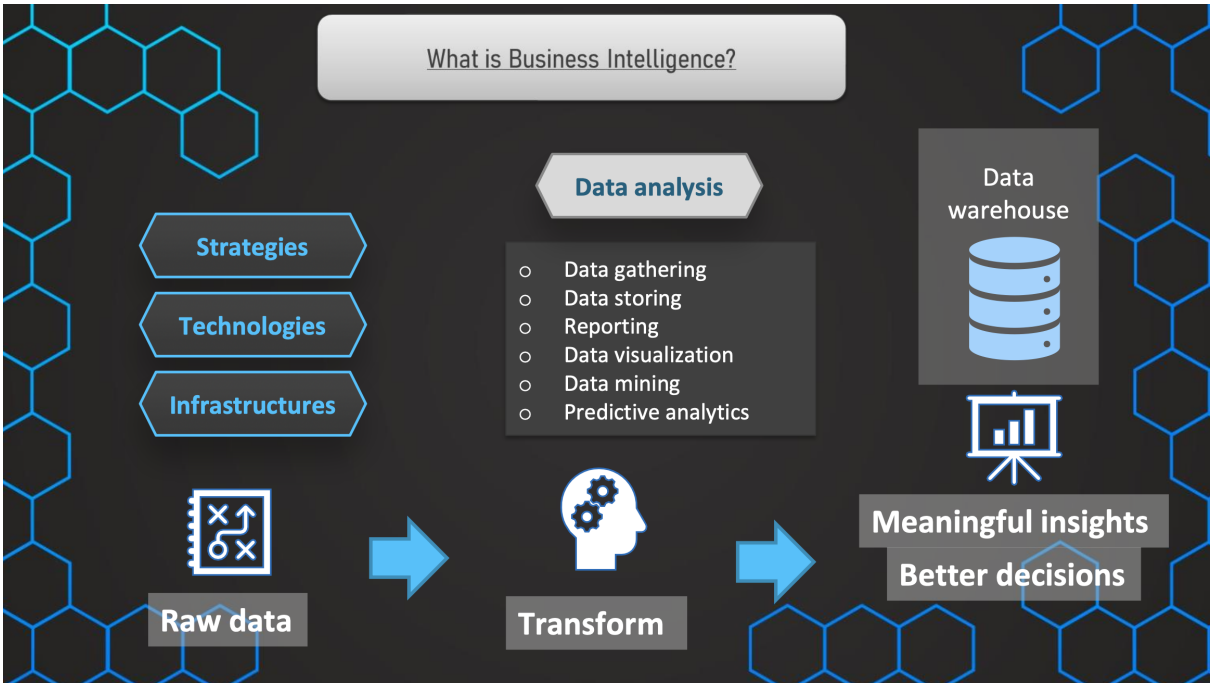
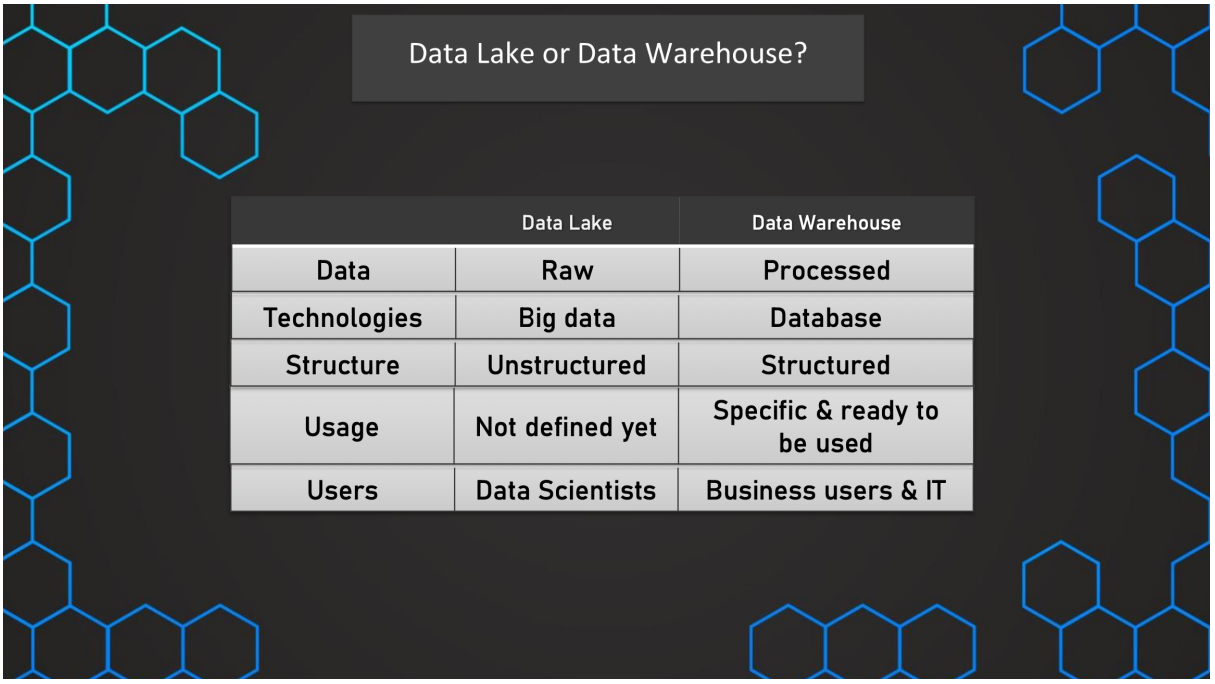
02. BI, Data Lake, Data Warehouse



