DHBW
Duale Hochschule
Baden-Württemberg
Mannheim

Kurs
Einführung
In das Thema
Data Ware House
&
Business Intelligence

Organisatorisches

Warm werden: Vorstellungsrunde

- Name: Joachim Brock
- Baujahr 1965, Verheiratet und zwei Töchter
- Ausbildung im Handwerk
- Studium: Kommunikationsinformatik
- Kerngebiete: Datenbanken, Business Intelligence, Schnittstellen und Software Entwicklung
- Werdegang: Softwarehaus, Eigene GmbH, Big-Date&KI, Expertensysteme, aktuell: Head of SWE bei AUVESY-MDT GmbH in Landau
- Erwartung: Interessierte lernbereiter Kurs mit engagierter Zusammenarbeit

Warm werden: Vorstellungsrunde

- Vorstellungsrunde
  - Name
  - Werdegang
  - Erwartungshaltung an den Kurs

Warm werden: Geimeinsame Regeln

- Regeln der Vorlesung:

1) Der aktuelle Redner darf ausreden
2) Fragen sind von allen erwünscht und nicht nur von wenigen
    - Manche Blöcke muss ich am Stück erklären. Keine Panik, Zeit zum Fragen kommt am Ende des Blocks; Dies Blöcke benenne ich extra
3) Pausen bitte einfordern, ich vergesse manchmal die Zeit
4) Bei Online: Wenn die Technik mit macht sollten wir die Kameras einsetzen

Warm werden: Orga

- Organisation der Vorlesung
  - Theorie
  - Beispiele aus der Praxis
  - Aufgaben und Übungen
    - Vorstellung durch Teilnehmer
  - Wiederholung der vorherigen Vorlesung im Schnelldurchgang
  - Prüfung KW 51/2023

- **Fragen vorab oder noch offene Punkte?**

DHBW
Duale Hochschule
Baden-Württemberg
Mannheim

Kurs
Einführung
In das Thema
Data Ware House
&
Business Intelligence

Kapitel 1: Erste Grundlagen

Business Intelligence

Was umfasst BI?

- OLAP (Online Analytical Processing)
  - Umfasst auch das DWH samt Dantenbanken
  - Weitere Themen sind Verteilung, Datenbeschaffung, Daten Vorbereitungen
- Analyse
  - Auswertunge, Statistiken, zyklisch & adhoc, Entscheidungsvorlagen
- Data Mining
  - Korelationen, Kausalitäten, Wissenbasiertes Lernen und Prognosen
- Projektorganisation
  - Planung, Aufbau, Pflege und Betrieb on BI-Systemen

OLAP

12 Regeln nach Edward F. Codd aus dem Jahr 1993

1) Multidimensionale Sicht auf die Daten
2) Transparenz (Trennung von UI und Architektur)
3) Zugriffsmöglichkeiten (Daten aus Operativen Systemen)
4) Konsistente Leistungsfähigkeit der Berichterstattung
5) Client-Server-Architektur mit Lasterverteilung
6) Generische Dimensionalität mit einheitlicher Dimensionierung
7) Dynamische Handhabung dünn besetzter Matrizen
8) Mehrbenutzerunterstützung
9) Einheitliche dimensionsübergreifende Operationen
10) Intuitive Datenanalyse
11) Flexibles Berichtswesen
12) Unbegrenzte Anzahl von Dimensionen und Konsolidierungsebenen

