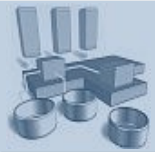
Servicebasierte Datenintegration

Präsentation zur Seminararbeit

Christoph Aßmann



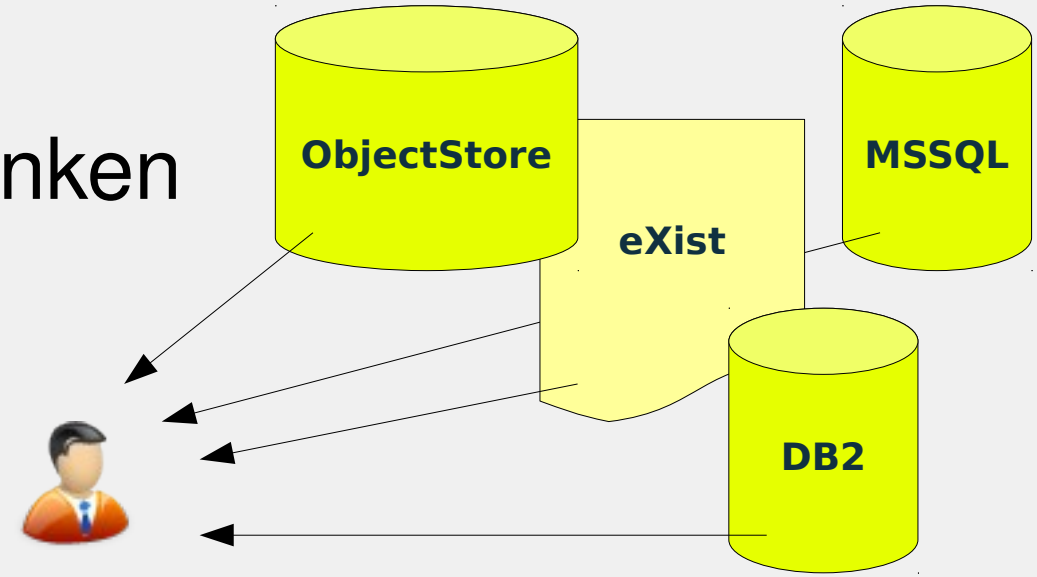
Inhalt

- Begriffe
- Motivation
- Abgrenzung Grid – Cloud
- OGSA: Architektur servicebasierter Grids
  - Standardisierung
  - Evolution / Erweiterungen
- Informatica Cloud Data Integration Solutions
- Zusammenfassung

Begriffsbestimmung: Datenintegration

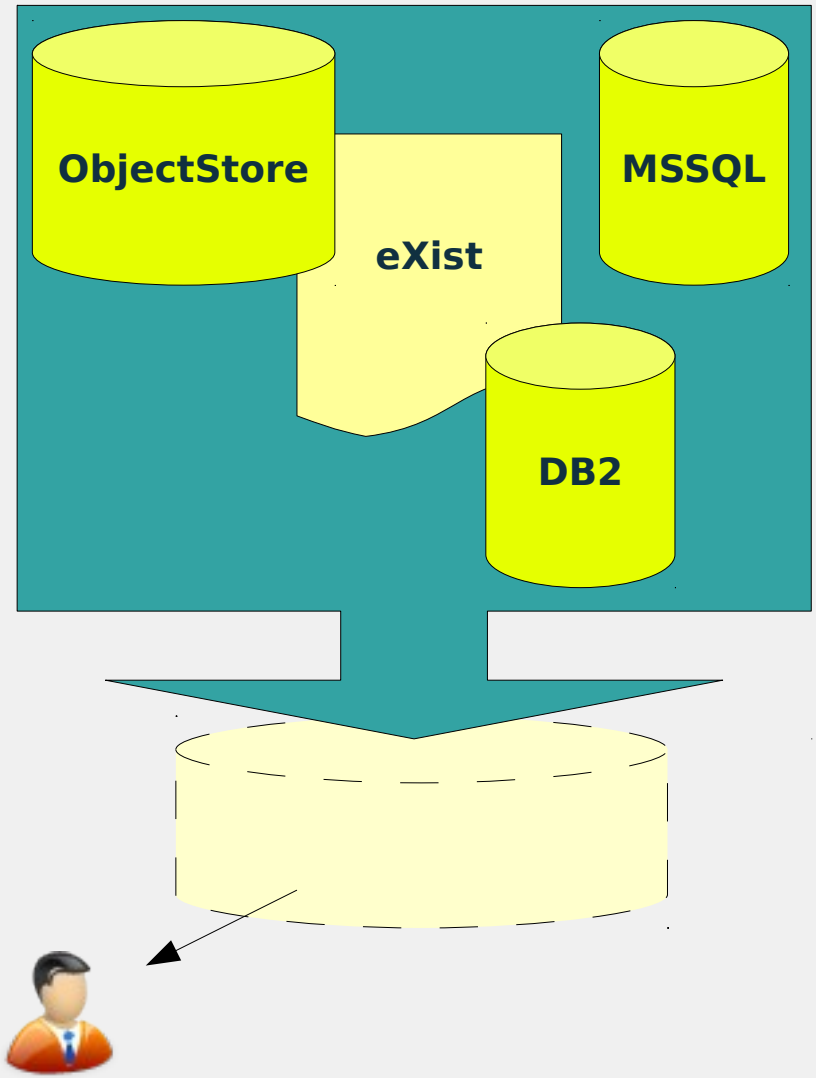
- Ziel: einheitlicher Zugriff auf mehrere DBs
  - Verteilte Datenbanken
  - Dezentral verwaltete Datenbanken

