



Information Integration
Materialisierte vs. Virtuelle Integration

4.11.2019
Felix Naumann

Überblick

1. Szenarien der Informationsintegration

- **Data Warehouse**
- Föderierte Datenbanken

2. Einführung

3. Materialisiert

- Data Warehouse

4. Virtuell

- Mediator-Wrapper System

5. Vergleich

- Flexibilität
- Antwortzeiten
- Aktualität
- etc.



Felix Naumann
Information Integration
Winter 2019/20

Real-life Informationsintegration

- Überblick: Zwei wesentliche Modelle

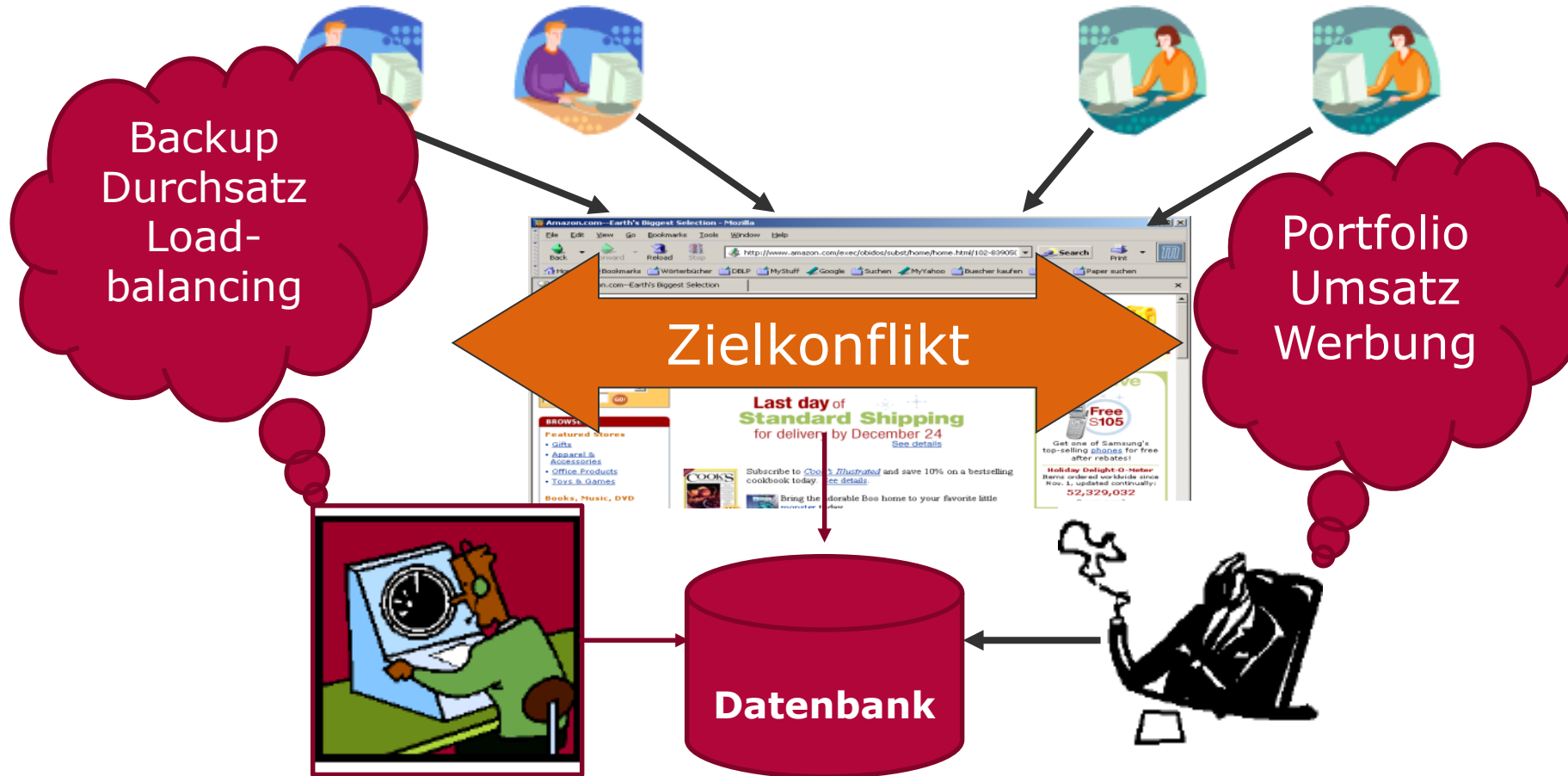
- Data Warehouses
 - Materialisierte Integration
 - Am Beispiel Buchhändler (Folien von Prof. Leser)

- Föderierte Datenbanken
 - Virtuelle Integration
 - Am Beispiel einer Life Sciences DB (DiscoveryLink)
 - Weitere Beispiele

Data Warehouse

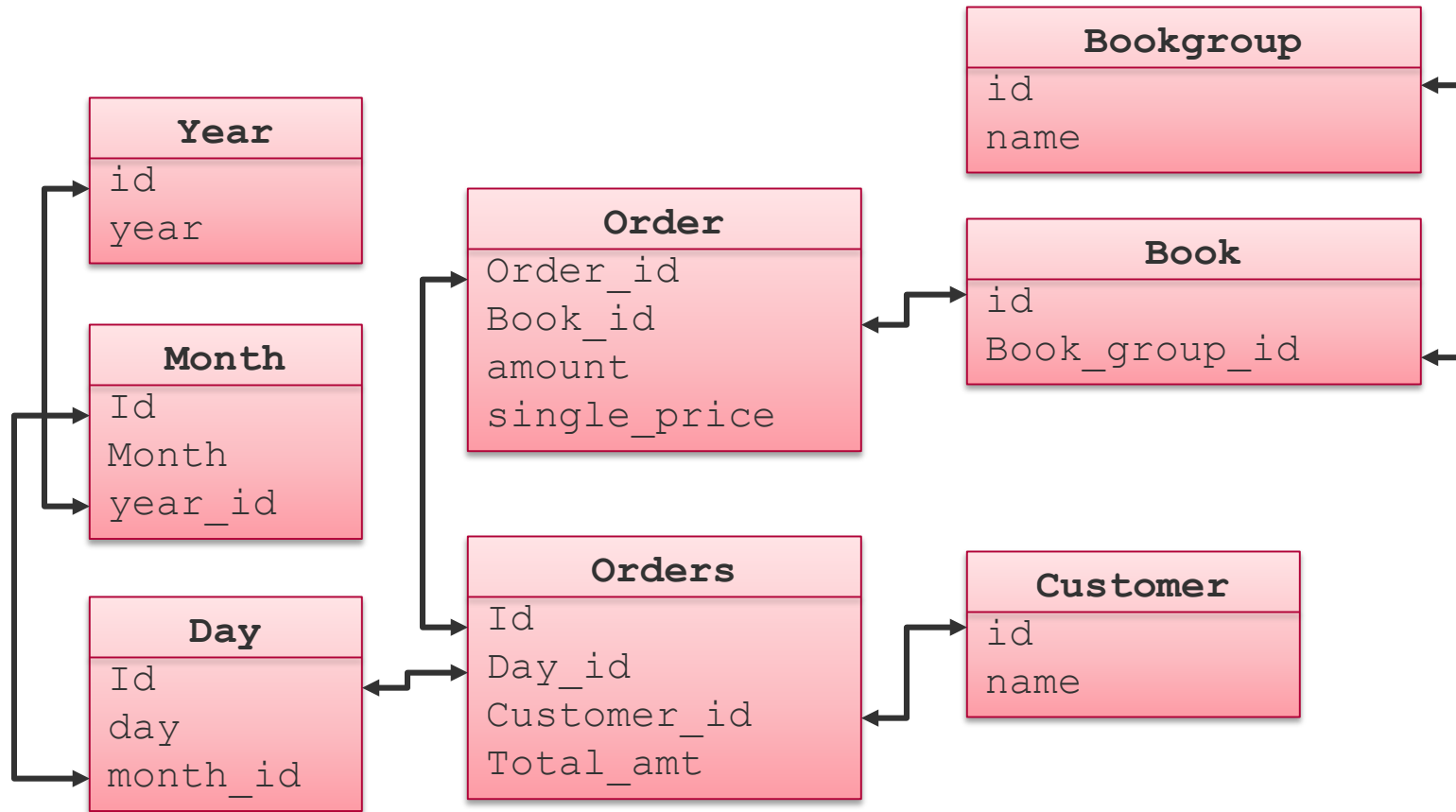
- Eine oder mehrere (ähnliche) Datenbanken mit Bücherverkaufsinformationen
- Daten werden oft aktualisiert
 - Jede Bestellung einzeln
 - Katalog Updates täglich
- Management benötigt Entscheidungshilfen (decision support)
- Komplexe Anfragen

Bücher im Internet bestellen



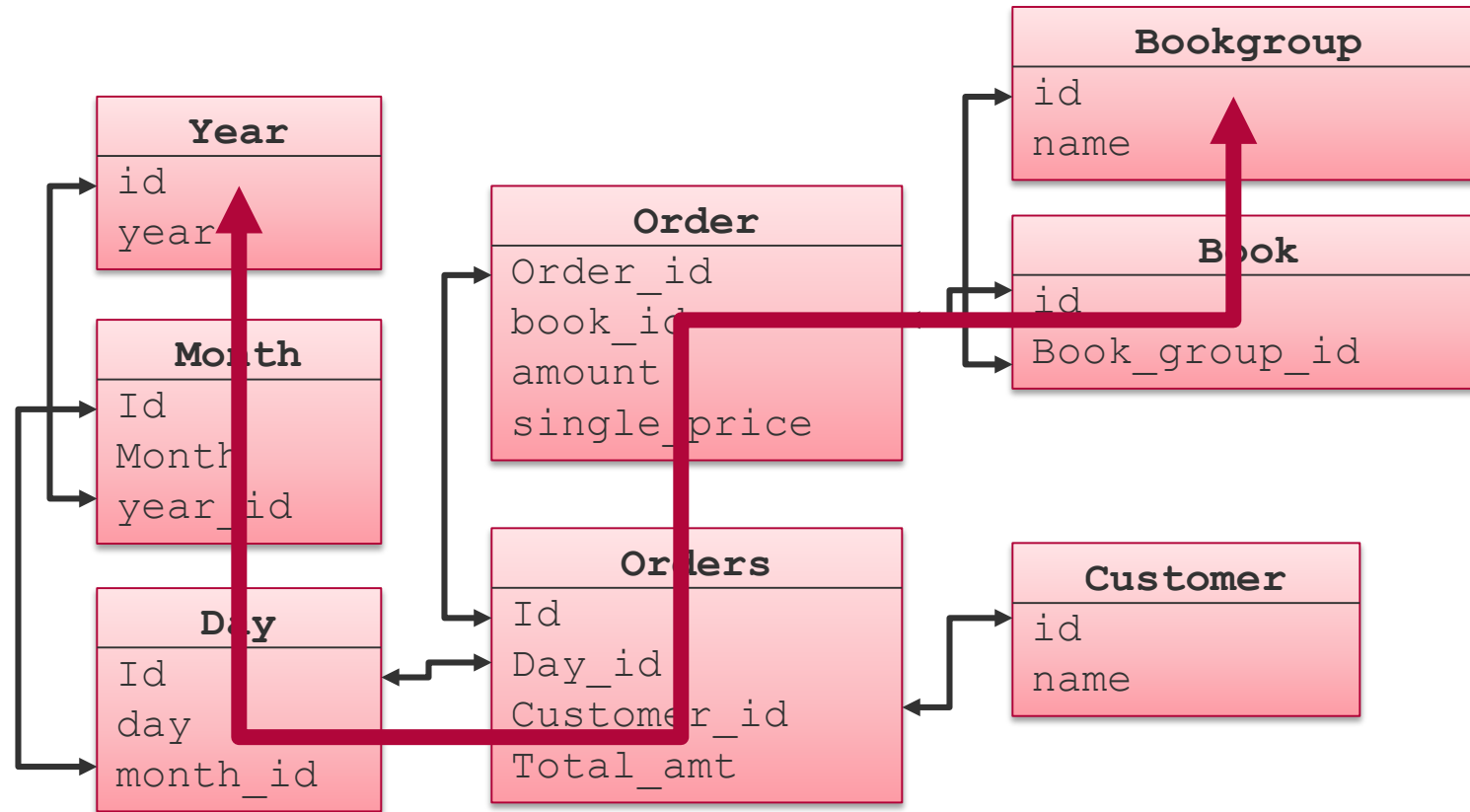
Felix Naumann
Information Integration
Winter 2019/20

Die Datenbank dazu



Fragen eines Marketingleiters

Wie viele Bestellungen haben wir jeweils im Monat vor Weihnachten, aufgeschlüsselt nach Produktgruppen?

