Data Analysis and Integration

Welcome

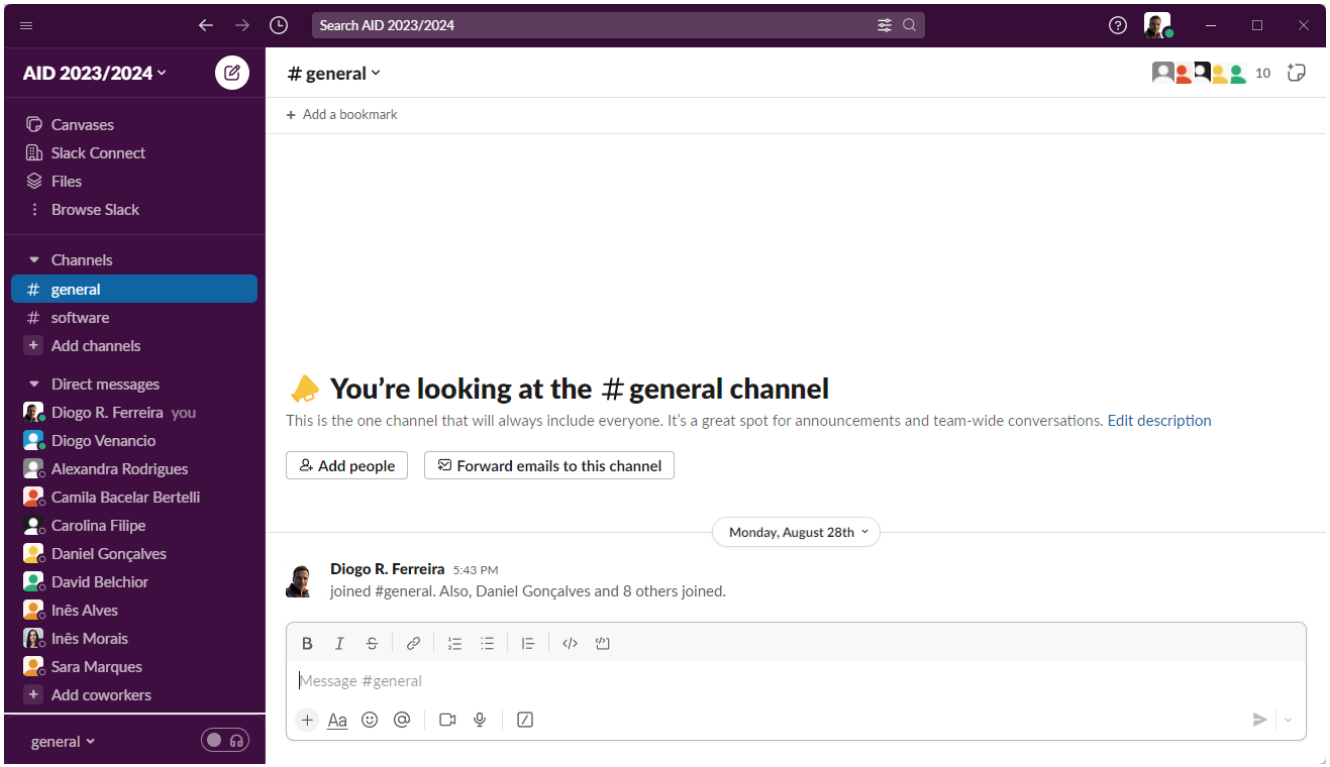
Teaching staff

- Prof. Diogo Ferreira
  - Lectures + Labs@Tagus
- Asst. Daniel Gonçalves
  - Labs@Alameda
- Asst. Inês Morais
  - Labs@Alameda



How to reach us

- Slack workspace
  - sign up with your e-mail address **@tecnico.ulisboa.pt**
  - https://join.slack.com/t/aid2023/signup