→ Problem:
Heterogene Datenbanken



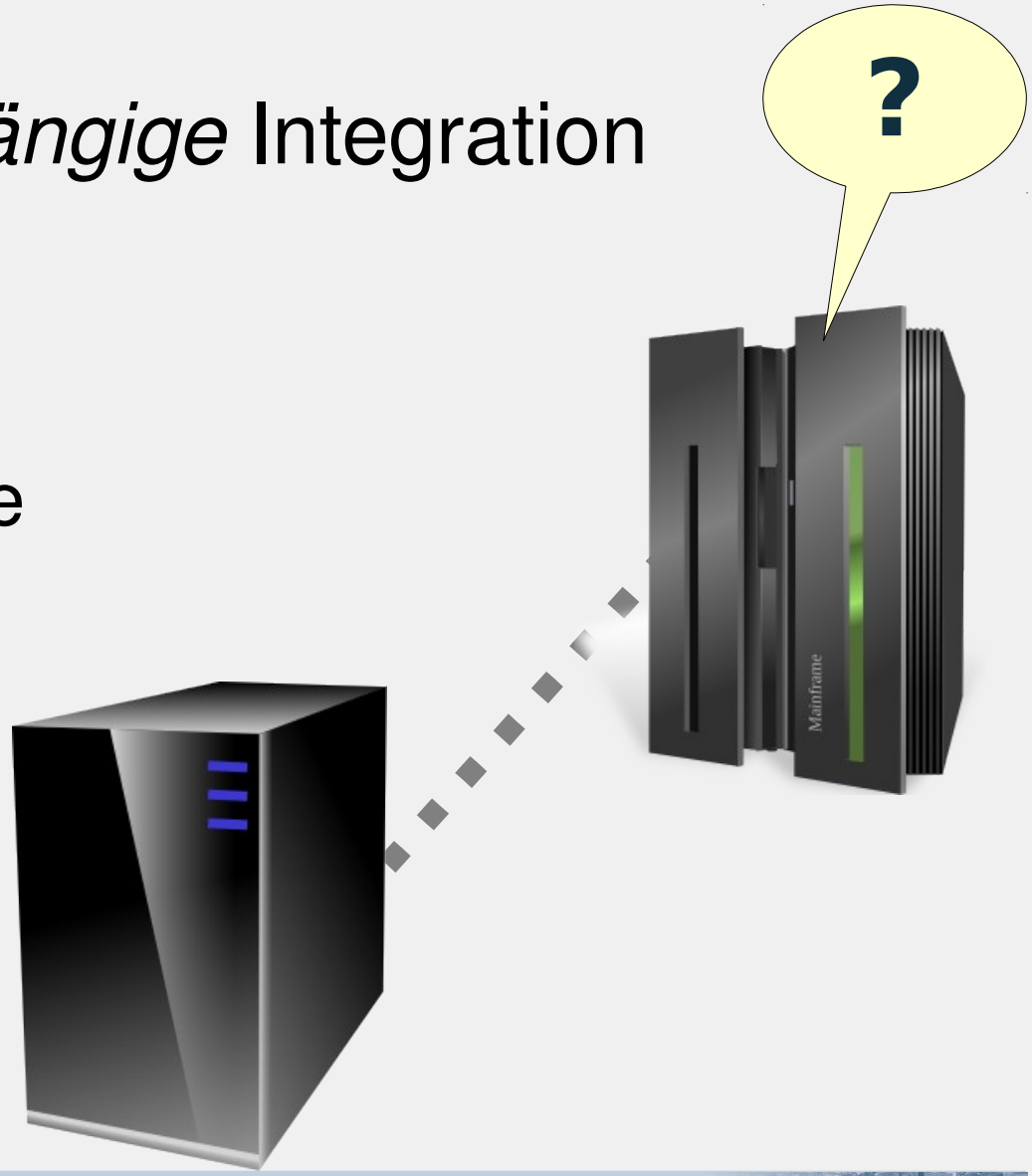
Begriffsbestimmung: Datenintegration

- Heterogenität: Ausprägungen
  - Zugriff
  - Syntax
  - Datenmodell
  - Schema
  - …
- Data Cleaning
- Erstellung einheitlicher Sicht



