Von der Datenquelle ins Data Warehouse

Die reale Welt der Datenanalyse

Wieso dieser Vortrag?

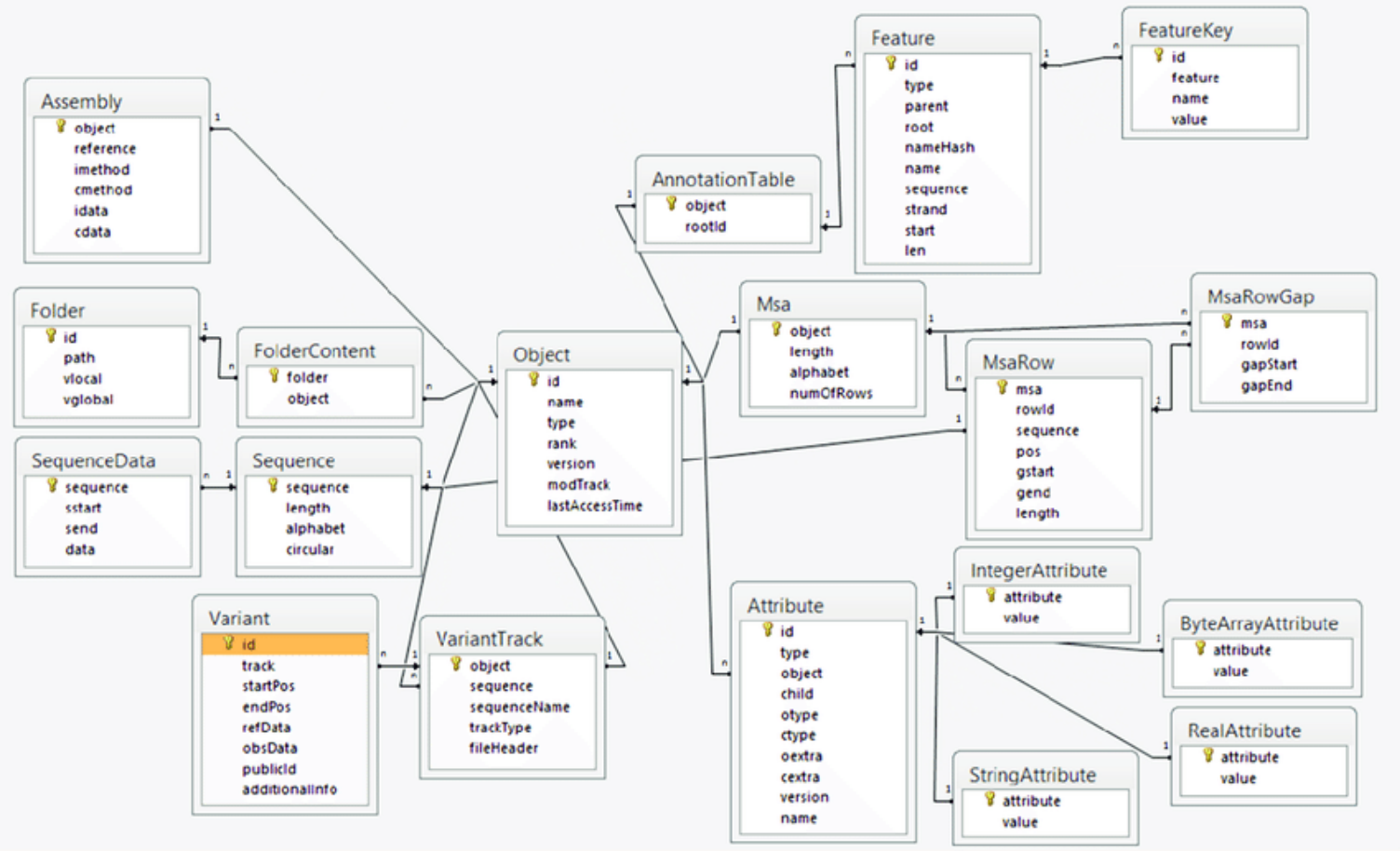
- Buzzwords: Data Science & Engineering, Machine Learning, Artificial Intelligence, Neural Networks, Big Data, …
- Alles basiert auf Daten
- Diese Daten müssen zuerst «langweilig» akquiriert und aufbereitet werden
- **-> Ohne gute Daten keine datenbasierten Applikationen!**

- *Und hinter jeder guten Visualisierung steckt viel Arbeit in der Datenaufbereitung*

Wo sind die Daten vorhanden?

