Zusammenfassung – Kapitel 1

Data Warehouse VS Data Lake

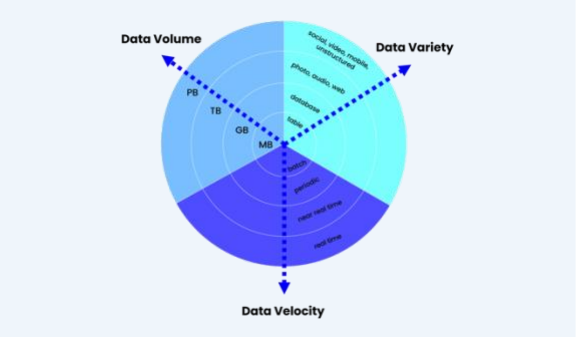
	Data Warehouse	Data Lake
Datenstruktur	Verarbeitet	Roh
Zweck der Daten	Aktuell in Gebrauch	Noch nicht festgelegt
Benutzer	Business-Anwender	Data Scientists
Flexibilität	feste Struktur, kompliziert & teuer zu ändern	Flexibel, einfach zu ändern

Business Intelligence (BI)

Business Intelligence (BI) bezeichnet Methoden und Technologien, um das eigene Unternehmen systematisch zu analysieren und bei der Entscheidungsfindung zu unterstützen. BI beschreibt den Prozess von der Datensammlung über die Datenbereitstellung und Analyse bis zur Informationsproduktion.

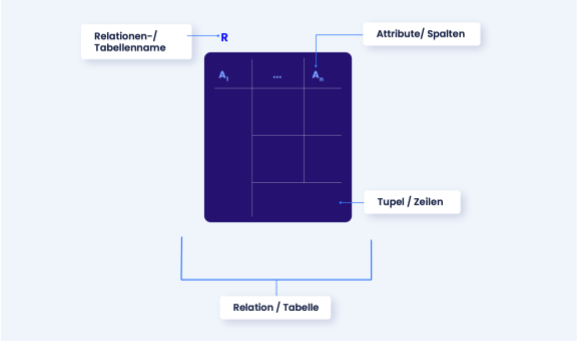
Big Data

Für Big Data sind stets die drei Vs von zentraler Bedeutung: *Volume*, *Velocity* und *Variety*. Es geht um große Datenmengen (*Volume*), die in kurzer Zeit anfallen und verarbeitet werden sollen (*Velocity*) und außerdem sehr vielfältig sind (*Variety*). Big Data ist dabei kein völlig eigenständiger Bereich, sondern eine Erweiterung des BI-Prozesses.



Relationale Datenbanken

Im relationalen Datenbankmodell bestehen Datenbanken aus drei Teilen: Tabellen, Attributen und Beziehungen. Eine Datenbank ist eine Sammlung von Tabellen. Jede Zeile steht für einen konkreten Datensatz, die Spalten bestimmen die Attribute dieser Daten. Beziehungen zwischen den Tabellen werden über Primär- und Fremdschlüssel realisiert. Ein Primärschlüssel identifiziert eindeutig einen Datensatz. Ein Fremdschlüssel verweist auf ein Attribut einer anderen Tabelle und stellt so eine Verknüpfung zwischen Daten her. Siehe dazu auch die Integritätsbedingungen weiter unten.



Datenbankschema

Das Datenbankschema umfasst alle strukturellen Informationen über eine Datenbank: Welche Datenbankobjekte sie enthält, wie diese beschaffen sind und die Beziehungen zwischen ihnen, den logischen sowie physischen Aufbau, welche Benutzer es gibt, welche Zugriffsrechte diese haben und weitere Details.

Integritätsbedingungen

Bereichsintegrität: Die Werte eines Datensatzes müssen zum jeweiligen Datentyp der Attribute passen.
Entitätsintegrität: Jeder Datensatz muss über sein Primärschlüsselattribut datenbankweit eindeutig identifizierbar sein.
Referentielle Integrität: Ein Fremdschlüssel muss entweder leer sein oder sich auf ein tatsächlich existierendes, eindeutiges Attribut beziehen.

SQL

SQL ist die Sprache, über die mit relationalen Datenbanken kommuniziert wird. Dabei wird die Anfrage an das DBMS (Datenbankmanagementsystem) geschickt und dort geprüft. SQL ist deklarativ, das heißt es reicht, einen Zielzustand zu deklarieren. Die Zwischenschritte zum erwünschten Ergebnis müssen nicht beschrieben werden. SQL hat vier Bestandteile:

- Data Query Language (DQL): Abfragen an die Datenbank
- Data Manipulation Language (DML): Ändern von Daten
- Data Definition Language (DDL): Änderungen am Datenbankschema
- Data Control Language (DCL): Rechteverwaltung des DBMS