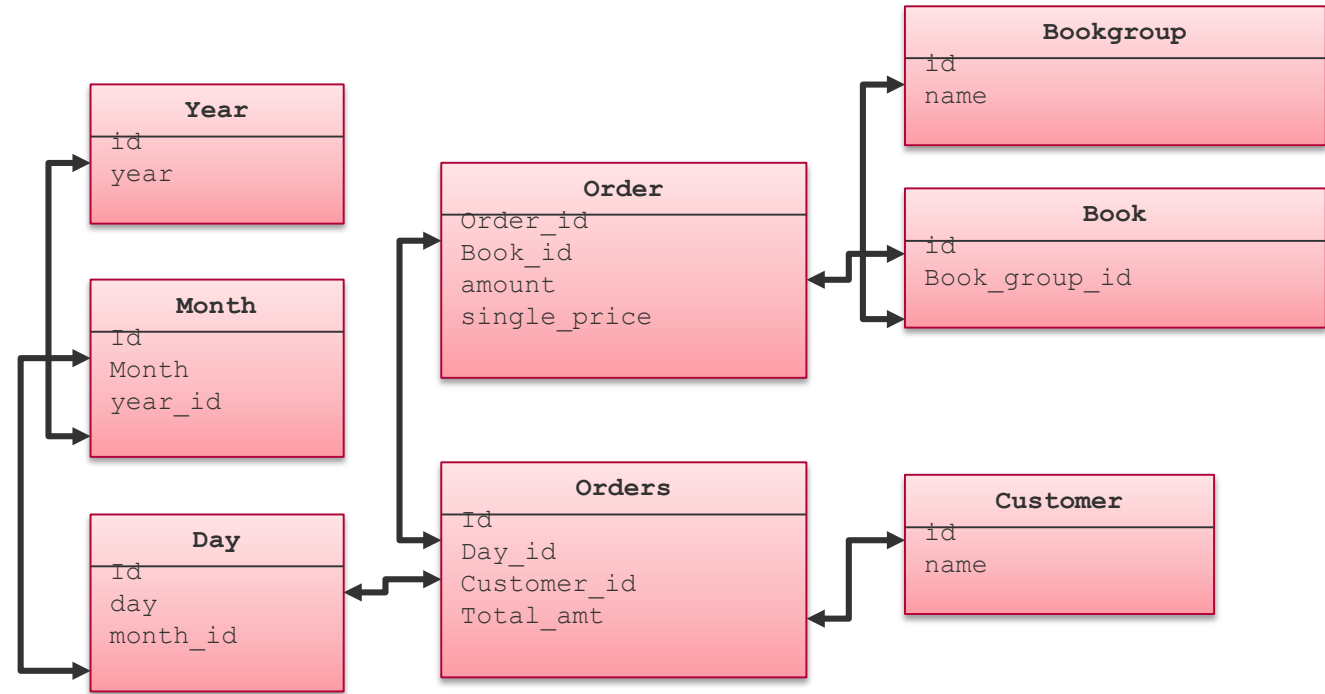


Technisch

```

SELECT Y.year, BG.name, count(B.id)
FROM year Y, month M, day D, order O,
orders OS, book B, bookgroup BG
WHERE M.year = Y.id
AND M.id = D.month
AND O.day_id = D.id
AND OS.order_id = O.id
AND B.id = O.book_id
AND B.book_group_id = BG.id
AND day < 24 and month = 12
GROUP BY Y.year, BG.name
ORDER BY Y.year

```



Felix Naumann
Information Integration
Winter 2019/20

Technisch

```
SELECT Y.year, BG.name, count(B.id)
FROM   year Y, month M, day D, order O,
       orders OS, book B, bookgroup BG
WHERE  M.year = Y.id
AND    M.id = D.month
AND    O.day_id = D.id
AND    OS.order_id = O.id
AND    B.id = O.book_id
AND    B.book_group_id = BG.id
AND    day < 24 and month = 12
GROUP BY Y.year, BG.name
ORDER BY Y.year
```

6 Joins

- Year: 10 Records
- Month: 120 Records
- Day: 3650 Records
- Orders: 36.000.000
- Order: 72.000.000
- Books: 200.000
- Bookgroups: 100

Problem!

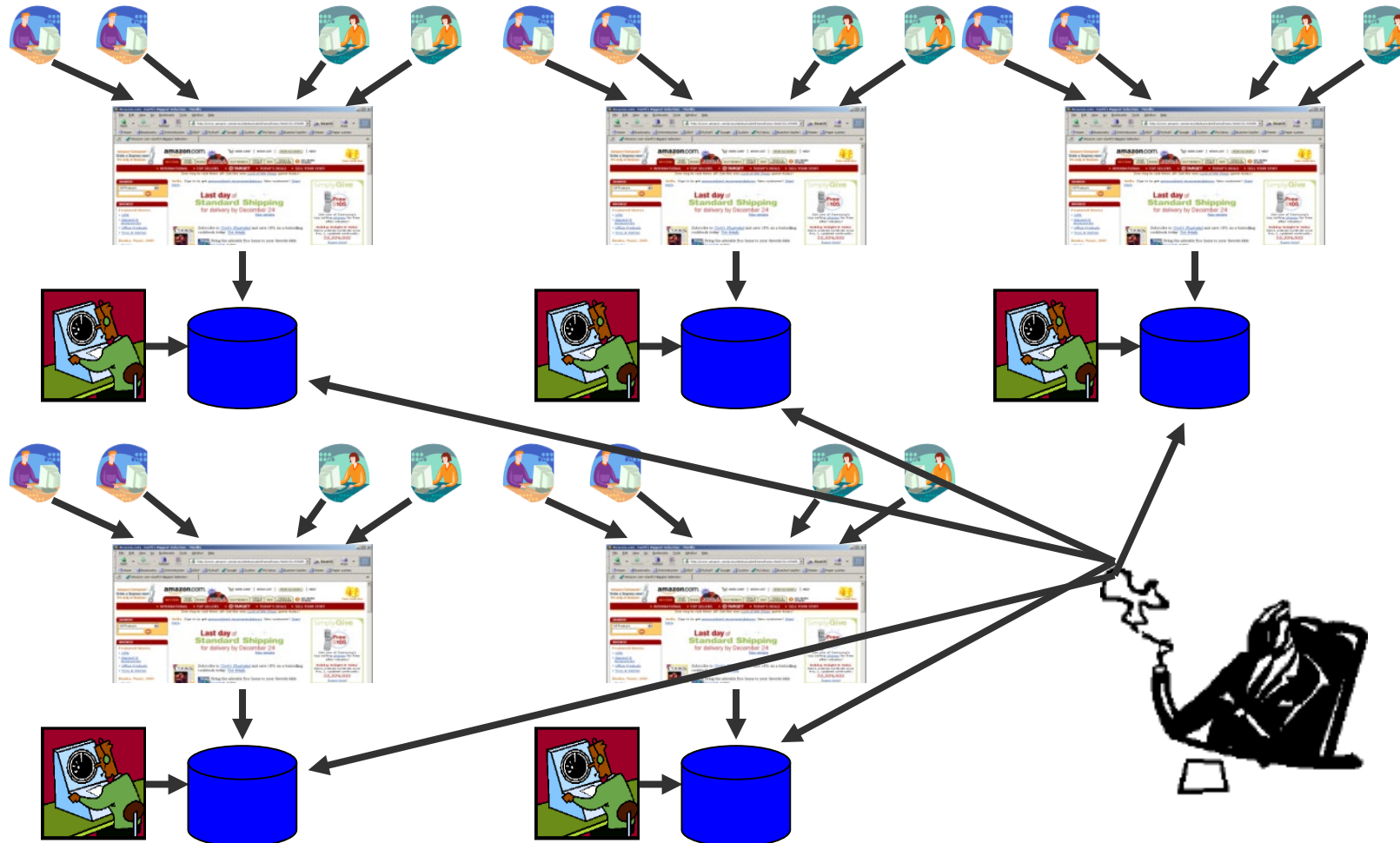
- Schwierig zu optimieren (Join-Reihenfolge)
- Je nach Ausführungsplan riesige Zwischenergebnisse
- Ähnliche Anfragen – ähnlich riesige Zwischenergebnisse

In Wahrheit ... noch schlimmer

- Es gibt noch:
 - Amazon.de
 - Amazon.fr
 - Amazon.it
 - ...
- Verteilte Ausführung
 - Count über Union mehrerer gleicher Anfragen in unterschiedlichen Datenbanken

HILFE!

In Wahrheit ...



Felix Naumann
Information Integration
Winter 2019/20

Technisch: Eine VIEW

```
CREATE VIEW christmas AS
    SELECT Y.year, PG.name, count(B.id)
FROM      DE.year Y, DE.month M, DE.day D, DE.order O, ...
WHERE     M.year = Y.id and
...
GROUP BY Y.year, PG.product_name
ORDER BY  Y.year
UNION
    SELECT Y.year, PG.name, count(B.id)
FROM      EN.year Y, EN.month M, EN.day D, EN.order O, ...
WHERE     M.year = Y.id and
...
```

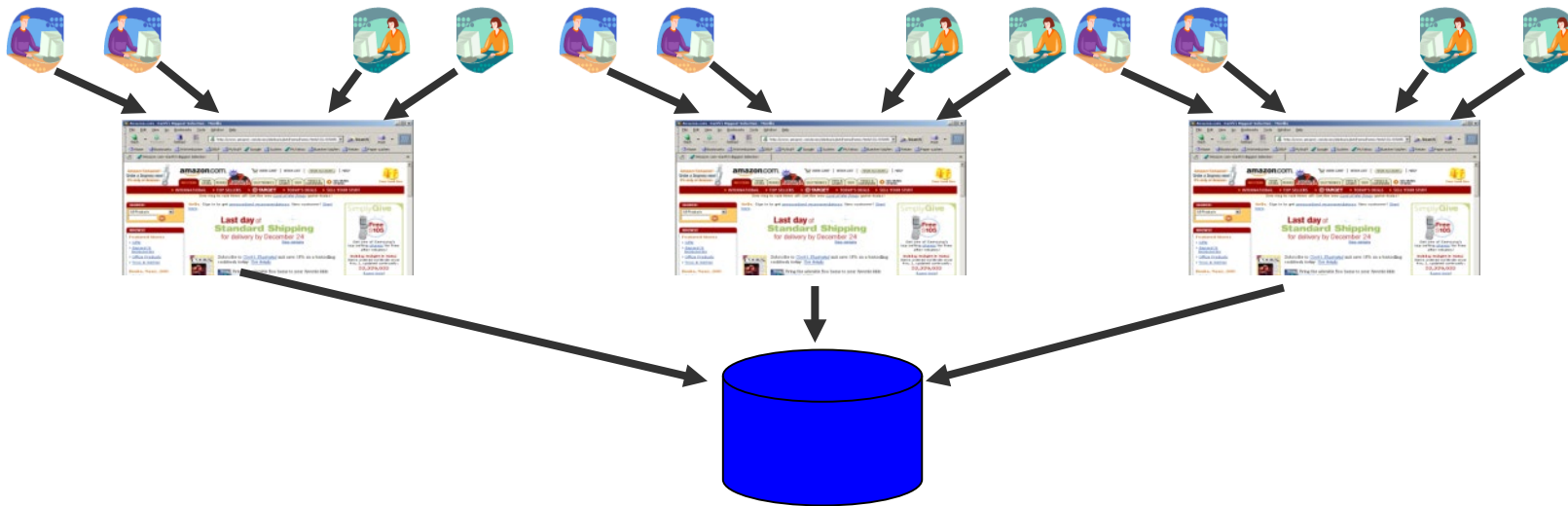
```
SELECT year, name, count(B.id)
FROM christmas
GROUP BY year, name
ORDER BY year
```

Probleme

- Count über Union über verteilte Datenbanken?
 - Integrationsproblem
- Berechnung riesiger Zwischenergebnisse bei jeder Anfrage?
 - Datenmengenproblem

Lösung des Integrationsproblems?

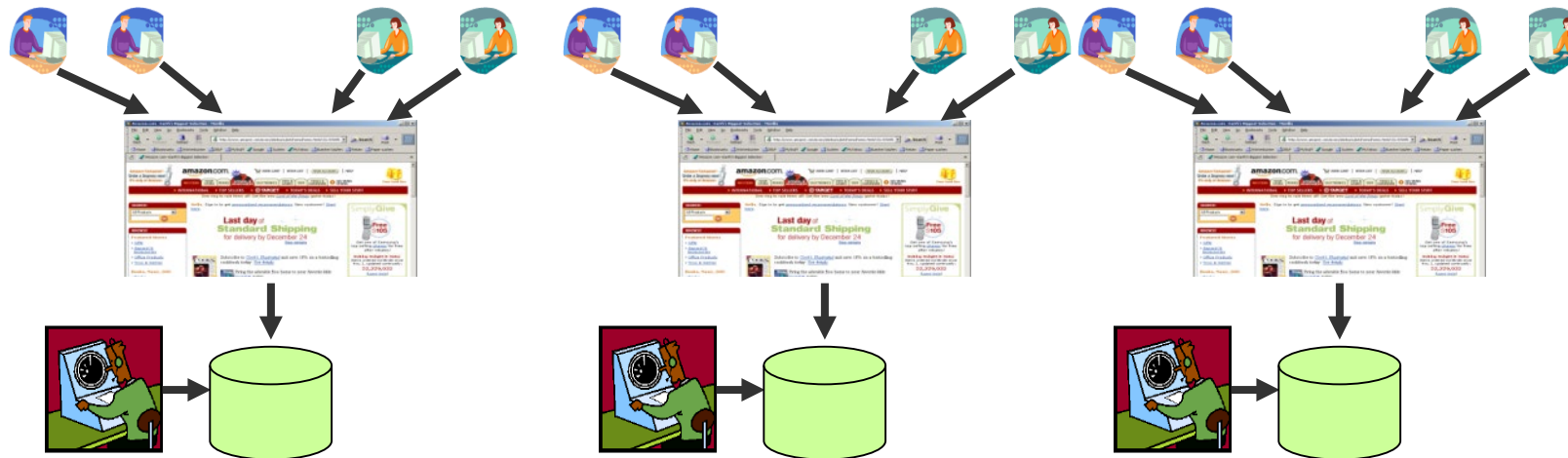
Zentrale Datenbank



- Aber Probleme:
 - Zweigstellen schreiben übers Netz
 - Schlechter Durchsatz
 - Lange Antwortzeiten im operativen Betrieb

Lösung Datenmengenproblem?