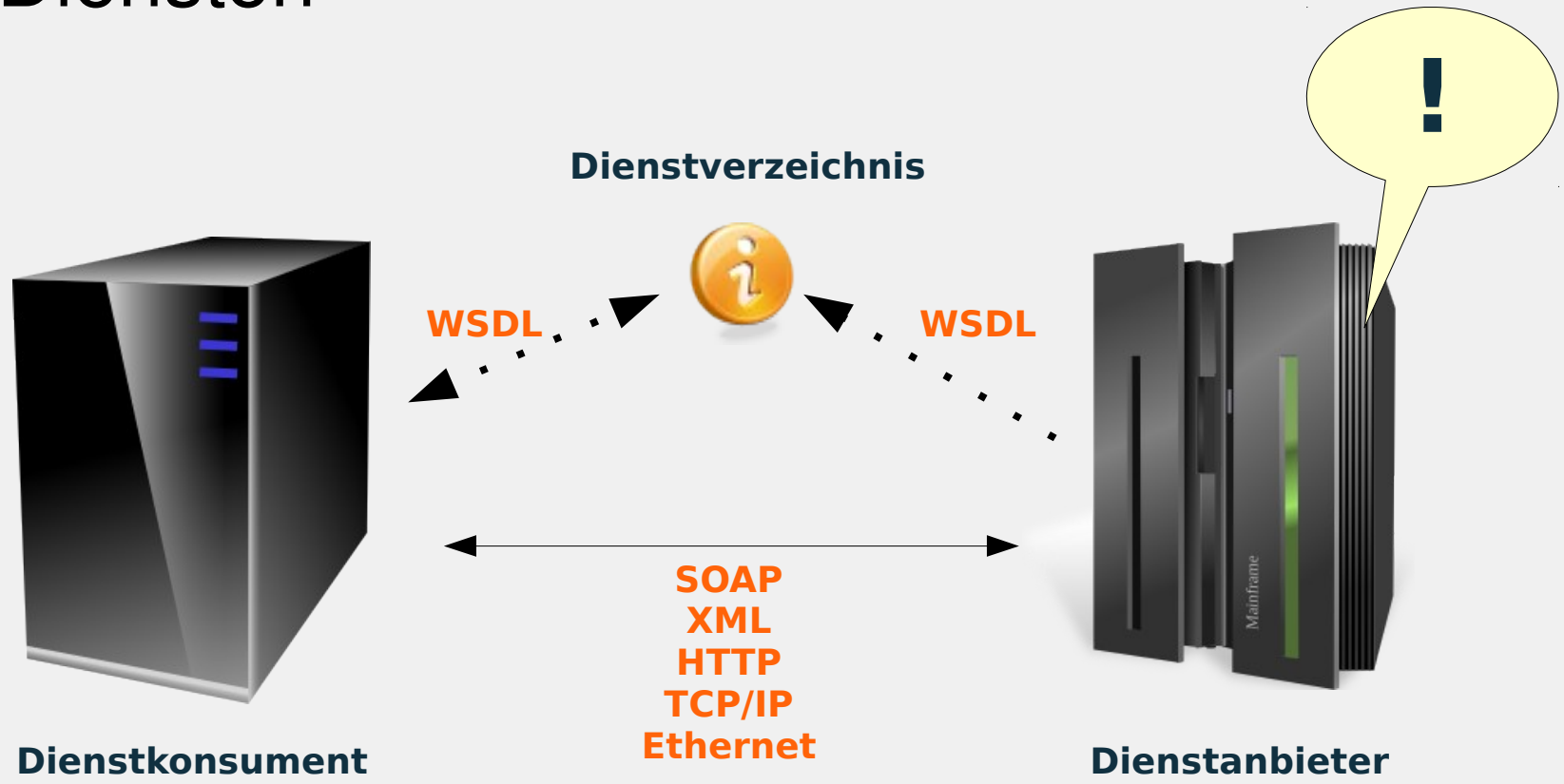
Begriffsbestimmung: Dienstbasierte Architekturen

- Ziel: *technologieunabhängige* Integration heterogener Systeme
  - Transportprotokoll
  - Programmiersprache
  - Plattform



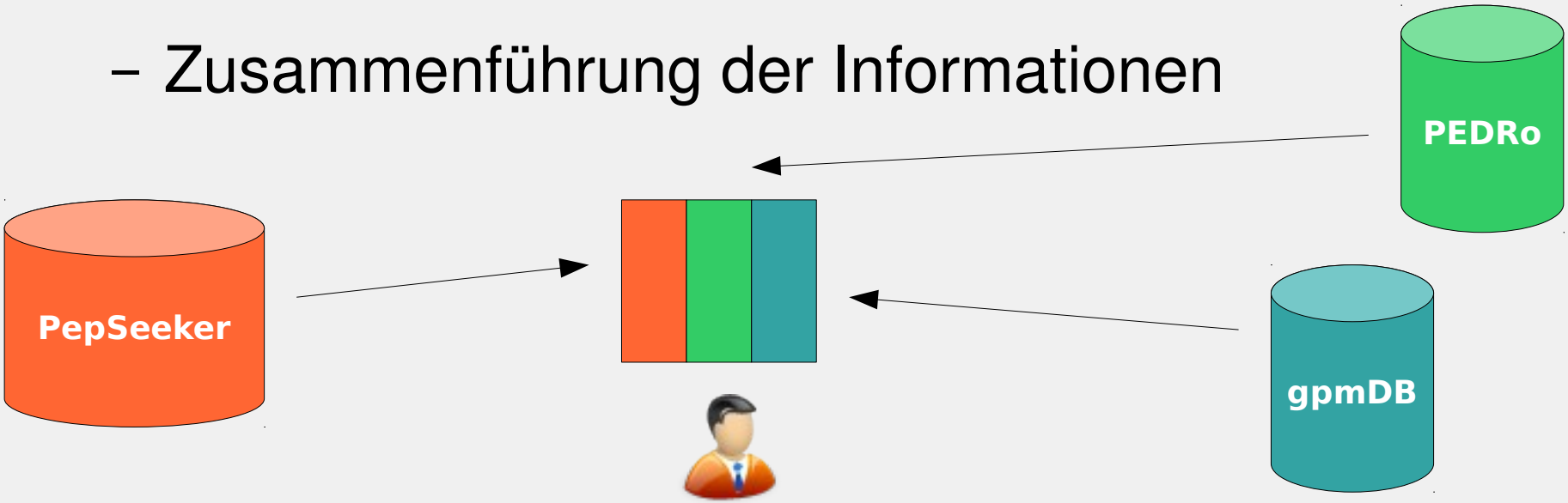
Begriffsbestimmung: Dienstbasierte Architekturen

- Beschreibung, Veröffentlichung, Suche, Nutzung von Diensten



Motivation

- ISPIDER Project
  - Erfassung biochemischer Daten
  - Identifizierung von Proteomen an verschiedenen Standorten
  - Zusammenführung der Informationen

