

Seminar CS715

Large-Scale Data Integration

Data Lakes



- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
 - Web Data Integration
 - Data and Web Mining
 - Data Web Technologies
- Room: B6 - B1.15
- eMail: chris@informatik.uni-mannheim.de
- Consultation: Wednesday, 13:30-14:30



- **Ralph Peeters**
- Graduate Research Associate
- Research Interests:
 - Entity Matching using Deep Learning
 - Product Data Integration
- Room: B6, 26, C 1.04
- eMail: ralph@informatik.uni-mannheim.de



- **M. Sc. Wi-Inf. Alexander Brinkmann**
- Graduate Research Associate
- Research Interests:
 - Data Search using Deep Learning
 - Product Data Categorization
- Room: B6, 26, C 1.03
- eMail: alex.brinkmann@informatik.uni-mannheim.de



You and Your Experience

- A Short Round of Introductions
 - What are you studying?
 - Which DWS courses did you attend?
 - What kind of experience do you have with Data Science/Data Engineering projects?
- Participants
 1. Schmidt, Thomas Günther
 2. Garza, José
 3. Omogha, Oghenekeno Utomudo
 4. Agarwal, Mayank

Agenda of Today's Kickoff Meeting

1. Seminar organization
2. Seminar topics
3. How to structure your seminar paper / presentation?
4. Your Questions

1. Organization

Learning Targets

- Writing a seminar thesis as an exercise for your master thesis
- Understanding and presenting state-of-the-art scientific work
- Searching and citing scientific papers / journal articles
- How to structure your thesis and presentation
- How to argue, how to explain, how to write!
- How to write a nicely formatted paper using LaTeX

Schedule

Date	Session
Friday, 17.03.2021, 9:00	Kick-off meeting
	Read papers about your topic and search for additional literature Prepare outline and argumentation line for the presentation
Until Friday 23.04.2022	Meet with your mentor to discuss your presentation
	Prepare draft of your presentation
Until Sun. 08.05.2022	Send draft presentation to your mentor
	Finalize your presentation
Friday, 20.05.2021 (10:00-12:30)	Presentation and discussion of your topic (30 % of your final grade)
	Write seminar thesis
Sunday, 10.07.2021	Submission of your seminar thesis (70 % of your final grade)

Formal Requirements

- Presentation
 - 15 minutes + 10 minutes discussion
 - should be 100% understandable for all participants
- Written report (paper)
 - 12-15 pages single column
 - including abstract and appendixes
 - not including bibliography
 - every additional page reduces your grade by 0.3
 - written in English language
 - use latex template of Springer Computer Science Proceedings
 - <http://www.springer.com/de/it-informatik/Incs/conference-proceedings-guidelines>
- Final grade
 - 70% written report
 - 30% presentation

Which template to use?

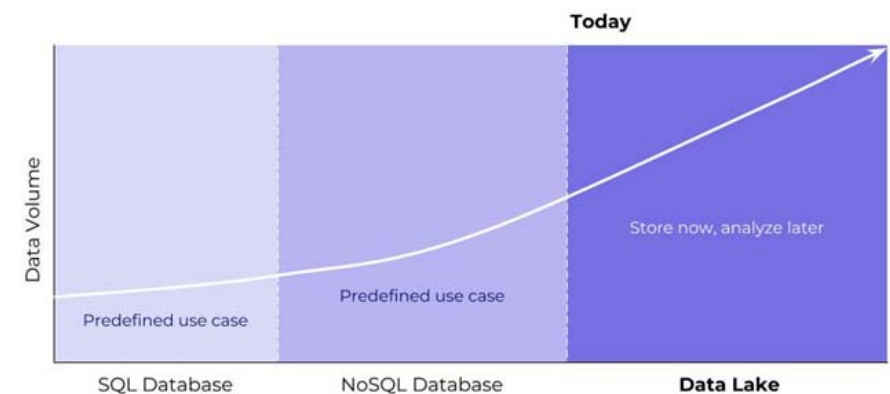


<http://www.springer.com/de/it-informatik/Incs/conference-proceedings-guidelines>

2. Seminar Topics

Data Lakes

- are repositories of raw data in different formats
- collect or generate metadata about the datasets
- provide a common access interface
- target users: data scientists
- different, not yet known use cases
- are used in a schema-on-read fashion



The Dataspace Vision

Alternative to classic data integration systems in order to cope with growing number of data sources.