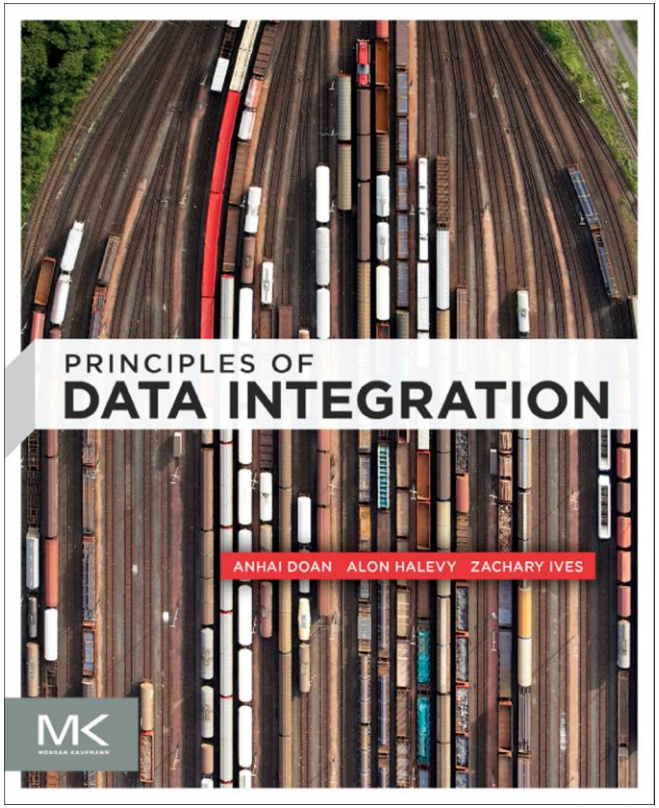


Topics

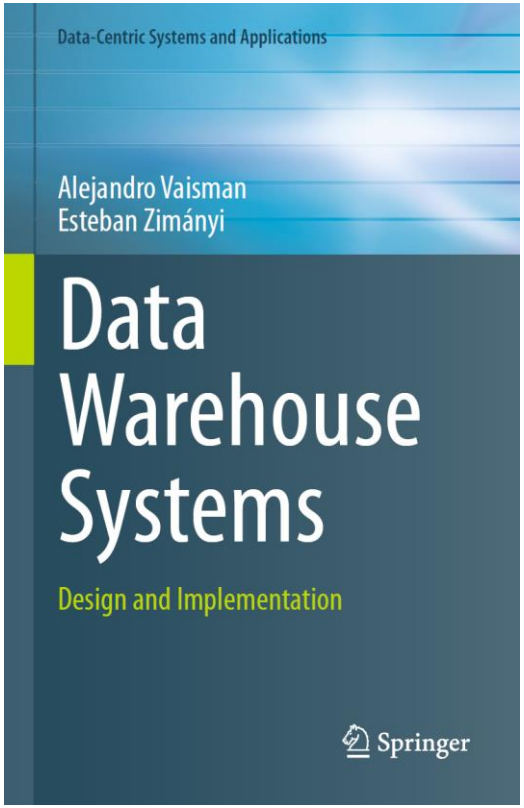
- Data integration
  - combine data from multiple sources (files, databases, etc.)
  - create mappings between different data structures
  - compare records and detect potential duplicates
  - analyze data quality
- Data analysis
  - design a data warehouse for multi-dimensional analysis
  - develop an ETL process to populate the DW tables
  - analyze data with OLAP cubes and operations
  - explore the data with MDX queries and reporting

Bibliography

- Principles of Data Integration, Doan/Halevy/Ives, Morgan Kaufmann (2012)
- Data Warehouse Systems, Vaisman/Zimányi, Springer (2014)



(selected topics)



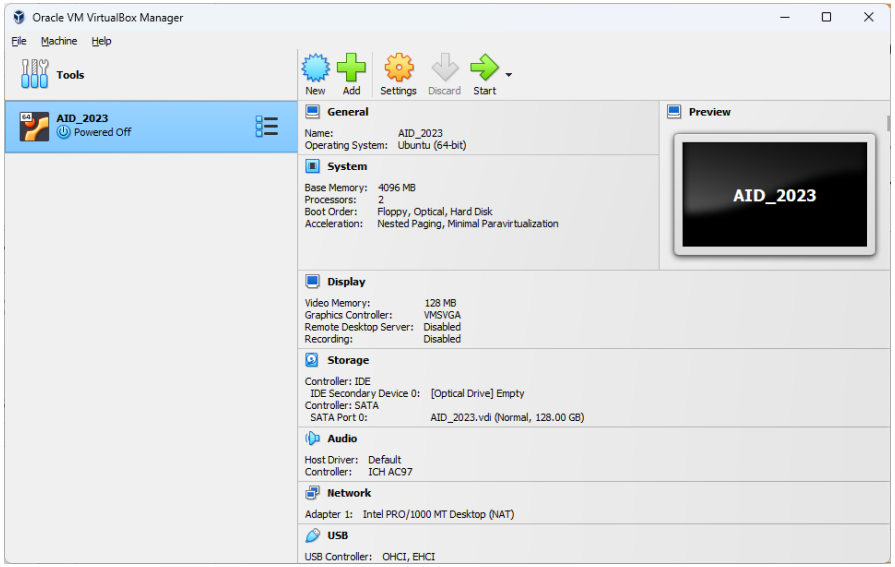
(selected topics)

