1 Grundlagen

In diesem Kapitel wird motiviert, welche Rolle analytische Datenbankanwendungen in der Praxis spielen und warum sie andere Anforderungen an die Repräsentation und Verarbeitung von Daten stellen als die typischen Anwendungen im Tagesgeschäft.

Daraus leiten wir die Notwendigkeit für sogenannte **Data Warehouses (Datenlager)** ab, in denen Daten speziell für Analysen gesammelt, integriert und aufbereitet werden.

Data Warehouse

1.1 Lernziele dieses Kapitels

- Sie können die Begriffe **OLAP** und **OLTP** voneinander abgrenzen und wissen um die unterschiedlichen Anforderungen von OLAP- und OLTP-Anwendungen an die zugrunde liegenden DBS.
- Sie kennen **typische Anwendungsfälle** für OLAP.
- Sie kennen die Definition eines **Data Warehouse nach Inmon** mit dessen vier wesentlichen Eigenschaften.
- Sie verstehen die Notwendigkeit einer **geplanten Datenarchitektur**, um nachhaltig verschiedene Abnehmer mit Daten für analytische Anwendungsfälle versorgen zu können.

1.2 OLAP und OLTP

Die ursprünglichen Anwendungen für die Datenverarbeitung mit automatischen Rechenmaschinen gehen bis ins 19. Jahrhundert zurück. So wurden z. B. Rechenmaschinen des amerikanischen Ingenieurs Herman Hollerith Ende des 19. Jahrhunderts eingesetzt, um die Massenerfassung medizinischer Daten und Volkszählungen in den USA zu unterstützen.

Diese Art von Anwendungen würde man heute **analytische Anwendungen** nennen: es geht dabei darum, aus großen Datenmengen zusammengefasste Ergebnisse zu errechnen, beispielsweise Anzahlen, Mittelwerte oder Verteilungen. Die einzelnen Datensätze – also z. B. die Krankenakte einer bestimmten Patientin oder die Einkäufe eines einzelnen Kunden – treten dabei in den Hintergrund.

Analytische Anwendungen

Im Gegensatz dazu stehen Anwendungen, bei denen einzelne Geschäftsvorfälle – **Transaktionen** genannt – verarbeitet werden: ein Kunde kauft einen Artikel, ein Produkt durchläuft eine Fertigungsstraße, eine Patientin wird behandelt. Hierbei werden anstatt umfangreicher Analysen viele kleine Änderungen an Datenbanken vorgenommen. Historisch ist diese Art der Datenverarbeitung jünger und etwa seit den 1970er Jahren üblich, da man hierfür leistungsfähige Rechner im täglichen Geschäftsprozess benötigt (z. B. die elektronische Supermarktkasse, die mit einem Warenwirtschaftssystem gekoppelt ist).

Diese unterschiedlichen Klassen von Anwendungen werden heute üblicherweise mit den Abkürzungen **OLTP** und **OLAP** bezeichnet:

- **OLTP** steht für **On-Line Transaction Processing** und beschreibt die Verarbeitung einzelner Geschäftsvorfälle.
- **OLAP** steht für **On-Line Analytical Processing** und bezieht sich auf analytische Anwendungen auf großen Datenmengen.

Dabei stellen OLAP und OLTP sehr unterschiedliche Anforderungen an Datenbanksysteme. Im Folgenden sind die wichtigsten Unterschiede aufgeführt:

Tabelle 1: OLAP und OLTP

	OLTP	**OLAP**
Zweck	Abbilden des Tagesgeschäfts	Gewinnen von Erkenntnissen aus historischen Daten
Nutzerkreis	Parametrische Nutzer im operativen Geschäft	Führungskräfte, Analysten, Produktmanager
Art der Anfragen	Einfügen, Ändern, Löschen einzelner Datensätze	Lesen und Aggregieren großer Datenmengen
Lastverhalten	Kurze Anfragen in hoher Frequenz	Wenige, aber aufwändige Anfragen
Antwortzeiten	Sekunden oder schneller	oft Minuten oder auch Stunden
Datenvolumen	relativ gering, wenn ältere Daten archiviert werden können	sehr groß
Datenbankentwurf	Normalisiert, z. B. 3NF, BCNF	Denormalisiert
Datenbewirtschaftung	Laufende Aktualisierungen	Periodisches Einfügen größerer Datenmengen aus den OLTP-Systemen