Properties of dataspaces

- may contain any kind of data (structured, semi-structured, unstructured)
- require no upfront investment into a global schema
- provide for data-coexistence
- provide give best effort answers to queries
- rely on pay-as-you-go data integration

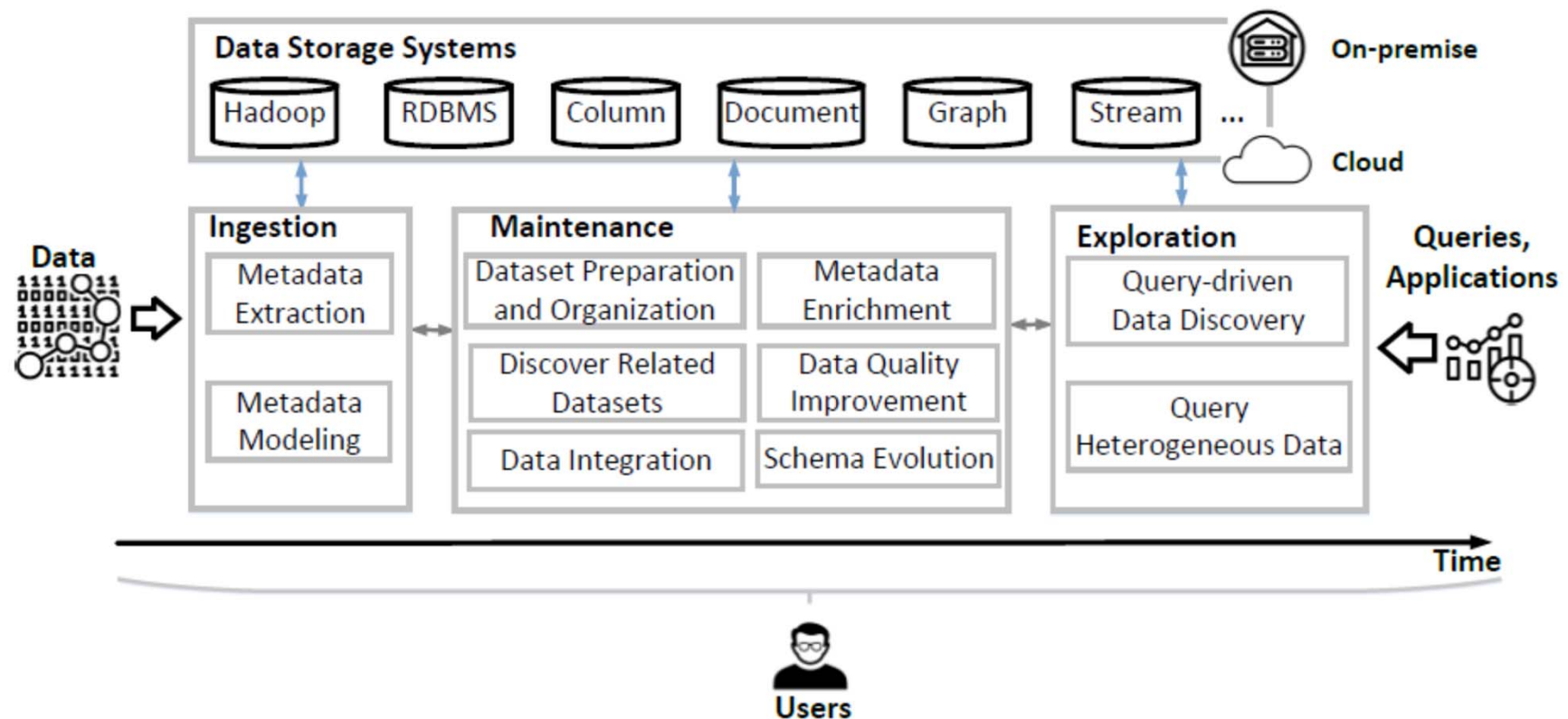


Franklin, M., Halevy, A., and Maier, D.: From Databases to Dataspaces
A new Abstraction for Information Management, SIGMOD Rec. 2005.

Madhavan, J., et al.: Web-scale Data Integration: You Can Only Afford to Pay As You Go, CIDR 2007.

Data Lakes

Criteria	Data Warehouses	Data Lakes
<i>Data ingestion</i>	ETL	Load-as-is
<i>Ingested data format</i>	Structured	Heterogeneous (structured, semi-structured, and unstructured)
<i>Data storage</i>	Relational databases	Hadoop, Relational databases, NoSQL data stores, etc
<i>Data access</i>	SQL queries (OLTP, OLAP)	Different query languages (e.g., SQL, Cypher), programming languages (e.g., Java, Python, R)



Kdl/#Ulkdq/#Fkulwrsk#Txl{/#dgg#P dwkldv#Mdnh=Gdw#Odnh#FrgfhsW#dgg#V |vwhp v=#D#Vxuyh|153541

Goals of the Seminar

1. review the concept of data lakes and their use cases
2. review the functionality of current data lake management systems
3. dig deeper into specific types of functionality, e.g. search

- **1. Data Lakes: Concepts, Functionalities, Examples**
- Student: Garza, José
- Mentor: Ralph Peeters
- Hai, Rihan, Christoph Quix, and Matthias Jarke. “Data lake concept and systems: a survey.” arXiv preprint arXiv:2106.09592 (2021).
- Sawadogo, Pegdwendé, and Jérôme Darmont. “On data lake architectures and metadata management.” *Journal of Intelligent Information Systems* 56.1 (2021): 97-120.
- Nargesian, Fatemeh, et al. “Data lake management: challenges and opportunities.” *Proceedings of the VLDB Endowment* 12.12 (2019): 1986-1989.

