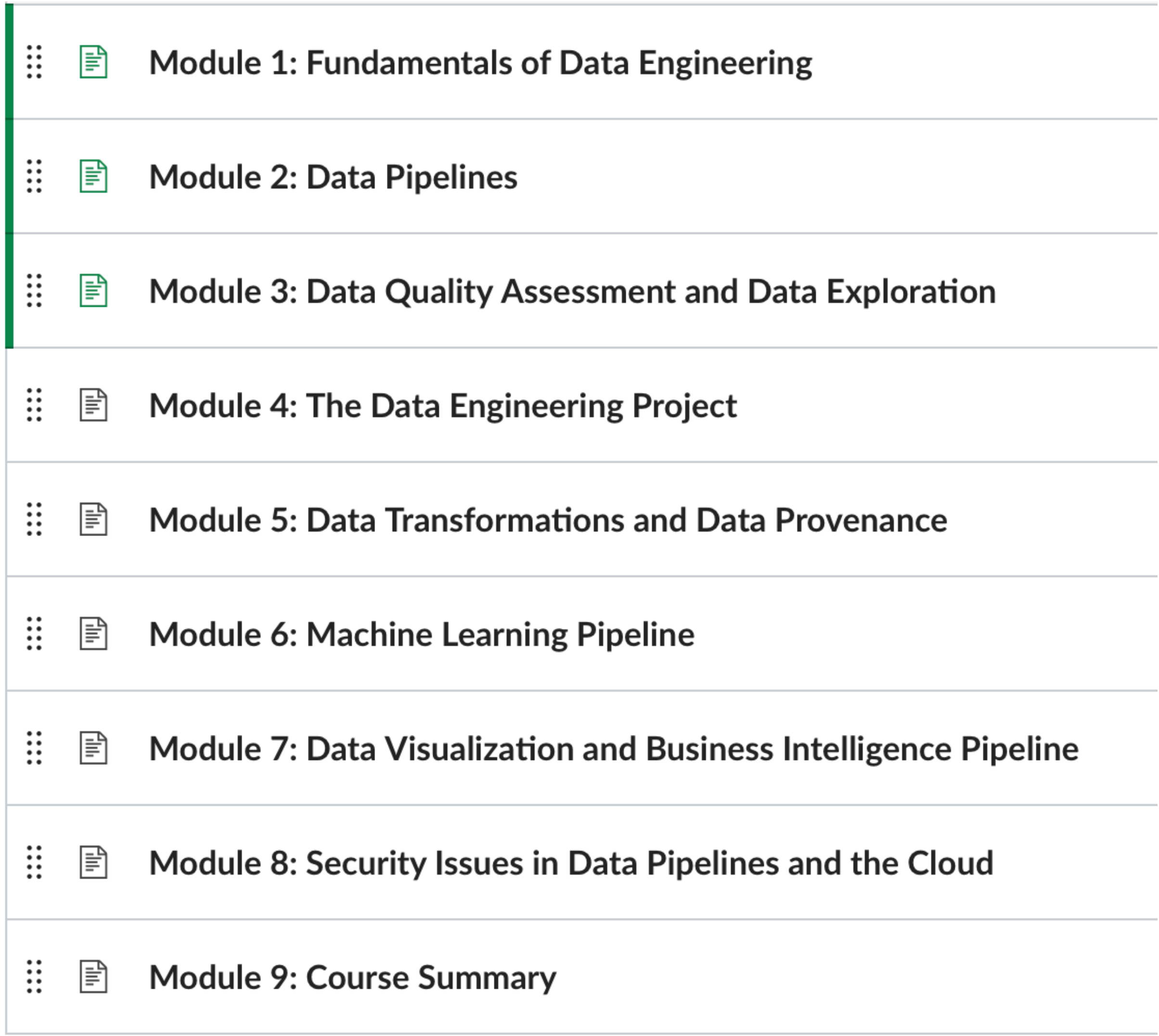
Objectives

- Project description – major points
- Pipelines to be used
- Project ideas and data sets

Project Description – key points

- Address a data engineering project from requirements to data-driven solution
- Project ideas & datasets
- Go through all the tasks of the pipeline
- Deliverable: GitHub