- We create data warehouse for Business Intelligence
- So Business Intelligence is made up of different strategies, procedures, and technologies and infrastructures, such as, for example, different tools that we use to create meaningful insights with data analysis.
- This means that we need to gather data, we need to manage and store data, so that we can create meaningful insights with reporting, data visualisations, but also some more complex tasks, such as data mining or predictive analytics.
- So what we are doing in Business Intelligence is we want to use and find raw data and transform this data in such a way that we can turn it into meaningful insights.
- So for example, as mentioned, we want to do data visualizations, reportings, and those meaningful insights we want to use to understand our company better and make better decisions in the future.
- The data warehouse here is basically a very important component of Business Intelligence because we use this data warehouse as our data

storage, but not only that, it's specifically for a centralized location of the structured data and the transformed data that we then can use to do our data visualizations, our reportings, and therefore it is one of the most important components of Business Intelligence.

Data Lake vs Data Warehouse



- **Both data lake and data warehouse are used as centralized data storage**
- A data warehouse is a centralized location for data storage and now the same thing is also true for a data lake.
- In Data Lake we usually store RAW data which comes from various systems and no processing is done to this data, in data warehouse we process the raw data through ETL process and then store in data warehouse with specific use case in mind like reporting and data analysis and we have the clean data sitting in a database
- So a database is just contained of tables with structured data. So this is helping us to create business intelligence solutions like reportings, data analysis very quickly and easily. So high query performance and high user friendliness.
- in Datalake the volume of data is so large that we use different technologies to use that huge data like Spark

- Datalake can have unstrcutured data like csv, json, parquet, videos, images, etc. Datawarehouse have data in a database with table format with rows and columns
- So should we use either one of them or maybe it is even better to use both? the answer is NO, they are not exclusive. So they are very different and we can have both of them.
- A data lake can have many disadvantages
    - Risks that the data quality is not ensured
    - People are not really using it and they are not sure and things are not working out so well.
- With the advancement of these Cloud technologies, a data lake can be of great usage because it's very scalable for large amounts of data. But if we want to get the most out of our data and make it easy to turn it into insights, then we can just use a data warehouse on top of some parts of our data lake.
- And we use an ETL process to get the data out of a data lake if a data lake is even necessary, and then we have this very user-friendly data warehouse in the end, which we can use then for our business intelligence strategies.
- Data Lake and Data Warehouse are not mutually exclusive and data lake is not replacing a data warehouse.

Quiz

Good job!
A data lake would be a good solution for the large volume of unstructured data.

Question 1:

You are working in a logistics company that uses different IoT devices in unstructured formats. What would you rather choose as a centralized location for the hundreds of millions of datasets?

- (●) **Data Lake**
- ( ) **Data Warehouse**

Good job!
A datawarehouse is a good fit for that. This is because a report usually needs to have structured data from tables and ideally they have a fast query performance to visualize the data quickly and a high user-friendliness. Therefore a data warehouse is the better fit.

Question 2:

You want to create a report for the finance department. The data comes from different sources. What would you rather create and establish as a centralized data source for this report?

- ○ Data lake
- ◉ **Data warehouse**