Comparison of database systems based on their handling of spatial big data

Adrian Roser
Mat.Nr.: 47938
Hochschule Karlsruhe Technik und Wirtschaft
Studiengang Geomatics

disy

Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

Content

1. General Topic
2. Milestones
3. Database systems
4. Used data
5. Comparison of the database systems
    1. Creating a benchmark
    2. Used Queries
    3. Soft factors
6. Outlook on new developments
7. Summary

General Topic

- Thesis is about handling **spatial big data**
- Tries to **define** the term **big data** for specific use cases
- Highlights problems with geospatial data
- Gives an **overview** about **database systems** which handle geospatial data
- Tests different database systems in a **real world work environment**
- **Analysis** of the test

General Topic

The thesis tries to answer the question :

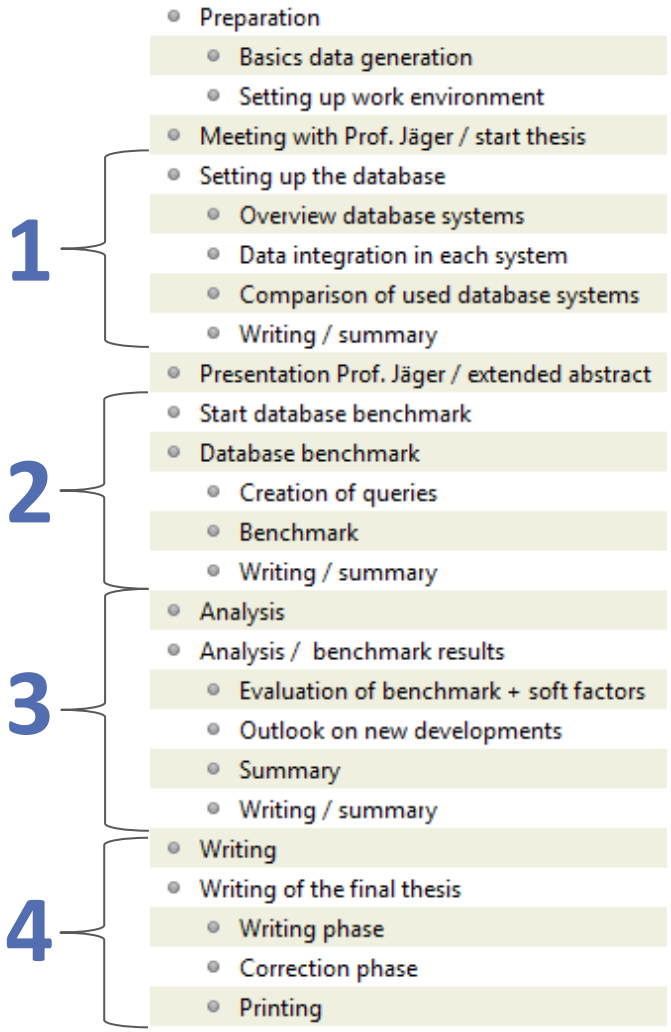
“How efficient are database systems, based on different technologies, managing big spatial data?“

- Especially in a real world work environment
- With different kind of data: real and synthetic

