Data Lakes Workshop

Julian Kunkel, Hendrik Nolte, Piotr Kasprzak

Agenda

- Welcome
- Short introductory round of all attendees
- Presentation of individual use cases (~10 min per user presentation)
  - CryoEM facility at UMG – Tat Cheng (UMG, Göttingen)
  - UMG-MeDIC: Establishing a Medical Research Data Sercive Unit – Markus Suhr (UMG, Göttingen)
  - Prediction of neurodevelopmental disorders in young children using multi sensory data analysis – Tomas Kulvicius (Uni Göttingen)
- GWDG data lake services and future plans
  - Activities at the GWDG – Julian Kunkel
  - Approaches for the GWDG data lake – Hendrik Nolte
  - Outlook GWDG Infrastructure Development – Piotr Kasprzak

15:30 Break and networking

- Groupwork: similarities of use cases and potential approaches
  - Concluding group discussion
- Conclusions

The workshop is considered to be highly interactive and bring us together as community
Please let's make good use of questions and interact and bring us all forward!
Motto: Let's build a bridge to the data lake together

Prof. Dr. Julian Kunkel

1) Personal details

- Since June/21 in Göttingen; Deputy Head GWDG for High-Performance Computing
- PhD Computer Science, University of Hamburg/DKRZ
- 2018-2021 Lecturer at the University of Reading

2) Research focus / field of research

- Data-driven workflows
- Parallel file systems
- Application of machine learning
- Performance portability
- Data reduction techniques
- Management of cluster systems
- Performance analysis of parallel applications&I/O
- Software engineering of scientific software
- Personalized teaching

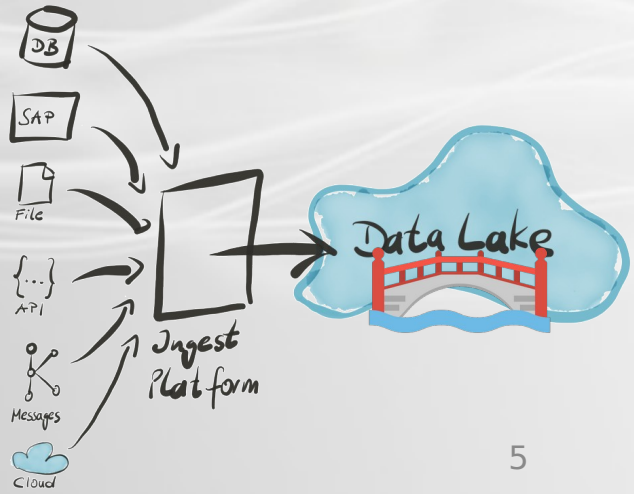
3) **Contact: julian.kunkel@gwdg.de**

Hendrik Nolte

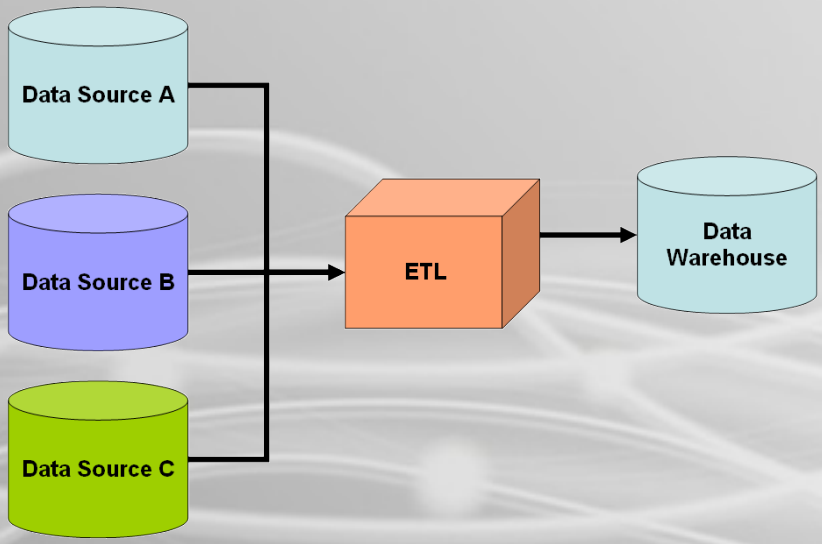
- Personal Details
  - Since November/19 at GWDG in the HPC-Team
  - B.Sc./M.Sc. Physics at University of Göttingen
- Research Focus
  - Data Lakes (in the broadest sense); focus:
    - General scalability
    - Agreement with good scientific practice (e.g. reproducibility)
    - Support of the full lifecycle of a data driven project
    - Processing of sensitive data

