universität koblenz
weiter:denken

Big Data
Module Overview (2023)

Frank Hopfgartner
Institute for Web Science and Technologies

Frank Hopfgartner

universität koblenz
weiter:denken

Professor for
Data Science

MY ROLES

- Head of Data Science Research Group
- Head of Institute for Web Science & Technologies
- Coordinator of MSc in Web and Data Science programme

MY MODULES

- Big Data
- Artificial Intelligence 1
- Seminar: Algorithmic and Data Bias
- Research Lab: Data Analysis in the Cloud

MY INTERESTS

- Intersection of Information Access and Data Science
- Focus on personal data analysis, e.g., interaction log files, heterogeneous sensor data

MY BACKGROUND

- PhD in Computing Science (University of Glasgow)
- Previous positions in Sheffield, Glasgow, Berlin, Dublin, Berkeley, London

Welcome from the teaching team

universität koblenz
weiter:denken

Tutorial Coordinator

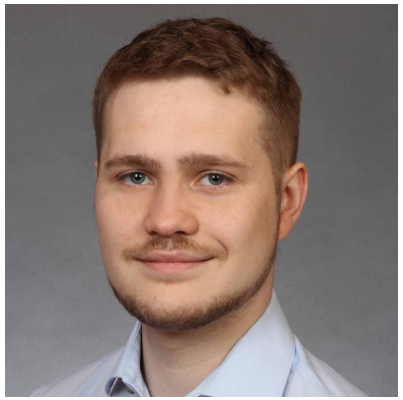
- Marina Ernst

Databricks Team

- Tania Sennikova
- Evgeny Chernyi
- Alan Mazankiewicz

Tutors

- Avishek Pathania
- Burge Vaishali
- Srikanata Sumanth
- Ritik Gupta



Module Aim

universität koblenz
weiter:denken

The main aim of this module is to give knowledge about big data analytics architectures and help students to understand **when and how** to appropriately use such **scalable data processing** solutions.

Module Objectives

universität koblenz
weiter:denken

- Provide an introduction to different big data computational **architectures and algorithms**;
- Provide an overview of existing big data analytics **products for volume, velocity, and variety of data**;
- Show how big data analytics is used in industry by means of **use cases**;
- Provide practical hands-on experience through use of cloud-based **software for big data processing**