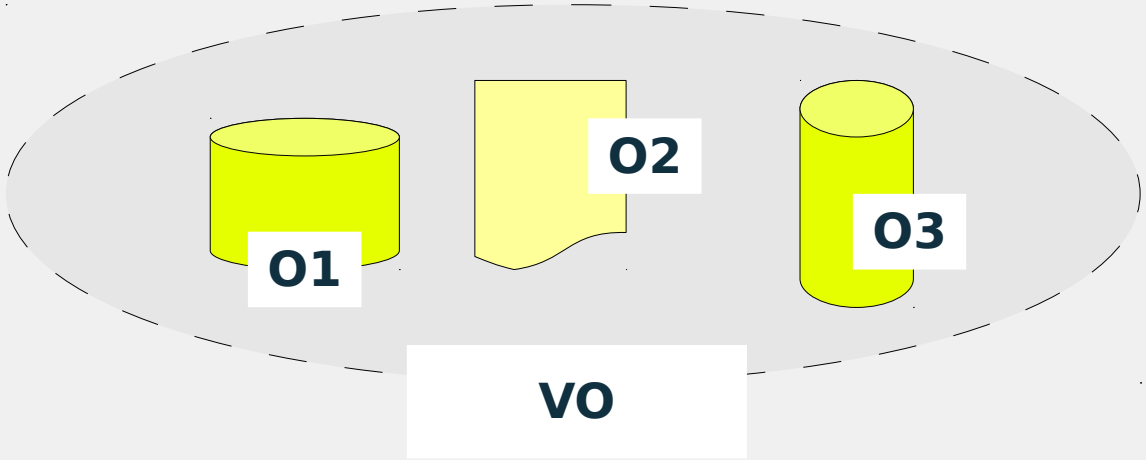


Abgrenzung Grid - Cloud

- Gemeinsamkeit: Dynamische Bereitstellung von Speicher und Rechenkapazität über Netzwerk
- Unterschiede:
  - Cloud
    - Zentraler Anbieter
    - Ökonomischer Ansatz
  - Grid
    - Virtuelle Organisation (VO) / dezentrale Admin.
    - Wissenschaftlicher Kontext

Abgrenzung Grid - Cloud

Grid

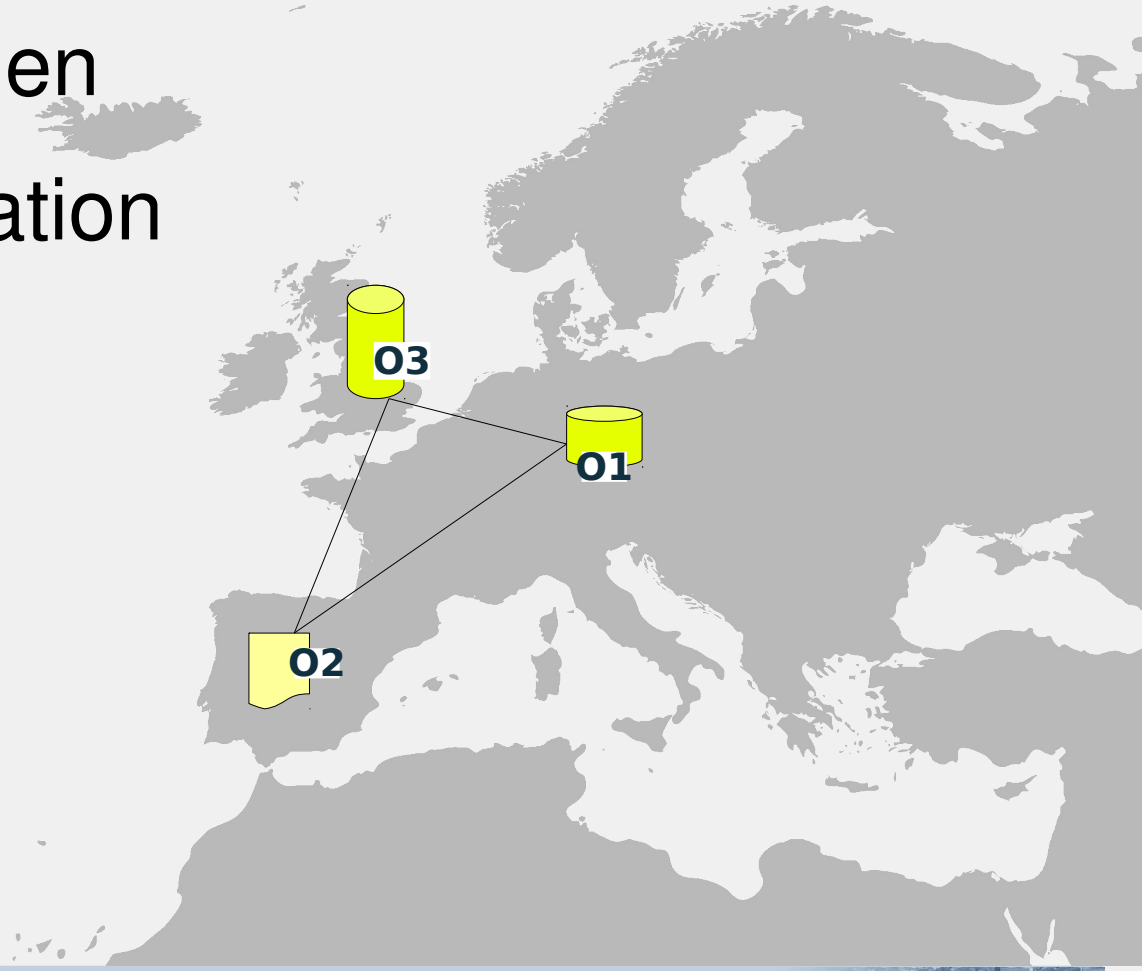


Cloud

OGSA: Architektur servicebasierter Grids

- Hoher Grad an Heterogenität
- Autonome Datenquellen
- Dezentrale Administration

→ Standardisierung



OGSA: Architektur servicebasierter Grids

- Standardisierungsgremium:
  Open Grid Forum (OGF)
- Diverse Arbeitsgruppen, u.a.:
  Database Access and Integration Services (DAIS-WG)
- Architektur:
  Open Grid Services Architecture (OGSA)