Aufgrund dieser unterschiedlichen Charakteristika von OLAP- und OLTP-Datenbanksystemen werden diese oft getrennt voneinander implementiert. Zwar kommen verstärkt Systeme auf den Markt, die diese Trennung aufzuheben versprechen. Der Standard in betrieblichen Informationssystemen ist allerdings, dass die beiden Arten von Datenbanksystemen getrennt voneinander betrieben werden. Die Daten des OLAP-Systems werden dann laufend oder regelmäßig aus den OLTP-Systemen bezogen.

1.3 Anwendungsfälle für OLAP

„Wir brauchen Zahlen" oder „die Zahlen zeigen, dass …" sind Formulierungen, die in Firmen und anderen Organisationen häufig zu hören sind, wenn basierend auf Analysen aktueller oder historischer Ereignisse Entscheidungen getroffen werden oder Pläne für die Zukunft erarbeitet werden müssen.

Die „Zahlen", von denen dabei die Rede ist, sind oft **aggregierte,** also zusammengefasste Messwerte über Geschäftsvorfälle, die über die **Zeit** erfasst wurden. Die Geschäftsvorfälle liefern dabei **Messwerte (Metriken)**, z. B. den Umsatz, den eine Bestellung ausmacht, oder die Körpertemperatur einer Patientin. Außerdem gibt es beschreibende **Dimensionen**, z. B. die Patientin, die Station im Krankenhaus, den Wochentag oder die Art der Erkrankung.

Die Metriken bestimmen dabei das Ergebnis der Analyse, während die Dimensionen zum Filtern, Gruppieren und Beschreiben der Ergebnisse dienen. Eine Anfrage im Beispiel eines Krankenhauses könnte lauten:

„Welche drei **Behandlungen** sind in jeder **Station** diejenigen, in denen **pro Quartal** jeweils der größte **Umsatz** und der größte **Gewinn** erwirtschaftet wird?"

Dabei sind Umsatz und Gewinn die Messwerte und Behandlung, Station und Quartal die beschreibenden Dimensionen.

Weitere typische **Anwendungsfälle** für OLAP sind etwa die Folgenden:

- Eine Produktmanagerin analysiert die Verkäufe ihrer Produkte über Regionen und Jahreszeiten, um Marketingkampagnen zu planen.
- Ein Fachbereich einer Hochschule entscheidet anhand der Entwicklung der Einschreibezahlen der vergangenen Semester, welche Studiengänge ausgebaut, eingestellt oder verstärkt beworben werden sollen.

- Eine Agentur analysiert den Verkehr auf den Webseiten der von ihr betreuten Kunden und beschließt, mit welchen Maßnahmen die Popularität der Seiten gefördert werden kann.
- Der Vorstand eines Konzerns entscheidet anhand der Kennzahlen der verschiedenen Unternehmensbereiche, welche Bereiche in Zukunft mehr gefördert werden sollen.

