DATA WAREHOUSE

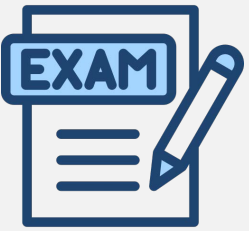
Chapter 1 – Basics Data Warehouse

CHAPTER 1 - AIMS

- Learn what a Data Warehouse is.
- Classification of the terms OLAP and OLTP.
- Definition of the terms Business Intelligence,
- Data Warehouse, and Data Lake. Tool Presentation:
  - PostgreSQL
  - PowerBI
  - Pentaho

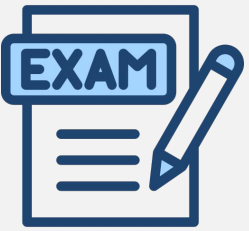
TASK
WHAT IS DATA WAREHOUSE

- Search for the term on the internet.
- Explain what you understand by it.
- Time: 5 minutes.



WHAT IS A DATA WARHOUSE?

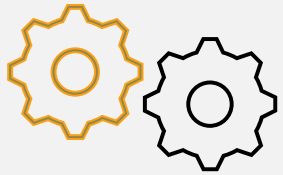
- A data warehouse is a database that stores and integrates large amounts of various data sources to enable comprehensive insights and informed decisions.



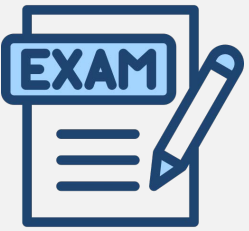
WHY DATA WAREHOUSE? 2 REASONS

- Two different purposes of data usage in a company:
  - Operational purposes of data storage
  - Decision-making and business understanding

Analytical Decisions

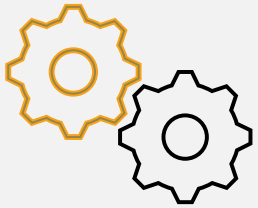


Operational data storage

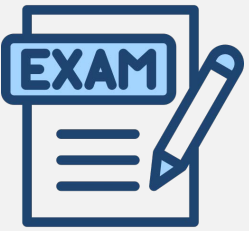


WHY DATA WAREHOUSE? 2 REASONS

- Operational purposes
  - Data usage for order processing and fulfillment
  - Handling complaints and reacting to them Utilization of data for inventory replenishment and other operational tasks
  - Necessary to keep the business running



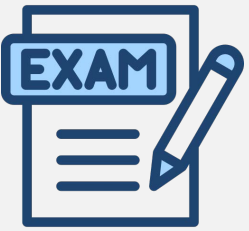
Operational data storage



WHY DATA WAREHOUSE? 2 REASONS

- Decision-making and business understanding
  - Improvement of future decisions
  - Answers to questions such as the best product category, current sales figures compared to previous months, improvement opportunities in the company
  - Performance evaluation and decision-making for the future

Analytical Decisions

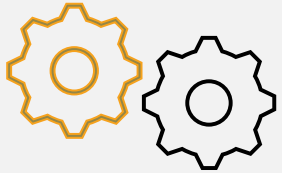


WHY DATA WAREHOUSE?
2 REASONS

- What is the best category?
- How are sales compared to last month?
- What can we improve?

➢ Evaluation of own performance
➢ Analytical decision

Analytical Decisions



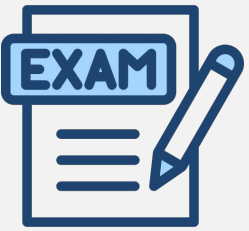
Operational data storage

- Take orders
- Respond to complaints
- Replenish inventory

➢ Keep the wheels turning

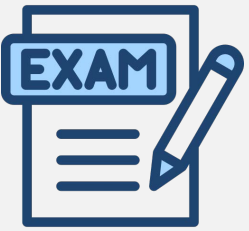
- Analytical data preparation (OLAP)
  - Observation of operations and identification of improvement opportunities

- Operational data storage (OLTP)
  - Used to maintain operations



TYPICAL STATEMENTS

- A lack of a data warehouse is often highlighted by statements like these:
  - We have a lot of data, but we can't really use it.
  - It's very complicated to access and analyze it because it's spread across different systems.
  - We just want to see what's relevant and have this data easily accessible and quickly available.
  - Ultimately, we want to make decision-based decisions and stop arguing over numbers.
- These statements emphasize the need and possibly the lack of a data warehouse.