OGSA: Architektur servicebasierter Grids

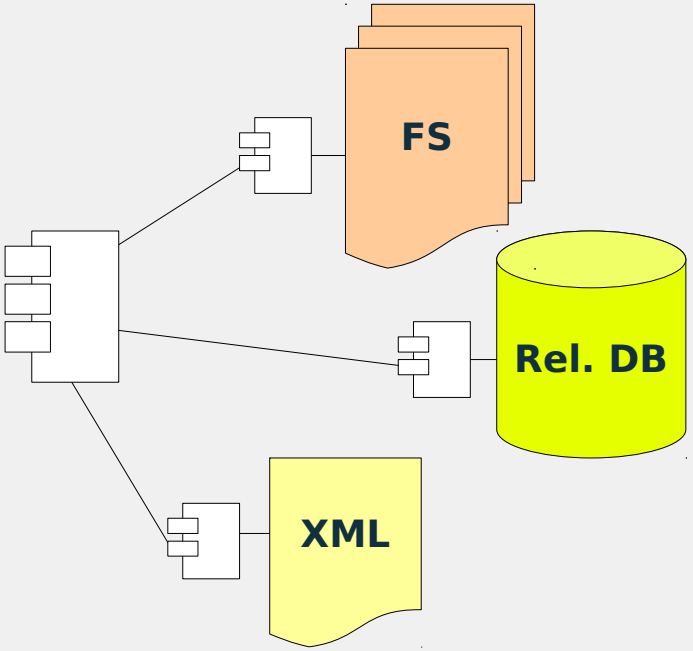
- Repräsentation von Ressourcen durch Dienste
  → Grid Services
- Standardisierung von Schnittstellen
  - Auffinden von Grid Services
  - Erzeugen / Beenden von Grid-Dienstinstanzen
  - Nachrichtenaustausch

OGSA-DAI / OGSA-DQP: Übersicht

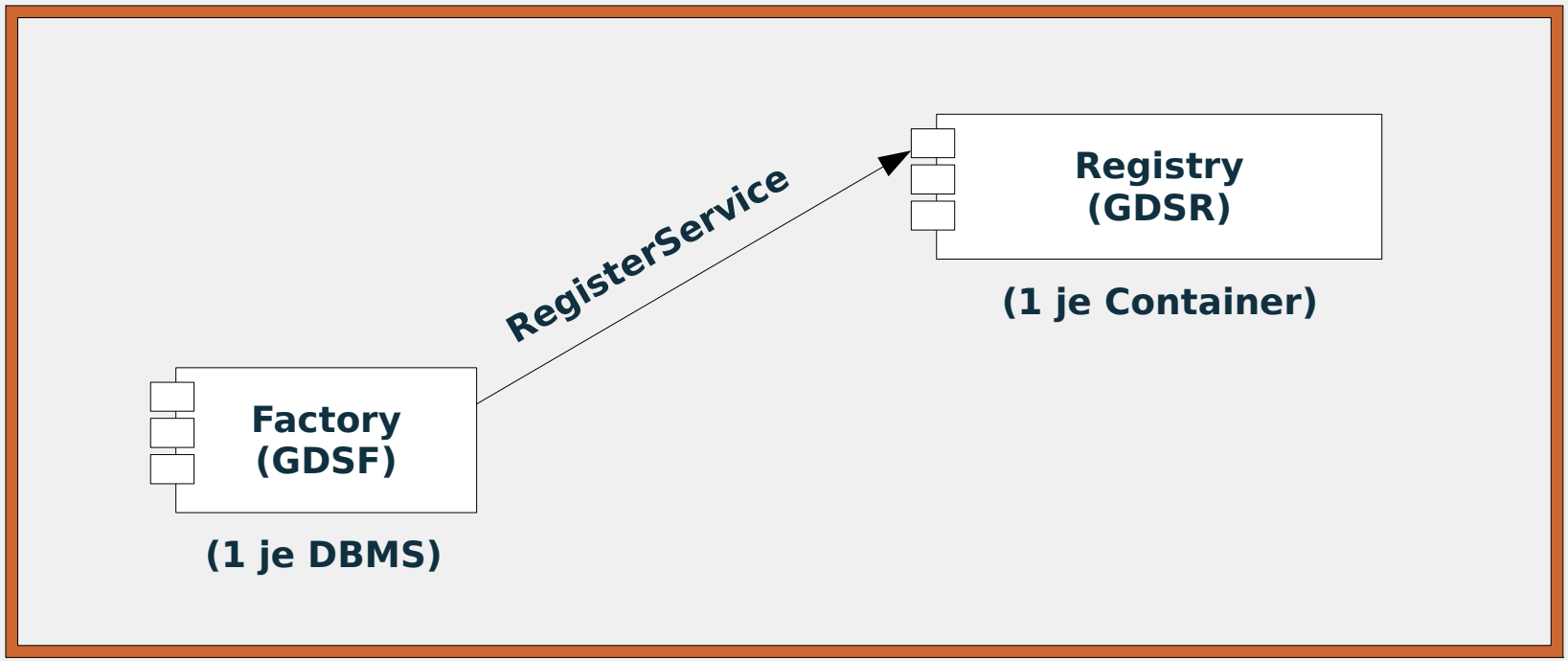
- OGSA-DAI:
  dienstbasierter Zugriff auf Datenbanken
- OGSA-DQP:
  Koordinierung des Zugriffs auf mehrere OGSA-DAI-Ressourcen

