Healthcare Data Analytics

Introduction to Big Data

Dr. Michael Strobel

27.06.2022

Inhalt

Letzte Vorlesung

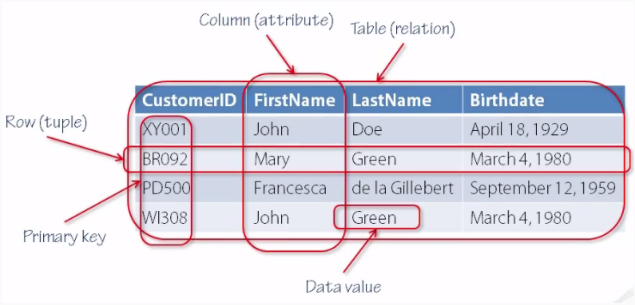
- Motivation durch visuellen Cortex
- Convolution (Faltung)
- Pooling
- Deep Convolution Neural Networks

Diese Woche

- Datenbanksysteme
- Data Warehousing
  - OLAP vs OLTP
  - Sterne und Schneeflocken
- Intro zu Big Data

Erinnerung: Relationale Datenbanken

- Die meist verwendeten Datenbanken arbeiten relational (PostgreSQL, MySQL, Oracle. . .)
- Daten werden in Zeilen gespeichert
- Operationen sind Transaktionen und folgen den Regeln von ACID
  - atomicity
  - consistency
  - isolation
  - durability



https://www.c-sharpcorner.com/article/sql-server-and-relational-database-part-one/

Use Cases von Relationalen Datenbanken

Typische Anfrage an relationale Datenbank

- Anfragen sind in der Regel eng begrenzt und beziehen sich auf eine einzelne Zeilen der Datenbank
- Performance wird auf Transaktionen / Requests optimiert

https://www.c-sharpcorner.com/article/sql-server-and-relational-database-part-one/

Online Transaction Processing (OLTP)

Ein Datenbank-Zugriffsparadigma bei denen Transaktionen ad-hoc und ohne größere Zeitverzögerung durchgeführt nennen wir *Online Transaction Processing (OLTP)*.

Eigenschaft	Online Transaction Processing System (OLTP)
Häufigste Leseoperation	Kleine Anzahl von Records pro query, Anfrage über key
Häufigste Schreiboperation	Random Zugriff, geringe Antwortzeit von User Input
Häufigster Anwendungszweck	Endkunde über Web Applikation
Typische Datengröße	Gigabyte bis Terabyte

Wie steht es aber mit Analytics auf relationalen Datenbanken?

- Typische Fragestellungen von Analytics sind statistische Kennzahlen von Gruppierten Daten
  - Wie viele Patient:innen haben wir im Juni behandelt?
  - Was sind die häufigst auftretenden Krankheitsbilder?
  - Welche Medikamente werden am wenigsten verordnet?
- Analytics kann durchaus auf Online Transaction Processing Systemen erfolgen, aber
- Typischerweise sind die Workloads **nicht** auf Analytics Fragestellungen optimiert

Online Analytics Processing (OLAP)