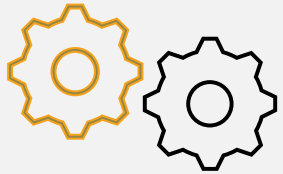


2 REASONS BUT ALSO 2 TYPES OF REQUIREMENTS

- Thousands of records at once
- Fast query performance
- Historical context

Analytical Decisions

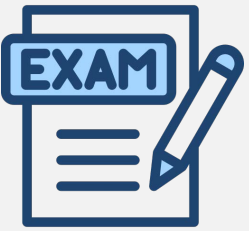
- Analytical data processing
  - Analysis and processing of large amounts of data simultaneously
  - Need for fast query performance for quick information delivery
  - Necessity of context for data interpretation over time or categories



- One record at a time
- Data input
- No long history

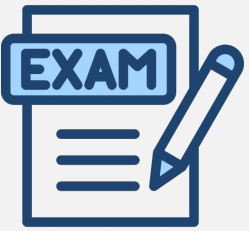
Operational data storage

- Operational data processing
  - Processing individual records to maintain operational functioning
  - Input or editing of data
  - Mainly interested in current data, hence no need for long data history



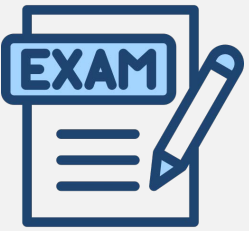
2 REASONS BUT ALSO 2 TYPES OF REQUIREMENTS

- Due to different requirements, it makes sense to keep operational and analytical systems separate.
- A data warehouse meets the needs of analytical data processing and serves as a data location for reporting and data analysis.
- It is a place for more comprehensive analysis and reporting compared to operational data processing.



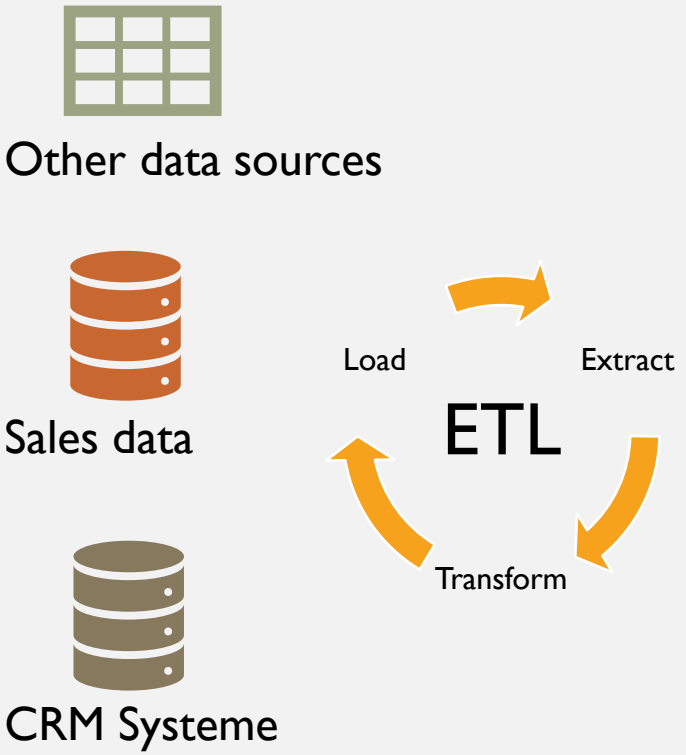
DATA WAREHOUSE

- A data warehouse is a database designed and optimized for analytical purposes.
- Key features of a data warehouse include:
  - User-friendliness:
    - Simple and understandable user interface for data analysts
    - Technical complexity should be minimized
  - Fast query performance:
    - Required to quickly retrieve and process large amounts of data
  - Optimized for best data analysis:
    - All aspects of the database are aimed at facilitating and optimizing data analysis.