- Zu 99% kommen Daten aus einem ERP oder CRM
- «Coole Datenanalysen» macht man nur mit einem PhD in Statistik ☺ (Oder Epidemiologie ;))
- Gleich wie bei Programmierjobs, das meiste ist basierend auf Geschäftsprozessen

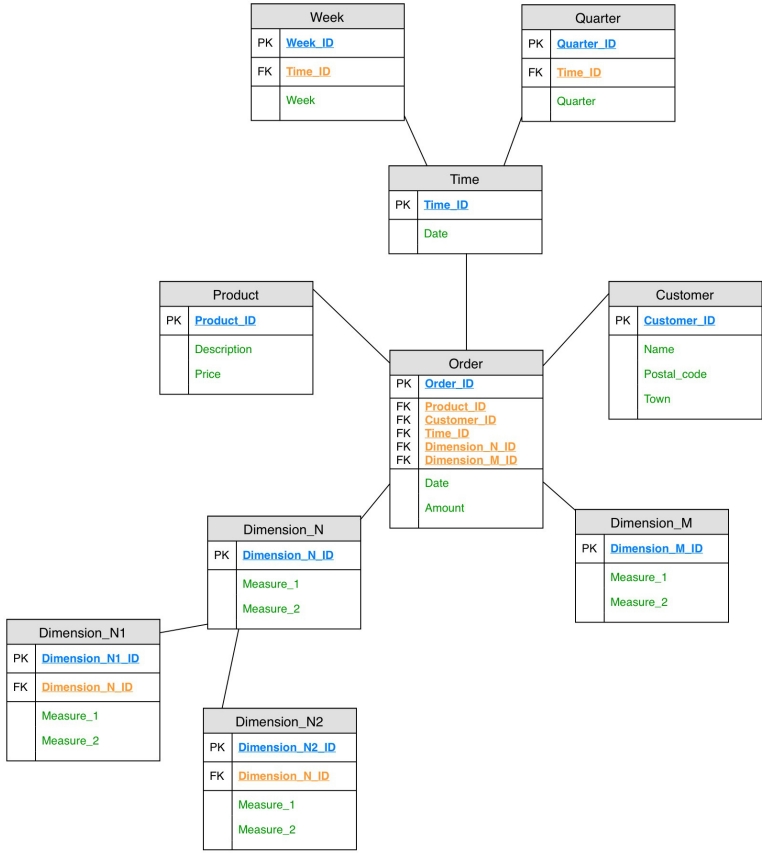
Wie sind die Daten?



(Kompliziert)

Quelle: https://www.researchgate.net/figure/ER-diagram-of-the-shared-data-storage-for-complex-data-types-The-central-entity-here_fig2_321663051

Wie sollten die Daten sein?



(schön)

Quelle: https://www.bbht.de/blog/data-vault-ein-agiles-data-warehouse.html

Wie sollten die Daten sein?

- Fakten und Dimensionen
- Fakten beinhalten FK's die auf Dimensionen verweisen
- Ein Fact beschreibt einen Geschäftsfall:

«Kunde Meier hat am 01.04.2020 ein Brot eingekauft in der Filiale Luzern, er hat dafür 2 Franken gezahlt mit seiner EC Karte. Der Verkäufer war Huber.»

Dimensionen: Kunde, Datum, Produkt, Standort, Zahlungsart, Verkäufer.

«Die Taxifahrt von Kunde Meier war am 01.04.2020, er hat vorher angerufen. Es ist Fahrer Huber ausgerückt und hat Herrn Meier von Luzern nach Kriens gefahren. Die Fahrt kostete 20 Franken, Meier hat noch 3 Franken Trinkgeld gegeben»

Dimensionen: Art der Dienstleistung, Kunde, Datum, Buchungsart, Fahrer, Startort, Zielort.