OGSA-DAI

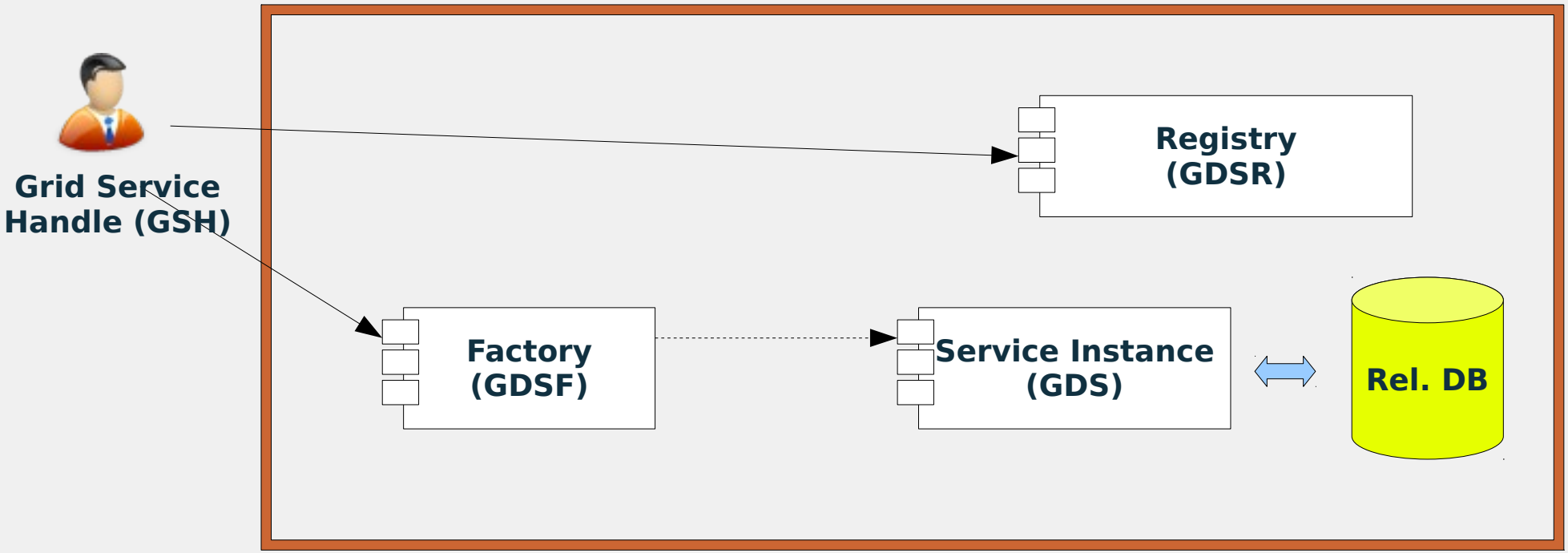
- Data Access and Integration (OGSA-DAI)
- Java-basierte Middleware zur Integration heterogener Datenquellen auf Basis der OGSA



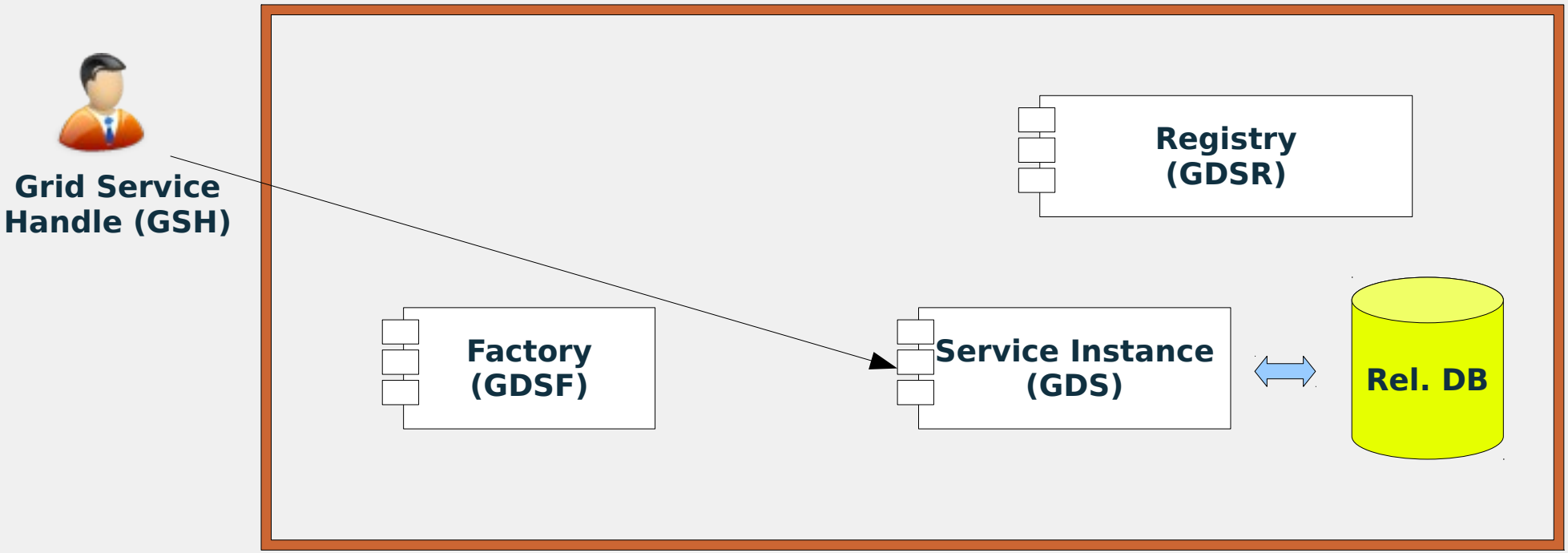
- Ablauf einer Anfrage gegen das Grid
  (a) Container-Start