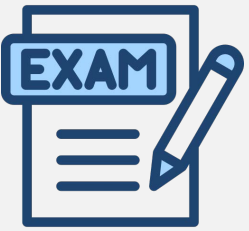


TYPICAL ENVIRONMENT OF A DATA WAREHOUSE

- Operational data systems
  - Different data sources such as sales data, HR system, CRM system
  - Different data formats and structures

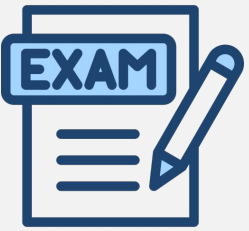


- Process of data warehousing:
  - Merge relevant data from various sources and store it centrally
  - This process is called the ETL process (Extract, Transform, Load)
  - ETL process is crucial and takes the majority of the time when creating a data warehouse (approximately 80-90%)



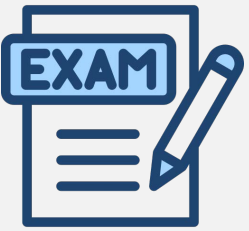
ETL PROZESS

- Extraktion
  - Data is extracted from various sources to avoid impacting their query performance and resources.
- Transformation
  - Data is integrated and structured into a unified format to prepare it for further processing. Data may be aggregated or formatted.
- Load
  - The transformed data is loaded into the central data warehouse, optimized for data analysis.