Quelle: https://www.hdm-stuttgart.de/~riekert/lehre/db-kelz/chap6.htm

OLAP

FASMI-Regeln nach Pendse und Creeth

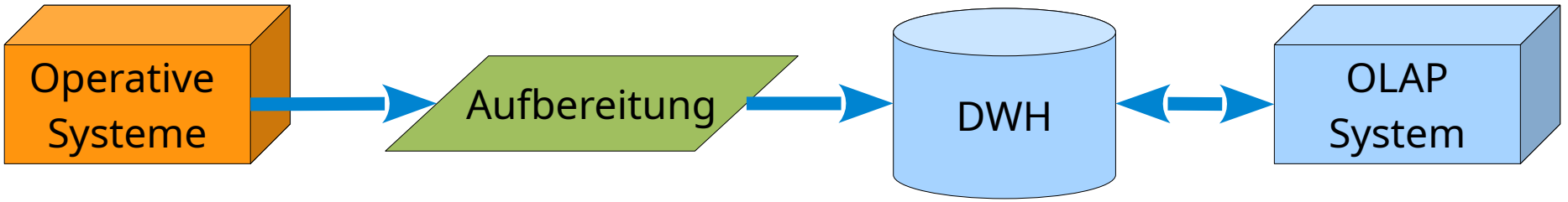
1) **F**ast = Schnelle Abfragen mit durchschnittlich 5 s bis max. 20 s
2) **A**nalysis = Einfache Analyse der Daten ermöglichen möglichst ohne Programmieraufwand
3) **S**hared = Mehrbenutzerbetrieb ermöglichen mit entsprechenden Schutzmaßnahmen
4) **M**ultidimensional = Struktur der Daten ermöglich beliebige Dimensionshierarchien
5) **I**nformation = Die Daten dürfen nicht durch das Systems in ihrer Transparenz beschränkt werden

Quelle: https://www.datenbanken-verstehen.de/business-intelligence/business-intelligence-grundlagen/anforderungen-business-intelligence/fasmi-regeln-pendse-creeth/