Die Daten werden dabei oftmals in Form sogenannter **Dashboards** (deutsch „Armaturenbretter", sinngemäß eher „Leitstände") aufbereitet. Dies sind grafische Überblicksdarstellungen, in denen die erhobenen Kennzahlen in verschiedenen Diagrammen aufbereitet werden. Abbildung 1 zeigt ein Beispiel:

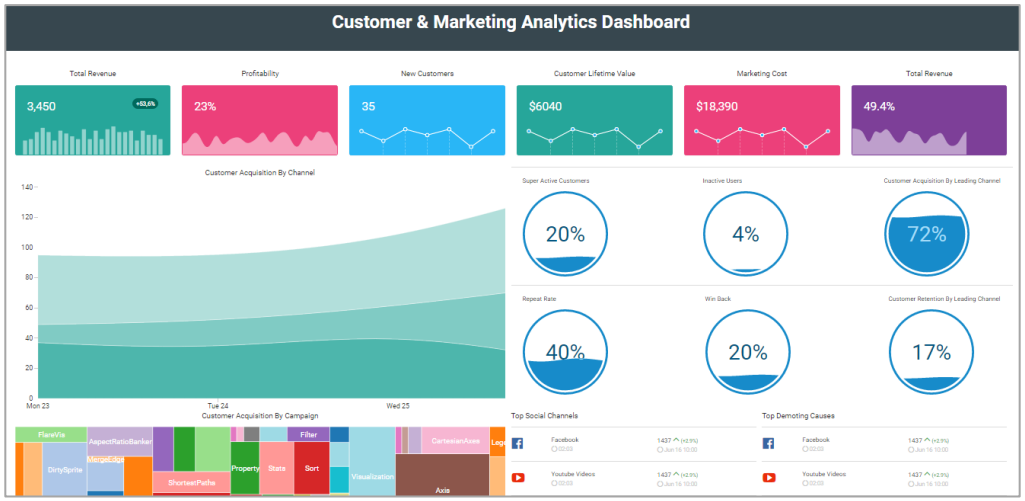

Abbildung 1: Beispiel eines Dashboards (Quelle: Wikipedia/HelicalInsight OpenSourceBI)

Hier sind verschiedene Aspekte von Kundenakquise-Aktivitäten eines Unternehmens gezeigt. Diese werden im Beispiel in Zeitreihen (oben, links), aktuellen aggregierten Ständen (rechts) sowie einer sogenannten Treemap (unten), die Größenverhältnisse in Flächen darstellt, angezeigt. Auffällig ist, dass alle Zahlen, die im Dashboard dargestellt sind, aus **Aggregatfunktionen** stammen. Es geht hier also um Summen, Anzahlen oder Mittelwerte.

Neben der Darstellung von Dashboards gibt es oft noch die Möglichkeit, entweder mit entsprechenden **Business-Intelligence**-Werkzeugen oder über frei formulierte Anfragen z. B. in SQL detailliertere Erkenntnisse aus den Daten zu gewinnen.

Für diese Art von analytischen Aktivitäten gibt es eine Reihe verschiedener Bezeichnungen. Im Detail unterscheiden sich diese darin, wer die Adressaten sind (z. B. operative Ebene oder Management), ob es eher um eine vorausschauende

oder historische Betrachtung geht, oder ob Steuerung oder Analyse im Vordergrund steht:

- Business Intelligence („Geschäftsverständnis“)
- Business Analytics („Geschäftsanalysen“)
- Reporting („Berichtswesen“)
- Decision Support Systems („Entscheidungsunterstützung“)
- Management Information Systems
- Enterprise Information Systems („Führungsinformationssystem“)

Aus unserer Datenbankperspektive haben aber alle diese Prozesse ähnliche Anforderungen, nämliche **aggregierte Auswertungen über große, historisierte und integrierte Datenbestände** bereitzustellen.

1.4 Gewachsene und geplante Datenarchitektur

Inmon [Inm05] beschreibt die historische Entwicklung von Datenbankanwendungen im Unternehmen in mehreren typischen Evolutionsstufen.

Geschäftsanwendungen beginnen oft wie in Abbildung 2 gezeigt mit eigener Datenhaltung. Diese kann datei- oder datenbankbasiert sein. Die Anwendung kapselt dabei den Zugriff auf die Daten, und Benutzer*innen greifen auf die Daten nur über die jeweilige Anwendung zu.