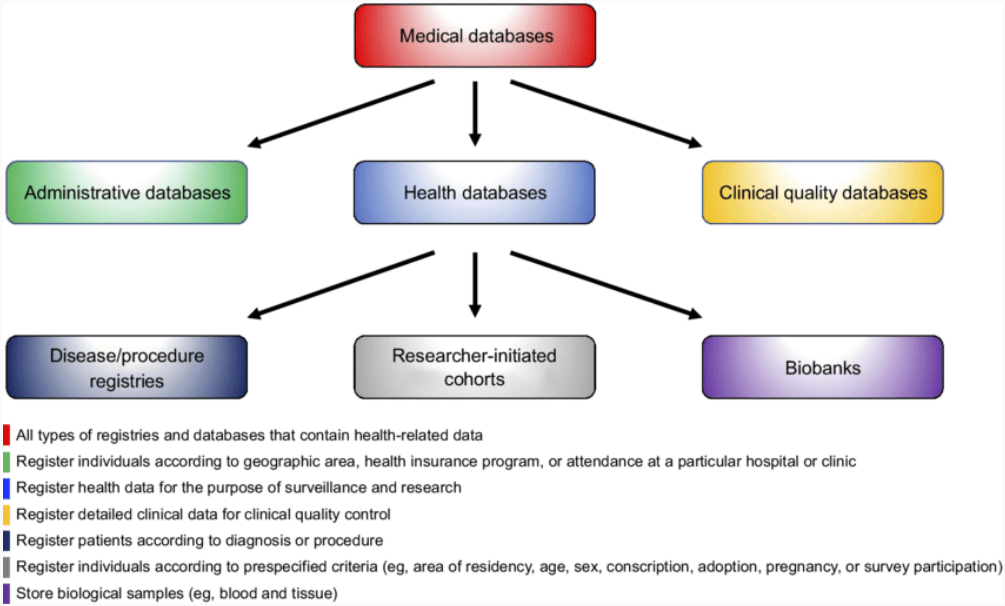
Ein Datenbank-Zugriffsparadigma bei denen Transaktionen komplexe und / oder zeitintensive Anfragen und Berechnungen durchgeführt werden nennen wir *Online Analytical Processing System (OLAP)*

Eigenschaft	Online Analytical Processing System (OLAP)
Häufigste Leseoperation	Aggregation über eine große Anzahl von Records
Häufigste Schreiboperation	Bulk / Batch Processing oder Event Stream
Häufigster Anwendungszweck	Analyst / Data Scientist für Entscheidungsunterstützung
Was stellen die Daten da?	Historie von Ereignissen
Typische Datengröße	Terabyte bis Petabyte

In medizinischen Einrichtungen sind Zahlreiche Datenbanken vorhanden

In der Medizin gibt es zahlreiche OLTP Systeme wie z.B.

- Bilder (PACS)
- Krankenhausinformationsystem (KIS)
- Radiologieinformationssystem (RIS)
- Laborinformationssystem (LIS)
- ...
- Abrechnungssystem
- Verwaltungssystem



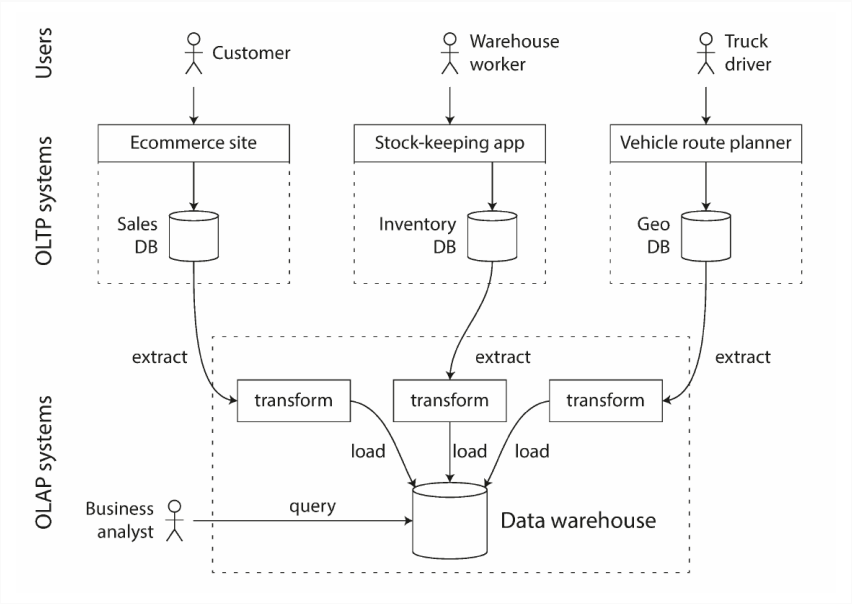
Data Warehousing

- Medizinische Einrichtungen haben Systeme die Ärzt:innen bzw. Patient:innen zur Verfügung stehen, diese sind meistens OLTP Systeme
- Die OLTP System müssen hoch-verfügbar sein und sind meistens kritisch für die Versorgung
- Aufgrund der Verfügbarkeitsanforderungen stehen diese Systeme nicht direkt für Analytics zur Verfügung, da diese mindestens Performance oder auch Verfügbarkeit beeinträchtigen
- Daher wird oft ein Zusätzlich System eingeführt auf dem Analytics Anwendungen zugreifen können, das **Data Warehouse**