OLAP DWH-Architektur

DWH Systeme samt Aufbau und Betrieb

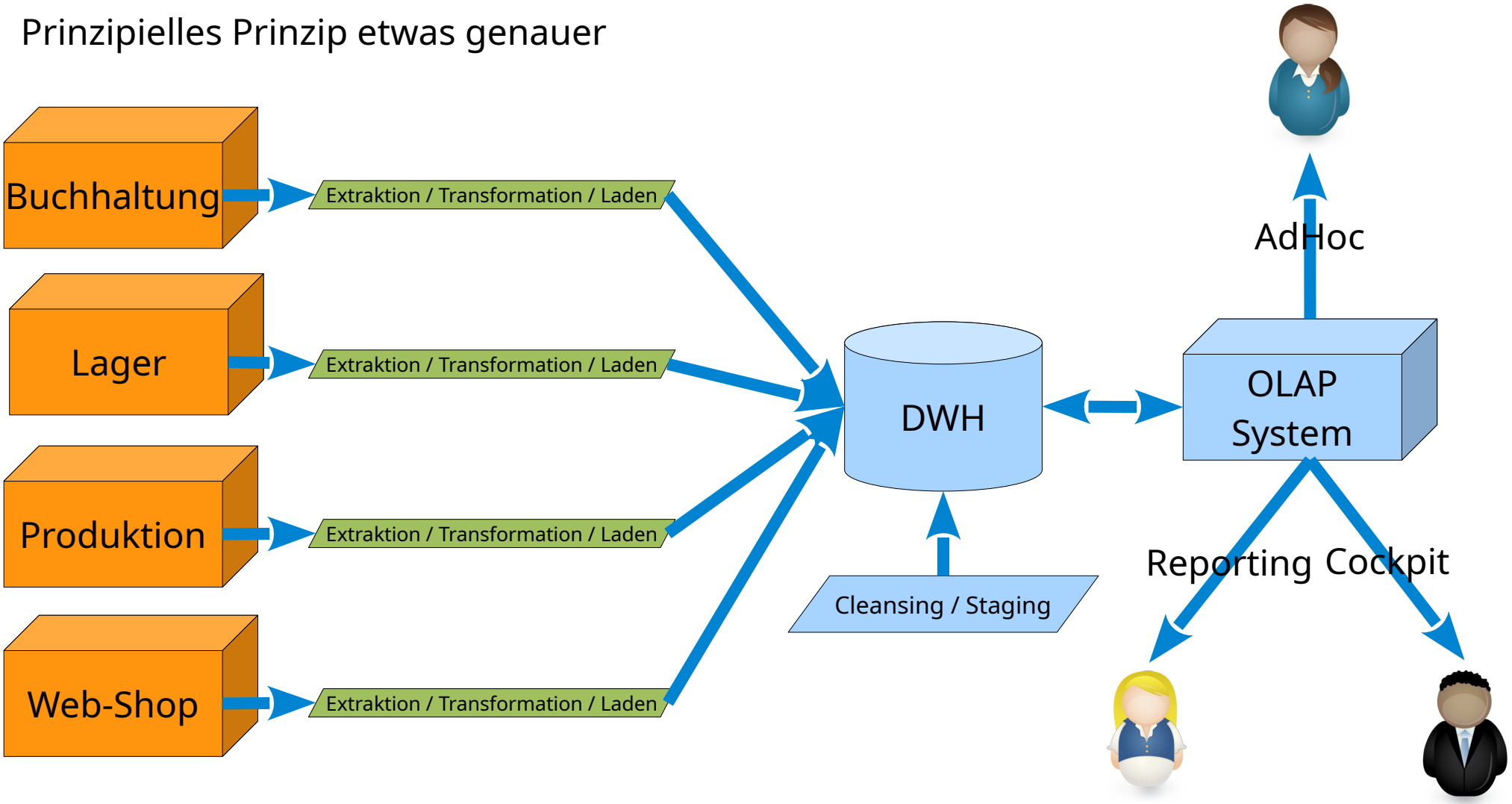
Prinzipielles Prinzip



OLAP DWH-Architektur

DWH Systeme samt Aufbau und Betrieb

Prinzipielles Prinzip etwas genauer



OLAP

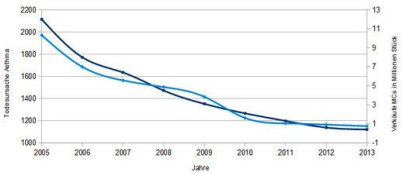
OLAP Anwendung

- Konzentration unterschiedliche Datenquellen
  - Datenreihen, Produktionsanlagen, Wirschaftssysteme, Statistiken, ...
- Ermöglicht globale Sicht auf Daten
  - Knzentration und Agregation von unübersichtliche Datenmengen
- Ermitteln von Korrelationen und Kausalität
  - Bitte nicht verwechseln!
  - Beispiel: Selbstmordrate und Nasa-Investitionen
- Beobachtung von Daten-Entwicklungen, auch von temporalen
  - Bis hin zu echzeitbeachtungen in Monitoring-Systemen
- Entscheidungsunterlage bieten
  - Analyse-Ergebnisse als Basis für Entscheidungen
  - Beispiel: Vorhersage Papierdicke zur Produktionssteuerung

OLAP

OLAP Risiken

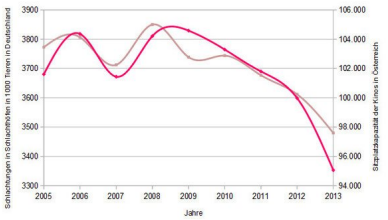
- Unvollständige Dimensionen
  - Korrelationen hängen an Dimesnsionen die nicht enthalten sind
- Fehlerhafte Daten
  - Verfäschung von Analyseergebnissen
- Große Lücken in den Faktendaten
  - Lücken in den Daten reduzieren Aussagekraft
- Dupletten in den Daten oder fehlende single Point of Truth
  - Wiedersprechende oder falsch verstärkende uneindeutige Wirkungen
- Datenschutzverletzungen
  - Einfache personenbezogene Daten
  - Besonders geschützte personenbezogene Daten, z.B. Medizin
- Scheinkorrelationen ohne Kausalitäten
  - Beispiel: MC<=>Asthma, siehe: https://scheinkorrelation.jimdofree.com/



