

Deka SmartDigital

Informations- & Datenkompetenz

Datenarten, Quellen und deren Strukturierung

.Deka

Informations- & Datenkompetenz

1. Arten von Daten und deren Quellen

Daten ist nicht gleich Daten – Struktur und Herkunft der Daten spielen eine große Rolle dafür, wie wir sie speichern und nutzen können. Zunächst unterscheiden wir häufig strukturierte von unstrukturierten Daten, und wir schauen auf interne vs. externe Datenquellen.

- **Strukturierte Daten**

Diese liegen in geordnetem Format vor, meist in Tabellen (Zeilen und Spalten) oder Datenbanken. Beispiele: Excel-Tabellen mit Finanzaufzeichnungen, relationale Datenbanken mit Kundeninformationen, CSV-Dateien. Strukturierte Daten folgen einem festen Schema (z. B. bestimmte Spaltennamen und Datentypen), was die Verarbeitung erleichtert.

- **Unstrukturierte Daten**

Diese Daten liegen nicht in vorgefertigten Tabellenformaten vor. Beispiele: Freitext (E-Mails, Dokumente), Audio-/Videodateien, Bilder oder PDF-Berichte. Sie enthalten wertvolle Informationen, sind aber schwerer automatisch auszuwerten, weil keine einheitliche Struktur vorliegt.

(Hinweis: Daneben gibt es auch halbstrukturierte Daten wie XML/JSON-Dateien, die gewisse Struktur tragen, aber flexibler als Tabellen sind.)

- **Interne Datenquellen**

Daten, die innerhalb des Unternehmens entstehen. Beispiele: Transaktionsdaten aus dem Core-Banking-System, Kundendaten aus dem CRM, Daten aus internen Umfragen oder Berichten. Interne Daten sind oft leicht zugreifbar, aber man muss auf Datenschutz und berechtigten Zugriff achten.

- **Externe Datenquellen**

Daten, die von außerhalb bezogen werden. Beispiele: Marktdaten (z. B. Börsenkurse, Wirtschaftsdaten), Open Data (offene öffentliche Datensätze, z. B. von statistischen Bundesämtern), Social-Media-Daten oder Daten von Dienstleistern/Partnern. Solche Daten können neue Perspektiven liefern, müssen aber auf Kompatibilität und Qualität geprüft werden, bevor man sie im eigenen Kontext nutzt.

- **Datenformate und Beispiele**

Zahlen und Kategorien (z. B. Umsatzstatistiken), Textdaten (Kundenfeedback, Dokumentationen), Zeitreihen (Kursverläufe, historische Leistungsdaten), geografische Daten (Standortkoordinaten) u.v.m. Unterschiedliche Datentypen erfordern teils unterschiedliche Methoden der Speicherung und Analyse.

2. Datenstrukturierung und Qualitätsmanagement

Damit Daten effektiv genutzt werden können, müssen sie gut organisiert und von hoher Qualität sein. Es geht darum, wie wir Daten strukturieren (ordnen, formatieren, speichern) und warum Datenqualität kritisch für verlässliche Ergebnisse ist.

Ohne ein Mindestmaß an Ordnung und Qualität können selbst die besten Analyse-Tools nicht weiterhelfen
→ Unordnung in Daten führt zu Ineffizienz, und schlechte Qualität führt zu irreführenden Analysen.

Daten strukturieren

Rohdaten liegen oft in verschiedensten Formaten vor. Der erste Schritt besteht darin, sie in einheitliche Strukturen zu bringen. Das bedeutet z. B. ein geeignetes Datenformat wählen (Excel-Tabelle, Datenbank, CSV) und konsistente Strukturmerkmale festlegen (Spaltennamen, Datentypen, Kodierungen).

Gut strukturierte Daten erleichtern das Wiederfinden und Analysieren. Beispiel: Anstatt frei formulierte Texte für ein Feld "Produktkategorie" zu sammeln, legt man eine feste Liste von Kategorien an – so entstehen konsistente Angaben anstelle eines Chaos an Begriffen.

Datenablage und -organisation

Überlegen Sie, wo und wie Daten gespeichert werden. Zentralisierte Ablage (z. B. in einer Datenbank oder einem Data Warehouse) verhindert Datensilos und Doppelarbeit. Zugriffsrechte und Datenschutz spielen hier ebenfalls eine Rolle. → Sensible Daten müssen sicher gespeichert und vor unbefugtem Zugriff geschützt werden.

→ Klare Benennungs- und Ablageregeln (z. B. Dateinamen, Ordnerstrukturen oder Datenbank-Schemata) helfen allen Beteiligten, Daten schneller zu finden und richtig zu interpretieren.

Datenbereinigung

Bei nahezu allen Datensätzen sind Bereinigungen nötig. Typische Aufgaben: Duplikate entfernen, Tippfehler korrigieren, fehlende Werte auffüllen oder kenntlich machen, ungültige Werte (z. B. "31.11." als Datum) anpassen.

→ Diese Aufräumarbeit ist zwar zeitaufwändig, aber essentiell, da Fehler in den Daten direkt zu Fehlern in der Analyse führen würden (siehe „Garbage in, garbage out“ Grundsatz). Oft nimmt die Datenaufbereitung den größten Teil der Zeit in einem Datenprojekt ein – das ist normal und gut investierte Zeit für Qualität.