Das Data Warehouse

- Data Warehouses sind Systeme die aus aus verschiedenen Datenquellen, meist OLTP Systeme, speisen
- Die Daten werden über den **Extract-Transform-Load (ETL)** aus den OLTP System geladen
  - **Extract**: lädt die Daten aus den OLTP Systemen
  - **Transform** : die Daten werden in für Analytics optimierte Datenformate überführt
  - **Load**: die Daten werden ins das Data Warehouse geladen
- Die ETL Prozesse können *periodisch* (z.B. täglicher DB dump) oder als *Stream* stattfinden
- Aufgrund der Flexibilität von SQL benutzen Data Warehouses meist ebenso SQL um Anfragen durchzuführen

Das Data Warehouse, Visualisierung



Kleppmann, M. (2017). Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems.

Stern- und Schneeflocken Schemas

Ein Schema für Datenbanken das sich für Analytics Zwecke gut eignet ist das **Star Schema**:

Im Zentrum steht die Fakten Tabelle (fact table):

- in diesem werden Events und deren Zeitpunkte gesammelt (zeilenweise), dies kann z.B. eine Behandlung eines Patienten sein
- Fakten sind in der Regel individuelle Event, da diese maximale Flexibilität für spätere Analysen erlaubt, können entsprechend aber sehr groß (petabyte oder mehr)
- Zu den eigentlichen Fakten im Fact table gibt es Referenzen (foreign keys) zu den *Dimensions Tabellen* (dimension tables)

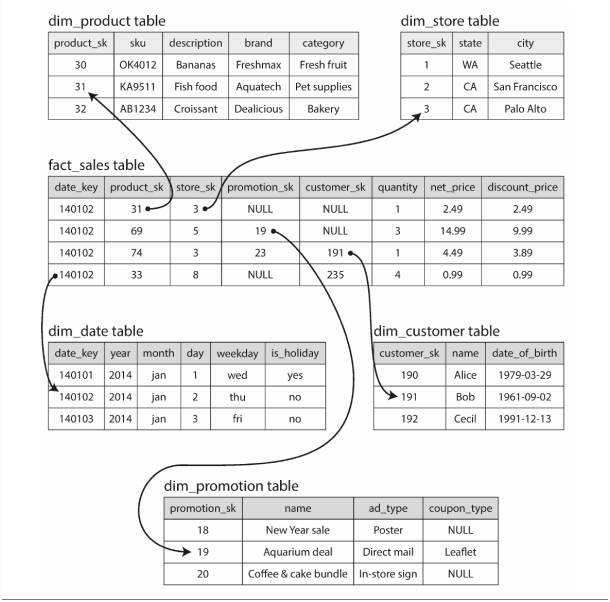
Stern- und Schneeflocken Schemas, cont'd

Dimensions Tabellen (dimension tables)

- In den *Dimensions Tabellen* werden weitere Daten wie das *wer, was, wo, wann, wie und warum* gespeichert
- Typische Dimensionstabellen sind z.B. Nutzer, Behandlungen, Standorte, Datum-Details, ...

Da ausgehend von der Fakten Tabelle sich mehrere Dimensions Tabellen sich Sternförmig ausbreiten wird dies auch das *Stern Schema* genannt. Wenn dies iteriert wird (also wieder Dimensionstabellen auch wieder Dimensionstabellen haben) wird dies *Schneeflocken Schema* genannt.