OLAP

Verständnisfragen

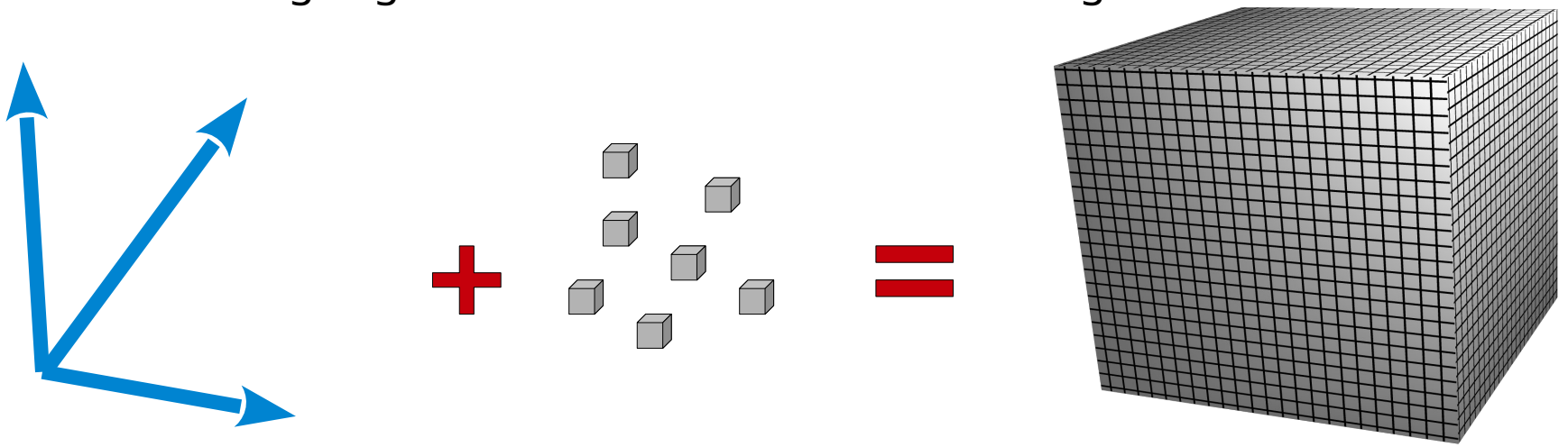
- Warum werden operative Daten von Auswertungsdaten getrennt?
- Welche Bereiche in einem Unternehmen haben Interesse an OLAP?
  Bitte nennen Sie min. 4 Bereiche und deren Nutzen.
- Zwei Vertriebsstellen mit überschneidenden Rechnungskreise benötigen ein gemeinsames OLAP, was empfehlen Sie als ersten Schritt?
- Ein Pharmakonzern bietet Ihnen hohe Summen für Ihre medizinischen Labordaten. Was ist zu beachten?
- In den Jahren 2005 bis 2013 korrelieren die Schlachtungen in deutschen Schlachthöfen und die Sitzplatzkapazität der österreichischen Kinos.
  Was leiten Sie daraus ab?



OLAP DWH-Architektur

DWH gundsätzliche logische Architektur

- Mehrdimensionale logische Datenarchitektur
- Fakten werden in Form einer Matrize gehalten
- Gruppierung der Stammdaten bestimmen die Dimensionen
- Symbolische Ähnlichkeit mit einem Würfel
- Normalisierung zugunsten Performance vernachlässigt