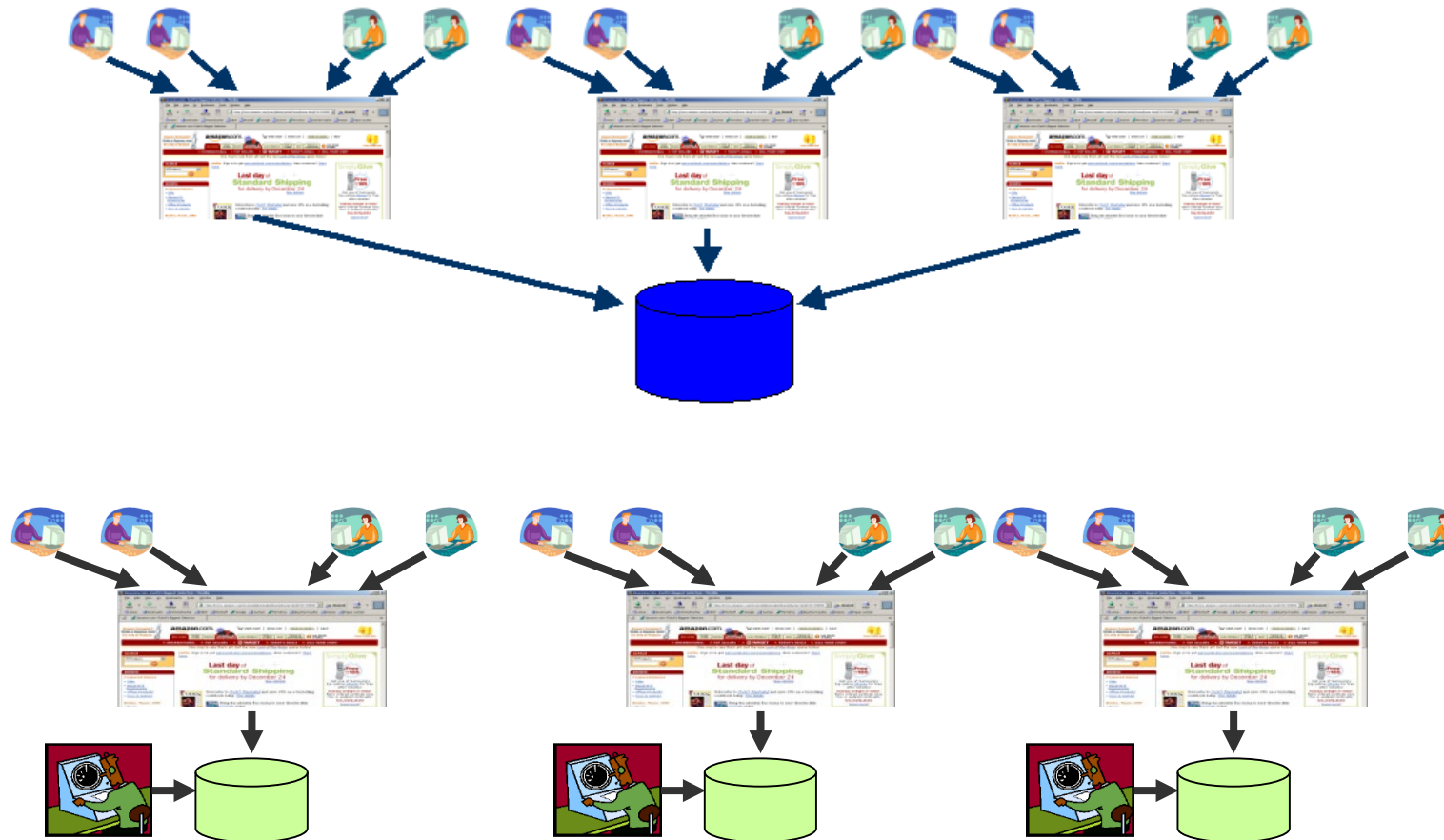
- Denormalisierte Schema zur Vermeidung großer Zwischenergebnisse



- Aber Probleme:

- Jeder lesende / schreibende Zugriff erfolgt auf eine Tabelle mit 72 Mill. Records
- Lange Antwortzeiten im operativen Betrieb

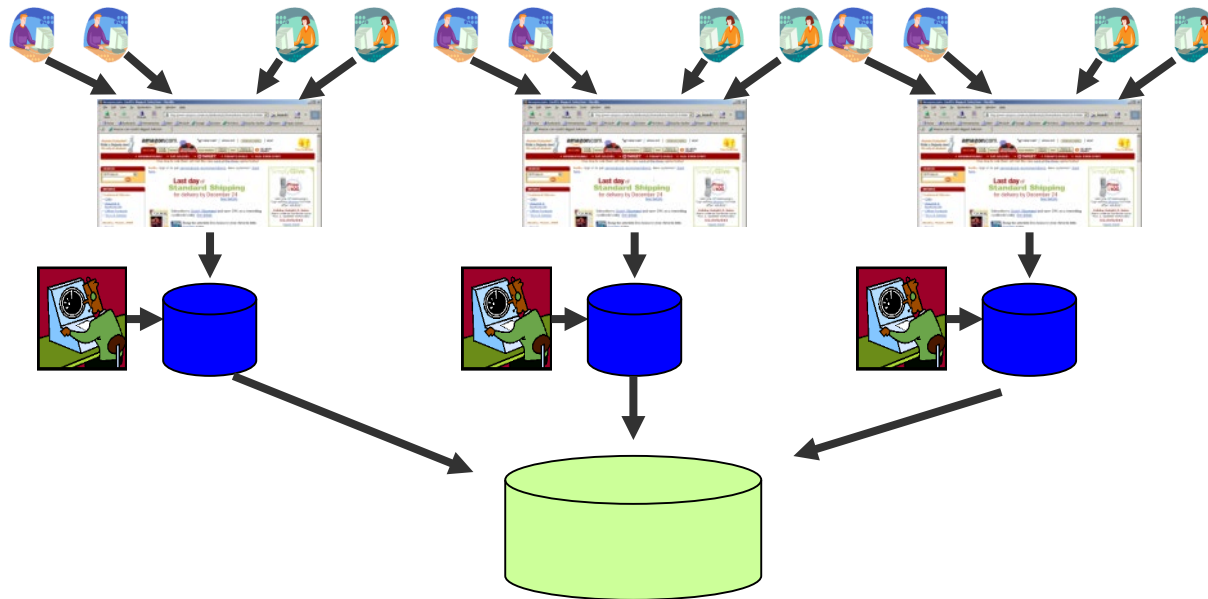
Zielkonflikt



Felix Naumann
Information Integration
Winter 2019/20

Tatsächliche Lösung

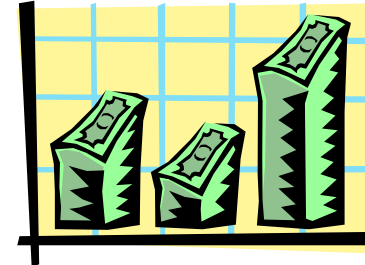
■ Aufbau eines Data Warehouse



- Redundante, transformierte Datenhaltung
- Asynchrone Aktualisierung

Weitere Anwendungsgebiete: Data Warehouses

- „Customer Relationship Management“ (CRM)
 - Identifikation von Premiumkunden
 - Personalisierung / Automatische Kundenberatung
 - Gezielte Massen-Mailings (Direktvertrieb)
- Controlling / Rechnungswesen
 - Kostenstellen
 - Organisationseinheiten
 - Personalmanagement
- Logistik
 - Flottenmanagement, Tracking
- Gesundheitswesen
 - Studienüberwachung, Patiententracking



Überblick

1. Szenarien der Informationsintegration
 - Data Warehouse
 - **Föderierte Datenbanken**
2. Einführung
3. Materialisiert
 - Data Warehouse
4. Virtuell
 - Mediator-Wrapper System
5. Vergleich
 - Flexibilität
 - Antwortzeiten
 - Aktualität
 - etc.



Felix Naumann
Information Integration
Winter 2019/20

Föderierte Datenbanken

- Mehrere autonome Informationsquellen
- Mit unterschiedlichsten Inhalten
 - Gene, Proteine, BLAST, etc.
- Und unterschiedlichsten Schnittstellen
 - HTML-Form, flat file, SQL, etc.
- Wissenschaftler (Biologe) benötigt z.B. möglichst viele Informationen über ein bestimmtes Protein
 - Funktion, Veröffentlichungen, verwandte Proteine usw.
- Sehr komplexe Anfragen
- Üblicher Ansatz: Browsing, Note-Taking, Copy & Paste
- Föderierte Datenbanken (wie DiscoveryLink) helfen.

Frage eines Biologen

- Finde alle menschlichen ESTs (DNA-Sequenzen), die nach BLAST zu mindestens 60% über mindestens 50 Aminosäuren identisch sind mit mouse-channel Genen im Gewebe des zentralen Nervensystems.



Quelle für das komplette Beispiel: *A Practitioner's Guide to Data Management and Data Integration in Bioinformatics*, Barbara A. Eckman in *Bioinformatics* by Zoe Lacroix and Terence Critchlow, 2003, Morgan Kaufmann.

Felix Naumann
Information Integration
Winter 2019/20

Verschiedene Informationsquellen

- Beteiligte Informationsquellen
 - Mouse Genome Database (MGD) @ Jackson Labs
 - SwissProt @ EBI
 - BLAST tool @ NCBI
 - GenBank nucleotide sequence database @ NCBI



- Alle Quellen sind/waren frei verfügbar

Herkömmlicher Ansatz: Browsing

1. Suche „channel“ Sequenzen im Gewebe des zentralen Nervensystems durch MGD HTML Formular



MG 2.7 - Gene Expression Data Query Form - Netscape

File Edit View Go Communicator Help

Gene Expression Data Query Form

Retrieve Reset Form

Sort by: Gene symbol Age Anatomical structure Assay Type Author

Max number of items returned: 10 100 500 No limit

Return: Assays Assay Results

Gene Symbol/Name:
 NOT contains Search

Gene Classifications: (You can browse the Gene Ontology (GO) Classifications)

 Molecular Function Biological Process Cellular Component

Chromosomal Location:
 Chromosome:
 Restrict search to a chromosomal region? (specify one of the following)
 Between and (Enter cM positions or locus symbols) endpoints.
 Within cM of locus locus.

Expression:
 detected not detected either

Developmental Stage(s): (You can browse Stage descriptions)

 TS 1 (0.0-2.5 dpc)
 TS 2 (1.0-2.5 dpc)
 TS 3 (1.0-3.5 dpc)
 TS 4 (2.0-4.0 dpc)

Anatomical Structure(s): (You can browse the Anatomical Dictionary)