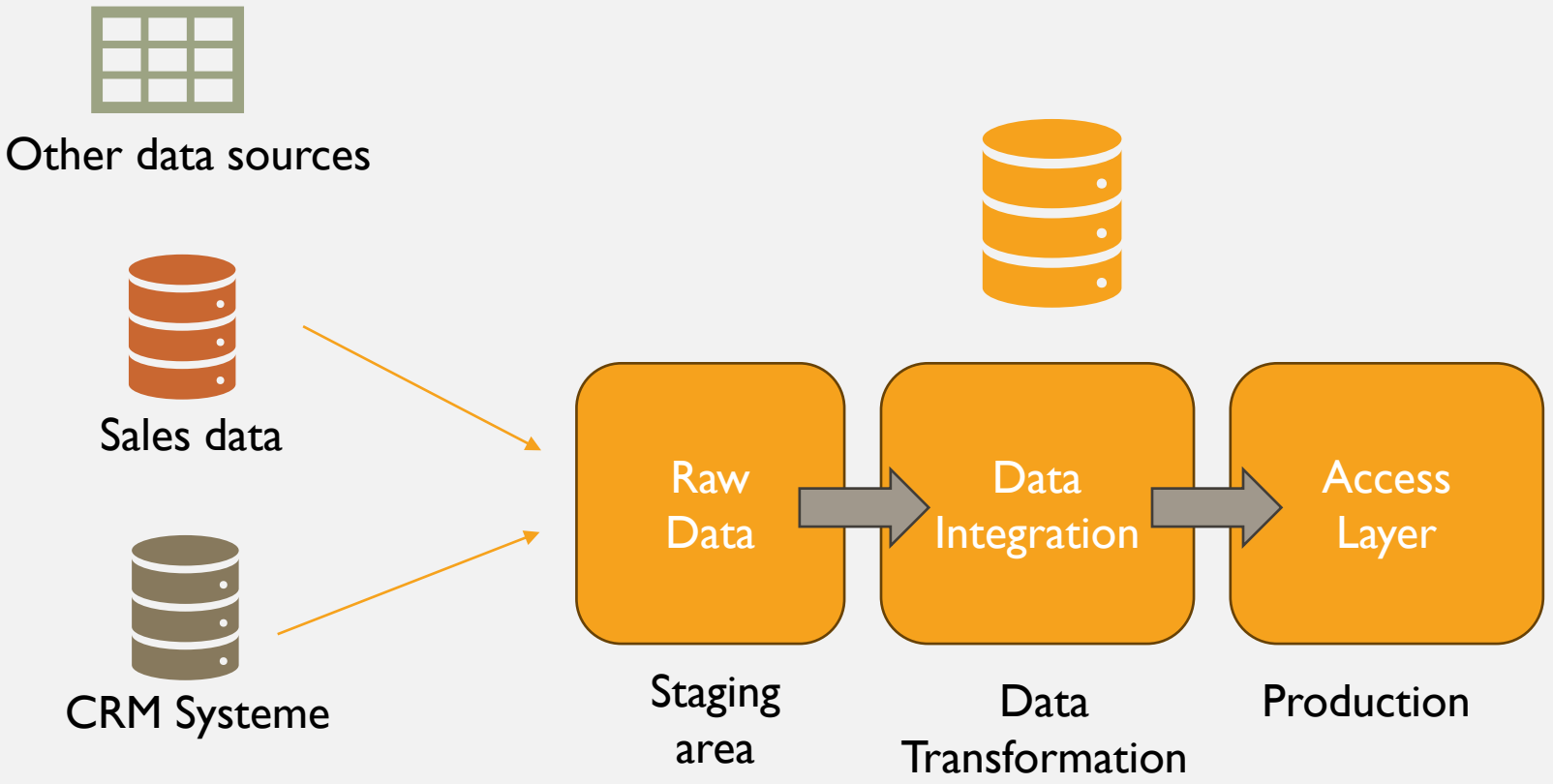


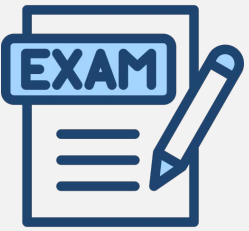
OBJECTIVES OF A DATA WAREHOUSE

- Centralized location for data from various sources in a consistent form
- Fast access to the data for quick query results User-friendly data structure for easy analysis and utilization
- Necessity of extracting, copying, and transforming data in the ETL process
- Objective: Building reports and data visualizations based on the data warehouse



WAS MACHT DAS DATA WARHOUSE?





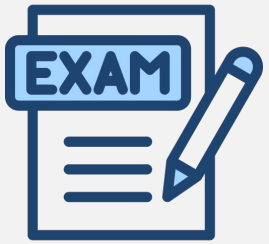
QUESTIONS CHAPTER 1- 1

WHICH TERM IS USED TO DESCRIBE OPERATIONAL DATA PROCESSING?

- OLAP
- OLTP

WHAT IS NOT AN ESSENTIAL CHARACTERISTIC OF A DATA WAREHOUSE?

- Central location for data
- Fast query performance
- User-friendly
- Data backup system



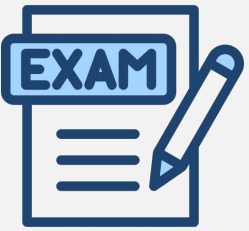
QUESTIONS CHAPTER 1- 1

WHICH TERM IS USED TO DESCRIBE OPERATIONAL DATA PROCESSING?

- OLAP
- **OLTP**

WHAT IS NOT AN ESSENTIAL CHARACTERISTIC OF A DATA WAREHOUSE?

- Central location for data
- Fast query performance
- User-friendly
- **Data backup system**



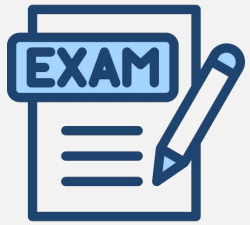
QUESTIONS CHAPTER 1- 1

WHICH TERM IS USED TO DESCRIBE OPERATIONAL DATA PROCESSING?

- Transactional and operational
- Analytical and operational
- Analytical and transactional

WHAT IS NOT AN EXAMPLE OF OPERATIONAL DATA PROCESSING?