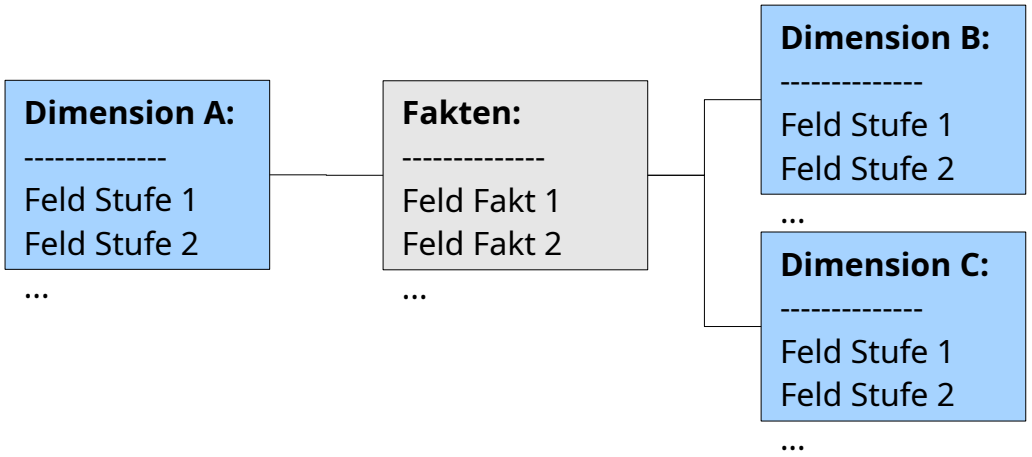


Dimensionen Fakten Data Warehouse Würfel

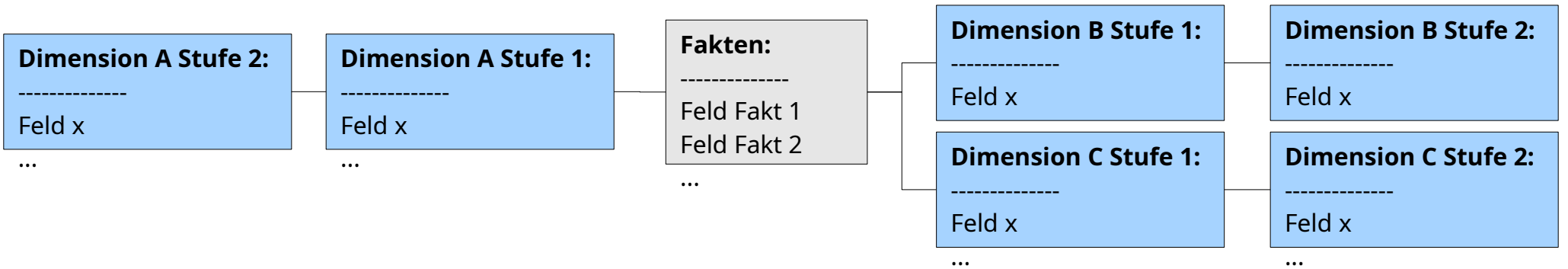
OLAP DWH-Architektur

DWH logische Architektur der Dimensionen

- Star Schema
  - Jede Dimension wird in einer Tabelle/Objekt zusammengefasst



- Snowflake Schema
  - Jede Hierarchie einer Dimension wird in einer Tabelle/Objekt gehalten



OLAP DWH-Architektur

DWH logische Architektur der Dimensionen

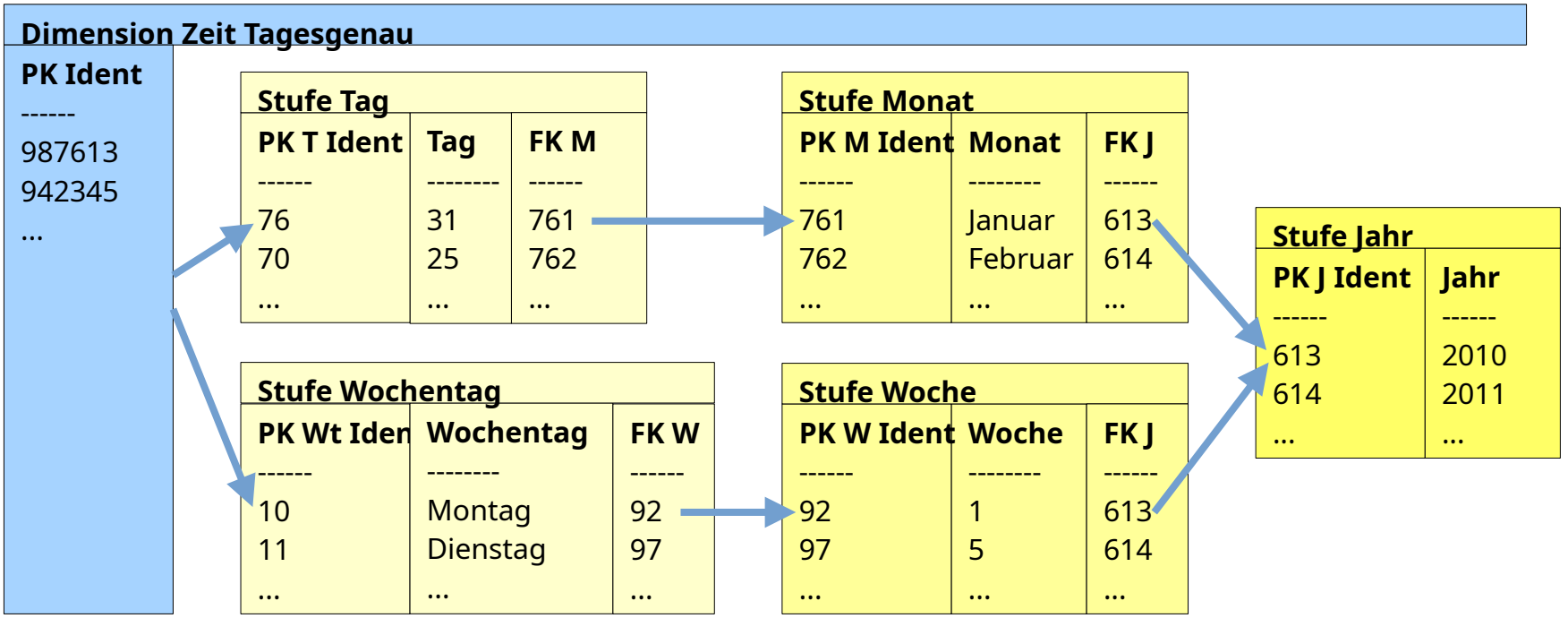
- Beispiel Star Schema für Zeitreihen oder Temporalisierung
  - Pro Tag einen Eintrag
  - Mehrere Pfade der Hierarchie gleichzeitig möglich
  - Einen künstlichen Primärschlüssel ist empfehlenswert