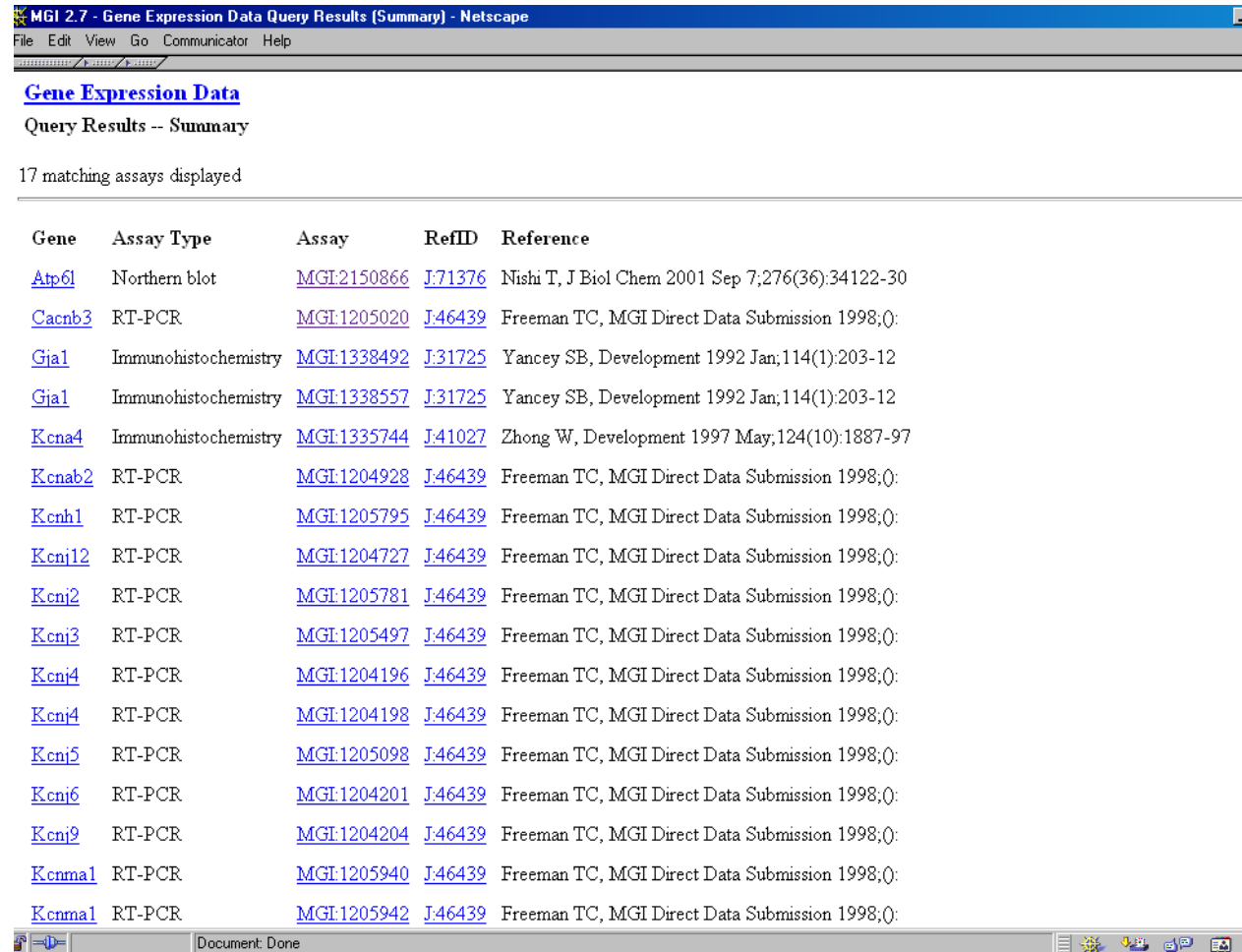
 Include: substructures superstructures

Document: Done

Felix Naumann
 Information Integration
 Winter 2019/20

Herkömmlicher Ansatz: Browsing

- MGD Resultat
 - 14 Gene aus 17 Experimenten

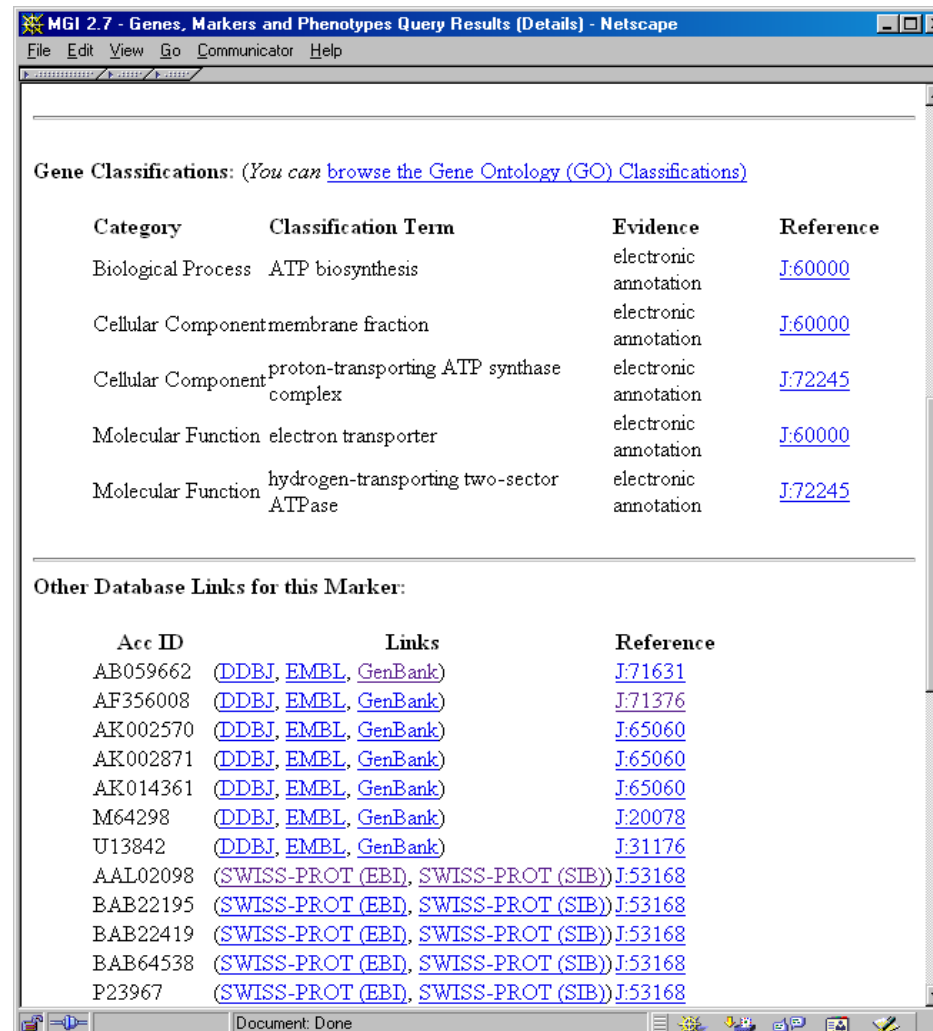


Gene	Assay Type	Assay	RefID	Reference
Atp6l	Northern blot	MGI2150866	J:71376	Nishi T, J Biol Chem 2001 Sep 7;276(36):34122-30
Cacnb3	RT-PCR	MGI1205020	J:46439	Freeman TC, MGI Direct Data Submission 1998;0:
Gja1	Immunohistochemistry	MGI1338492	J:31725	Yancey SB, Development 1992 Jan;114(1):203-12
Gja1	Immunohistochemistry	MGI1338557	J:31725	Yancey SB, Development 1992 Jan;114(1):203-12
Kcna4	Immunohistochemistry	MGI1335744	J:41027	Zhong W, Development 1997 May;124(10):1887-97
Kcnab2	RT-PCR	MGI1204928	J:46439	Freeman TC, MGI Direct Data Submission 1998;0:
Kcnh1	RT-PCR	MGI1205795	J:46439	Freeman TC, MGI Direct Data Submission 1998;0:
Kcni12	RT-PCR	MGI1204727	J:46439	Freeman TC, MGI Direct Data Submission 1998;0:
Kcni2	RT-PCR	MGI1205781	J:46439	Freeman TC, MGI Direct Data Submission 1998;0:
Kcni3	RT-PCR	MGI1205497	J:46439	Freeman TC, MGI Direct Data Submission 1998;0:
Kcni4	RT-PCR	MGI1204196	J:46439	Freeman TC, MGI Direct Data Submission 1998;0:
Kcni4	RT-PCR	MGI1204198	J:46439	Freeman TC, MGI Direct Data Submission 1998;0:
Kcni5	RT-PCR	MGI1205098	J:46439	Freeman TC, MGI Direct Data Submission 1998;0:
Kcni6	RT-PCR	MGI1204201	J:46439	Freeman TC, MGI Direct Data Submission 1998;0:
Kcni9	RT-PCR	MGI1204204	J:46439	Freeman TC, MGI Direct Data Submission 1998;0:
Kcnma1	RT-PCR	MGI1205940	J:46439	Freeman TC, MGI Direct Data Submission 1998;0:
Kcnma1	RT-PCR	MGI1205942	J:46439	Freeman TC, MGI Direct Data Submission 1998;0:

Felix Naumann
Information Integration
Winter 2019/20

Herkömmlicher Ansatz: Browsing

- Details zu jedem der 14 Gene ansehen
- Durchschnittlich fünf SwissProt Links pro Gen



Gene Classifications: (You can [browse the Gene Ontology \(GO\) Classifications](#))

Category	Classification Term	Evidence	Reference
Biological Process	ATP biosynthesis	electronic annotation	J:60000
Cellular Component	membrane fraction	electronic annotation	J:60000
Cellular Component	proton-transporting ATP synthase complex	electronic annotation	J:72245
Molecular Function	electron transporter	electronic annotation	J:60000
Molecular Function	hydrogen-transporting two-sector ATPase	electronic annotation	J:72245

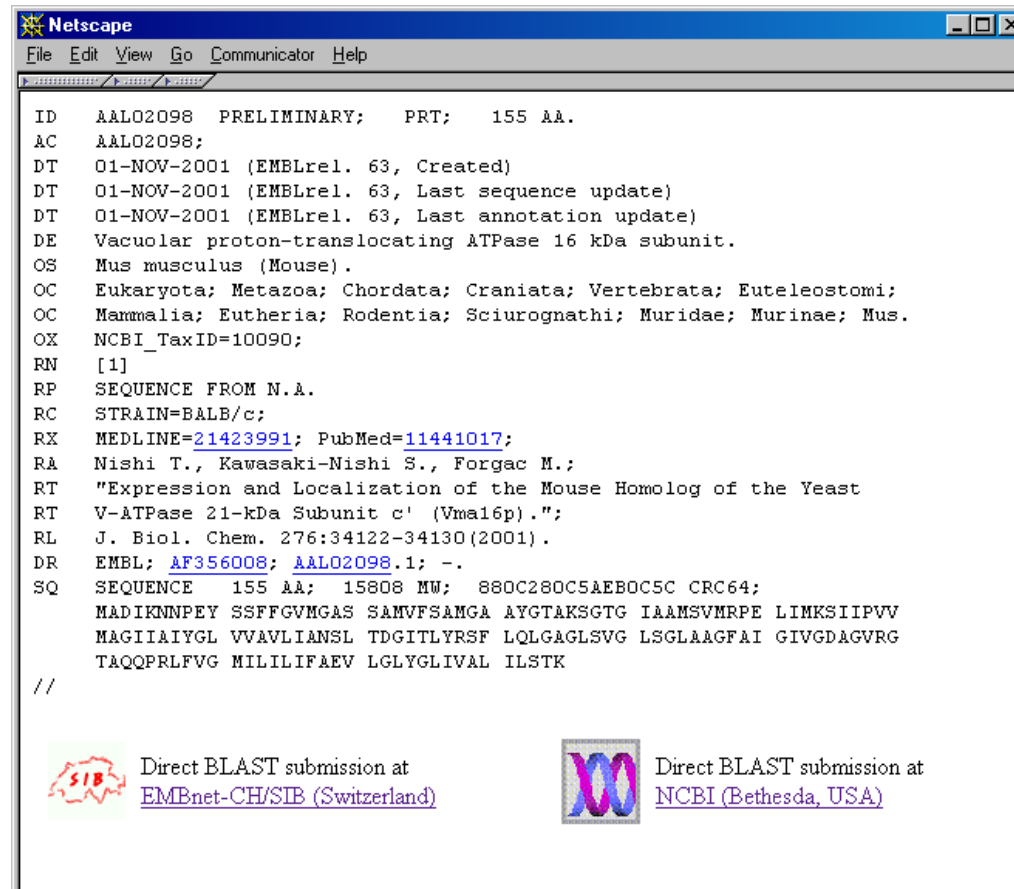
Other Database Links for this Marker:

Acc ID	Links	Reference
AB059662	(DDBJ, EMBL, GenBank)	J:71631
AF356008	(DDBJ, EMBL, GenBank)	J:71376
AK002570	(DDBJ, EMBL, GenBank)	J:65060
AK002871	(DDBJ, EMBL, GenBank)	J:65060
AK014361	(DDBJ, EMBL, GenBank)	J:65060
M64298	(DDBJ, EMBL, GenBank)	J:20078
U13842	(DDBJ, EMBL, GenBank)	J:31176
AAL02098	(SWISS-PROT (EBI), SWISS-PROT (SIB))	J:53168
BAB22195	(SWISS-PROT (EBI), SWISS-PROT (SIB))	J:53168
BAB22419	(SWISS-PROT (EBI), SWISS-PROT (SIB))	J:53168
BAB64538	(SWISS-PROT (EBI), SWISS-PROT (SIB))	J:53168
P23967	(SWISS-PROT (EBI), SWISS-PROT (SIB))	J:53168

elix Naumann
nformation Integration
Vinter 2019/20

Herkömmlicher Ansatz: Browsing

- Betrachten jedes SwissProt Eintrages
- Durch Klick BLAST Algorithmus anwerfen



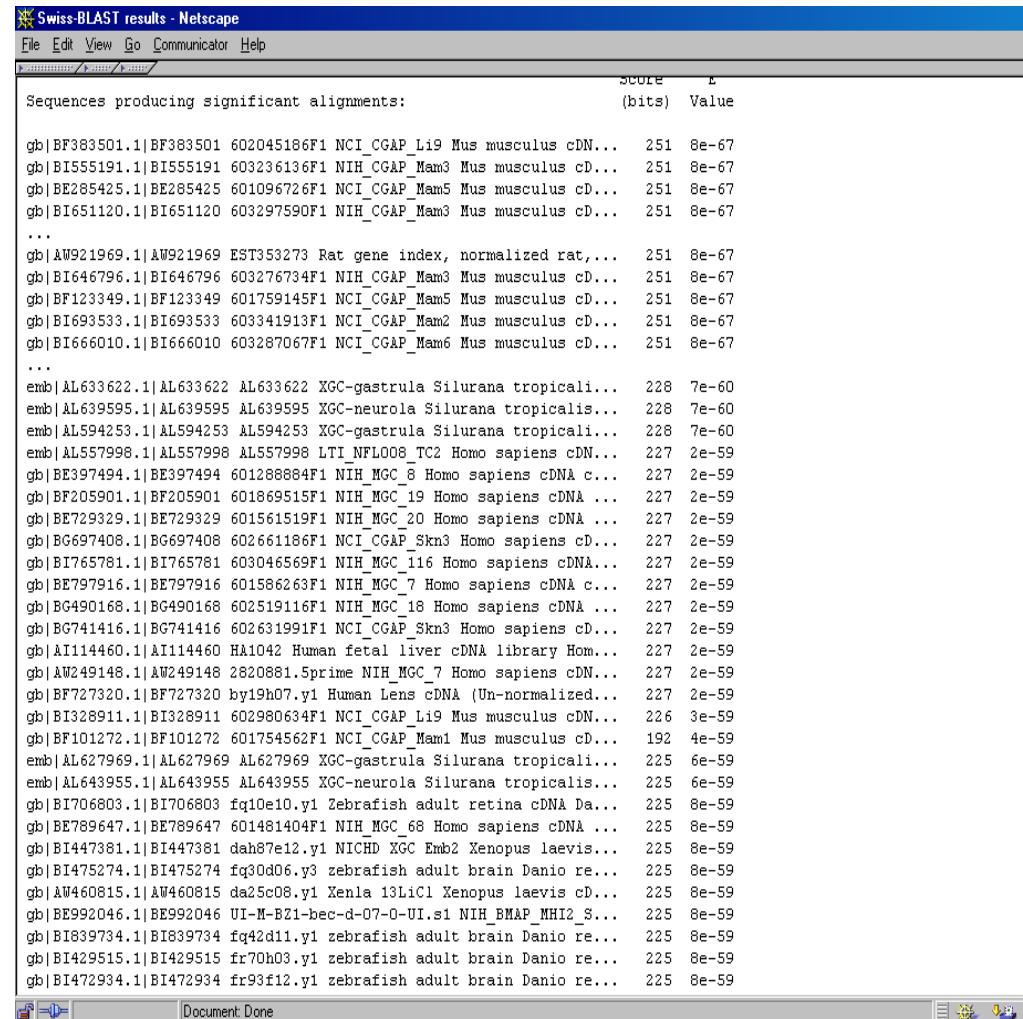
```

Netscape
File Edit View Go Communicator Help
ID AAL02098 PRELIMINARY; PRT; 155 AA.
AC AAL02098;
DT 01-NOV-2001 (EMBLrel. 63, Created)
DT 01-NOV-2001 (EMBLrel. 63, Last sequence update)
DT 01-NOV-2001 (EMBLrel. 63, Last annotation update)
DE Vacuolar proton-translocating ATPase 16 kDa subunit.
OS Mus musculus (Mouse).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX NCBI_TaxID=10090;
RN [1]
RP SEQUENCE FROM N.A.
RC STRAIN=BALB/c;
RX MEDLINE=21423991; PubMed=11441017;
RA Nishi T., Kawasaki-Nishi S., Forgac M.;
RT "Expression and Localization of the Mouse Homolog of the Yeast
RT V-ATPase 21-kDa Subunit c' (Vma16p).";
RL J. Biol. Chem. 276:34122-34130(2001).
DR EMBL; AF356008; AAL02098.1; -.
SQ SEQUENCE 155 AA; 15808 MW; 880C280C5AEB0C5C CRC64;
MADIKNNPEY SSFFGVMGAS SAMVFSAMGA AYGTAKSGTG IAAMSVMRPE LIMKSIIIPVV
MAGIIAIYGL VVAVLIANSL TDGITLYRSF LQLGAGLSVG LSGLAAGFAI GIVGDAGVRG
TAQQPRLFVG MILILIFAEV LGLYGLIVAL ILSTK
//
SIB Direct BLAST submission at
EMBLnet-CH/SIB (Switzerland)
NCBI Direct BLAST submission at
NCBI (Bethesda, USA)