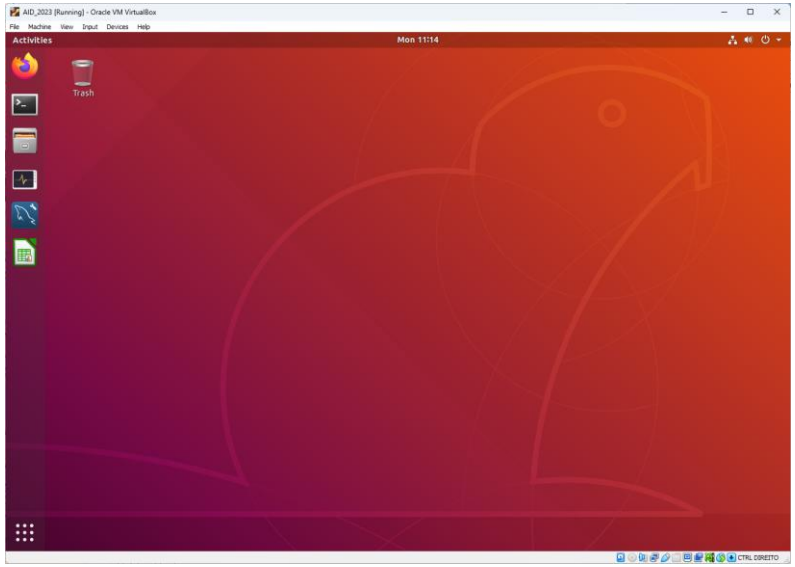
Tools

- Software
  - MySQL Server & Workbench
  - Pentaho Data Integration (PDI)
  - DataCleaner
  - Pentaho Schema Workbench (PSW)
  - Pentaho Server & Saiku Analytics
  - Pentaho Report Designer (PRD)

Tools

- Software is provided as a Virtual Machine (VM)
  - VirtualBox runs on Windows, Linux, Mac (Intel)
  - http://groups.tecnico.ulisboa.pt/aid-meic/virtualbox/
  - requirements: 8 GB of RAM, to assign 4 GB to the VM





Labs

- Each lab has a **lab guide**:
  - Lab 0: Preparing the virtual machine
  - Lab 1: Review of SQL
  - Lab 2: SQL Views and Data Integration
  - Lab 3: Introduction to ETL tools
  - Lab 4: String matching
  - Lab 5: Duplicate detection
  - Lab 6: Data profiling
  - Lab 7: Creating a data warehouse
  - Lab 8: OLAP cubes and business analytics
  - Lab 9: The MDX query language
  - Lab 10: Reporting

Labs

- Labs: take a screenshot (example)

34. Find the sum of salaries by department. The results should look like this:

```
+---------+--------------------+------------+
| dept_no | dept_name          | sum_salary |
+---------+--------------------+------------+
| d005    | Development        |    4434974 |
| d007    | Sales              |    3715959 |
| d004    | Production         |    2928341 |
| d009    | Customer Service   |    1914195 |
| d002    | Finance            |    1492870 |
| d001    | Marketing          |    1249477 |
| d006    | Quality Management |    1212103 |
| d008    | Research           |    1064935 |
| d003    | Human Resources    |     643182 |
+---------+--------------------+------------+
```

Take a screenshot of your query and results and submit it on Fénix for lab credit.

