


# Course Syllabus

A .pdf version of the syllabus can be downloaded from [here](https://canvas.vt.edu/courses/176740/files/29032883?wrap=1)  
(<https://canvas.vt.edu/courses/176740/files/29032883?wrap=1>).

## ECE 5984 - Data Engineering Project

### Course Meta

- Instructor: Dr. Nektaria Tryfona
- Email: [tryfona@vt.edu](mailto:tryfona@vt.edu) (<mailto:tryfona@vt.edu>)
- Zoom Day/Time: Wednesday 7PM Eastern Time
- Zoom URL: <https://viriniatech.zoom.us/j/5718583104>   
(<https://viriniatech.zoom.us/j/5718583104>)

### Course Catalog Description



Fundamentals of data engineering. The role of data engineer. Data engineering lifecycle. Data quality and valuation. Data provenance. Data generation, ingestion, transformation, storage, serving Artificial Intelligence and Machine Learning (AI/ML), visualization and business analytics. Automation and task orchestration. Data systems. E(xtract)T(ransform)L(oad) data. Build, test, maintain data pipelines. Data lakes. Real-world problems with an emphasis on end-to-end engineering solutions. Cloud services and open-source data engines/platforms. Engineering portfolio. Master of Information Technology (MIT) students only. Pre: Graduate Standing.

### Course Description

The goal of this course is to equip you with the engineering knowledge and hands-on experience to address data engineering environments, platforms and projects. Thus, you will have access to a custom data engineering pipeline comprising open-source tools hosted on the AWS commercial platform, in order to complete a data engineering project. At the end of the semester, you will have an Engineering Portfolio in GitHub, highlighting your data engineering knowledge and skills.

### Learning Objectives

Having successfully completed this course, the student will be able to:

-  Identify the component of the data engineering lifecycle
-  Illustrate data quality issues
- Describe provenance issues
- Apply data validation methods

- Design data pipeline orchestration environments
- Evaluate data storage environments
- Analyze the role of data lakes in data engineering
- Propose end-to-end engineering solutions using cloud services and open-source data platforms
- Construct an engineering deliverable portfolio

## Course Structure

This is an asynchronous online class. Pre-recorded lectures are available from the course website. There will be a weekly synchronous (live) Zoom session that will be for Q&A, help with assignments, project introductions, etc. Attending the synchronous sessions is optional. These sessions will be recorded on a best-effort basis for later viewing if you cannot attend the live session. While you are not required to attend the live Zoom session, it is expected to watch the Zoom recording sometime during the week since important announcements or clarifications will be made there.

The course is divided into 9 Modules; some Modules are bi-weekly. Each Module typically consists of readings and lecture videos.

The course has 4 Assignments and 1 Project. Each Assignment has 2 parts, Lab and Homework, to help you gain hands-on experience:

(a) In the **Lab**, you will follow detailed guidelines to connect data engineering concepts discussed in the lectures with the custom pipeline provided to address parts/phases of the data pipeline.

(b) In the **Homework** part, you “fly solo”, following the Lab process; you will have to edit the provided code in certain areas to achieve the task being asked for.

The Project is "the sum of it all", following the sequence of the data engineering phases as addressed in the course. Based on the gained experience and techniques learned from Assignments (and the included Labs), the Project sums the data engineering phases resulting in the final product. Projects will be executed in teams of 1-3 students.

It is important to note that projects will focus on the Machine Learning and/or Data Visualization/BI pipeline, to cover possible backgrounds and interests. In case you are not familiar with ML projects, we will provide guidelines for the project.

The programming language to be used is Python. Basic knowledge in Python or other programming language is required coming for this class. Guided labs will build on your existing understanding of basic programming concepts that apply to any programming language.

Note that because the instructor and Virginia Tech share copyright, the materials on this course website are only for the use of students enrolled in this course.




## Course Materials

We will be using scientific articles from selected journals and conferences focusing on state-of-the-art practices and solutions. All readings will be made available on the Canvas course site.