Introduction: Data Lake

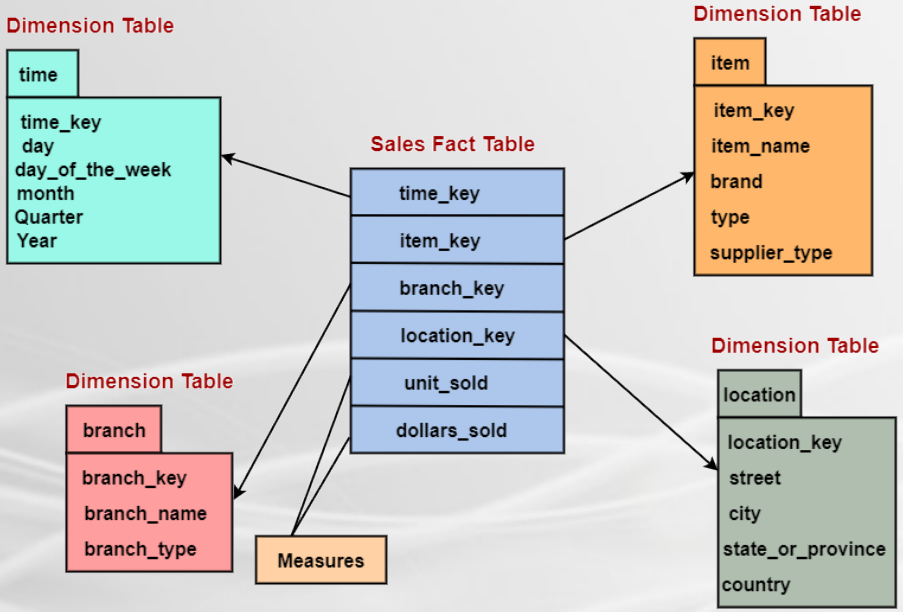
- With cheap storage costs, people promote the concept of data lake
  - Combines data from many sources and of any type
  - Allows for conducting future analysis and not miss any opportunity
- Attributes of the data lake
  - **Dump** everything: all time all data: raw sources and processed data
    - Decide during analysis which data is important, e.g., no “schema“ until read
  - **Dive** in anywhere: enable users across multiple business units to
    - Refine, explore and enrich data on their terms
  - **Fishing** for data, i.e., flexible access
    - Shared infrastructure supports various patterns
    - Usage: Batch, interactive, online, search