```

Herkömmlicher Ansatz: Browsing

- Betrachten jedes BLAST Resultats um
 - nicht-menschliche Treffer zu eliminieren,
 - andere Bedingungen zu prüfen (>60% Identität, etc.).



Sequences producing significant alignments:	SCORE (bits)	E Value
gb BF383501.1 BF383501 602045186F1 NCI_CGAP_Li9 Mus musculus cDN...	251	8e-67
gb BI555191.1 BI555191 603236136F1 NIH_CGAP_Mam3 Mus musculus cd...	251	8e-67
gb BE285425.1 BE285425 601096726F1 NCI_CGAP_Mam5 Mus musculus cd...	251	8e-67
gb BI651120.1 BI651120 603297590F1 NIH_CGAP_Mam3 Mus musculus cd...	251	8e-67
...		
gb AW921969.1 AW921969 EST353273 Rat gene index, normalized rat,...	251	8e-67
gb BI646796.1 BI646796 603276734F1 NIH_CGAP_Mam3 Mus musculus cd...	251	8e-67
gb BF123349.1 BF123349 601759145F1 NCI_CGAP_Mam5 Mus musculus cd...	251	8e-67
gb BI693533.1 BI693533 603341913F1 NCI_CGAP_Mam2 Mus musculus cd...	251	8e-67
gb BI666010.1 BI666010 603287067F1 NCI_CGAP_Mam6 Mus musculus cd...	251	8e-67
...		
emb AL633622.1 AL633622 AL633622 XGC-gastrula Silurana tropicali...	228	7e-60
emb AL639595.1 AL639595 AL639595 XGC-neurola Silurana tropicalis...	228	7e-60
emb AL594253.1 AL594253 AL594253 XGC-gastrula Silurana tropicali...	228	7e-60
emb AL557998.1 AL557998 AL557998 LTI_NFL008_TC2 Homo sapiens cDN...	227	2e-59
gb BE397494.1 BE397494 601288884F1 NIH_MGC_8 Homo sapiens cDNA c...	227	2e-59
gb BF205901.1 BF205901 601869515F1 NIH_MGC_19 Homo sapiens cDNA ...	227	2e-59
gb BE729329.1 BE729329 601561519F1 NIH_MGC_20 Homo sapiens cDNA ...	227	2e-59
gb BG697408.1 BG697408 602661186F1 NCI_CGAP_Skn3 Homo sapiens cd...	227	2e-59
gb BI765781.1 BI765781 603046569F1 NIH_MGC_116 Homo sapiens cDNA...	227	2e-59
gb BE797916.1 BE797916 601586263F1 NIH_MGC_7 Homo sapiens cDNA c...	227	2e-59
gb BG490168.1 BG490168 602519116F1 NIH_MGC_18 Homo sapiens cDNA ...	227	2e-59
gb BG741416.1 BG741416 602631991F1 NCI_CGAP_Skn3 Homo sapiens cd...	227	2e-59
gb AI114460.1 AI114460 HA1042 Human fetal liver cDNA library Hom...	227	2e-59
gb AW249148.1 AW249148 2820881.Sprime NIH_MGC_7 Homo sapiens cDN...	227	2e-59
gb BF727320.1 BF727320 by19h07.y1 Human Lens cDNA (Un-normalized...	227	2e-59
gb BI328911.1 BI328911 602980634F1 NCI_CGAP_Li9 Mus musculus cDN...	226	3e-59
gb BF101272.1 BF101272 601754562F1 NCI_CGAP_Mam1 Mus musculus cd...	192	4e-59
emb AL627969.1 AL627969 AL627969 XGC-gastrula Silurana tropicali...	225	6e-59
emb AL643955.1 AL643955 AL643955 XGC-neurola Silurana tropicalis...	225	6e-59
gb BI706803.1 BI706803 fq10e10.y1 Zebrafish adult retina cDNA Da...	225	8e-59
gb BE789647.1 BE789647 601481404F1 NIH_MGC_68 Homo sapiens cDNA ...	225	8e-59
gb BI447381.1 BI447381 dah87e12.y1 NICHD XGC Emb2 Xenopus laevis...	225	8e-59
gb BI475274.1 BI475274 fq30d06.y3 zebrafish adult brain Danio re...	225	8e-59
gb AW460815.1 AW460815 da25c08.y1 Xenla 13LiCl Xenopus laevis cd...	225	8e-59
gb BE992046.1 BE992046 UI-M-BZ1-bec-d-07-0-UI.s1 NIH_BMAP_MHI2_S...	225	8e-59
gb BI839734.1 BI839734 fq42d11.y1 zebrafish adult brain Danio re...	225	8e-59
gb BI429515.1 BI429515 fr70h03.y1 zebrafish adult brain Danio re...	225	8e-59
gb BI472934.1 BI472934 fr93f12.y1 zebrafish adult brain Danio re...	225	8e-59

Felix Naumann
Information Integration
Winter 2019/20

Herkömmlicher Ansatz: Browsing

- Für jeden verbleibenden Eintrag
 - Komplette EST Sequenz bei GenBank holen



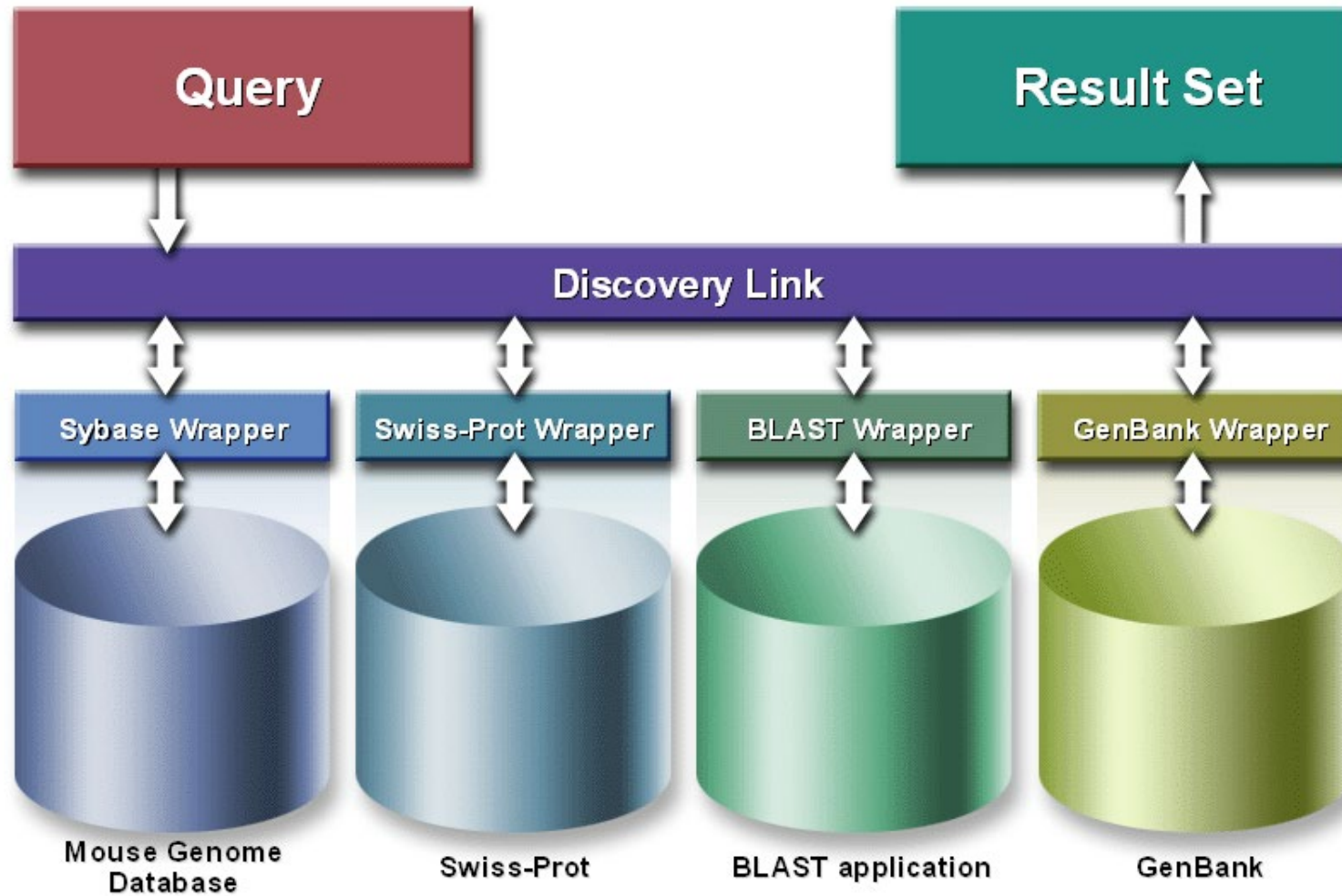
Alles sehr mühselig!

Felix Naumann
Information Integration
Winter 2019/20

Idee der Integration

- Bildung eines globalen Schemas (Schemaintegration)
 - Gespeichert als Datenbankschema (z.B. in DiscoveryLink)
- Generierung von *Wrappern* für jede Datenquelle
 - Softwarekomponente
 - Mapping von lokalen Schemata auf globales Schema
 - Kennt Anfragefähigkeiten der Quellen

DiscoveryLink Architektur



Felix Naumann
Information Integration
Winter 2019/20

Eigenschaften föderierter IS

- Daten bleiben vor Ort.
- Informationsquellen sind autonom (und wissen oft nicht von ihrer Integration).
- Anfragen werden deklarativ an das globale Schema gestellt.
- Anfrage wird so verteilt wie möglich ausgeführt.
 - Je nach Mächtigkeit der Quellen
 - Föderiertes System gleicht etwaige mangelnder Fähigkeiten aus.

Finde alle menschlichen ESTs (DNA-Sequenzen), die nach BLAST zu mindestens 60% über mindestens 50 Aminosäuren identisch sind mit mouse-channel Genen im Gewebe des zentralen Nervensystems.

„Einfache“ SQL-Anfrage um alle vorigen Schritte zu vereinen:

```
SELECT g.accnum, g.sequence
FROM   genbank g, blast b, swissprot s, mgd m
WHERE  m.exp = "CNS"
AND    m.defn LIKE "%channel%"
AND    m.spid = s.id AND s.seq = b.query
AND    b.hit = g.accnum
AND    b.percentid >= 60 AND b.alignlen >= 50
```

Föderierter DBMS Ansatz

- Effiziente Ausführung durch Optimierer
 - Herkömmliche Optimierung
 - Wrapper helfen mit
 - Kostenmodell und
 - domänenspezifischen Funktionen

- Sichere Ausführung
 - Wiederholbar
 - Transaktional

Weitere Anwendungsgebiete: Föderierte Datenbanken

- Meta-Suchmaschinen
- Unternehmensfusionen
 - Kundendatenbanken
 - Personaldatenbanken
- Grid
- Krankenhausinformationssysteme
 - Röntgenbilder
 - Krankheitsverlauf (Akte)
 - Verwaltung
 - Krankenkasse...
- Verteiltes Arbeiten („groupware“)
- Peer Data Management und P2P



Felix Naumann
Information Integration
Winter 2019/20

Überblick

1. Szenarien der Informationsintegration
 - Data Warehouse
 - Föderierte Datenbanken
2. **Einführung**
3. Materialisiert
 - Data Warehouse
4. Virtuell
 - Mediator-Wrapper System
5. Vergleich
 - Flexibilität
 - Antwortzeiten
 - Aktualität
 - etc.