How the schedule looks like



Module 1: Fundamentals of Data Engineering

Module 2: Data Pipelines

Module 3: Data Quality Assessment and Data Exploration

Module 4: The Data Engineering Project

Module 5: Data Transformations and Data Provenance

Module 6: Machine Learning Pipeline

Module 7: Data Visualization and Business Intelligence Pipeline

Module 8: Security Issues in Data Pipelines and the Cloud

Module 9: Course Summary

Addressed in Project Milestone

Assignment 3 (only Lab 3) Oct 16-Nov 5

Assignment 4 (only Lab 4) Oct 30-Nov 19

Implement a Pipeline

There are 4 options

Batch → ML → Visualization

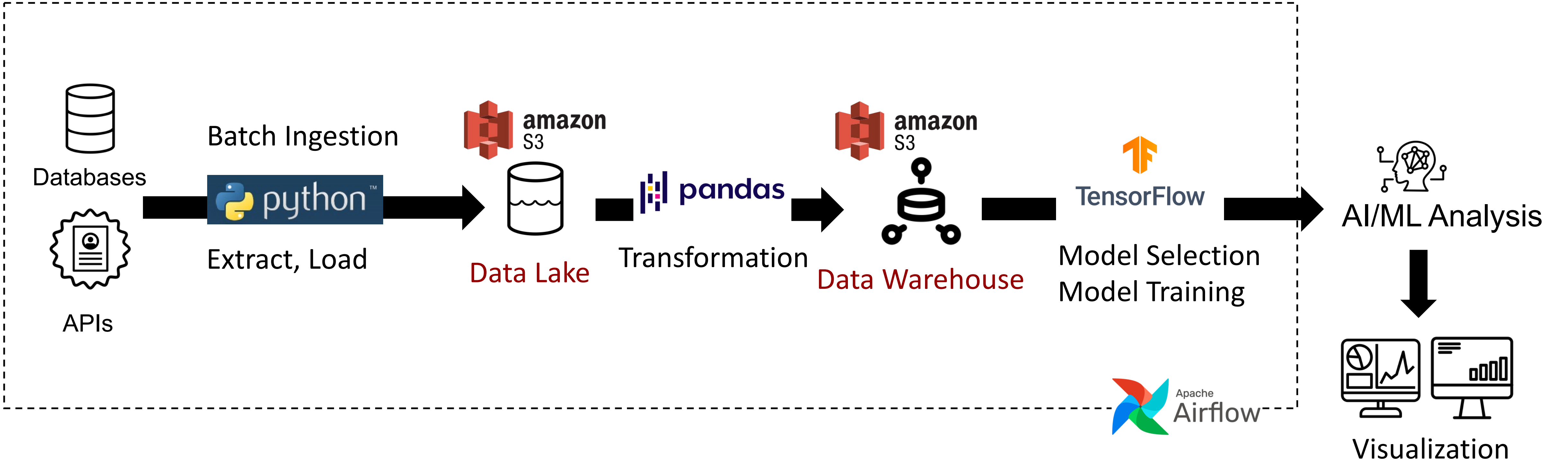
Batch → Visualization (very easy, probably no added value)

Stream → ML (difficult)

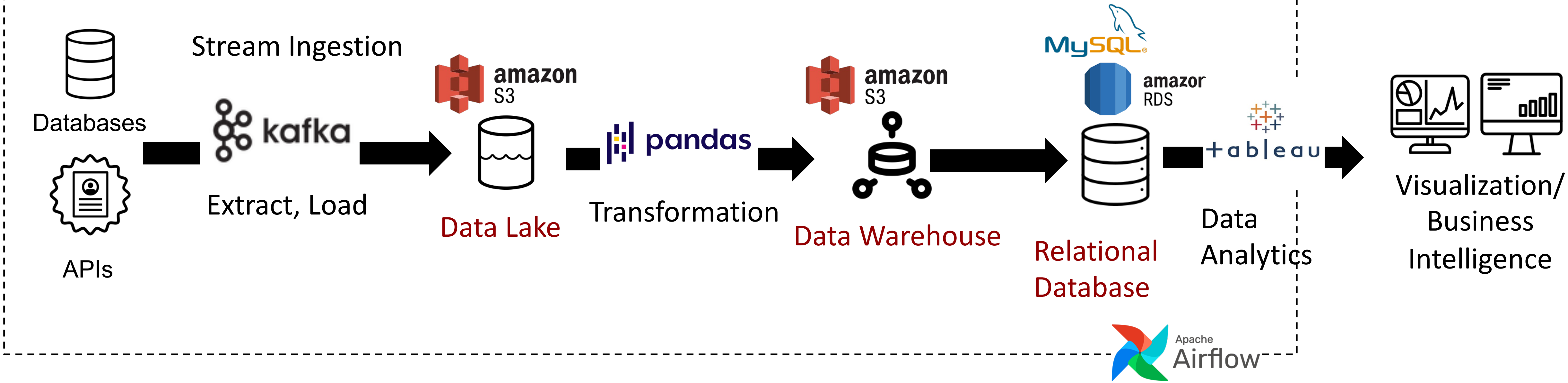
Stream → Visualization

- Batch → Visualization
- Stream → ML

Data Pipeline 1 Batch – ML - Vizualization



Data Pipeline 2 Stream-Visualization



How?