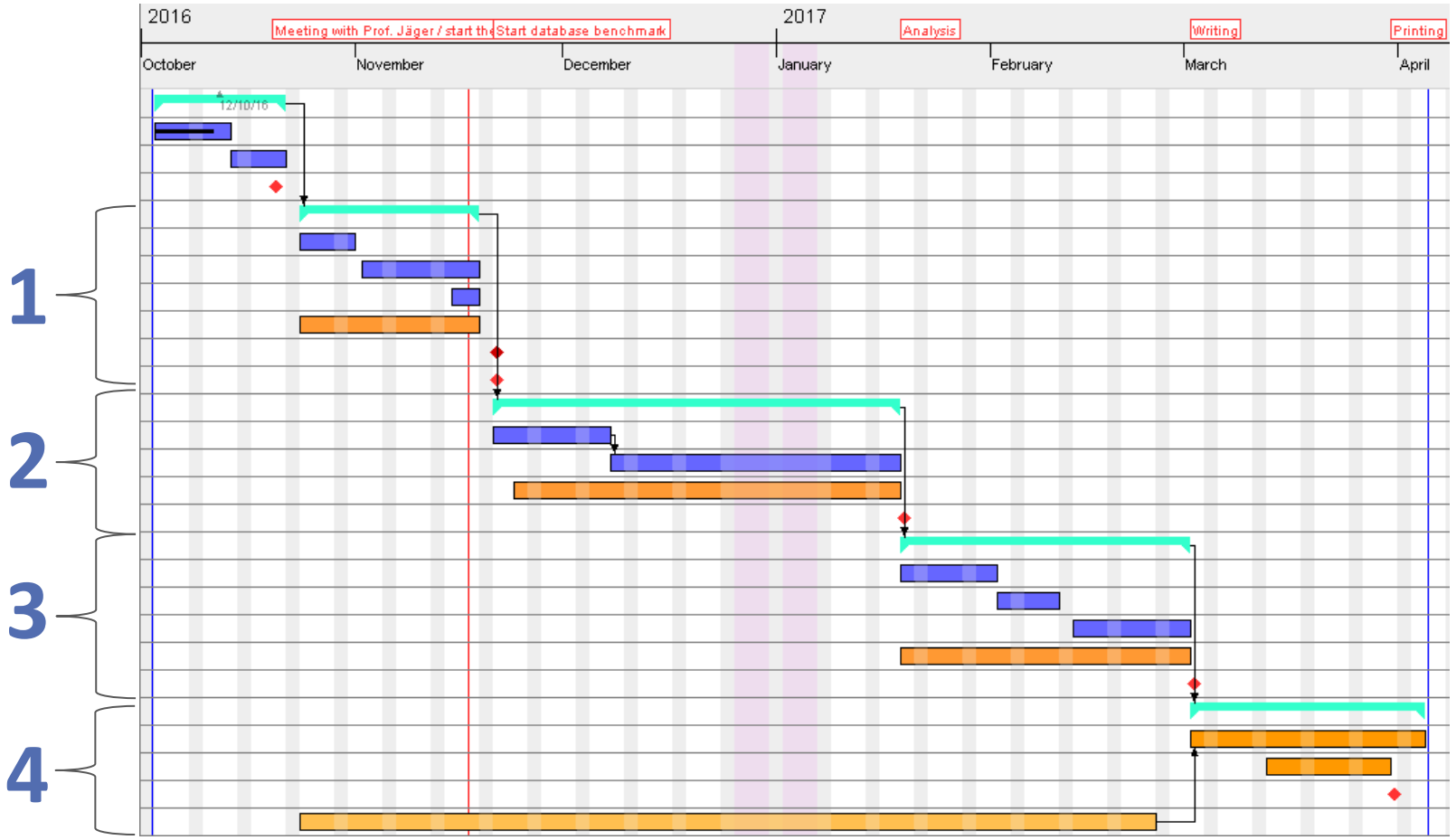
Milestones

There are 4 main milestones:

1. Setting up the databases
2. Benchmark
3. Analyze results
4. Writing phase



Milestones


2016
2017
Meeting with Prof. Jäger / start th
Start database benchmark
Analysis
Writing
Printing
October
November
December
January
February
March
April
12/10/16
1
2
3
4

Database systems

- There are a lot of different database systems
- Thesis aims to classify those
- Arguments for the database systems used are given

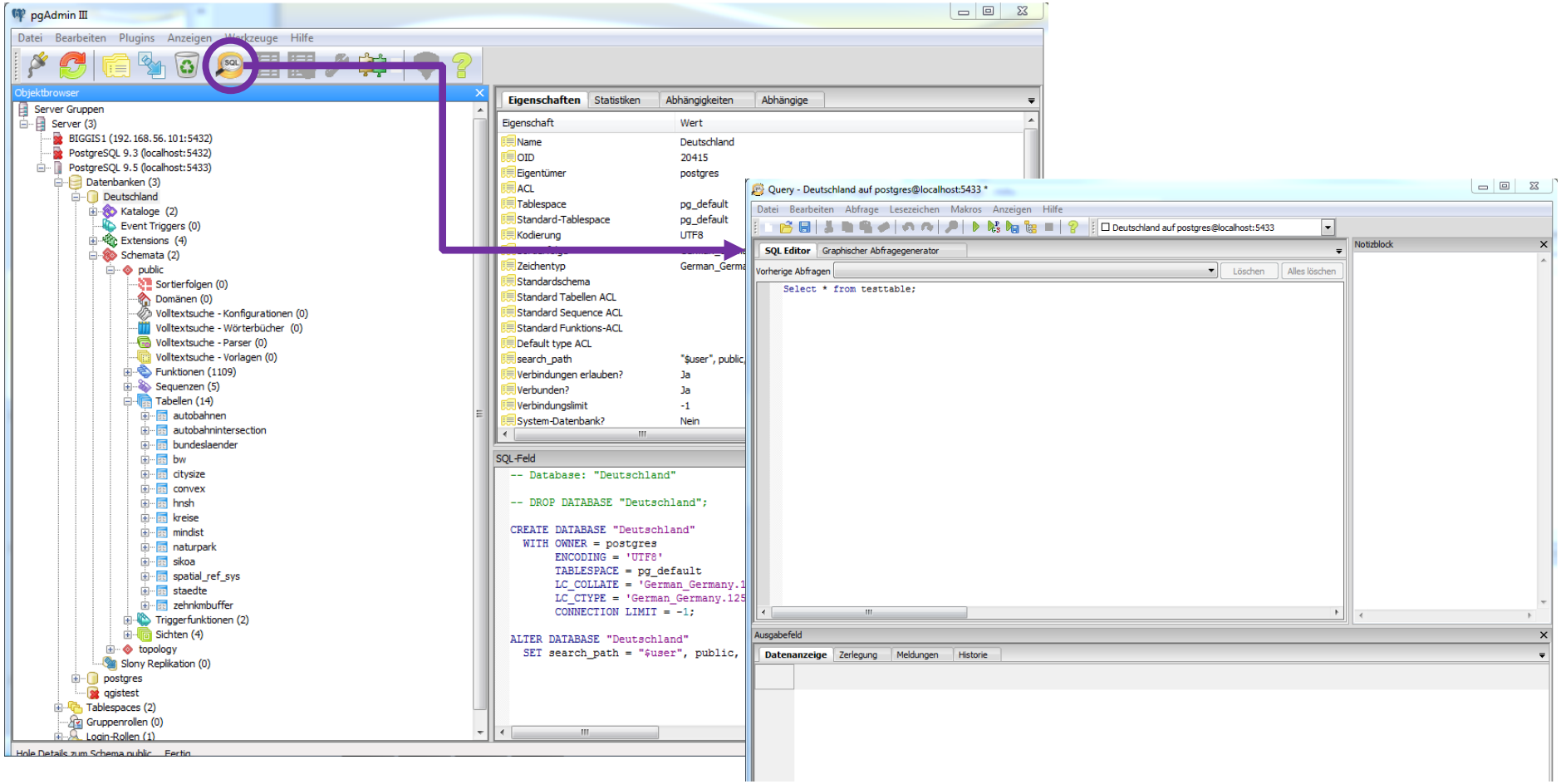
3 used Systems are:

- **PostgreSQL 9.6**
- **Oracle Database 12c**
- **EXASolo 5.0**

Database systems: PostgreSQL

- **Open source**
- One of the most popular open source systems
- Provides a wide range of analytical features especially regarding spatial data
- Lot of extensions e.g. **PostGIS**
- Technology comparable to Oracle Database
- Straightforward GUI with **pgAdmin3**

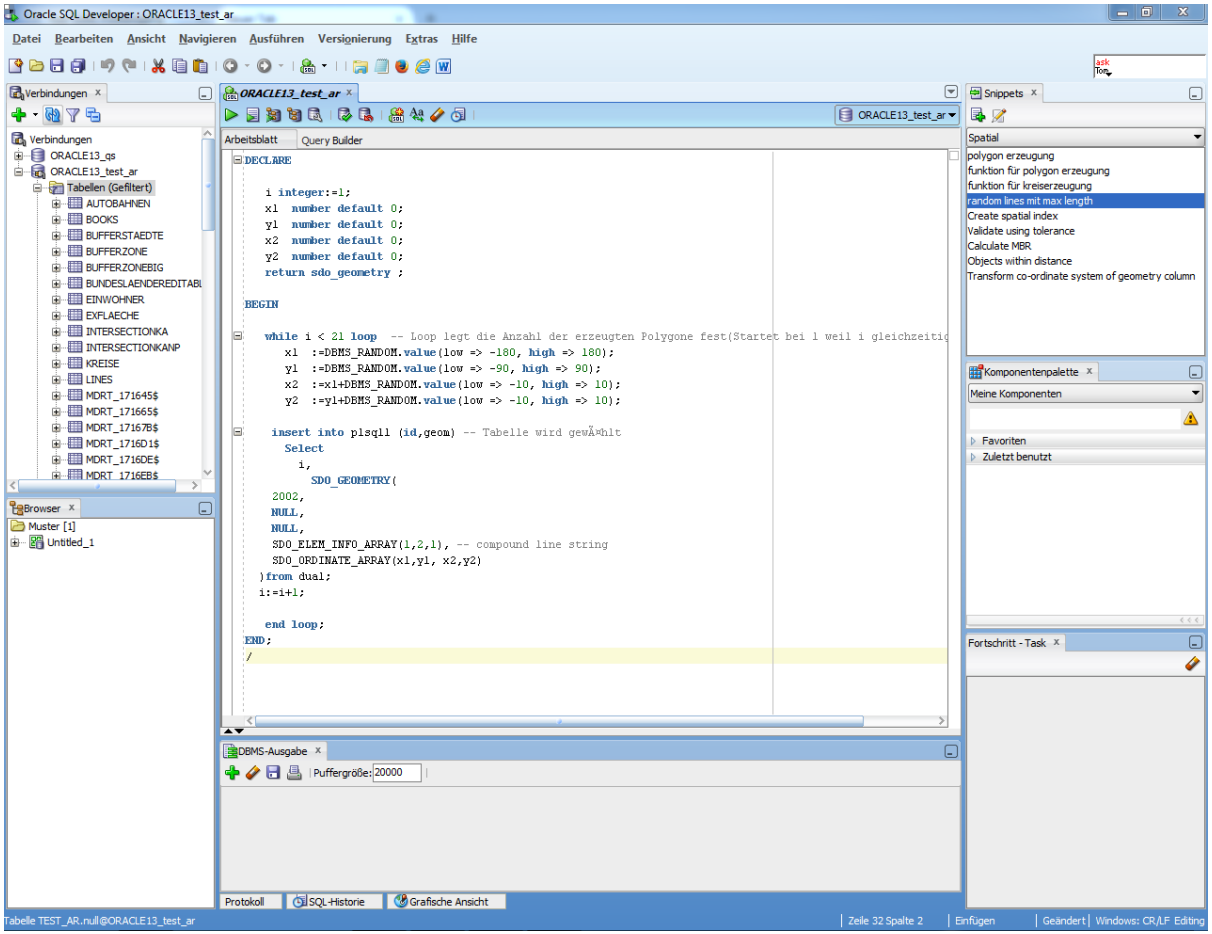
Database systems: PostgreSQL



Database systems: Oracle

- **Commercial database**
- Biggest market share
- Expensive
- Large amount of analytical tools
- Separate extensions like **Oracle Spatial**
- Free plug-ins e.g. **GeoRaptor**
- Old version of SQL Developer
- More complex GUI, **SQL Developer 3**