- A banking app receives and processes a transfer request from a user.
- An employee in a warehouse scans a product that has been ordered and is being picked.
- A regional manager calculates how many items an employee has selected on average per hour.
- A billing system calculates the monthly fees for a customer and sends the invoice via email.ail.



QUESTIONS CHAPTER 1- 1

WHICH TERM IS USED TO DESCRIBE OPERATIONAL DATA PROCESSING?

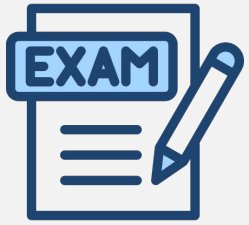
- Transactional and operational
- **Analytical and operational**
- Analytical and transactional

WHAT IS NOT AN EXAMPLE OF OPERATIONAL DATA PROCESSING?

- A banking app receives and processes a transfer request from a user.
- An employee in a warehouse scans a product that has been ordered and is being picked.
- **A regional manager calculates how many items an employee has selected on average per hour.**
- A billing system calculates the monthly fees for a customer and sends the invoice via email.ail.

TASK
WHAT IS BUSINESS INTELLIGENCE

- Search for the term on the internet.
- Explain what you understand by it.
- Time: 5 minutes.

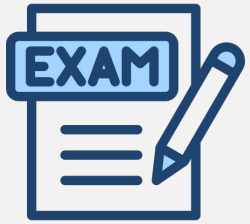


BUSINESS INTELLIGENCE

- Business Intelligence (BI) is the use of data analytics to support business decisions and improve company performance.

- Data warehouses are created for business intelligence.

BUSINESS INTELLIGENCE



Strategies

Technologies

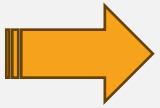
Infrastructures

Data analysis

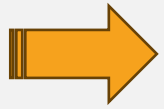
- Data gathering
- Data storing
- Reporting
- Data visualization
- Data mining
- Predictive analytics

Data Warehouse

Raw data

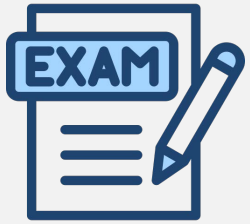


Transformation



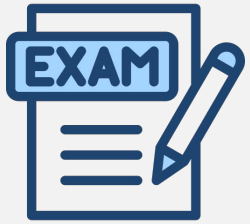
Visualization as a basis for decision-making

BUSINESS INTELLIGENCE



- Business Intelligence consists of various strategies, processes, and technologies:
  - Data analysis tools:
    - Utilization of different tools to create meaningful insights
  - Data management and storage:
    - Capturing, managing, and storing data
  - Generating meaningful insights:
    - Reporting, data visualizations, data analysis such as data mining or predictive analytics
- Goal of Business Intelligence:
  - Utilize raw data and transform it into meaningful insights
  - Use the insights gained to better understand the business and improve future decisions

BUSINESS INTELLIGENCE
ROLLE DES DATA WAREHOUSE

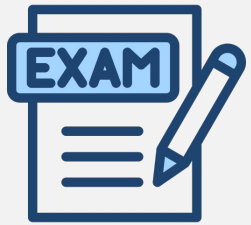


- The role of a data warehouse in Business Intelligence:
  - Key component for data storage and management
  - Central location for structured and transformed data
  - Utilization of data for data visualizations and reporting

- Difference between Data Lake and Data Warehouse:
  - Data Lake: Also used for data storage, but differences from the data warehouse
  - Desire to distinguish between both concepts

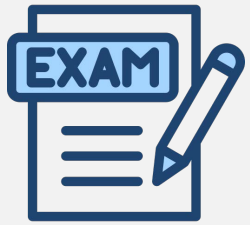
TASK
WHAT IS DATA LAKE

- Search for the term on the internet.
- Explain what you understand by it.
- Time: 5 minutes.