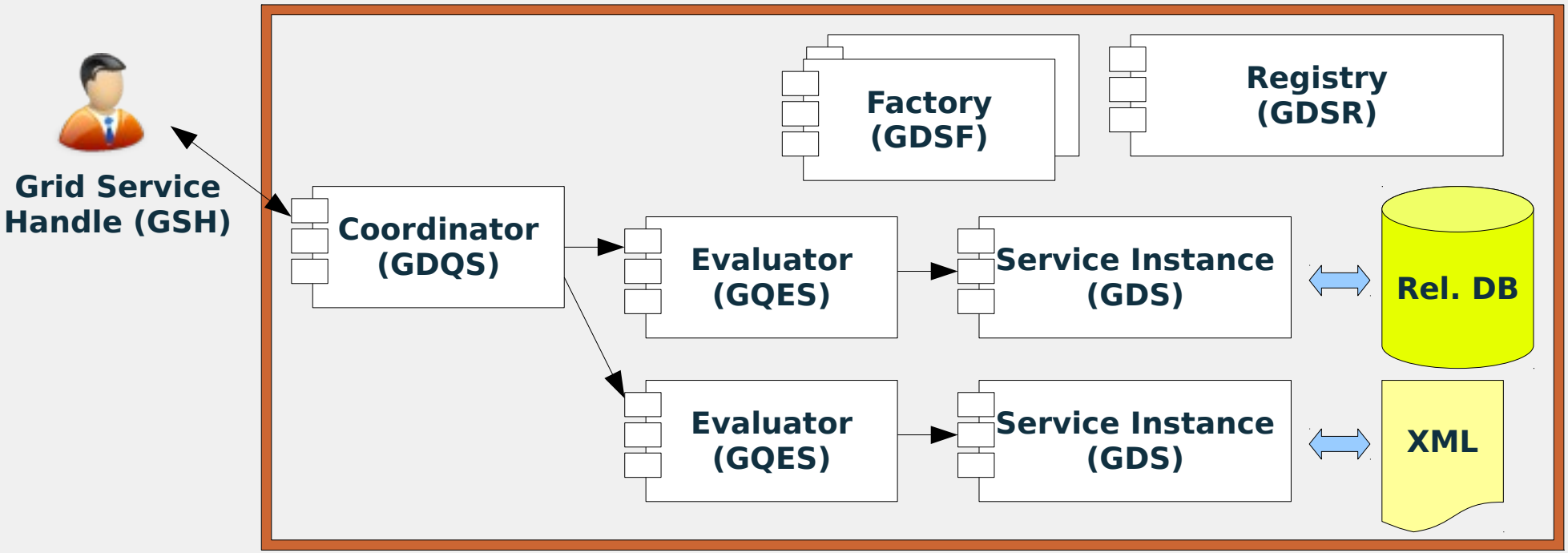
- Ablauf einer Anfrage gegen das Grid
  (b) Dienstlokalisierung



- Ablauf einer Anfrage gegen das Grid
  (c) Durchführen einer Anfrage



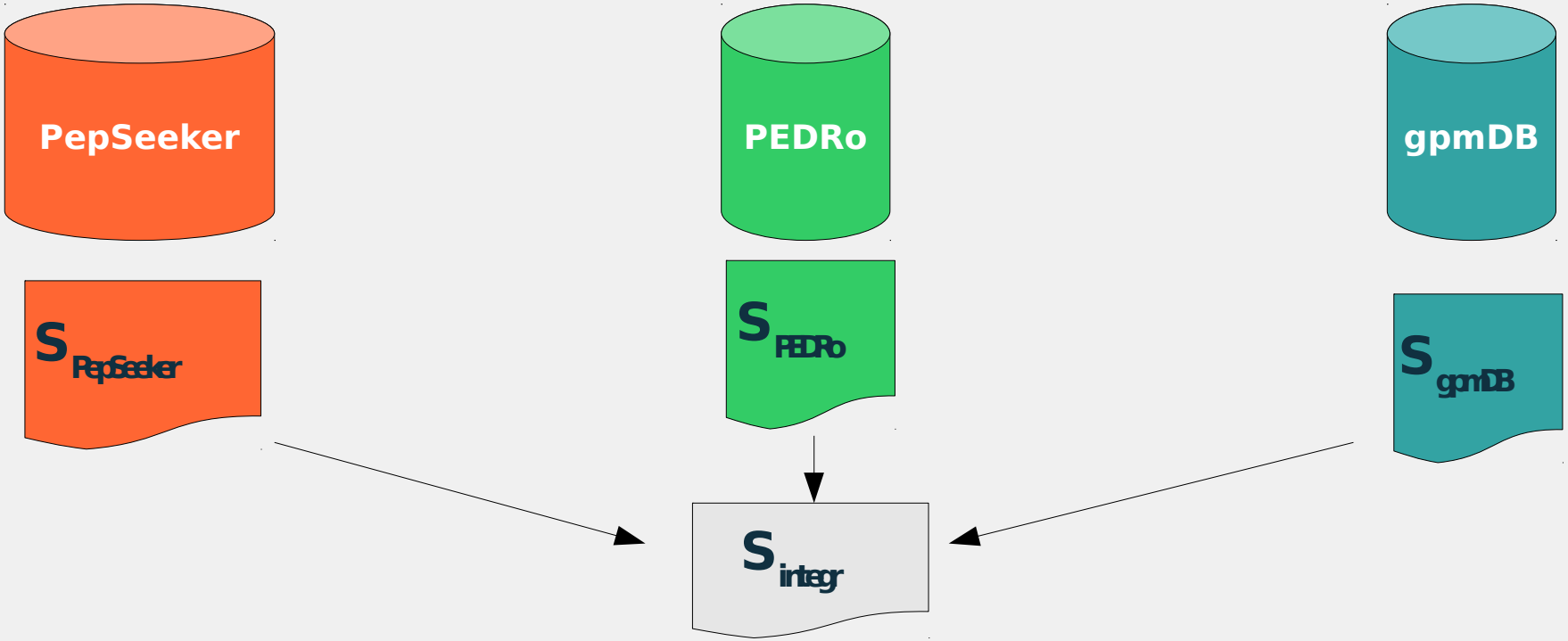
- Ablauf einer Anfrage gegen das Grid
  (c) Durchführen einer verteilten Anfrage via DQP