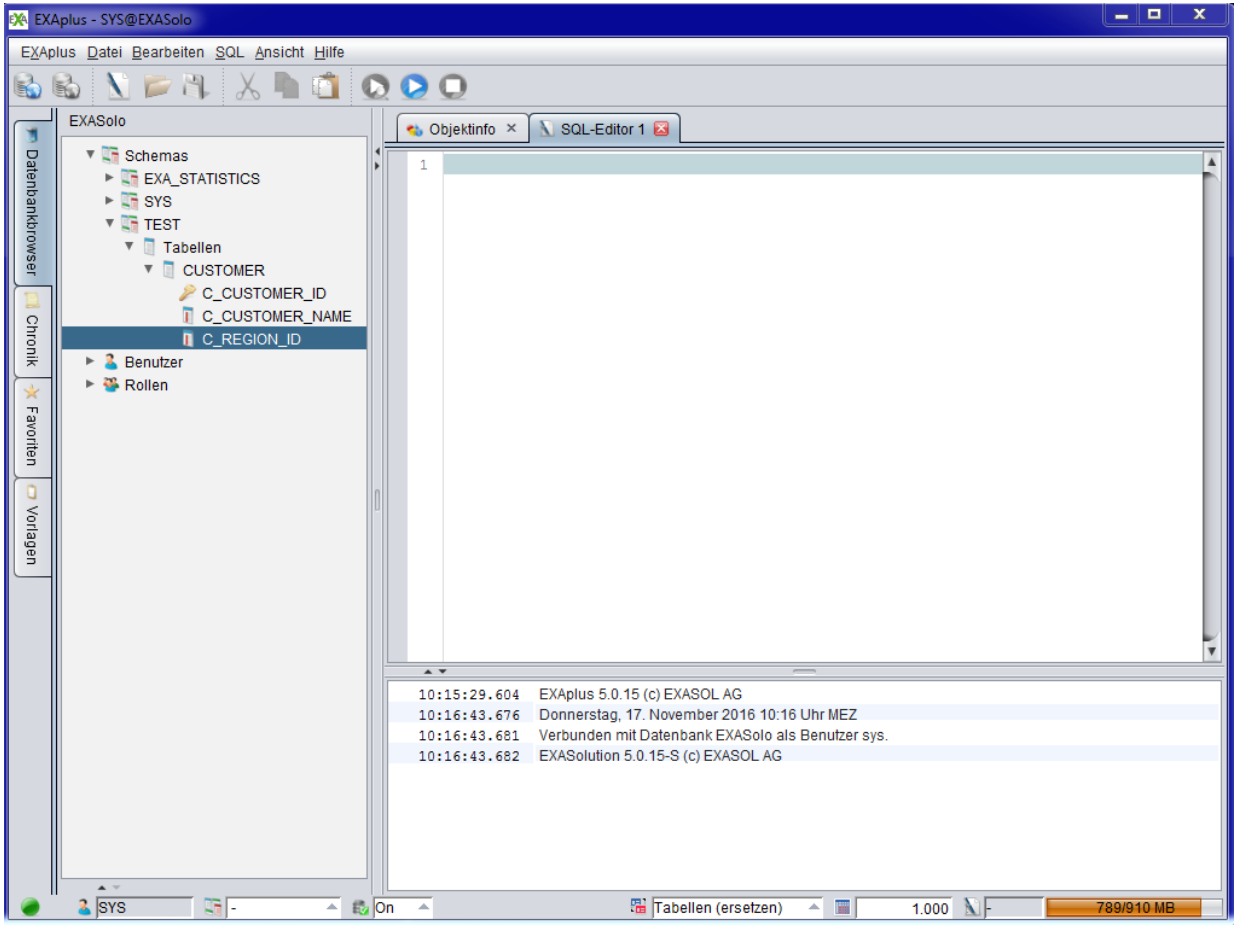
Database systems: Oracle



Database systems: EXASol

- **Commercial database**
- Uses **in-memory** technology
- License and hardware are expensive
- Very fast
- Relatively new database
- Not as many functionalities as the other test systems
- GUI **EXAPlus 5.0**