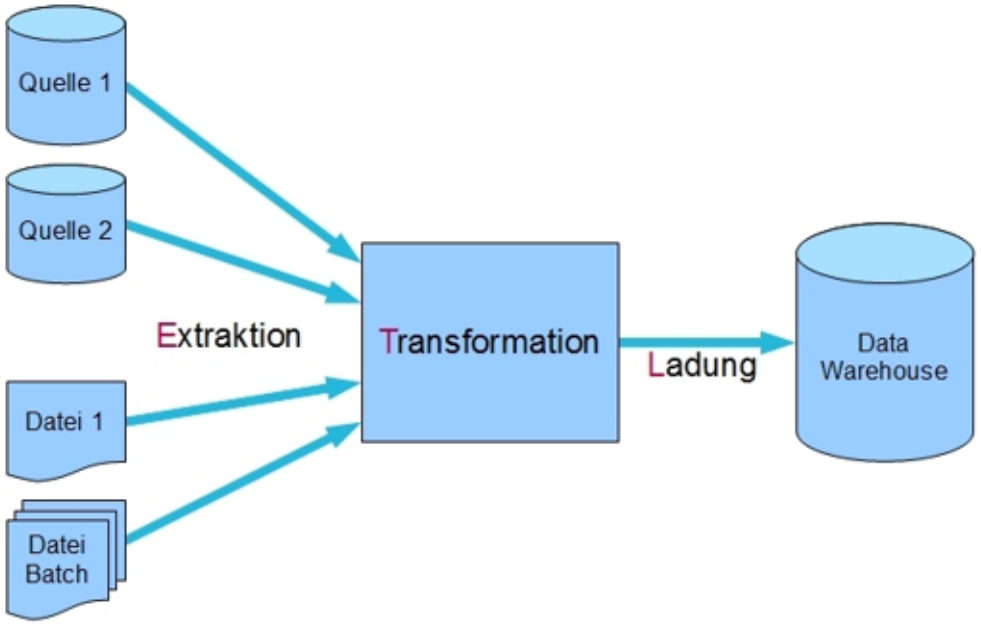
Bezahlte Beträge sind meistens direkt am Fact angehängt.

OLTP vs. OLAP

Merkmal	OLTP	OLAP
Grundlegende Orientierung	transaktionsorientiert	analyse- und themenorientiert
Anwendungsbereich	operative Systeme (Administrations- und Dispositionssysteme)	Data Warehouse Systeme
Nutzer	Administration, Sachbearbeiter	Entscheidungs- und Führungskräfte, Manager
Datenstruktur	zweidimensional, anwendungsbezogen	multidimensional, subjektbezogen
Dateninhalt	detaillierte, nicht verdichtete Einzeldaten	verdichtete und abgeleitete Daten
Datenverwaltungsziele	transaktionale Konsistenzerhaltung	zeitbasierte Versionierung
Datenaktualität	aktuelle Geschäftsdaten	historische Verlaufsdaten
Datenaktualisierung	durch laufende Geschäftsvorfälle	periodische Datenaktualisierung („Snapshot“)
Zugriffsform	lesen/schreiben/löschen	lesen/verdichten
Zugriffsmuster	vorhersehbar, repetitiv	ad hoc, heuristisch
Zugriffshäufigkeit	hoch	mittel bis niedrig
Transaktionsart und Dauer	kurze Lese und Schreiboperationen	lange Lesetransaktionen

https://www.tecchannel.de/a/bi-methoden-teil-1-ad-hoc-analysen-mit-olap,1751285,2

Wie werden die Daten so?