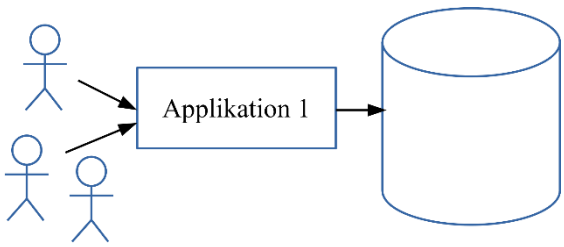

Abbildung 2: Anwendung mit eigener Datenhaltung

Mit der Einführung universeller Datenbanksysteme, die Transaktionen, Mehrbenutzersynchronisation sowie Rollen und Rechte mitbringen, wurde es ermöglicht, ein übergreifendes Datenmodell zu erstellen, das mehrere Anwendungen unterstützt. Wie in Abbildung 3 gezeigt stützten sich die beteiligten Anwendungen auf einen gemeinsamen Datenbestand und es wurden Redundanzen und Widersprüche vermieden.

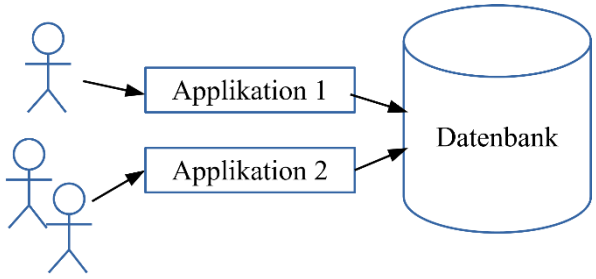

Abbildung 3: Zentrales Datenbanksystem

Grundsätzlich erlaubte es nun diese allgemeine, integrierte Datenhaltung, Analysen über den gemeinsamen Datenbestand zu erstellen. Hierdurch treten allerdings die in Tabelle 1 aufgeführten Unterschiede zu Tage: ein auf OLTP optimiertes Datenbanksystem ist nicht uneingeschränkt für Analysen zu gebrauchen und umgekehrt. Wie in Abbildung 4 gezeigt, wurden nun getrennte Datenbanksysteme für die beiden Anwendungsfälle umgesetzt.

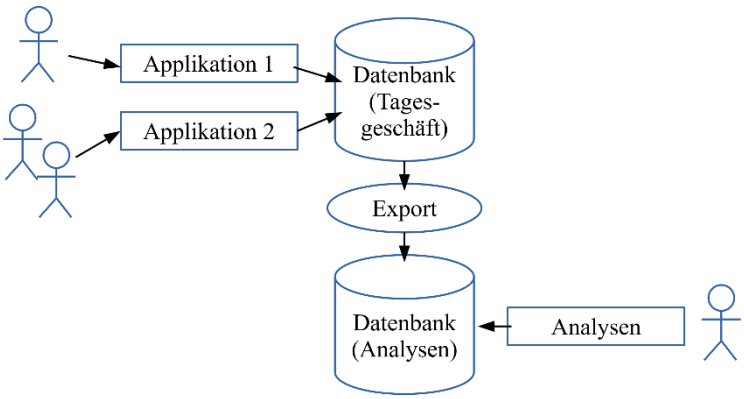

Abbildung 4: Getrennte Datenbanken für OLAP und OLTP

Ein Nachteil bei diesem Ansatz ist es, dass aufgrund der unterschiedlichen Datenmodelle für OLTP und OLAP ein Prozess eingerichtet werden muss, der Daten aus der Datenbank für das Tagesgeschäft exportiert, aufbereitet und in die Analyse-Datenbank einfügt. Dieser Export wird typischerweise periodisch für größere Datenmengen angestoßen, z. B. als täglicher Export und Import, der jede Nacht stattfindet.