Database systems: EXASol



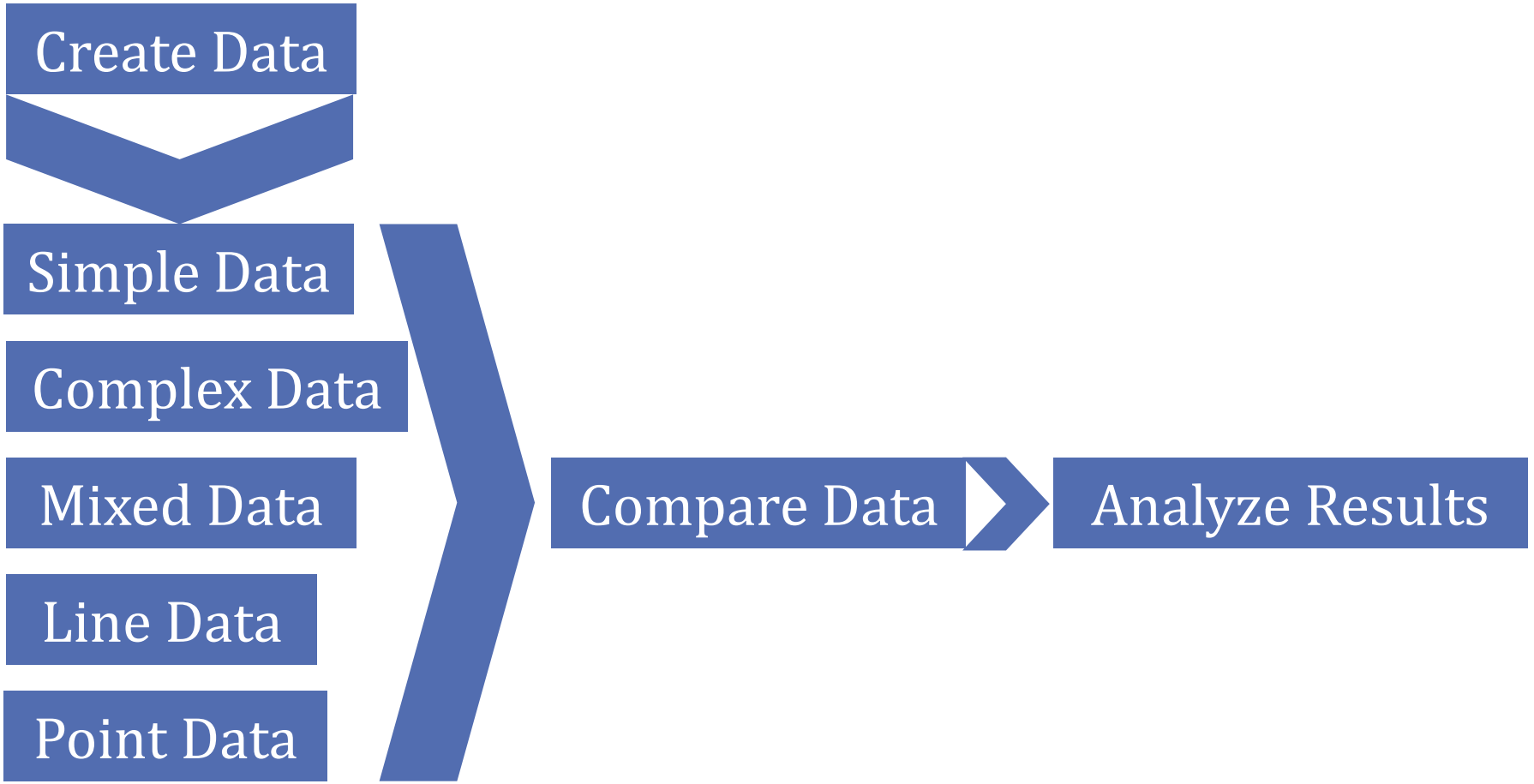
Used data

- **Two** different **data sets** are used in the comparison
- The difference is that one set is **synthetic data** and the other is **real world data**
- Both are **comparable** regarding **data volume**
- The data volume fits the limitations of the test environment

Used data: Synthetic data

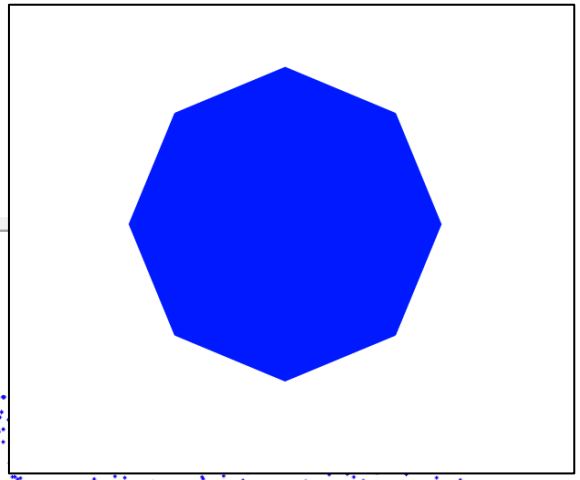
- Synthetic data is especially good for testing
- Created specifically for the thesis
- Realized in **PL/SQL**
- No underlying logic in the data, makes it interesting for benchmarks
- It is possible to generate **simple** or **complex data**
- This helps figuring out weaknesses of data base systems