Intended Learning Outcomes

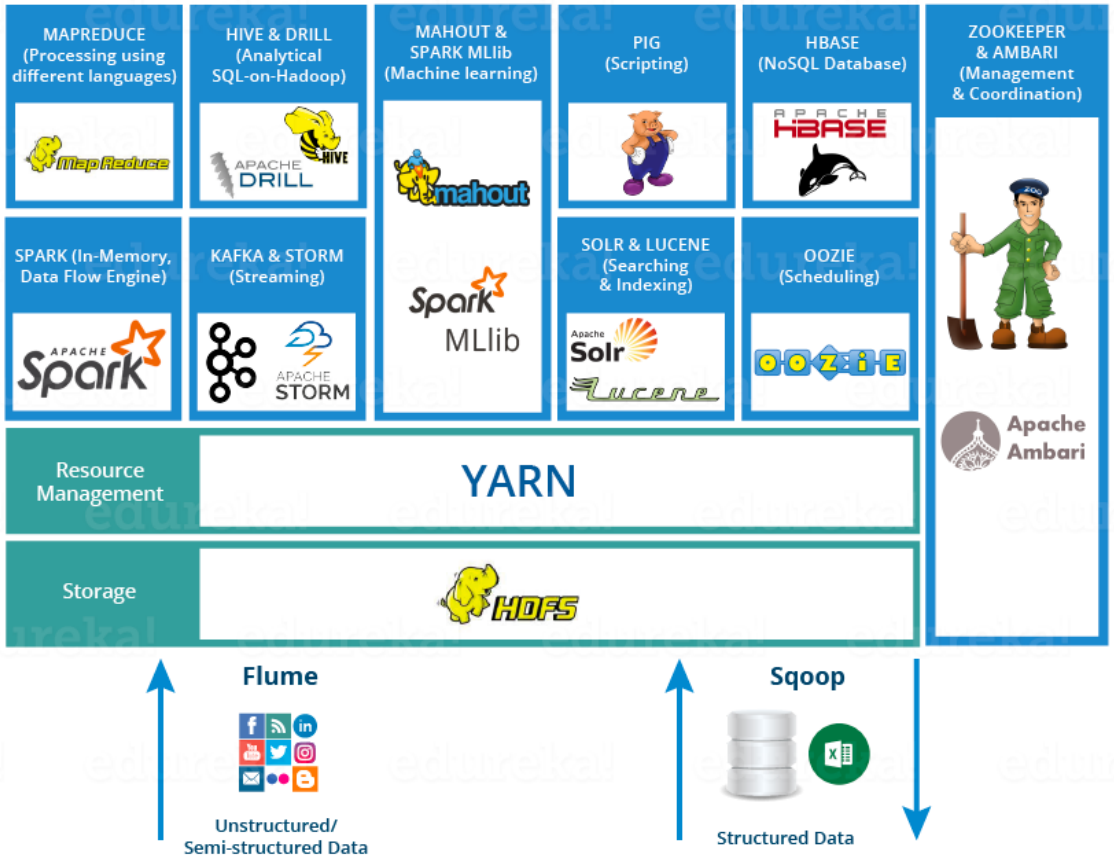
universität koblenz
weiter:denken

At the end of this module, you will be able to:

- Understand the challenges and opportunities in dealing with Big Data including **which situations are more or less appropriate** for Big Data analytics.
- Have an understanding of the on-going development of **Big Data infrastructure solutions** for Volume, Variety, and Velocity including industry-driven and open-source solutions.
- Have an understanding of the knowledge required **to use such infrastructure** and how to best support data science practices on top of such architectures for non-technical stakeholders (e.g., executives)