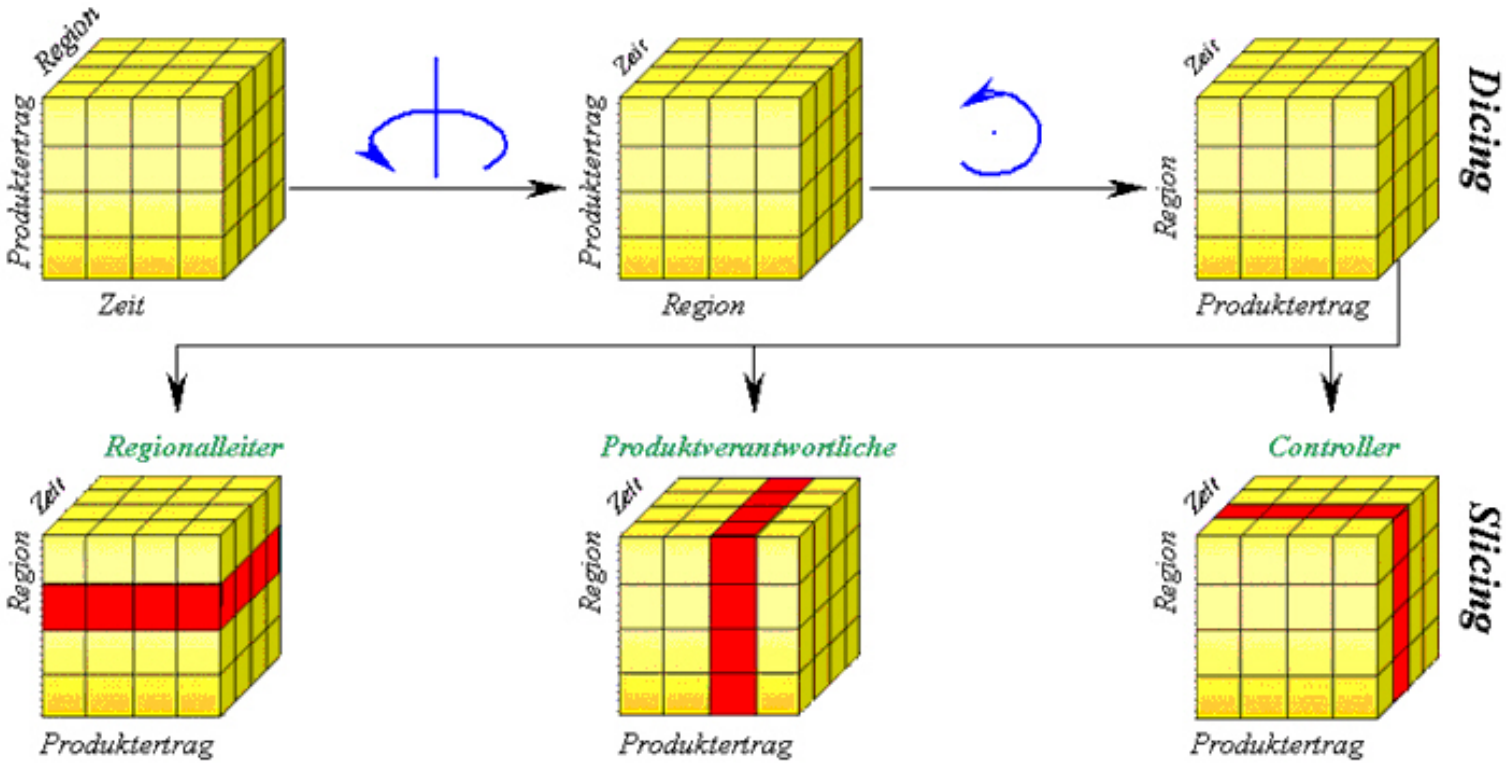


Extract: Die Daten werden über eine Schnittstelle vom Quellsystem in eine Staging Area geladen.
Transform: Die Daten werden bearbeitet (Daten werden berechnet, Keys ausgetauscht, Verdichtet, Ergänzt)
Load: Die Daten kommen in ein DataWarehouse (Datenbank) und sind bereit analysiert zu werden.

https://www.itwissen.info/ETL-extract-transform-load-Extrakt-Transformation-Ladung.html

Was jetzt?