Contrast to: Data Warehouse



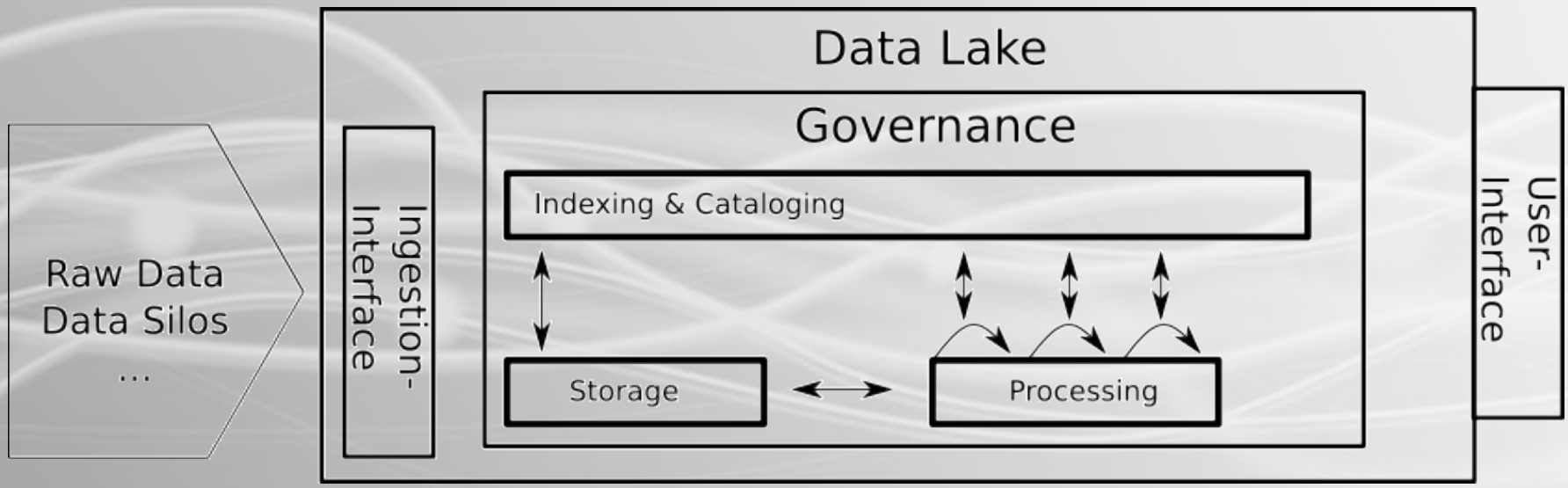
https://en.wikipedia.org/wiki/Data_integration#/media/File:Datawarehouse.png



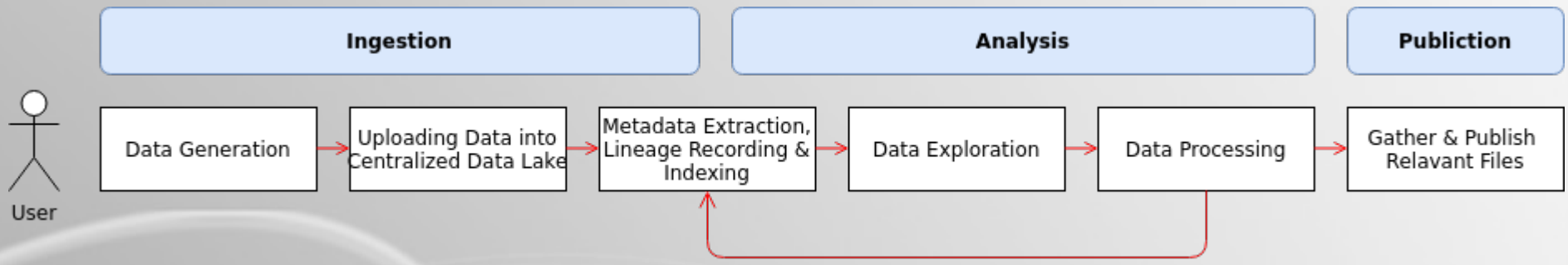
https://www.javatpoint.com/data-warehouse-what-is-star-schema

- Strict structured data
- Motivation: Business Use Case

Data Lake
Basic Architecture



User Workflow



Multiple users are typically involved

Data Lake Projects at GWDG

- Advising users in data management procedures & procurements
- Deployment of a data and processing infrastructure
  - Ceph based S3 storage
  - Virtual machines for smale-scale processing
  - HPC enabled for large-scale processing
- Development of a data lake solution
  - Combines various tools under one management and control
  - Additional software layers, e.g., for governance
- Secure data processing for highest privacy levels
  - Hardened workflow
- We will organize a follow up meeting for the NHR