Stern- und Schneeflocken Schemas, Visualisierung



dim_product table

product_sk	sku	description	brand	category
30	OK4012	Bananas	Freshmax	Fresh fruit
31	KA9511	Fish food	Aquatech	Pet supplies
32	AB1234	Croissant	Dealicious	Bakery

dim_store table

store_sk	state	city
1	WA	Seattle
2	CA	San Francisco
3	CA	Palo Alto

fact_sales table

date_key	product_sk	store_sk	promotion_sk	customer_sk	quantity	net_price	discount_price
140102	31	3	NULL	NULL	1	2.49	2.49
140102	69	5	19	NULL	3	14.99	9.99
140102	74	3	23	191	1	4.49	3.89
140102	33	8	NULL	235	4	0.99	0.99

dim_date table

date_key	year	month	day	weekday	is_holiday
140101	2014	jan	1	wed	yes
140102	2014	jan	2	thu	no
140103	2014	jan	3	fri	no

dim_customer table

customer_sk	name	date_of_birth
190	Alice	1979-03-29
191	Bob	1961-09-02
192	Cecil	1991-12-13

dim_promotion table

promotion_sk	name	ad_type	coupon_type
18	New Year sale	Poster	NULL
19	Aquarium deal	Direct mail	Leaflet
20	Coffee & cake bundle	In-store sign	NULL

Kleppmann, M. (2017). Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems.

Speicherung in Spalten

Um möglichst effizient und mit SIMD Vektorisierung Aggregate wie SUM, MIN, MAX, AVG, COUNT zu berechnen bietet es sich an die Daten Spaltenweise statt zeilenweise zu speichern

fact_sales table

date_key	product_sk	store_sk	promotion_sk	customer_sk	quantity	net_price	discount_price
140102	69	4	NULL	NULL	1	13.99	13.99
140102	69	5	19	NULL	3	14.99	9.99
140102	69	5	NULL	191	1	14.99	14.99
140102	74	3	23	202	5	0.99	0.89
140103	31	2	NULL	NULL	1	2.49	2.49
140103	31	3	NULL	NULL	3	14.99	9.99
140103	31	3	21	123	1	49.99	39.99
140103	31	8	NULL	233	1	0.99	0.99

Columnar storage layout:

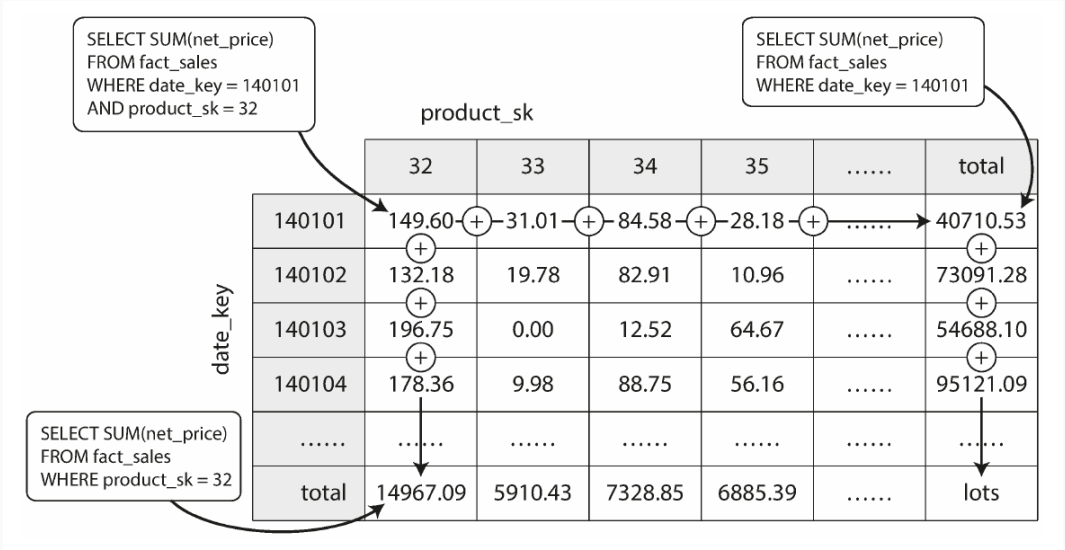
date_key file contents:	140102, 140102, 140102, 140102, 140103, 140103, 140103, 140103
product_sk file contents:	69, 69, 69, 74, 31, 31, 31, 31
store_sk file contents:	4, 5, 5, 3, 2, 3, 3, 8
promotion_sk file contents:	NULL, 19, NULL, 23, NULL, NULL, 21, NULL
customer_sk file contents:	NULL, NULL, 191, 202, NULL, NULL, 123, 233
quantity file contents:	1, 3, 1, 5, 1, 3, 1, 1
net_price file contents:	13.99, 14.99, 14.99, 0.99, 2.49, 14.99, 49.99, 0.99
discount_price file contents:	13.99, 9.99, 14.99, 0.89, 2.49, 9.99, 39.99, 0.99

Kleppmann, M. (2017). Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems.

