1. Describe the phases of data preparation in the data warehouse.

2. Name the steps taken for filling the data warehouse and characterize them.

3. What is the differential snapshot problem and record linkage? The trivial way is a complete evaluation which compares each tuple with all the others.
   
   (a) What are techniques to reduce the complexity of duplicate detection? For each technique, name benefits and drawbacks.
   
   (b) What are possible algorithms to solve the differential snapshot problem? State their algorithmic complexity.

4. Name typical data quality issues and how to locate them. Which database techniques may prevent which problems?

5. Determination of outliers: given the data in Table 1, construct the line of best fit. How can you determine outliers with the constructed line?

\[
\begin{array}{|c|c|}
\hline
X & Y \\
1 & 1 \\
2 & 10 \\
3 & 6 \\
4 & 4 \\
5 & 3 \\
\hline
\end{array}
\]

Tabelle 1: Regressionswerte