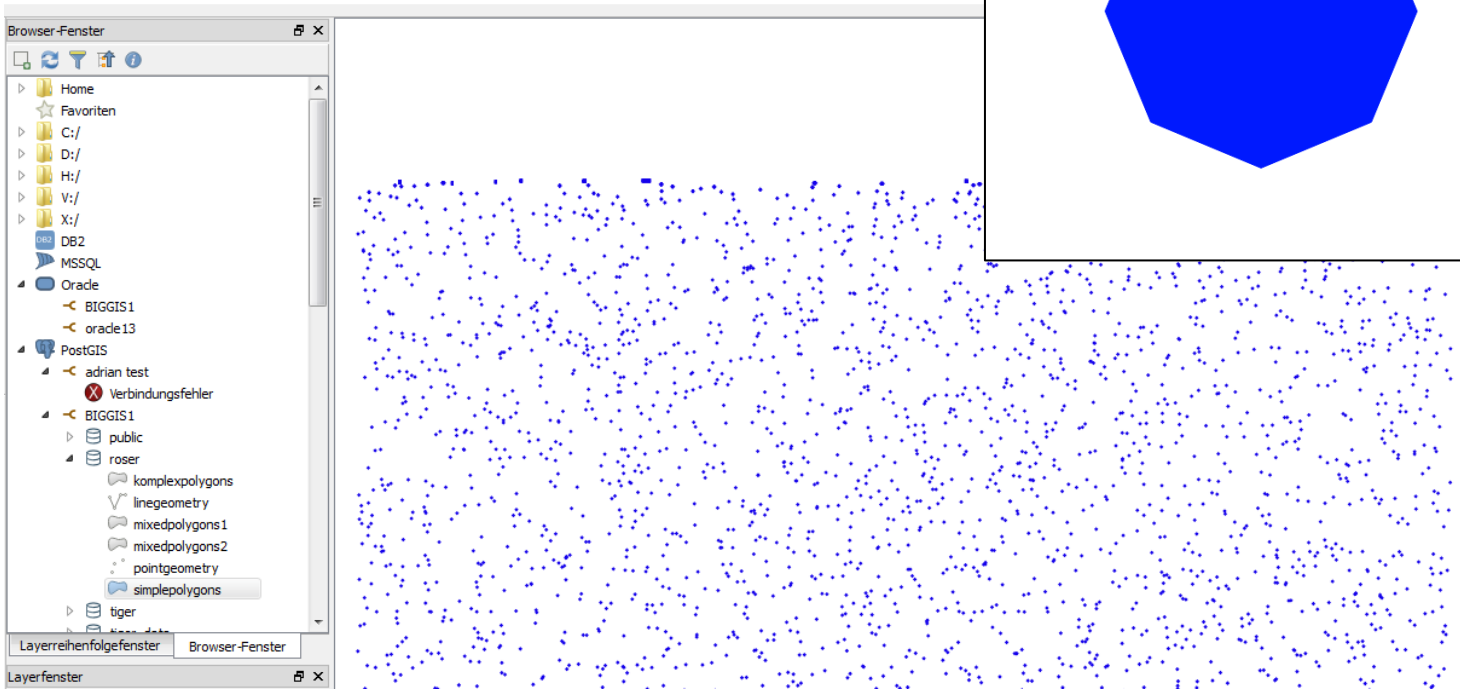
Used data: Synthetic data



Used data: Synthetic data

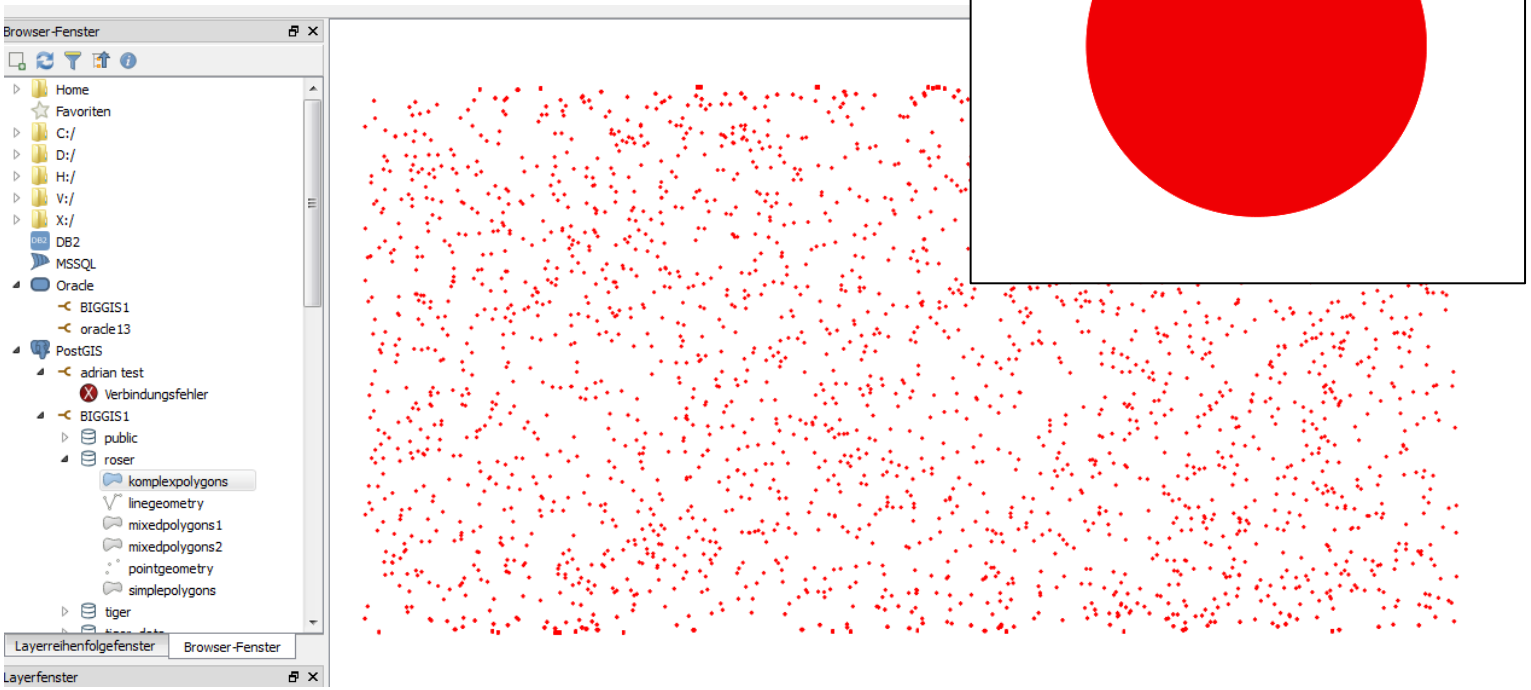
Simple Data





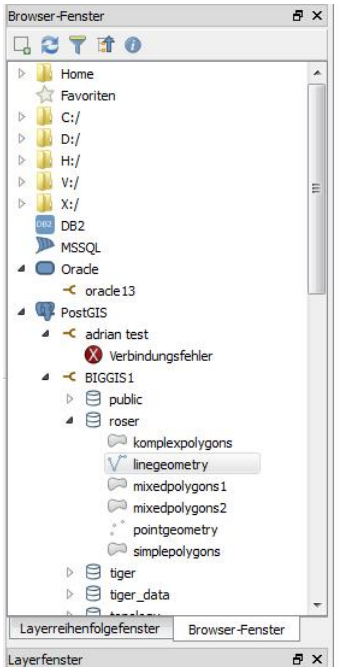
Used data: Synthetic data

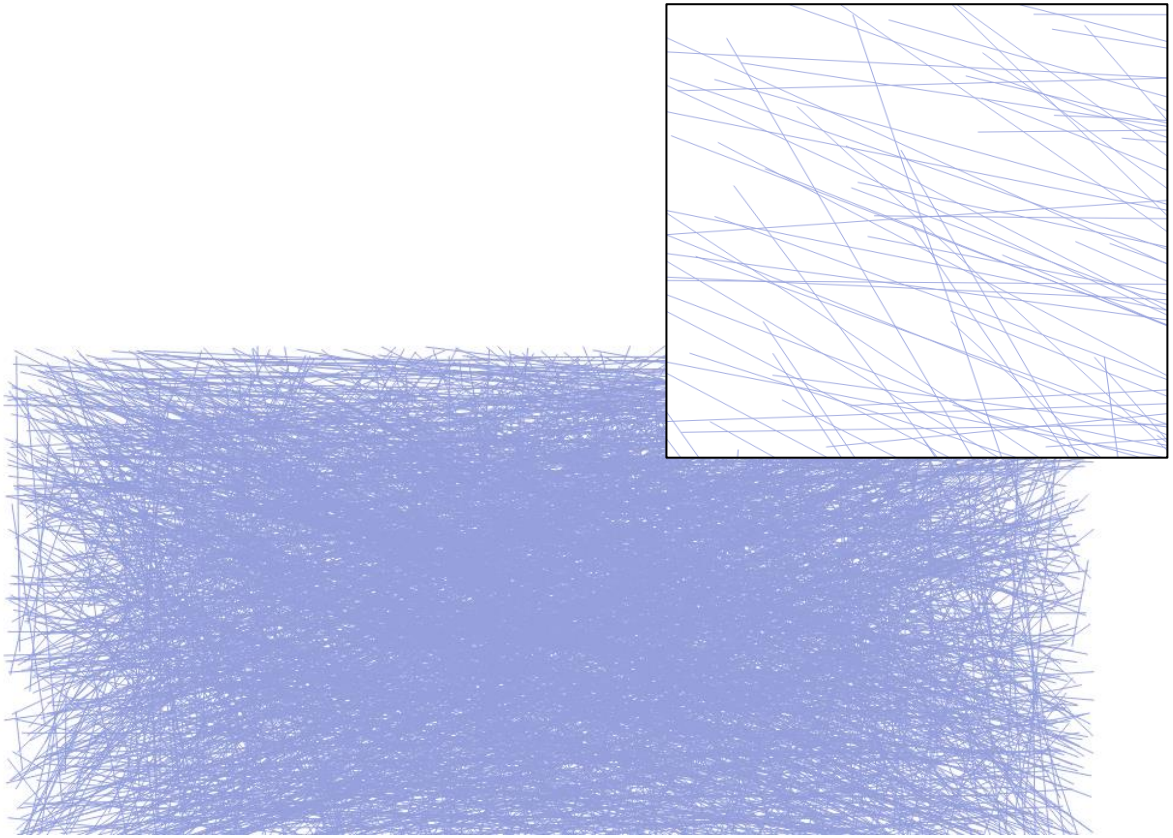
Complex Data



Used data: Synthetic data

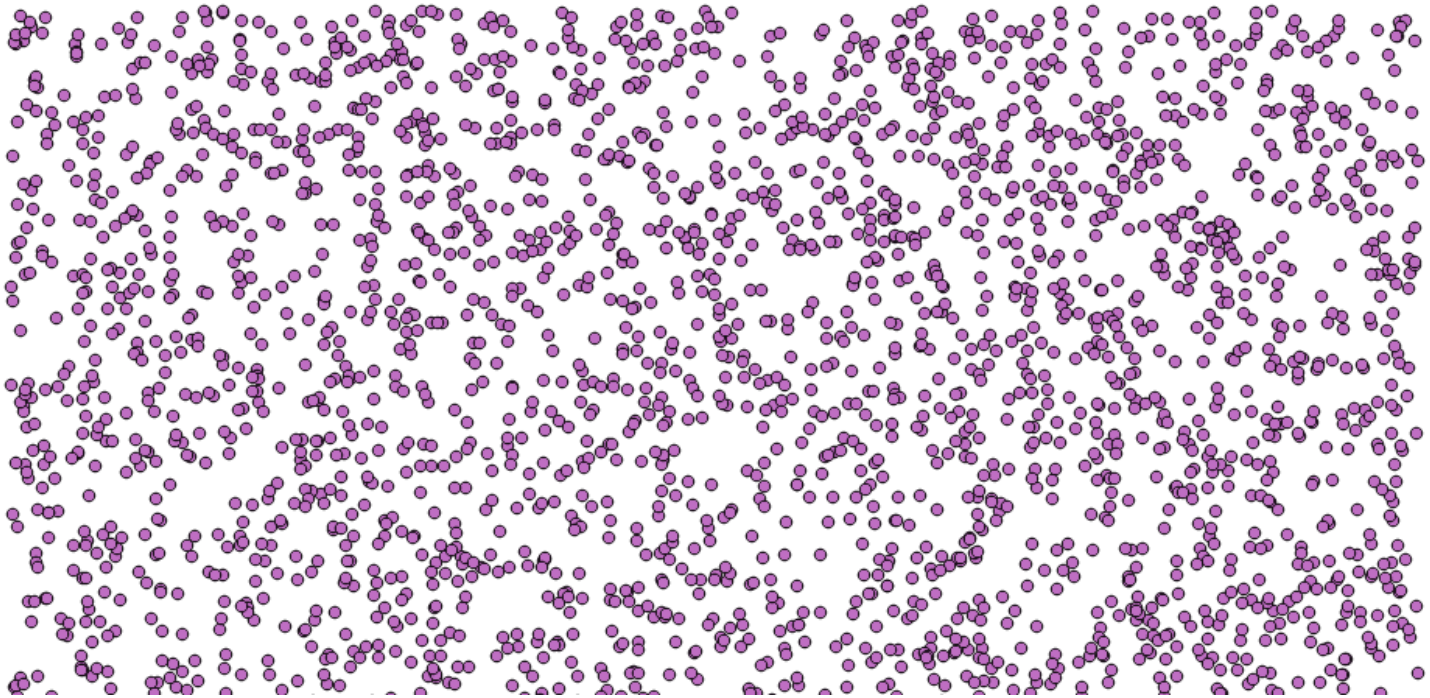
Line Data





Used data: Synthetic data

Point Data



Used data: Real world data

- Data from a **real world work environment**
- From **Disy Informationssysteme**
- With this data has already been worked with
- Good to use this data because the thesis aims to evaluate a realistic work environment
- Results with this data show how good a database software is optimized for work use

Why use 2 different data sets?

Comparison: Benchmark

- The benchmark consists of the **comparison** of computing time of different queries
- A list of queries has to be created