DATA LAKE VS. DATA WAREHOUSE

- Data Lake:
  - Data is extracted from various systems and stored in the data lake without prior processing.
- Data Warehouse:
  - Data is transformed through the ETL process and stored in a database with structured data.
  - Objective: User-friendly data for specific use cases.
- Advantages of a Data Warehouse:
  - Quick and easy creation of business intelligence solutions such as reporting and data analysis.
  - High query performance and user-friendliness.

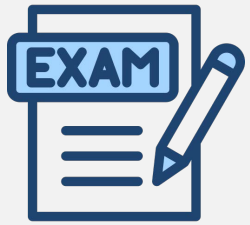


DATA LAKE VS. DATA WAREHOUSE

- Often these terms are used interchangeably.

	Data Lake	Data Warehouse
Data	Raw	Edited

- Data Lake:
  - Raw data is stored without prior processing
  - Data from various systems is stored directly in the data lake
- Data Warehouse:
  - Data is transformed through the ETL process to obtain clean data in a database
  - Structured data stored in tables in a database

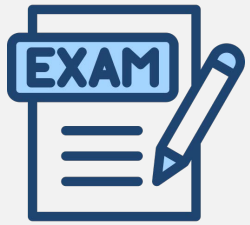


DATA LAKE VS. DATA WAREHOUSE

	Data Lake	Data Warehouse
Data	Raw	Edited
Technologies	Big data	Databases

- Data Warehouse:
  - Focus on user-friendliness and specific use cases
  - Creation of business intelligence solutions such as reporting and data analysis

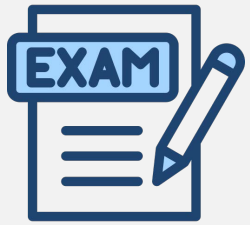
- Data Lake:
  - Large volumes of data and various data types
  - Utilization of big data technologies due to unstructured data like CSV, JSON files, images, and videos



DATA LAKE VS. DATA WAREHOUSE

	Data Lake	Data Warehouse
Data	Raw	Edited
Technologies	Big data	Databases
Structur	Non-structered	structured

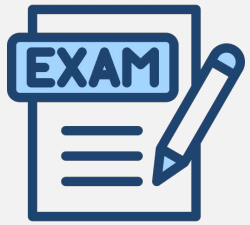
- In a Data Lake, the data is unstructured.
- This differs from a Data Warehouse, where structured data is stored in tables.
- Data types in the Data Lake:
  - CSV files
  - JSON files
  - Videos
  - Images
- Indeterminate use cases due to the diversity of data types.



DATA LAKE VS. DATA WAREHOUSE

	Data Lake	Data Warehouse
Data	Raw	Edited
Technologies	Big data	Databases
Structur	Non-structered	structured
Usage	Not defined	Specified and Available

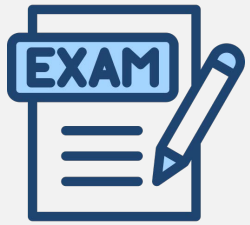
- Uncertainty about the exact use case:
  - All data is inserted into the Data Lake in raw format.
  - Doubt about how the data should be used.
  - Possibly the use cases are only intended for the future.
  - Different ideas about the use of the data.



DATA LAKE VS. DATA WAREHOUSE

	Data Lake	Data Warehouse
Data	Raw	Edited
Technologies	Big data	Databases
Structur	Non-structered	structured
Usage	Not defined	Specified and Available

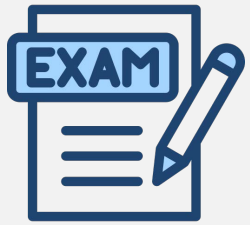
- Data access and usage in the Data Lake:
  - Individuals can extract data when they have specific use cases.
  - Central storage location for all available enterprise data.
- Compared to Data Warehouse:
  - Clear end goals or even multiple goals are set from the beginning.
  - Data is processed, cleaned, and inserted into the database for specific use cases.