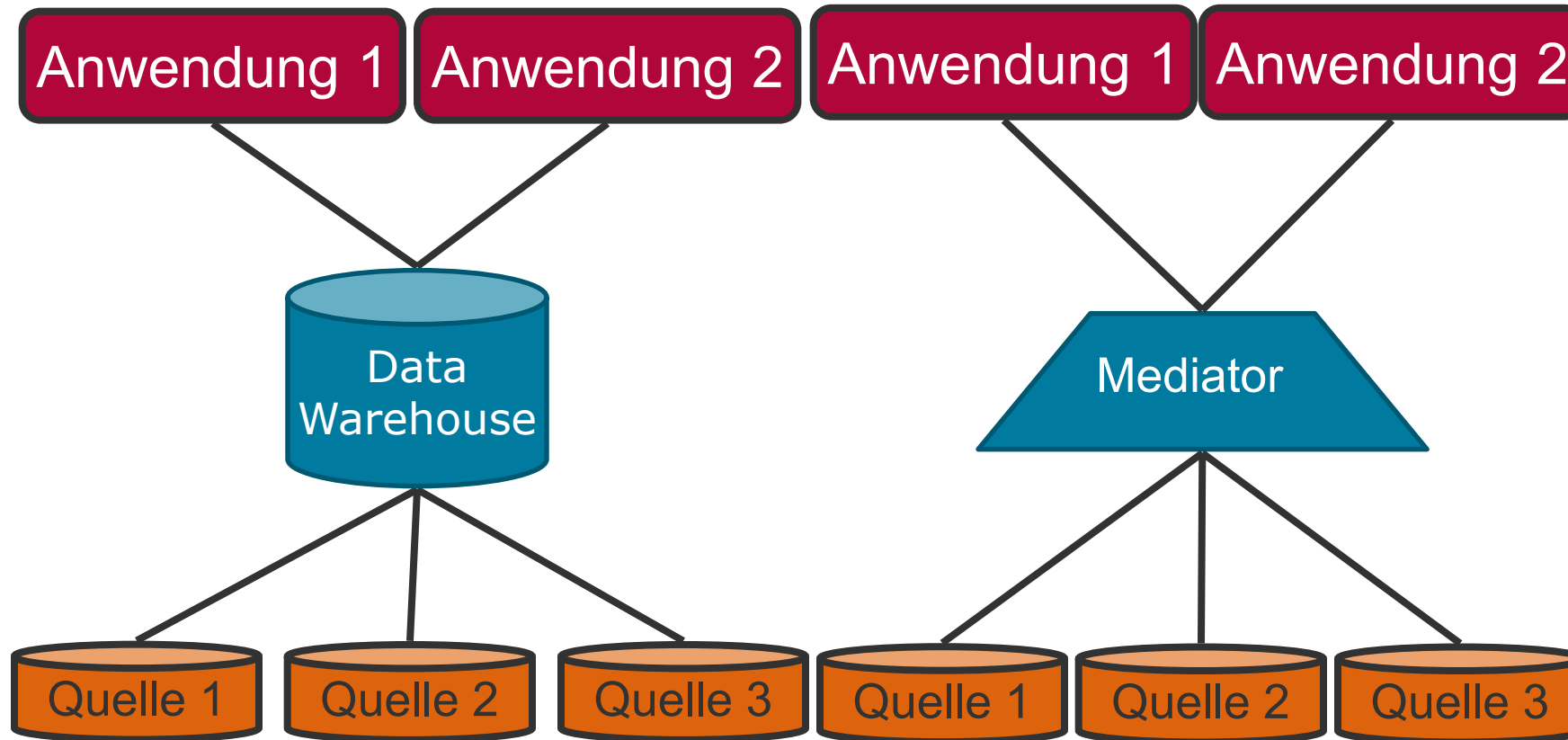
Felix Naumann
Information Integration
Winter 2019/20

Integration

- Materialisiert
 - A priori Integration
 - Zentrale Datenbasis
 - Zentrale Anfragebearbeitung
 - Typisches Beispiel: Data Warehouse

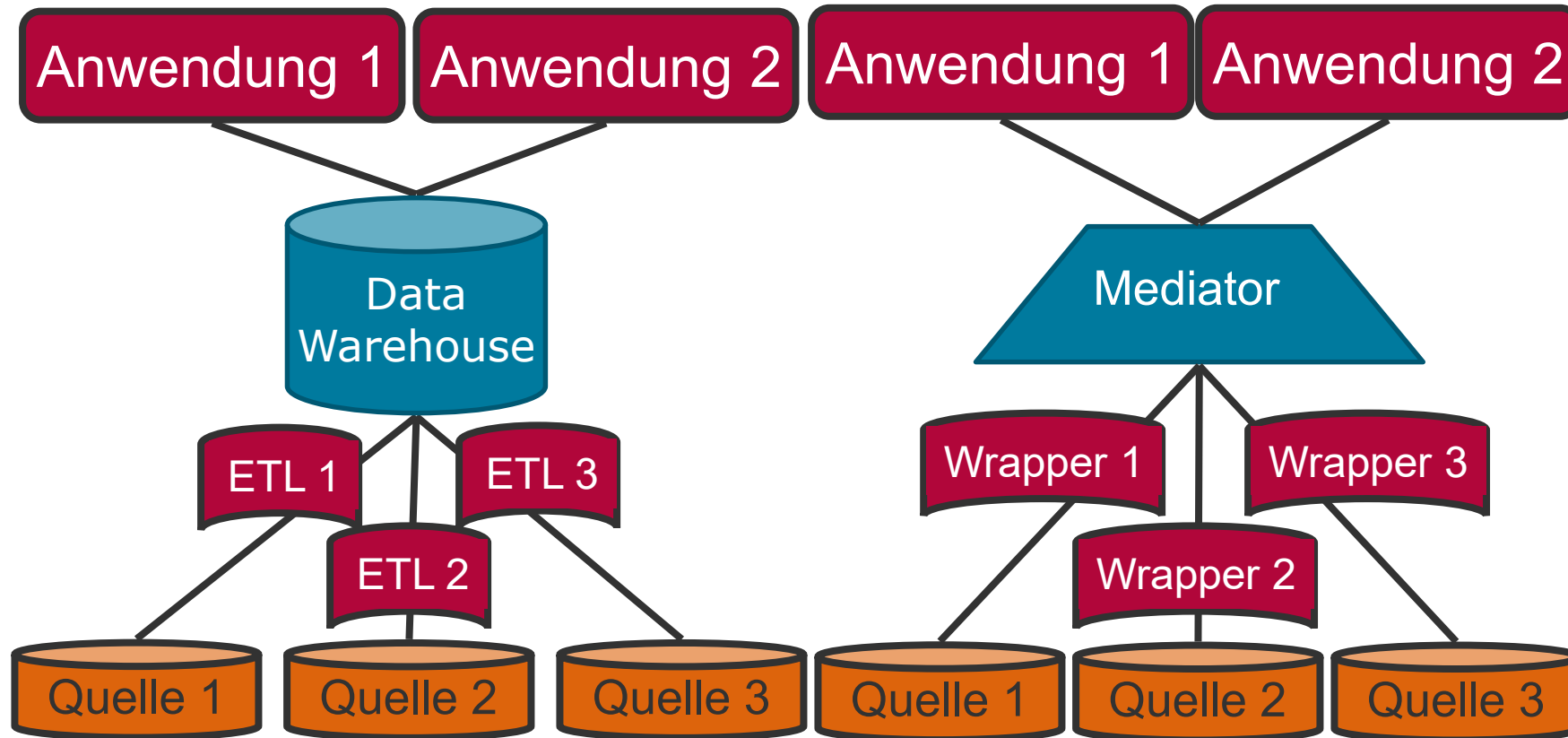
- Virtuell
 - On demand Integration
 - Dezentrale Daten
 - Dezentrale Anfragebearbeitung
 - Typisches Beispiel: Mediator-basiertes Informationssystem

Data Warehouse vs. Mediator-basiertes Informationssystem

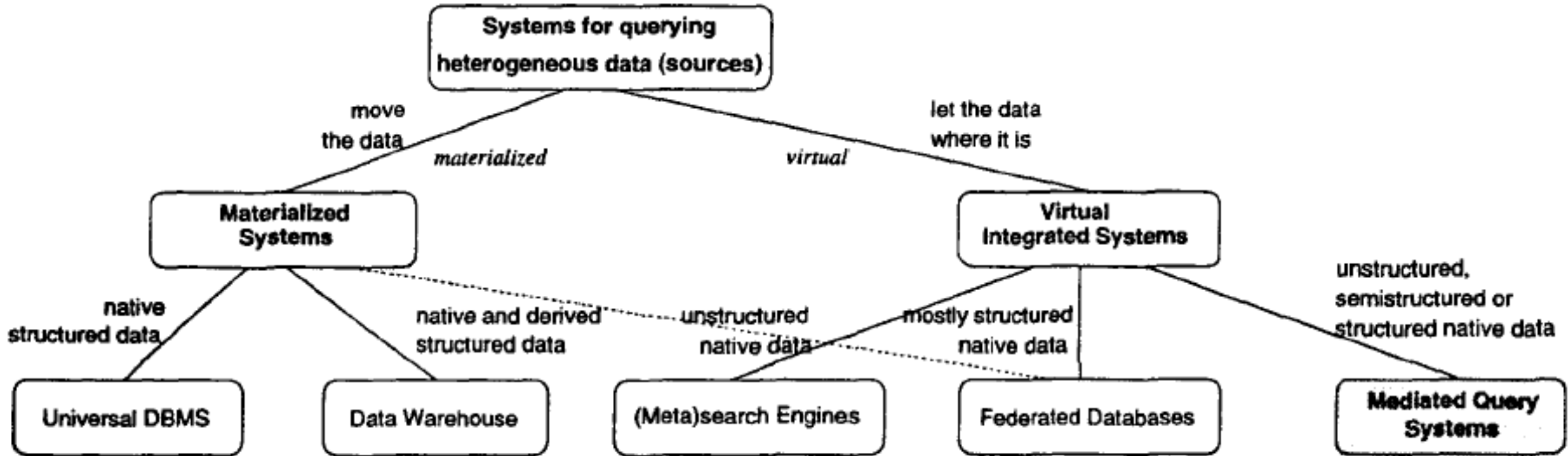


Felix Naumann
Information Integration
Winter 2019/20

Data Warehouse vs. Mediator



Taxonomie nach [DD99]



Data Warehouse vs. Mediator

- Jetzt jeweils kurzer Überblick
 - Datenfluss
 - Anfragebearbeitung
 - Entwurf und Entwicklung (Schema)

- Details in den folgenden Wochen

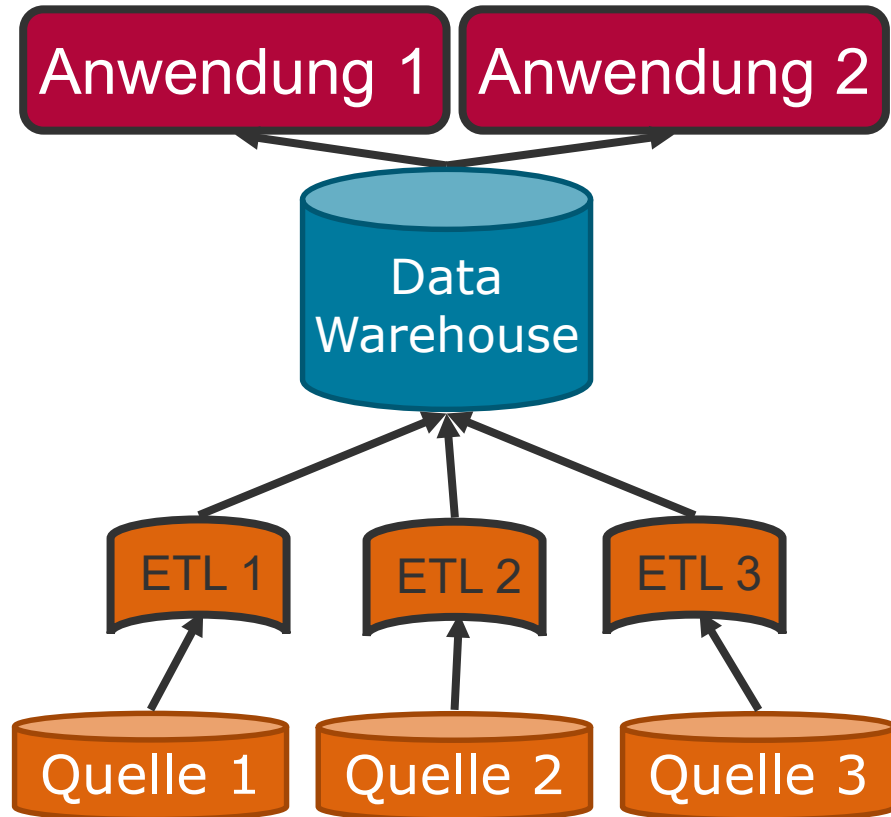
Überblick

1. Szenarien der Informationsintegration
 - Data Warehouse
 - Föderierte Datenbanken
2. Einführung
3. **Materialisiert**
 - **Data Warehouse**
4. Virtuell
 - Mediator-Wrapper System
5. Vergleich
 - Flexibilität
 - Antwortzeiten
 - Aktualität
 - etc.