Data Cubes

- Die verschiedenen Fact Table können als (hochdimensionaler) Würfel interpretiert werden, bei dem die Faktentabellen die Seiten des Würfel darstellen.
- Hierbei sprechen wir von *Data Cubes* oder auch *OLAP cubes*. Die Dimensionen sind natürlich nicht (wie im Beispiel) auf zwei Dimensionen beschränkt.
- Anhand verschiedener Dimensionen der Tabellen können Zeilen oder Spaltenweise z.B. SUM, MIN, MAX, AVG COUNT vorberechnet werden.
- Um möglichst effizient und mit SIMD Vektorisierung Aggregate wie SUM, MIN, MAX, AVG COUNT zu berechnen bietet es sich an die Daten Spaltenweise statt zeilenweise zu speichern

Data Cubes, Visualisierung



Kleppmann, M. (2017). Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems.

Big Data

Definition: Big Data

Unter *Big Data* versteht man Daten, die in großer Vielfalt, in großen Mengen und mit noch höherer Geschwindigkeit anfallen. Dies ist auch als die drei V-Begriffe bekannt (Variety, Volume, Velocity).

Quelle: https://www.oracle.com/de/big-data/what-is-big-data/

Big Data, Volumen

Die Menge an Daten ist wichtig. Bei Big Data müssen Sie große Mengen an unstrukturierten Daten mit geringer Dichte verarbeiten. Dabei kann es sich um Daten mit unbekanntem Wert handeln, z. B. Daten-Feeds von Twitter, Clickstreams von einer Webseite oder mobilen App oder Daten von Gerätesensoren. Für einige Unternehmen können das etliche Terabytes an Daten sein. Für andere Hunderte von Petabytes.

Big Data, Velocity (Geschwindigkeit)

Die Geschwindigkeit ist die Schnelligkeitsrate, mit der Daten empfangen werden und mit der (vielleicht) auf sie reagiert wird. Im Normalfall fließt die höchste Geschwindigkeit von Daten direkt in den Speicher und wird nicht auf eine Festplatte geschrieben. Einige internetfähige, intelligente Produkte arbeiten in Echtzeit oder beinahe in Echtzeit. Für sie sind Auswertungen und Aktionen in Echtzeit erforderlich.

Big Data, Variety (Vielfalt)

Vielfalt bezieht sich auf die zahlreichen verfügbaren Datentypen. Traditionelle Datentypen waren strukturiert und ideal für relationale Datenbanken geeignet. Durch die Zunahme von Big Data gibt es nun neue, unstrukturierte Datentypen. Unstrukturierte und semistrukturierte Datentypen wie Text, Audio und Video erfordern zusätzliche Vorabverarbeitung, um die Bedeutung und die unterstützenden Metadaten zu gewinnen.

Anwendungen von Big Data

Einige Anwendungen die unter Big Data fallen sind

- Large Hadron Collider am CERN $\rightarrow$ dieser hat bereits mehrere hundert Petabyte an Daten generiert
- Genom Daten wie z.B. GeneBank
- Der Google Suchindex

Nächstes Vorlesung

In der nächsten Vorlesung beschäftigen wir und mit Big Data Software, insbesondere für ETL Prozesse. Hierbei gehen wir auf den MapReduce Algorithmus ein und seine Implementierung bzw. Fortentwicklung. Dazu sehen wir insbesondere Apache Spark in Aktion.

Referenzen

- Kleppmann, M. (2017). Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems. " O'Reilly Media, Inc.".
- https://www.oracle.com/de/big-data/what-is-big-data/