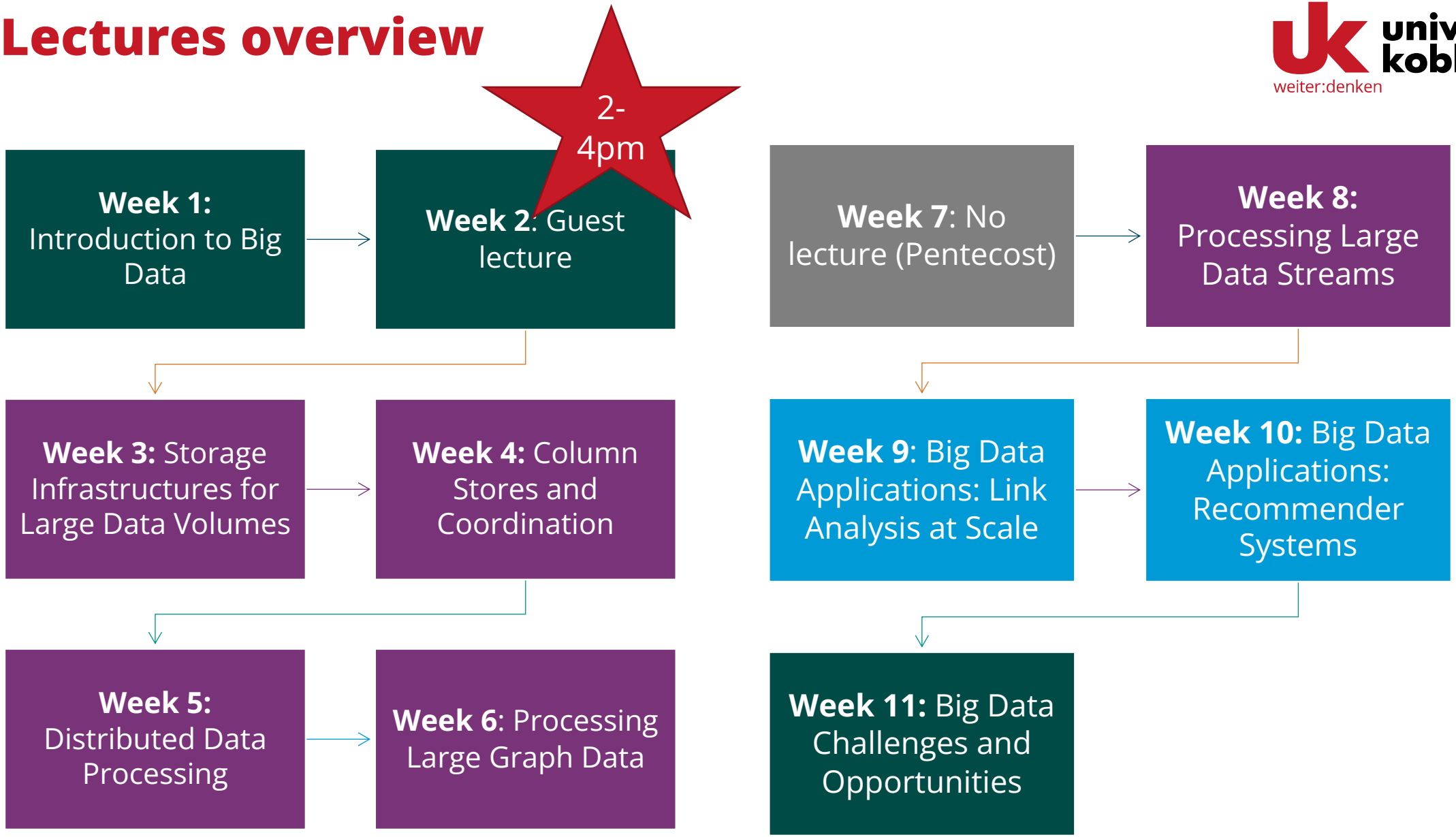
Core technology: The Hadoop ecosystem

universität koblenz
weiter:denken



Lectures overview

universität koblenz
weiter:denken



Assignments

universität koblenz
weiter:denken

- Weekly exercises
  - Released on Fridays after the tutorial sessions
  - To be completed individually
  - Submit by the following Thursday at 23:59 on OLAT
  - Score **60% or higher** in the weekly exercises and submit **80% of all assignments** to qualify for the exam

- Exam
  - Date: 03 August 2023 in D 028

Weekly exercises

universität koblenz
weiter:denken

You will learn how to use the cloud-based big data platform called Databricks Community Edition.

> “Databricks provides a unified, **open platform for all your data**. It empowers data scientists, data engineers and data analysts with a simple collaborative environment to run interactive and scheduled data analysis workloads.“