Evaluation

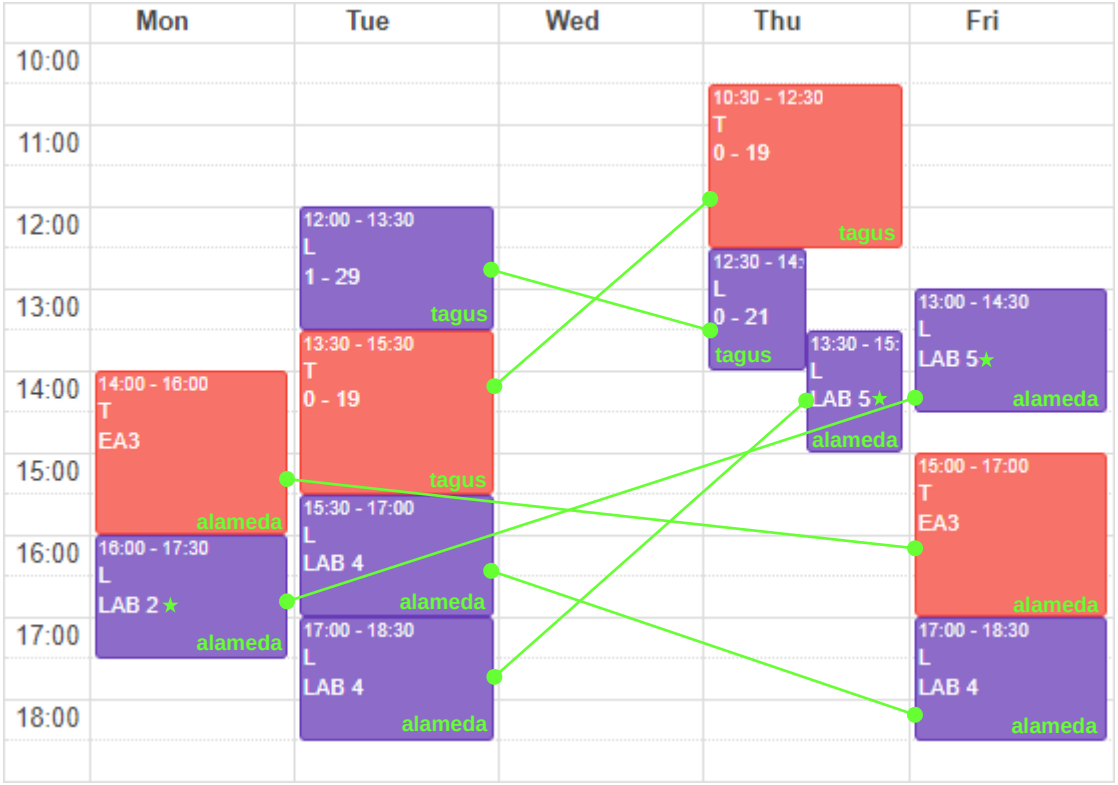
- Evaluation formula:

  Final grade = 15% Labs + 35% Project + 50% Exam

  - labs: submit screenshot on Fénix (weekly 2x)
  - project: submit report on Fénix
  - exam: minimum grade 8.0 (without rounding)

Groups

- Registration
  - groups of 2, registration begins after this lecture



★ big room

Important dates

- Labs & project

Projects	Beginning	End
Project: Lab 1	18/09/2023 00:01	22/09/2023 23:59
Project: Lab 2	18/09/2023 00:02	22/09/2023 23:59
Project: Lab 3	25/09/2023 00:01	29/09/2023 23:59
Project: Lab 4	25/09/2023 00:02	29/09/2023 23:59
Project: Lab 5	02/10/2023 00:01	06/10/2023 23:59
Project: Lab 6	09/10/2023 00:01	13/10/2023 23:59
Project: Lab 7	09/10/2023 00:02	13/10/2023 23:59
Project: Lab 8	16/10/2023 00:01	20/10/2023 23:59
Project: Lab 9	16/10/2023 00:02	20/10/2023 23:59
Project: Lab 10	23/10/2023 00:01	27/10/2023 23:59
Project: Project	23/10/2023 00:02	27/10/2023 23:59

Important dates

- Exams

Tests/Exams	Day	Beginning	End	
Exam: 1ª Época	07/11/2023	08:00	10:00	alameda
Exam: 1ª Época	07/11/2023	08:00	10:00	tagus
Exam: 2ª Época	30/01/2024	13:00	15:00	alameda
Exam: 2ª Época	30/01/2024	13:00	15:00	tagus
Exam: Época Especial	24/07/2024	13:00	15:00	alameda
Exam: Época Especial	24/07/2024	13:00	15:00	tagus

Hitachi prize

- To be awarded to a group of students in this course
- Sponsored by Hitachi, provider of Pentaho software
- Presentation at the company, prize awarded by jury