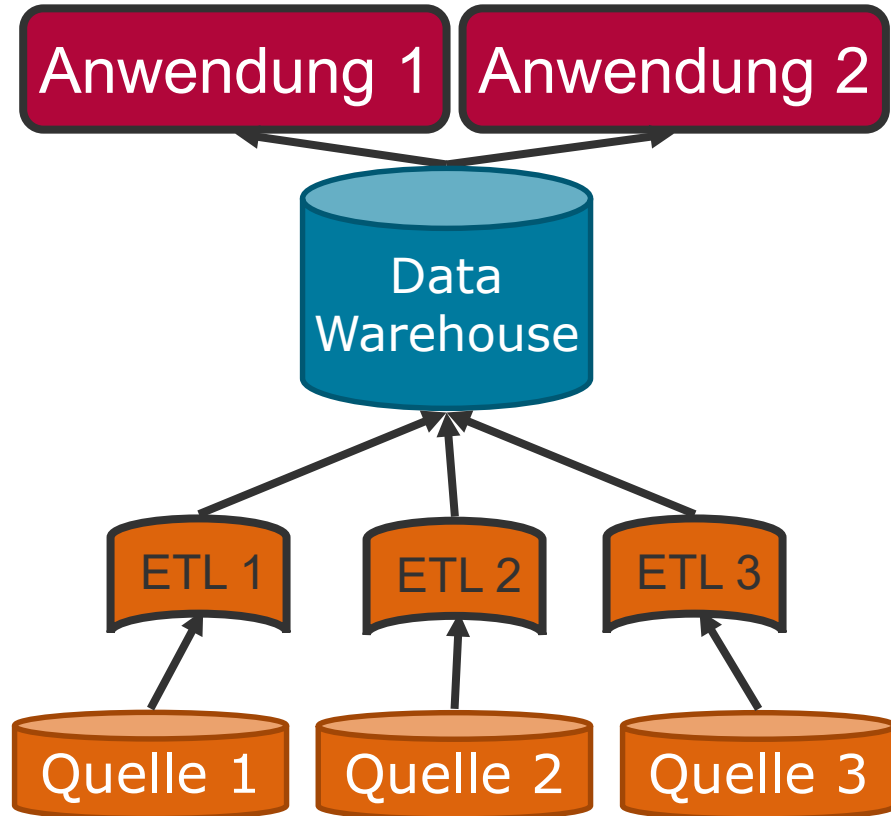
Felix Naumann
Information Integration
Winter 2019/20

Materialisierte Integration - Datenfluss



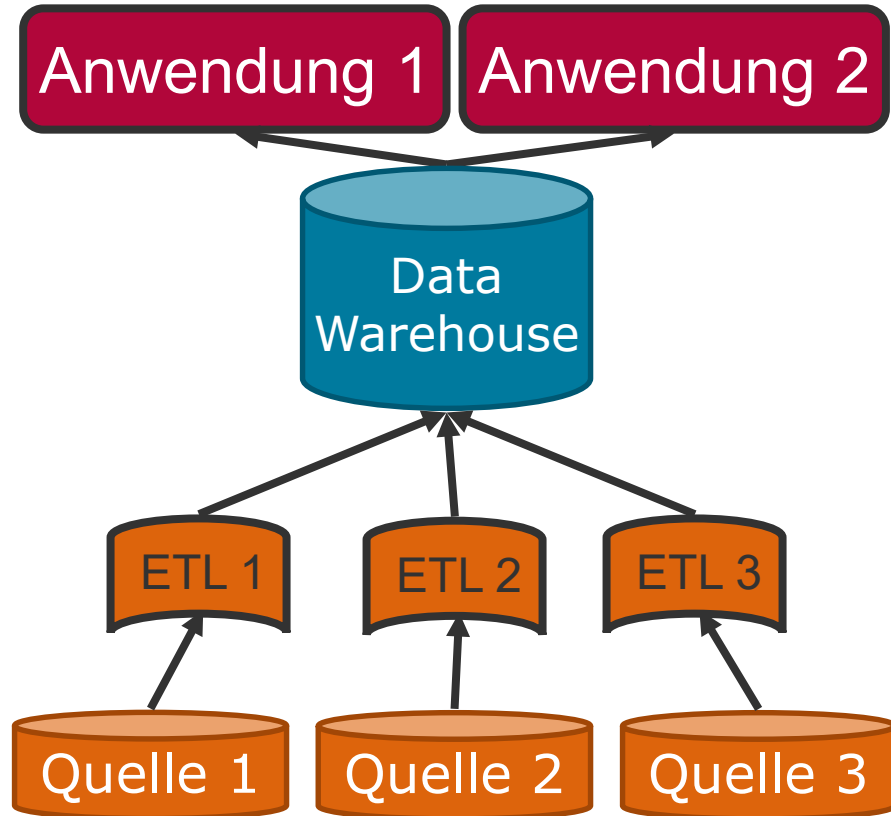
- Push
- Erstmalige „Bevölkerung“ (population) des DW
 - Data Cleansing
- Periodischer Datenimport
 - Stündlich / Täglich / Wöchentlich
 - Materialisierte Sichten / Sicht-Updates
- Redundante Datenhaltung
- Aggregation und Löschung alter Daten
 - Je älter, desto „aggregierter“

Materialisierte Integration - Anfragebearbeitung

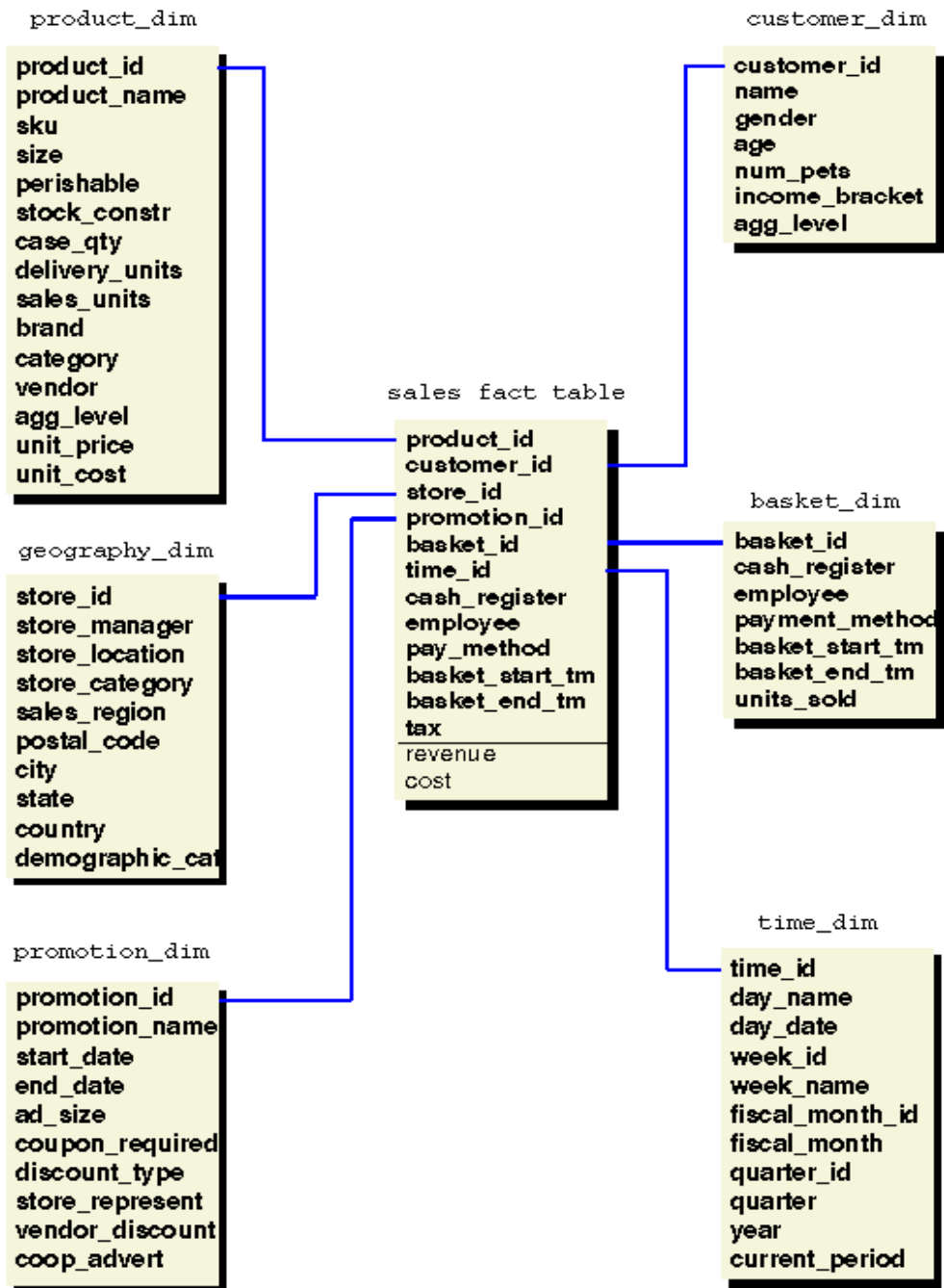


- Wie „normale“ DBMS
- Besonderheiten
 - Star Schema
 - Aggregation
 - Decision Support
- Siehe auch VL DWH

Materialisierte Integration - Schema



- Bottom-Up Entwurf
- Schemaintegration
- Star-Schema
 - *Fact-Table*
 - *Dimension Tables*



- Bottom-Up Entwurf
- Schemaintegration
- Star-Schema
 - *Fact-Table*
 - *Dimension Tables*

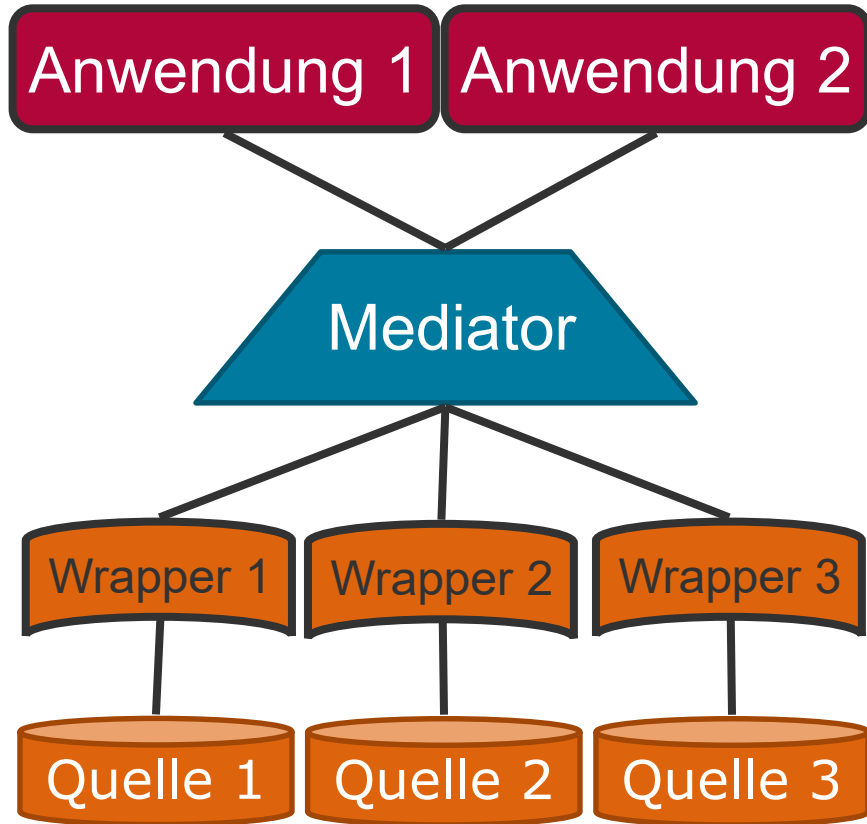
Überblick

1. Szenarien der Informationsintegration
 - Data Warehouse
 - Föderierte Datenbanken
2. Einführung
3. Materialisiert
 - Data Warehouse
4. **Virtuell**
 - **Mediator-Wrapper System**
5. Vergleich
 - Flexibilität
 - Antwortzeiten
 - Aktualität
 - etc.



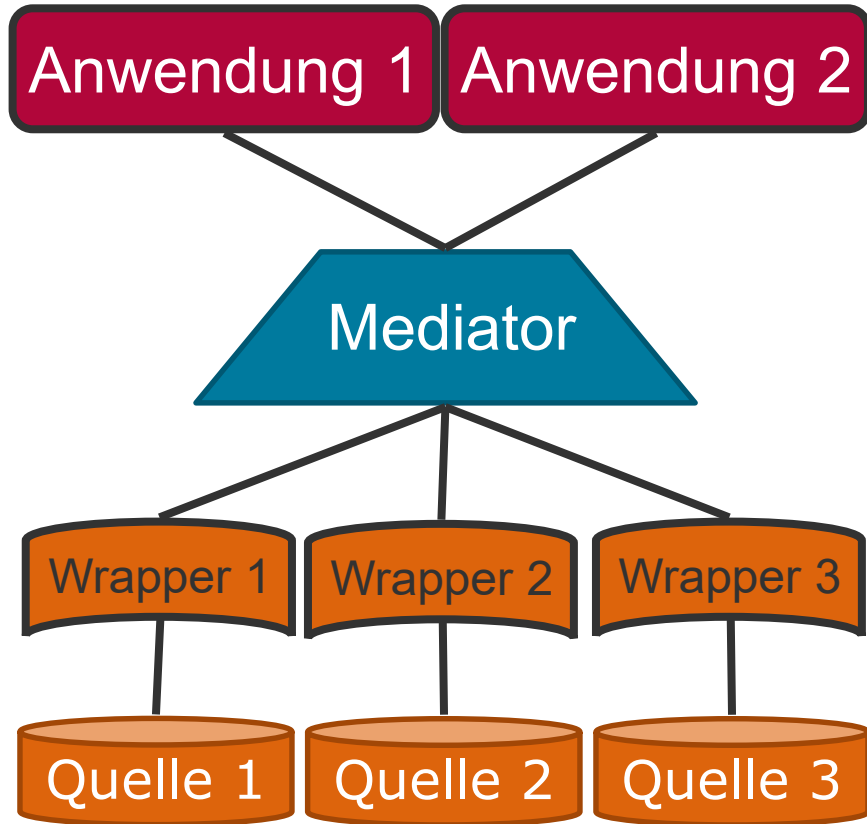
Felix Naumann
Information Integration
Winter 2019/20