- There will be teams of 1-3 students.
- Project ideas and datasets are provided
- You may bring your project idea and dataset
  - properly document its goal and outcomes
  - discuss it with us

Project Phases

- Set-up the pipeline (M2)
- Data Exploration and Transformation (M3)
- Machine Learning (M6)
- Business Analytics (M7)
- Data Transformation & Data Provenance (M5) in “report” (mandatory)

- Go through all the tasks of the pipeline

Project Examples (more in Canvas)

- Predict number of COVID-19 related deaths and graph the predicted deaths in the future
  Dataset: https://github.com/nytimes/covid-19-data

- Visualize real time traffic incidents based on location
  Dataset: https://developer.tomtom.com/traffic-api/documentation/product-information/introduction

- Visualize the live scores of multiple football matches
  Dataset: https://www.api-football.com/

- Detect whether a transaction is false or not and plot the number of faulty transactions compared to legitimate ones
  Dataset: https://www.kaggle.com/datasets/kartik2112/fraud-detection

- Built your own ChatGPT

Project Deliverables

- Project Proposal
- Project Milestone
- Final Project Submission

  - GitHub Portfolio (https://github.com/ankurchavda/streamify)
  - Code
  - Documentation (Title, Description/what the project does, Methods per phase, data provenance answers to questions, innovation and scalability, technical or platform concerns, difficulties, limitation of the pipeline for the project)

* think as data engineers → problem solver

1. Project Proposal

Submit a pdf description of your proposed project (about 1-2 pages). Include the following sections:

- **Title**: What is your project title?
- **Purpose**: What problem are you trying to solve? What questions will your project address? How will the project show/visualize the results?
- **Pipeline and Methods:** Which pipeline will you use? What dataset will you use? Depending on the pipeline, what methods and tools will you use?
- **The final result you are expecting to achieve**

2. Project Milestone (covers Modules 1-5)

This milestone acts as a progress marker

The project progress should follow the sequence of the course modules.

Submit a pdf report (around 3 pages) written as an *early draft* of your final report, including:

- What have you done so far?
- What pipeline, project, and dataset are you using?
- Does your data need data cleaning or preprocessing? If so, what?
- Will you perform Exploratory Data Analysis? Which methods are you going to use?
- What information about data provenance have you listed? Answer the characteristic data provenance questions addressed in Module 5
- Are there any untested assumptions or other reasons that would prevent you from completing your project?
- What are your next steps?
- What is left to do?

3. Project Deliverables and Deadlines

- October 8: Project Proposal due [10%]
- October 29: Project Milestone due [20%] (covers modules 1-5)
- December 8: Final project due (report pdf and GitHub Repository) [70%]

- Hard deadlines → mark your calendars

What course of action(s) should be taken if we decide on a dataset (e.g., topic, scope) -- then decide to change it?

What course of action(s) should be taken if the scope chosen is unable to work as expected (e.g., code, kafka issues, API versioning, unexpected errors)?

One of the questions I have with the approach to our exercises -- is that the AWS session(s) seems to "freeze" or "lag" if services are running for more than 45 min on idle. Is the project aimed to be completed in such a way where we run through it in one session (apologies ahead of time if I'm not wording it correctly)

The final deliverable is a GitHub Repository --- would be it correct to say that this would look like Assignment 1 but just on Github?

Will there be different/additional assignments on top of the project from now on? or is the remainder of the course solely focused on the project from now on?

How will be teams be chosen? Can we work individually or will we be assigned team mate(s). How will the grading be done if we are on a team? Are there specific roles for team mates?

If our project is pulling large amounts of data, will that impact the AWS quota? What is the cap on how much we can process/store in terms of this project?

How to

- Read carefully the assignment & address it
- Many questions are answered in the assignment
- We do not correct the code
- No troubleshooting beyond what is related to class

- Use Forum & office hours
- No email about "code is not working" etc.

- Friday evening before the deadline is late
- TA checks the discussion forum and email at least once a day and generally responds within 24 hours.
- On weekends, we will respond within 36 hours
- Deadlines matter