The following book is recommended but not mandatory:




- Harenslak, B.P., Rutger de Ruiter, J. (2021). *Data Pipelines with Apache Airflow*.

Publisher: Manning. ISBN-13: 978-1617296901. [https://www.amazon.com/Data-Pipelines-Apache-Airflow-Harenslak-dp-1617296902/dp/1617296902/ref=mt\\_other?\\_encoding=UTF8&me=&qid=1663171016](https://www.amazon.com/Data-Pipelines-Apache-Airflow-Harenslak-dp-1617296902/dp/1617296902/ref=mt_other?_encoding=UTF8&me=&qid=1663171016)  ([https://www.amazon.com/Data-Pipelines-Apache-Airflow-Harenslak-dp-1617296902/dp/1617296902/ref=mt\\_other?\\_encoding=UTF8&me=&qid=1663171016](https://www.amazon.com/Data-Pipelines-Apache-Airflow-Harenslak-dp-1617296902/dp/1617296902/ref=mt_other?_encoding=UTF8&me=&qid=1663171016))

or <https://www.oreilly.com/library/view/data-pipelines-with/9781617296901/>  (<https://www.oreilly.com/library/view/data-pipelines-with/9781617296901/>)

As this is a hands-on computing class, you will also need the following software:

We will be using the following software:

- Python + IDE (windows). To download and install Python, visit the official Python website and choose your 3.x version. (Usually, it is advised to use the latest stable version available, however, certain libraries might have specific Python version dependencies) <https://www.python.org/downloads/>  (<https://www.python.org/downloads/>)
- PyCharm <https://www.jetbrains.com/pycharm/download/>  (<https://www.jetbrains.com/pycharm/download/>)
- <https://github.com/>  (<https://github.com/>). For this, you will need a Gmail account; detailed guidelines will be provided
- An AWS account. We will take care of this.


We will discuss how to use the provided Virtual Machine and install the aforementioned materials in class. It is expected that you are able to admin and computationally use your own computer (install software, update system settings, *edit and run Python programs*, etc) as needed.

A working and reliable computer with Internet access (with sufficient bandwidth to watch lecture videos, attend Zoom sessions, and download course materials) is required, as well as access to

Canvas <https://canvas.vt.edu/> and Zoom <https://virginiatech.zoom.us/> 

(<https://virginiatech.zoom.us/>). Students are responsible for all materials on the Canvas course site. For assistance with IT, login, or related computing issues, contact <http://4help.vt.edu> (<http://4help.vt.edu/>) or call 540-231-HELP(4357). Please exhaust all resources available to you before contacting the instructor.

## Course Website

 Course website will be available on Canvas to registered students. Lectures, announcements, additional readings, assignments and the project will be available on the class website.

## Class Schedule

The course website will be available on Canvas to registered students. Lectures, announcements, additional readings, assignments and the project will be available on the class website.

## Grading Policy

### Total Score/Points

ITEM	PERCENTAGE
Assignments	40%
Final Project	60%

### Final Grades

Final course grades will be determined after all work is completed and graded. Final grades will be based on the following scale:

GRADE	RANGE
A	100% to 93%
A-	< 93% to 90%
B+	< 90% to 87%
B	< 87% to 83%
B-	< 83% to 80%
C+	< 80% to 77%
C	< 77% to 73%
C-	< 73% to 70%
D+	< 70% to 67%
D	< 67% to 63%
	< 63% to 60%



GRADE	RANGE
F	< 60% to 0%

If you have questions about your performance at any point during the semester, please contact the instructor.

## Participation

Participants in this course should expect to spend at least 8-10 hours per week involved in the activities and completion of labs, assignments, and other deliverables over the semester. Note however that, depending on your background, you may need to invest more time to understand the material and complete the assignments and semester project. This is similar to the time one would invest in a course in a traditional classroom setting. This time invested is an average with some weeks requiring more, some less time to complete all assignments and activities.