Dimension Zeit Tagesgenau					
PK Ident	**Jahr**	**Monat**	**Woche**	**Tag**	**Wochentag**
------	------	--------	--------	--------	--------
987613	2010	Januar	1	31	Montag
942345	2011	Februar	5	25	Dienstag
…	…	…	…	…	…

OLAP DWH-Architektur

DWH logische Architektur der Dimensionen

- Ein Beispiel Snowflake Schema für Zeitreihen oder Temporalisierung
  - Pro Tag mehrere gleichzeitige Pfade möglich
  - Jeweilige künstlichen Primärschlüssel ist empfehlenswert
  - Travesierung durch die Hierarchiestufen notwendig



OLAP DWH-Architektur

DWH logische Architektur der Dimensionen

Stufe Tag

PK T Ident	Tag	FK Monat
73	1	761
75	1	792
76	31	901
70	25	901
...	...	...

Stufe Monat

PK M Ident	Monat	FK Jahr
761	Januar	613
762	Februar	613
792	Januar	614
901	Januar	615
...	...	...

Stufe Jahr

PK J Ident	Jahr
613	2020
614	2021
...	...

Fakten Vertrieb

PK ID	FK T	FK N	Umsatz
5001	73	1001	10.333,05 €
5002	75	1001	42.123,44 €
5003	73	1003	72.042,42 €
5042	70	1042	1.099,01 €
...	...	...	...

- Beispiel Snowflake Schema Zeitreihen und Regionen
  - Fakten referenzieren nur auf die niedrigste Granularität
  - Redundanzen in den Stufen ggf. notwendigen

Umsatz 42.123,44 €
am 1. Januar 2021
in der deutschen Zentral in Mannheim

Stufe Niederlassung

PK N Ident	Name	FK S
1001	Zentrale	761
1002	Notre Dame	792
1003	Limmat	901
1042	Oerlikon	901
...	...	...

Stufe Stadt

PK S Ident	Stadt	FK Land
761	Mannheim	49
792	Paris	33
712	Luzern	41
901	Zürich	41
...	...	...

Stufe Land

PK L Ident	ISO2
49	DE
41	SW
33	FR
...	...

OLAP DWH-Architektur

Verständnisfragen

- Warum wird im DWH gerne gegen die Normalformen verstoßen?
- Nennen Sie min. zwei weitere Dimensionen, bei denen ein Verstoß gegen Normalform sinnvoll ist.
- Sie sollen ein DWH entwerfen mit extrem schnellen Zugriffzeiten. Welches Modell (Star oder Snowflake) wählen Sie?
- Bauen Sie ein Starschema für den Vertrieb mit folgenden Inhalten: Verkaufzeitstempel, Menge, Einzelpreis, MwSt, Produktname, Produktgruppe, Filiale, Ort, PLZ.
- Schreiben Sie eine SQL Abfrage, welche die Verkaufssumme der jeweiligen Produktgruppen in den Filialen zwischen 12:00 Uhr und 13:00 Uhr am 18.08.2022 ausgibt.