- Reports / Dashboards
- Cubes / Datamarts

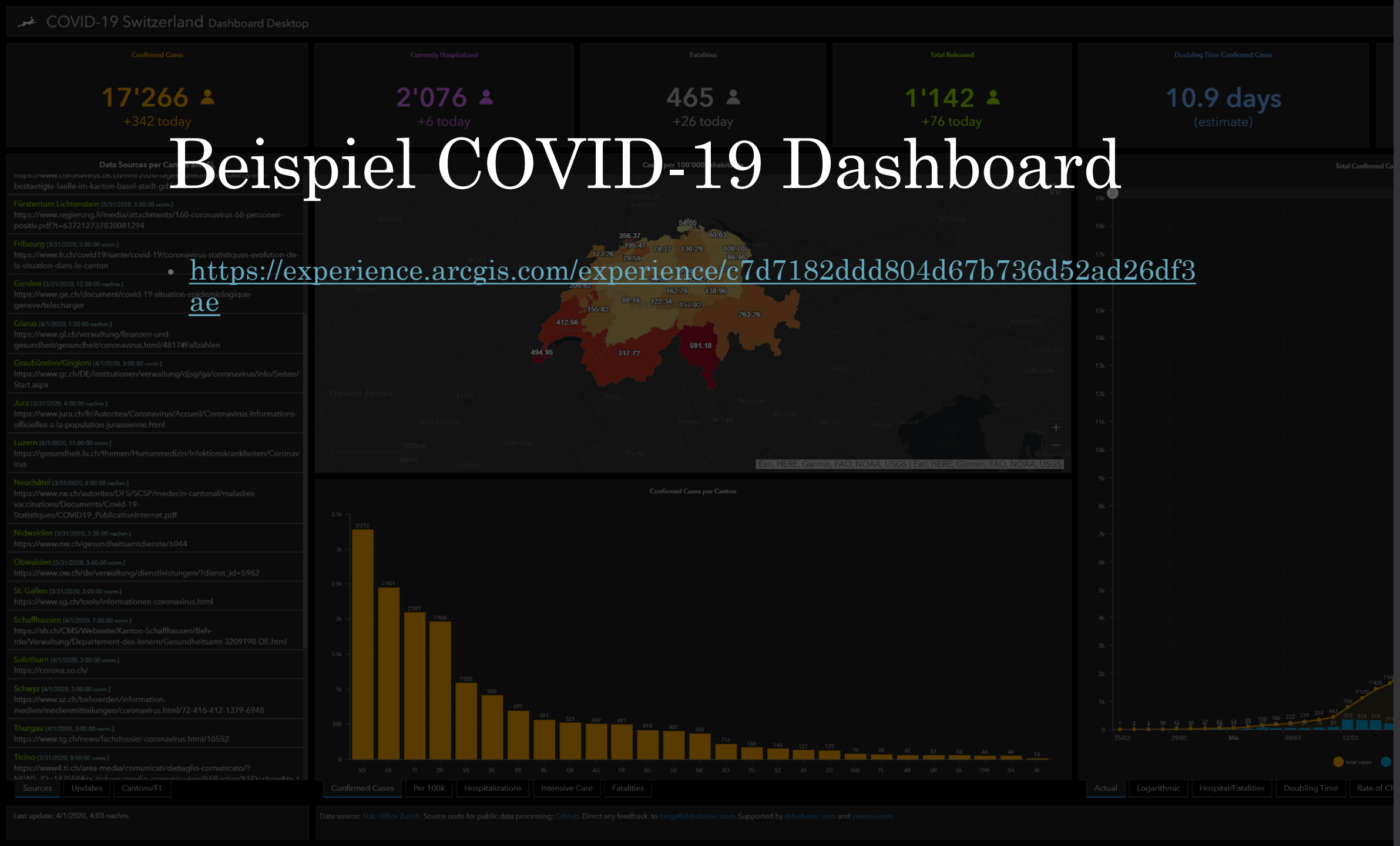


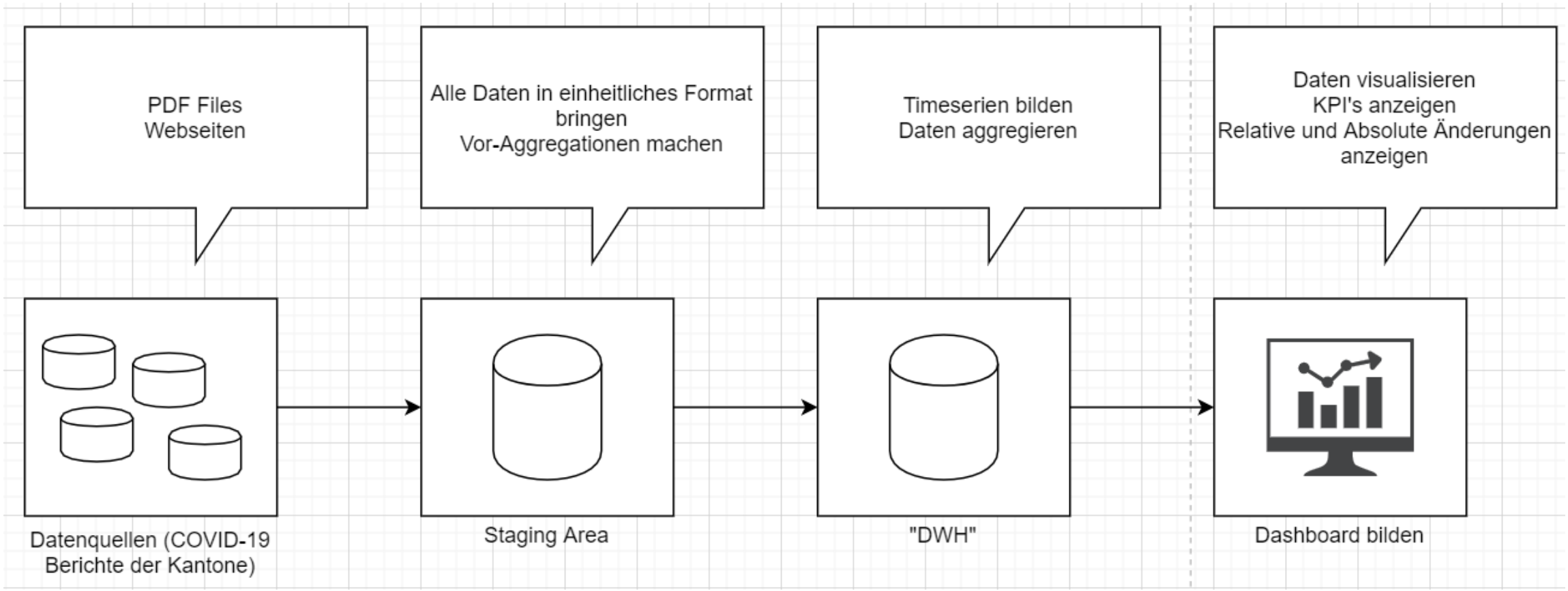
Was sind Cubes?

https://linearis.at/blog/2012/10/25/was-kann-ein-cube-eigentlich/

Beispiel COVID-19 Dashboard

- https://experience.arcgis.com/experience/c7d7182ddd804d67b736d52ad26df3ae




PDF Files
Webseiten
Alle Daten in einheitliches Format bringen
Vor-Aggregationen machen
Timeserien bilden
Daten aggregieren
Daten visualisieren
KPI's anzeigen
Relative und Absolute Änderungen anzeigen
Datenquellen (COVID-19 Berichte der Kantone)
Staging Area
"DWH"
Dashboard bilden

Was brauche ich damit ich als Data Engineer / Business Intelligence Specialist arbeiten kann?

- Grundlegendes Verständnis von Datenbanktechnologien (OLAP und OLTP)
- Gute Kenntnisse in SQL (nicht nur Selects, SQL kann noch viel mehr!)
- Domänenwissen der Branche
- IT Wissen
- Scriptsprachen wie PowerShell, R, Python von grossem Vorteil
- Hohe Problemlösungskompetenz (kein Problem ist gleich wie das andere)
- Lust täglich neue Rätsel zu lösen
- Lust täglich neue Dinge und Technologien zu lernen
- Gestalterisches Flair

Was brauche ich um als Data Scientist arbeiten zu können?

- Mathematik und Statistik
- … Mathematik und Statistik
- Python, Julia oder R
- **Breites Domänenwissen**
  - Ich habe letzte Woche ein statistisch korrektes Modell erstellt um Bestellungen vorauszusagen, nur leider weiss ich, dass der Onlineshop nicht 2'000'000 Bestellungen pro Tag verarbeiten kann ;-)

Fragen?