- **2. Comparison of Data Lake Management Platforms (Set 1)**
- Student: Schmidt, Thomas Günther
- Mentor: Alexander Brinkmann
- Hai, Rihan, Christoph Quix, and Matthias Jarke. “Data lake concept and systems: a survey.” arXiv preprint arXiv:2106.09592 (2021)
- Sawadogo, Pegdwendé, and Jérôme Darmont. “On data lake architectures and metadata management.” *Journal of Intelligent Information Systems* 56.1 (2021): 97-120.
- Platforms to compare:
 - Goods from Google
 - Azure Data Lake from Microsoft
 - AWS Lake from Amazon

- **2. Comparison of Data Lake Management Platforms (Set 2)**
- Student: Agarwal, Mayank
- Mentor: Alexander Brinkmann
- Hai, Rihan, Christoph Quix, and Matthias Jarke. “Data lake concept and systems: a survey.” arXiv preprint arXiv:2106.09592 (2021)
- Sawadogo, Pegdwendé, and Jérôme Darmont. “On data lake architectures and metadata management.” *Journal of Intelligent Information Systems* 56.1 (2021): 97-120.
- Platforms to compare:
 - Snowflake
 - Data Bricks Delta Lake
 - SAP VORA

- **4. Dataset Search within Data Lakes**
- Student: Omogha, Oghenekeno Utomudo
- Mentor: Alexander Brinkmann
- Chapman, A., Simperl, E., Koesten, L. *et al.* Dataset search: a survey. *The VLDB Journal* (2019)
- Trabelsi, et al.: Improved Table Retrieval Using Multiple Context Embeddings for Attributes. Big Data 2019.
- Chapman, Adriane, et al. “Dataset search: a survey.” *The VLDB Journal* 29.1 (2020): 251-272.
- Erkang Zhu, Dong Deng, Fatemeh Nargesian, and Renée J. Miller. 2019. JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes. SIGMOD 2019.

3. How to Structure Your Paper / Presentation

Goal of Seminar Paper

- A seminar paper differs significantly from a master thesis
 - The topic is already defined
 - No need to implement or develop algorithms
 - No need to perform experiments
 - Primarily: reproduction and re-organization of content that is already available
- Goals of seminar paper
 1. Describe the problem / task
 2. describe several existing methods/systems for handling the task,
 3. compare the methods/systems and their evaluation using a systematic set of comparison criteria

How to Structure Your Paper?

1. Introduction and Problem Statement

- Which problem/task is addressed? Why is the problem important?
- Structure of your paper

2. Description of Existing Approaches

- Overview of existing methods and features used by the methods
- Detailed description of **selected methods**
- Comparison of the selected methods using a **set of comparison criteria**

3. Evaluation

- Comparison and **discussion of the evaluation tasks**, metrics
- Comparison of the evaluation results

4. Conclusion

- What did the comparison of the methods and evaluation results show?
- Can something be concluded for future work?

5. Bibliography

Learn from Examples

- Read survey articles and identify the structure from the previous slide
 - Why can this paragraph be found at that position?
 - What is the purpose of some section / subsection?
- Important
 - Read survey articles!
 - Read conference or journal papers.
- Textbook on how to write a thesis
 - Zobel: Writing for Computer Science, 3rd Edition, Springer 2014.
- University Library: Academic Writing Consultancy
 - <https://www.bib.uni-mannheim.de/en/writing-consultancy/>
 - Open consulting hour: every Wednesday 10 am - noon

Citing different Types of Publications

- Journal article
 - Good to cite, current research results
 - Survey articles (very good for an overview)
- Conference and workshop paper
 - Good to cite, current research results
- Books (sometimes cited)
 - Textbooks
 - Collections of articles/papers => Cite specific paper in book
- Websites
 - better not cited, exceptions are, e.g., W3C Specifications
 - Wikipedia is not an exception!!! **Do not cite Wikipedia, ever!**
- Slide sets
 - **Never cite!**

How to Find Relevant Publications?

- Use Standard Search Engines
- **Use Google Scholar**
 - we use it a lot ourselves
- Search Engines of the University's library
 - see slides from the library course
- **Exploit references:** Given a relevant document x
 - Follow references in the past: papers y that x has cited
 - Follow references in the future: papers y that cited x („**cited by**” functionality in Google scholar)

4. Questions?