- Queries should run on at least **2 of the 3** of the test systems
- 2 kinds of queries simple and complex
- Simple queries are one function
- Complex queries are combinations of simple queries
- Number of supported queries gets analyzed in soft factors

Comparison: Excerpt of queries

Measuring functions	Comparison functions	Generation functions
Point x, y	Contain	Boundary
Length	Crosses	Buffer
Area	Difference	Convex hull
Geometry by number	Disjoint	Envelope
Number of geometries	Distance	Intersection
Centroid	Equals	Transform
Dimension	Intersects	Union
Geometry type	Overlaps	
Is empty	Touches	
Is simple	Within	

Comparison: Soft factors

- Additionally to a performance benchmark **soft factors** are reviewed
- They are important for the **practical** aspect of the thesis

Some important soft factors are:

- Pricing
- Required hardware
- List of functionalities
- Installation / Setup

Comparison

- After the benchmark has been completed the results are evaluated
- **Pros and cons** of each database system get listed
- There probably won't be a clear best system, it will depend on:
  - Target group
  - Field of application
  - Personal preferences

Outlook & new developments

- In the outlook future developments and upcoming important tasks for database systems are outlined
- Especially the handling of 2 important data types
  - **3D data** (e.g. CityGML)
  - **Geospatial-temporal data** (focus on big data and tools like GeoWave)

This chapter will only give a glimpse on what the new challenges are going to be.

Summary

- Focus of the thesis is an in depth **comparison of three database systems**
- It uses **different data** to ensure a more detailed result
- Also an **outlook** on future challenges and possibilities is included
- Thesis focuses on a **realistic work environment**