Virtuelle Integration - Datenfluss



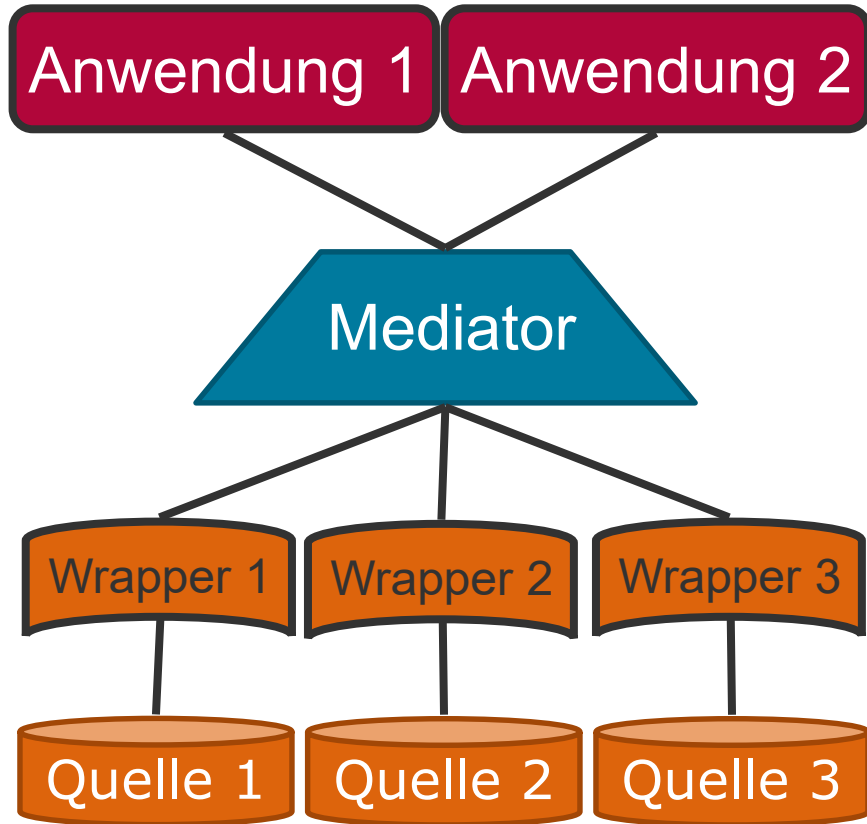
- Pull
- Daten sind in Quellen gespeichert.
- Nur die zur Anfragebeantwortung notwendigen Daten werden übertragen.
- Data Cleansing nur online möglich.

Virtuelle Integration - Anfragebearbeitung



- Optimierung schwierig
 - Fähigkeiten der Quellen
 - Geschwindigkeit der Quellen
- Viele mögliche Pläne
 - Redundante Quellen
 - Redundante Pläne
- Dynamisch, um ausfallende Quellen auszugleichen

Virtuelle Integration - Schema



- Top-Down Entwurf
- Leicht erweiterbar
 - Global: Neue Quellen suchen
 - Lokal: Nur ein *mapping* verändern.
- Schema Mapping statt Schema-integration

Überblick

1. Szenarien der Informationsintegration
 - Data Warehouse
 - Föderierte Datenbanken
2. Einführung
3. Materialisiert
 - Data Warehouse
4. Virtuell
 - Mediator-Wrapper System
- 5. Vergleich**
 - Flexibilität
 - Antwortzeiten
 - Aktualität
 - etc.



Felix Naumann
Information Integration
Winter 2019/20

Dimensionen des Vergleichs

- Aktualität
- Antwortzeit
- Flexibilität / Wartbarkeit
- Komplexität
- Autonomie
- Anfragebearbeitung / Mächtigkeit
- Read / Write
- Größe / Speicherbedarf
- Ressourcenbedarf
- Vollständigkeit
- Data Cleansing
- Informationsqualität

Aktualität (up-to-date-ness)

Materialisierte Integration

- Je nach Update-Frequenz
- In Unternehmen meist täglich (über Nacht)
- Beispiel SwissProt
 - Updates in SwissProt täglich
 - Aber: Release nur monatlich

Virtuelle Integration

- Sehr gut
- Abhängig von Aktualität der autonomen Quellen
- Manchmal: Caching

Antwortzeit (response time)

Materialisierte Integration

- Sehr gut
- Lokale Bearbeitung
- Wie DBMS
 - Optimierung
 - Materialisierte Sichten
 - Indices
 - ...
- Allerdings: Typische Anfragen sind komplex

Virtuelle Integration

- Nicht gut
- Daten sind entfernt
 - Übertragung durch das Netz
- Abhängig von Antwortzeit der Quellen
- Optimierung schwierig
- Komplexe Operatoren müssen naiv ausgeführt werden.
- Data Cleansing Operationen müssen nachgeholt werden.

Flexibilität / Wartbarkeit (flexibility / maintenance)

Materialisierte Integration

- Schwierig
- Entfernen / Ändern / Hinzufügen einer Quelle kann gesamte Integration verändern (bei GaV)
- Lokale Wartung eines großen und wachsenden Datenbestandes
 - Mit Indices etc.
- Tägliche Integration nötig

Virtuelle Integration

- Einfacher
- Entfernen / Ändern / Hinzufügen einer Quelle wirkt sich nur auf das mapping dieser Quelle aus (bei LaV)
- Quellen müssen Daten selbst warten.
 - Backups, DBMS Wartug etc.

Komplexität (complexity)

Materialisierte Integration

- Wie DBMS
- Komplexe Anfragen
- Anfrageplanung im GaV leicht
- Quellen sind oft untereinander ähnlich.
 - Oft sind es selbst DBMS

Virtuelle Integration

- Modellierung der Quellen wichtig
 - Fähigkeiten der Quellen
- Anfrageplanung in LaV schwierig
- Oft verschiedenste Quellen
 - Web Services
 - HTML Formulare
 - Flat Files
 - ...

Autonomie (autonomy)

Materialisierte Integration

- Quellen wenig autonom
 - Keine Kommunikations-autonomie
 - Geringe Ausführungs-autonomie
 - Geringe Design-autonomie
- Müssen bulk-read o.ä. zulassen
- Update notifications

Virtuelle Integration

- Quellen können autonom sein.
- Volle Design-Autonomie
- Fast volle Kommunikations-Autonomie
 - Gewisse Kommunikation ist nötig, sonst nicht Teilnehmer der Integration
- Fast volle Ausführungs-Autonomie
 - Nur: Anfragen müssen irgendwann beantwortet werden.

Anfragebearbeitung / Mächtigkeit (query planning / expressiveness)

Materialisierte Integration

- Anfragebearbeitung wie DBMS bzw. anderes globales System
- Anfragemächtigkeit wie globales System
 - z.B. volle SQL Mächtigkeit

Virtuelle Integration

- Anfragebearbeitung komplex
 - Verteilung
 - Autonomie
 - Heterogenität
- Mangelnde Fähigkeiten der Quellen können global eventuell ausgeglichen werden.
- Aber auch: Spezialfähigkeiten der Quellen können genutzt werden:
 - Image retrieval
 - Text Index

Materialisierte Integration

- Read immer möglich
- DW: Write oft nicht gewünscht, aber möglich
 - Kann zu Inkonsistenz mit Quellen führen

Virtuelle Integration

- Read meist möglich
- Verfügbarkeit!
- Write meist nicht möglich
 - Bei Redundanz: Wohin schreiben?
 - Transaktionen schwierig
 - Autonomie

Größe / Speicherbedarf (size / memory consumption)

Materialisierte Integration

- Hoch
 - Redundante Datenhaltung
 - DW: Historische Daten
- Wachstum
 - Stetig wachsend
 - Oder konstant durch zunehmende Aggregation im Laufe der Zeit
- Footprint: wie DBMS

Virtuelle Integration

- Gering
 - Metadaten
 - Cache
 - Zwischenergebnisse
- Footprint: wie DBMS

Ressourcenbedarf (resource consumption)

Materialisierte Integration

- Planbare Netzwerklast
- Daten werden eventl. unnötig übertragen
 - Abhängig von Anfrage
 - Aggregation
 - Pre-Aggregation

Virtuelle Integration

- Potentiell hohe Netzwerklast
- Daten werden mehrfach übertragen.
 - Cache kann helfen.
- Nur jeweils nötige Daten werden übertragen.

Je nach *Workload*.
Spannendes Optimierungsproblem!

Vollständigkeit (completeness)

Materialisierte Integration

- Gut
- Annahme: Materialisation ist vollständig

Virtuelle Integration

- Nur bei Verfügbarkeit aller nötigen Quellen
- Gegebenenfalls Anfrage unbeantwortbar oder nur unvollständig beantwortbar
 - Fuzzy Anfragesemantik:
 - Alle Tupel?
 - Alle Attribute?
- Definition der Vollständigkeit
 - Open World Assumption
 - Closed World Assumption

Datenreinigung (Data Cleansing)

Materialisierte Integration

- Viele Methoden
 - Aufwändig
- Offline (über Nacht)

Virtuelle Integration

- Online cleansing schwierig
 - Aufwand
 - Keine Interaktion mit Experten möglich

Informationsqualität (information quality)

Materialisierte Integration

- Hoch
- Kontrolliert
- Kann bei Bedarf verbessert werden.

Virtuelle Integration

- Abhängig von Quellen
- Oft zweifelhaft
- Autonomie

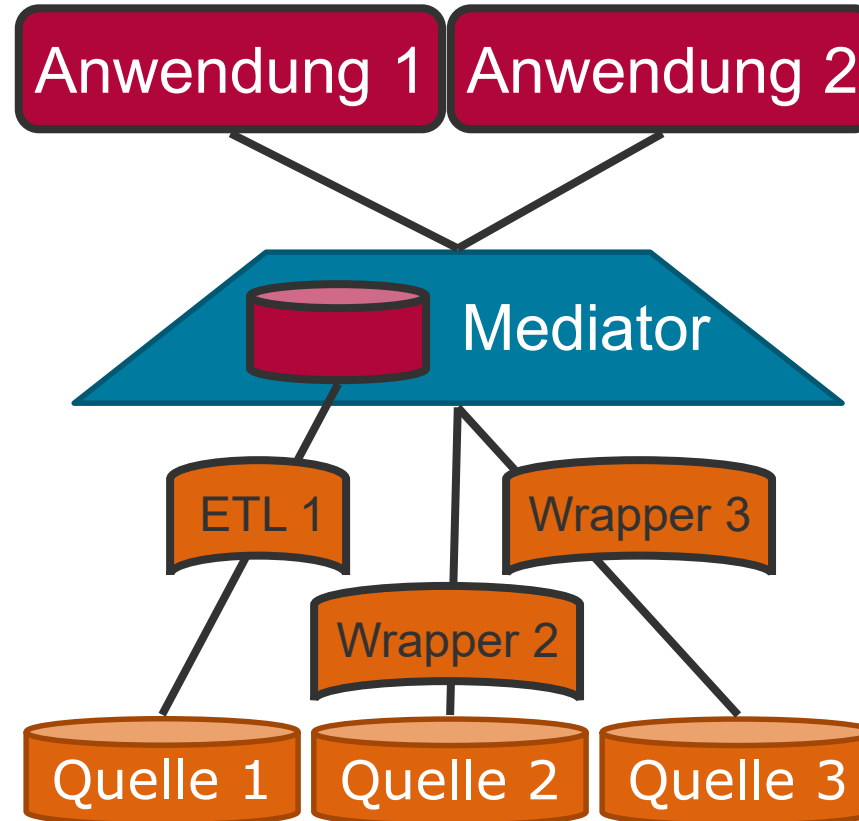
Zusammenfassung Vor- und Nachteile

	Materialisiert	Virtuell
Aktualität	- (Cache)	+
Antwortzeit	+	-
Flexibilität	- (GaV)	+ (LaV)
Komplexität Anfragebearbeitung	-	--
Quell-Autonomie	-	+
Anfragemächtigkeit	+	-
Read/Write	+/+	+/-
Größe	-	+
Ressourcenbedarf	? (workload)	? (workload)
Vollständigkeit	+	? (OWA, CWA)
Datenreinigung	+	-
Informationsqualität	+	-

Felix Naumann
 Information Integration
 Winter 2019/20

Hybrider Ansatz

- Teile der Daten werden materialisiert
 - Oft benötigte Daten (Cache)
 - Als bulk verfügbare Daten
 - Dump Files
 - SQL Zugang
 - ...
- Teile der Daten bleiben bei den Quellen
 - Oft aktualisierte Daten
 - Daten mit beschränktem Zugang
 - mind. eine gebundene Variable
 - Beschränkte Lizenzen
- Optimierung bevorzugt lokale Daten
 - Prüfung, ob Aktualisierung vorliegt



Überblick

1. Szenarien der Informationsintegration
 - Data Warehouse
 - Föderierte Datenbanken
2. Einführung
3. Materialisiert
 - Data Warehouse
4. Virtuell
 - Mediator-Wrapper System
5. Vergleich
 - Flexibilität
 - Antwortzeiten
 - Aktualität
 - etc.



Felix Naumann
Information Integration
Winter 2019/20

- [BKLW99] Busse, Kutsche, Leser, Weber, Federated Information Systems: Concepts, Terminology and Architectures. Forschungsbericht 99-9 des FB Informatik der TU Berlin, 1999.
Online: http://www.informatik.hu-berlin.de/~leser/publications/tr_terminology.ps
- [DD99] [Ruxandra Domenig](#), Klaus R. Dittrich: An Overview and Classification of Mediated Query Systems. [SIGMOD Record 28](#)(3): 63-72 (1999)