DATA LAKE VS. DATA WAREHOUSE

	Data Lake	Data Warehouse
Data	Raw	Edited
Technologies	Big data	Databases
Structur	Non-structered	structured
Usage	Not defined	Specified and Available
User	Data Scientists	Business User & IT

- Differences in data structure and data technologies:
  - More challenging to ensure data quality and manage data
  - Higher level of expertise required, especially for data scientists
- Data usage in the Data Lake:
  - Data scientists use the Data Lake for experiments and ideas



DATA LAKE VS. DATA WAREHOUSE

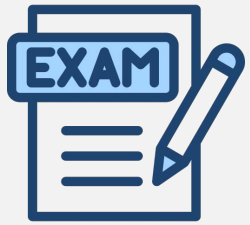
	Data Lake	Data Warehouse
Data	Raw	Edited
Technologies	Big data	Databases
Structur	Non-structered	structured
Usage	Not defined	Specified and Available
User	Data Scientists	Business User & IT

- Purpose of the Data Warehouse:
  - Providing user-friendly data for business users
- Processing and structuring data in the Data Warehouse:
  - Data is structured in a database and made available for easy use
  - The main differences between a Data Lake and a Data Warehouse lie in the data usage and accessibility for different user groups.

DATA LAKE VS. DATA WAREHOUSE FRAGEN

- When should we use which solution?
- Are Data Lake and Data Warehouse mutually exclusive?
- Is it better to use only one of the solutions or to use both?

- Answer: They are not mutually exclusive and can be used together.



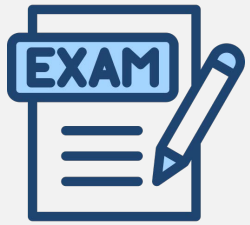
DATA LAKE VS. DATA WAREHOUSE

- Disadvantages of a Data Lake
  - Risks related to data quality and user utilization
  - Potential issues with data usage and performance

- Advantages of a Data Lake in the Cloud
  - Scalability for large volumes of data
- Utilization of a Data Warehouse based on a Data Lake
  - Using a Data Warehouse for data analysis and business intelligence

DATA LAKE VS. DATA WAREHOUSE FACTS

- Using an ETL process to extract data from the Data Lake
- Data Warehouse as a user-friendly solution for Business Intelligence strategies
- Summary:
  - Differences between Data Lake and Data Warehouse, their non-exclusivity, and the possibility of complementation by a Data Warehouse.



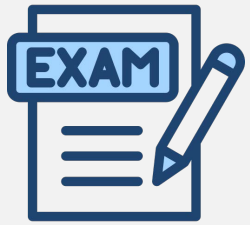
QUESTIONS CHAPTER I- III

YOU WORK IN A LOGISTICS COMPANY THAT USES VARIOUS IOT DEVICES IN UNSTRUCTURED FORMATS. WHAT WOULD YOU PREFER AS A CENTRAL STORAGE LOCATION FOR THE HUNDREDS OF MILLIONS OF RECORDS?

- Data lake
- Data warehouse

YOU WANT TO CREATE A REPORT FOR THE FINANCE DEPARTMENT. THE DATA COMES FROM VARIOUS SOURCES. WHAT WOULD YOU PREFER TO SET UP AS THE CENTRAL DATA SOURCE FOR THIS REPORT?

- Data lake
- Data warehouse



QUESTIONS CHAPTER I- III

YOU WORK IN A LOGISTICS COMPANY THAT USES VARIOUS IOT DEVICES IN UNSTRUCTURED FORMATS. WHAT WOULD YOU PREFER AS A CENTRAL STORAGE LOCATION FOR THE HUNDREDS OF MILLIONS OF RECORDS?

- **Data lake**
- Data warehouse

YOU WANT TO CREATE A REPORT FOR THE FINANCE DEPARTMENT. THE DATA COMES FROM VARIOUS SOURCES. WHAT WOULD YOU PREFER TO SET UP AS THE CENTRAL DATA SOURCE FOR THIS REPORT?

- Data lake
- **Data warehouse**