Datenintegration

Im Laufe der Zeit entstehen in einer größeren Organisation oft mehrere Datenbanksysteme, die die verschiedenen Anwendungen unterstützen. Aus einer Analysesicht ist es dann meistens sinnvoll, Datenbestände aus mehreren dieser Systeme zu **integrieren,** also zusammenzuführen und anzugleichen. So könnten etwa bei einem Versandhändler Informationen aus der Logistik mit denen

aus dem Online-Shop, dem Call Center und der Kundenverwaltung integriert werden, um herauszufinden, unter welchen Umständen es beim Versand von Bestellungen besonders oft zu Beschwerden kommt.

Problematisch dabei ist, dass die nun notwendigen Datentransporte zwischen verschiedenen Systemen oft nicht geplant, sondern ad-hoc eingerichtet werden, so wie es das aktuelle Projekt gerade erfordert. Daraus entsteht eine schwer verständliche Sammlung von Datentransport- und -abgleichprozessen, die Inmon als das „Spinnennetz“ bezeichnet, wie in Abbildung 5 gezeigt:

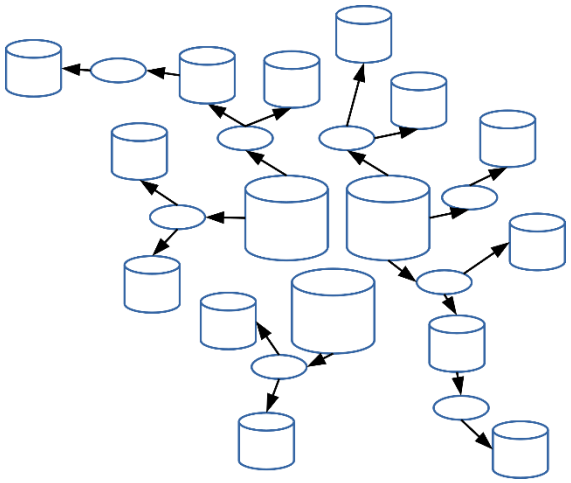

Abbildung 5: Das „Spinnennetz“

Diese Architektur ist gekennzeichnet von zahlreichen Abhängigkeiten, die im Laufe der Zeit durch die verschiedenen Datentransporte zwischen DBS aufgebaut wurden. Es ist meistens keine Struktur zu erkennen, und durch die Abhängigkeiten werden Änderungen an diesem Gesamtsystem sehr erschwert, da eine Änderung an einer Stelle oft Anpassungen bei den Abnehmern der jeweiligen Daten nach sich zieht, die weiter kaskadieren können.

Auch inhaltlich gibt es Schwierigkeiten bei dieser Art von Architektur. Die Daten werden möglicherweise aus den OLTP-Systemen als Ist-Stände zu unterschiedlichen Zeitpunkten exportiert. Ihnen fehlt damit ein **Zeitbezug**, so dass Daten aus unterschiedlichen Quellen zusammen keinen Sinn ergeben können, da sie zu abweichenden Zeitpunkten erhoben wurden.

Bei der fachlichen Interpretation der unterschiedlichen Datenbestände kann es ebenfalls Unterschiede geben: bedeutet die Kundennummer aus Abteilung 1 das Gleiche wie die Kundennummer in Abteilung 2? Oder ist dort vielleicht eine Vertragsnummer gemeint? Hier gibt es also Herausforderungen in Bezug auf die **Datenintegration**.

Schließlich gibt es allgemein das Problem des mangelnden Verständnisses und Überblicks der gewachsenen Architektur. Es ist schwierig, darin die **fachlich korrekten Quellen** für bestimmte Informationen auszumachen. Unter Umständen führt das dazu, das im gerade laufenden Projekt weitere Datentransporte ad-hoc eingeführt werden.

1.5 Data Warehouses

Der Begriff **Data Warehouse** (**DWH**, deutsch **Datenlager**) wurde von William Inmon geprägt. Er sieht das Data Warehouse im Zentrum einer geplanten Umgebung („architected environment“) als Gegenentwurf zum vorgenannten „Spinnennetz“.