https://databricks.com/

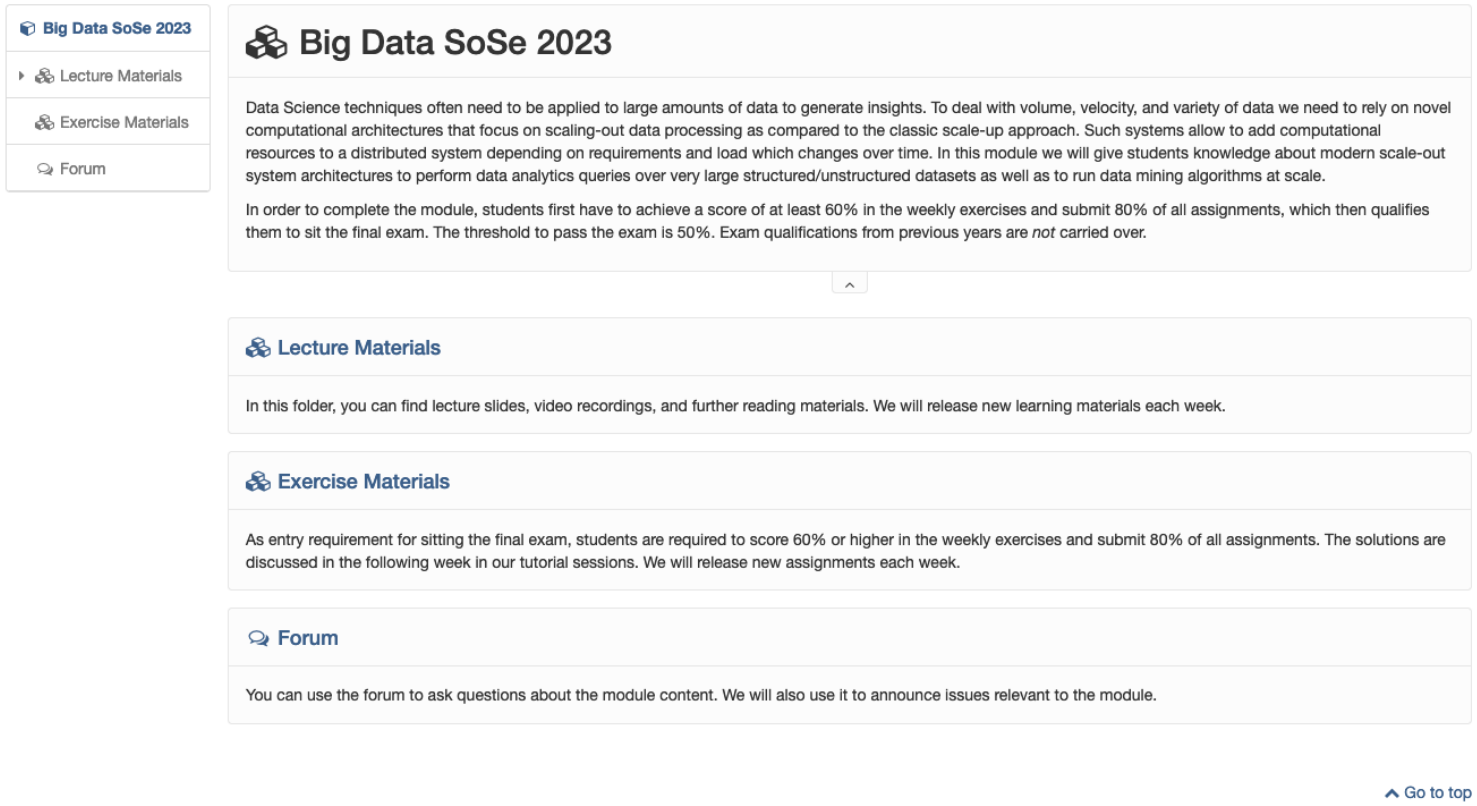
Weekly activities

universität koblenz
weiter:denken

- Lectures
  - Fridays, 12:00-14:00 (s.t.) in D028
- Tutorials
  - Fridays, 14:00-16:00 (c.t.) in E011
  - **First session: 28 April 2023 (starting s.t.)**
- Exercises
  - Assignments released on Fridays at 17:00
  - **First assignment: 19 April 2023**
  - Due on Thursday at 23:59

OLAT

universität koblenz
weiter:denken



https://olat.vcrp.de/url/RepositoryEntry/4061135343

Guest lecture (28 April 2023)

universität koblenz
weiter:denken

- Introduction to Databricks platform.
- Helpful to get started for the weekly assignments.
- Please note that KLIPS *wrongly* states that the 28 April session will not take place.
- **It will take place starting at 14:00 (s.t.) in E011.**

databricks

Plagiarism (1/2)

universität koblenz
weiter:denken

- Submitted solutions to exercises are part of the assessment process
- Solutions have to be prepared independently and must contain only the individual's own work
- You are allowed to discuss exercise sheets and potential solutions with other students, but it is explicitly forbidden to copy solutions and code of others
- Internet research is allowed but solutions must be phrased in one own's words and code has to be developed by yourself

Plagiarism (2/2)

universität koblenz
weiter:denken

- Also small changes of text and code (such as renaming of variables) still counts as a plagiarism
- Plagiarism is a severe academic misconduct and will be punished accordingly
- In case of plagiarism the student will be expelled from the course and the exam (you lose one year); severe cases of plagiarism may be criminally prosecuted
- If two students have (partially) identical solutions, both will be punished as outlined above (so do not share your solutions with others)

Use the tutor sessions

universität koblenz
weiter:denken

- You can ask any module-related questions
- There will be a dedicated thread in the **forum** on OLAT for questions for the tutor sessions
- During the sessions we aim to navigate through the questions in a plausible manner.
- However, you can (and should) ask more questions when they come up during the sessions