

Data Engineering – II, 1TD076 62016

Teachers

Salman Toor, salman.toor@it.uu.se

Teacher Assistant

Li Ju, li.ju@it.uu.se

Anand Mathew Muthukulam Simon (anand-mathew.muthukulam-simon.6015@student.uu.se)

Xiong Luo (xiong.luo.7609@student.uu.se)

Ella Johanna Schmidtbreick (ella-johanna.schmidtbreick.4283@student.uu.se)

Finn Vaughankraska (finn.vaughankraska.2674@student.uu.se)

Prerequisites

Data Engineering-I 10.0hp

Introduction

Data Engineering-II is an advanced level course offers in the international data science masters programme at the department of information technology, Uppsala University. We expect students to know the basics of distributed computing infrastructures, distributed analysis using batch processing, intermediate level practical and conceptual understanding with the frameworks like Spark and Hadoop and mature understanding with machine learning (ML) algorithms. Data Engineering-II course consists of four modules: (M1) Contextualization and containers, (M2) Data stream processing, (M3) Orchestration frameworks and production-grade computational workflows and (M4) Federated and distributed machine learning. M1 module will cover basic concepts related to contextualization and how to use containers. M2 module will cover the fundamentals of data stream processing, different architectures, available frameworks and one assignment based on the Apache Pulsar framework. M3 module will cover advanced techniques of contextualization, orchestration, building complex workflows, ML model serving and the role of

continues integration and continues deployment in production-grade ML workflows. M4 Module will focus on challenges related to computationally expensive ML model training and possible solutions based on distributed/federated model training.

Assessment Criteria

The assessment consists of successfully completing and handing in reports for computer assignments, the research seminar and the mini-project, each of which is associated with a number of points. There will be a total of 10 points to collect, distributed as follows

	Basic	Extra
C1	1	1
C2	1	1
C3	1	1
C4	1	1
S	1	
Mini-project:	1	3

Each **computer lab (C)** should be completed individually and will be divided in two parts (except C1). The first part will cover fundamental material and is mandatory for a passing grade. The second part, which will cover more advanced topics, is not mandatory but will count towards higher marks. Note that computer assignments are a major part of the course examination, so budget significant time (roughly 1 week of work) apart from the scheduled lab occasion to complete the labs.

S is a literature seminar. The students will be presented with a number of articles to choose from. Successful completion of S requires a written summary/review of





the selected articles to be handed in prior to the seminar, as well as active participation in the seminar. The particular deadlines that apply will be posted in the student portal.

The **Mini-project** will be conducted in groups of 4 students. Here, a small software project is to be completed and presented in a written report and orally to peers and teachers in a short presentation. More details regarding the mini-projects, as well as assessment criteria, will be posted in separate documents in the student portal.

The final grade on the course will be determined based on the number of collected points, according to the following limits. Note that irrespective of the total number of points, all basic parts (Part 1) of the computer labs must be completed in order to pass. It is also necessary to obtain at least one point (sufficient performance) on the mini-project.

(Pass) 3:	6 of which at least 1 point on the mini-project
4:	9 of which at least 2 points on the mini-project
5:	12

Colour scheme

Regular	
Compulsory	
Optional	
Deadlines	

Lecture Plan (tentative, can be subject to revision)

L - Lectures (2x45min)

C - Computer Labs (2x45min)

S - Seminar (2x45min)

P - Final presentation (oral): 10-min/group.

L1, -> Course Introduction

L2, -> (M1) Contextualization and containers

C1, -> Docker based container orchestration exercise. (Online)

C1, -> Docker based container orchestration exercise. (Online)

L4, -> (M2) Data Stream Processing (Part-1)

L5, -> (M2) Data Stream Processing (Part-2)

C2, -> A practical introduction to one of the popular Data stream processing framework Apache Pulsar. (Online)

C2, -> A practical introduction to one of the popular data stream processing frameworks Apache Pulsar. (Online)

L6, -> (M3) Distributed Computing Infrastructures

L7, -> (M3) Continuous Integration + Model serving

C3, -> Design and implementation of computational workflows using Kubernetes framework, and hands-on experience with CI/CD using Github Actions. (Online)

C3, -> Design and implementation of computational workflows using Kubernetes framework, and hands-on experience with CI/CD using Github Actions. (Online)

L8, -> Announcement of the projects and literature seminars

L9, -> (M4) Distributed Machine Learning

L10, -> (M3) Distributed Machine Learning

C4, -> Federated and distributed machine learning. (Online)

C4, -> Federated and distributed machine learning. (Online)

S, -> Literature Seminars (Onsite)

S, -> Literature Seminars (Onsite)

S, -> Literature Seminars (Onsite)

L11, -> Guest Lecture (Onsite)

CX, -> Extra lab session (Online)

Project presentations (Onsite)

Project presentations (Onsite)

Project presentations (Onsite)

Submission Deadlines

Note: You will only get extra points if the submission will be within the due deadline. All late submissions will only get a basic 1 passing point.

Computer Lab 1

- Announcement date: 28-03-2025
- **Deadline: 08-04-2025**

Computer Lab 2

- Announcement date: 07-04-2025
- **Deadline: 24-04-2025**

Computer Lab 3

- Announcement date: 24-04-2025
- **Deadline: 06-05-2025**

Computer Lab 4

- Announcement date: 06-05-2025
- **Deadline: 15-05-2025**

Literature Seminar

- Announcement date: 28-04-2025
- **Deadline: 12-05-2025**

Mini-project

- Announcement date: 28-04-2025
- **Deadline: 28-05-2025**

Welcome to the course!