Ein Data Warehouse ist dabei eine **„themenbezogene, integrierte, nichtflüchtige und historisierte Sammlung von Daten zur Unterstützung von Managemententscheidungen“** [Imn05, Kap. 2]. Diese Definition umfasst verschiedene Aspekte:

- **Themenbezug:** Das DWH ist rund um Geschäftsobjekte wie Kunden, Bestellungen, Artikel usw. organisiert. Anders als OLTP-Systeme dient es nicht dazu, bestimmte Geschäftsprozesse zu unterstützen, sondern es bietet umfassende Informationen über diese Geschäftsobjekte an.
- **Integration:** Die Daten im DWH werden aus mehreren Quellen bezogen und so konvertiert und angeglichen, dass sie ein gemeinsames und konsistentes Bild bieten. Dazu gehört beispielsweise, dass Kodierungen angeglichen werden und sichergestellt wird, dass Fremdschlüssel nicht auf fehlende Werte zeigen.
- **Nichtflüchtigkeit (Persistenz):** Daten im DWH werden nicht geändert, sondern üblicherweise genau einmal geschrieben und vielfach gelesen. Gelöscht werden Daten höchstens im Zuge planmäßiger Bewirtschaftungsprozesse, wobei sie dann nach einer bestimmten Aufbewahrungszeit bewusst entfernt werden.
- **Historisierung:** Die Datensätze im DWH sind mit Informationen über die Zeit versehen, zu der sie gegolten haben. Wenn es neue Erkenntnisse über die Geschäftsobjekte gibt, werden keine Daten geändert, sondern neue Daten geschrieben, die diese Veränderungen darstellen. Dadurch enthalten die Daten im DWH eine Historie, die es ermöglicht, Auswertungen über Zeitfenster zu berechnen, z. B. die Umsätze eines bestimmten Monats.

Es sind verschiedene Möglichkeiten denkbar, ein DWH gemäß der Definition von Inmon aufzubauen. Eine in der Praxis besonders verbreitete Vorgehensweise ist die **Dimensionale Modellierung**, die besonders von Ralph Kimball und Koautor*innen [Kim13] vorangetrieben wurde und wird. Mit dieser werden wir uns in Kapitel 2 befassen.

Übungsaufgaben

1.1 Wodurch unterscheiden sich die Anwendungsgebiete OLTP und OLAP? Welche Anforderungen stellen diese jeweils an ein Datenbanksystem?

1.2 Wie definiert Inmon den Begriff DWH? Welches sind die vier entscheidenden Eigenschaften?

1.3 Was sind die Probleme einer gewachsenen Datenarchitektur?

Zusammenfassung

In diesem Abschnitt haben wir die Begriffe **OLTP** und **OLAP** voneinander abgegrenzt. OLTP – On-Line Transaction Processing – bezeichnet dabei die typischen Anwendungsfälle für DBS im Tagesgeschäft, bei denen einzelne Geschäftsvorfälle verbucht werden. OLAP – On-Line Analytical Processing – dagegen bezeichnet die analytischen Anwendungsfälle, in denen häufig mit gruppierenden Anfragen mit Aggregatfunktionen auf historischen Daten gerechnet wird.

Durch die unterschiedlichen Anforderungen dieser beiden Familien von Anwendungsfällen werden hierfür oft getrennte DBS aufgesetzt, die unterschiedliche Datenmodelle verwenden – normalisierte Modelle für OLTP, auf Analysen optimierte Modelle für OLAP. Dadurch werden Datentransporte zwischen den verschiedenen DBS notwendig. Im Laufe der Zeit kann so eine **gewachsene Architektur** mit einem Wildwuchs an Datenübertragungen zwischen verschiedenen Systemen entstehen.

Data Warehouses sollen dieser eine geplante Architektur gegenüberstellen, in der Daten gezielt für Analysen aufbereitet werden. Sie halten die Daten themenbezogen, integriert, historisiert und persistent vor, um analytische Anwendungsfälle in der jeweiligen Geschäftsdomäne zu unterstützen.