## Make-up Policy

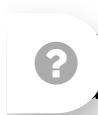
All assignments, and the project must be submitted by their posted due dates. Late submissions will not be accepted unless there are documented, exceptional circumstances. Such arrangements must be made prior to the due date if at all feasible. If circumstances prevent participation or timely completion of any assessment, students must contact the professor to arrange for adjustments in the schedule *in advance of the due date and not after the completion date has passed*.

## Communication Expectations and Netiquette

Since this is an asynchronous online course, email communication will be a recommended way to communicate with me –I will do my best to respond in a timely manner. That being said, if you email me after 5 pm or on the weekend, I may not be able to respond until the next day or the next Monday, respectively.

All members of the class are expected to follow rules of common courtesy in all email messages, threaded discussions and chats. When making a post or sending an email, please sign the email with your first name (or what you prefer to be called) so that others know whom they are talking to. It is not always apparent from your email address. I will send out announcements through Canvas periodically. These will go to your VT email address, so make sure you monitor it.

The Discussions Board section of the course website is a great way to ask general questions and to get to know your fellow students -I encourage you to participate. When posting or emailing in this class, you should:



Refer to the Core Rules of Netiquette for general guidelines of proper behavior.

Make posts that are on topic and within the scope of the course material.

- Take your posts seriously and review and edit your posts before sending.
- Be as brief as possible while still making a thorough comment.

- Always give proper credit when referencing or quoting another source.
- Be sure to read all messages in a thread before replying.
- Don't repeat someone else's post without adding something of your own to it.
- Avoid short, generic replies such as, "I agree." You should include why you agree or add to the previous point.
- Always be respectful of others' opinions even when they differ from your own.
- When you disagree with someone, you should express your differing opinion in a respectful, non-critical way.
- Do not make personal or insulting remarks.
- Be open-minded.

## Honor Code

"As a Hokie, I will conduct myself with honor and integrity at all times. I will not lie, cheat, or steal, nor will I accept the actions of those who do."

The tenets of the Virginia Tech Graduate Honor Code will be strictly enforced in this course, and all assessments shall be subject to the stipulations of the Graduate Honor Code. For more information on the Graduate Honor Code, please refer to the GHS Constitution

at [https://graduateschool.vt.edu/content/dam/graduateschool\\_vt\\_edu/graduate-honor-system/Constitution2018.pdf](https://graduateschool.vt.edu/content/dam/graduateschool_vt_edu/graduate-honor-system/Constitution2018.pdf)

([https://graduateschool.vt.edu/content/dam/graduateschool\\_vt\\_edu/graduate-honor-system/Constitution2018.pdf](https://graduateschool.vt.edu/content/dam/graduateschool_vt_edu/graduate-honor-system/Constitution2018.pdf)). In this course we may use services such as

turnitin <https://www.turnitin.com>  (<https://www.turnitin.com/>) to assess student submissions.

In general, discussion and cooperative learning on general topics is encouraged. Such discussion must be limited to general information such as lecture and text material or how to use the software. Sharing your assignment/project answers, or using another student's assignment/project solutions, design, implementation, or other specific results is strictly prohibited and is an honor code violation. Copying computer files, designs, or solutions from any source is strictly prohibited and is an honor code violation. Ask the instructor if you ever have a question about what constitutes acceptable or unacceptable sharing.


## Disabilities

Any student who feels that they may need an accommodation because of a disability (learning disability, attention deficit disorder, psychological, physical, etc.), should contact the Virginia Tech Services for Students with Disabilities (SSD) office <http://www.ssd.vt.edu/> (<http://www.ssd.vt.edu/>) at 540-231-3788 and schedule a confidential consultation.

## Course Evaluations

Course evaluations for this class will be administrated online at the end of the course. Your feedback is important. Please take the time to complete the SPOT survey when sent in Canvas.

# Course Summary:

Date	Details	Due
Wed Aug 30, 2023	 <u>Lab - Setting up the environment</u> ( <a href="https://canvas.vt.edu/courses/176740/assignments/1816315">https://canvas.vt.edu/courses/176740/assignments/1816315</a> )	due by 11:59pm