Einschub: Schemaintegration

- Dezentrale Administration
  → Schemaevolution
  → einmalige Schemaintegration unzureichend
- Grid Data Integration System (GDIS)
  - Mapping-Katalog
  - Änderung / Hinzufügen einer Ressource: Update
- Ausführliche Beschreibung s. Ausarbeitung

ISPIDER: virtuelle Datenintegration

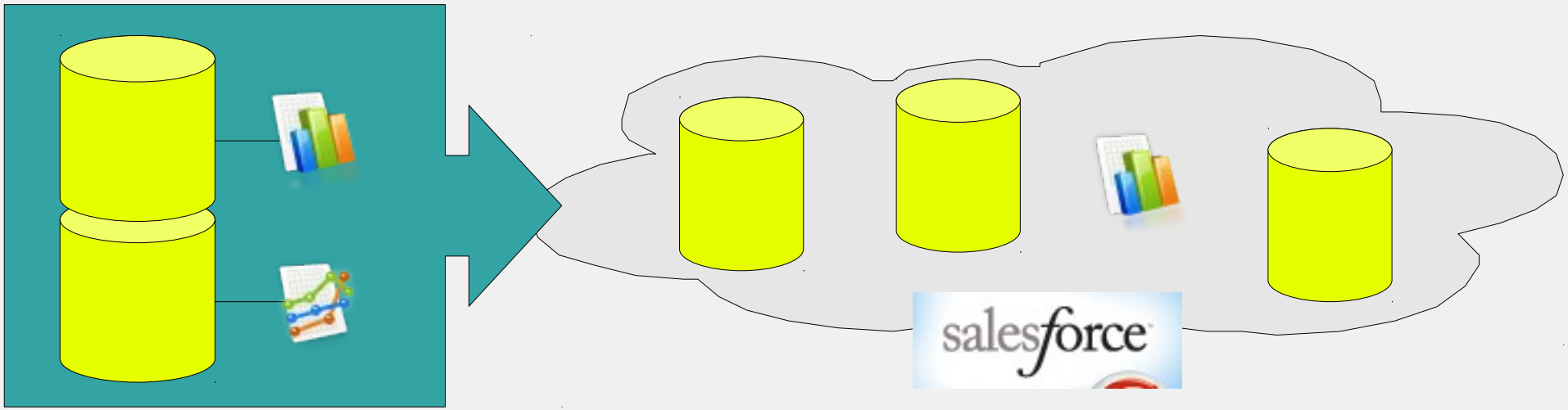


LSID	src_id	acc_nr	col_xy
URN:LSID:ispider.man.ac.uk:pepseeker.protein:1	1	ENSP00000339074	...
URN:LSID:ispider.man.ac.uk:pepseeker.protein:2	2	ENSP00000339074	...
URN:LSID:ispider.man.ac.uk:pedro.protein:1	1	ENSP00000339074	...
URN:LSID:ispider.man.ac.uk:gpmdb.protein:1	1	ENSP00000339074	...

- Nutzung von OGSA-DAI und OGSA-DQP
- Manuelles Erstellen des globalen Schemas
  - Basis: S_{PEDRo}
  - Beschreibung der Korrespondenzen $S_{local} \rightarrow S_{global}$
    → IQL, Transformation Pathways
- Speicherung in *Schemas & Transformations Repository*
  → Schemaevolution möglich

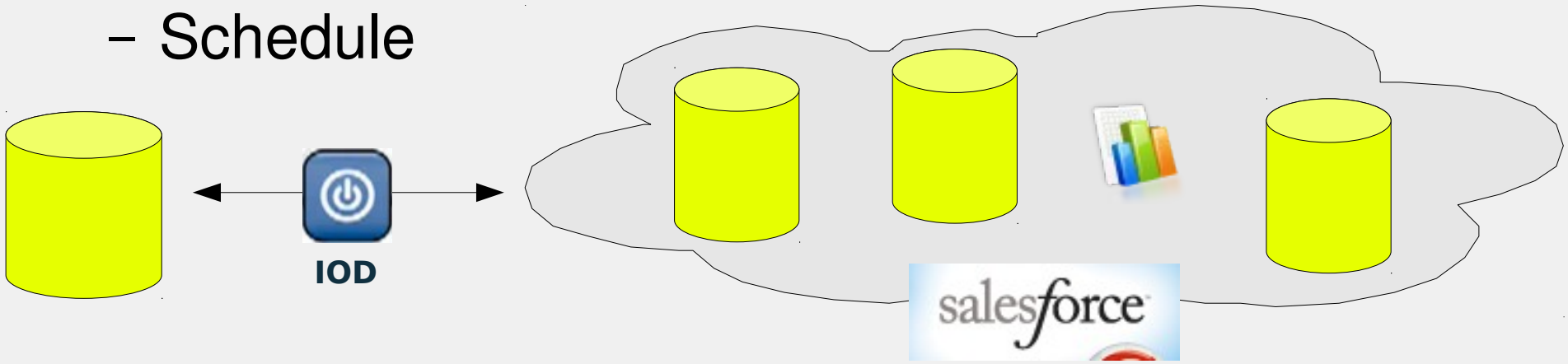
Informatica Cloud Data Integration Solutions

- SaaS Anwendung: Salesforce CRM
- Cloud Plattform: Force.com
- Problematik: Migration lokaler Datenbestände aus Legacy Software nach Salesforce CRM



Informatica Cloud Data Integration Solutions

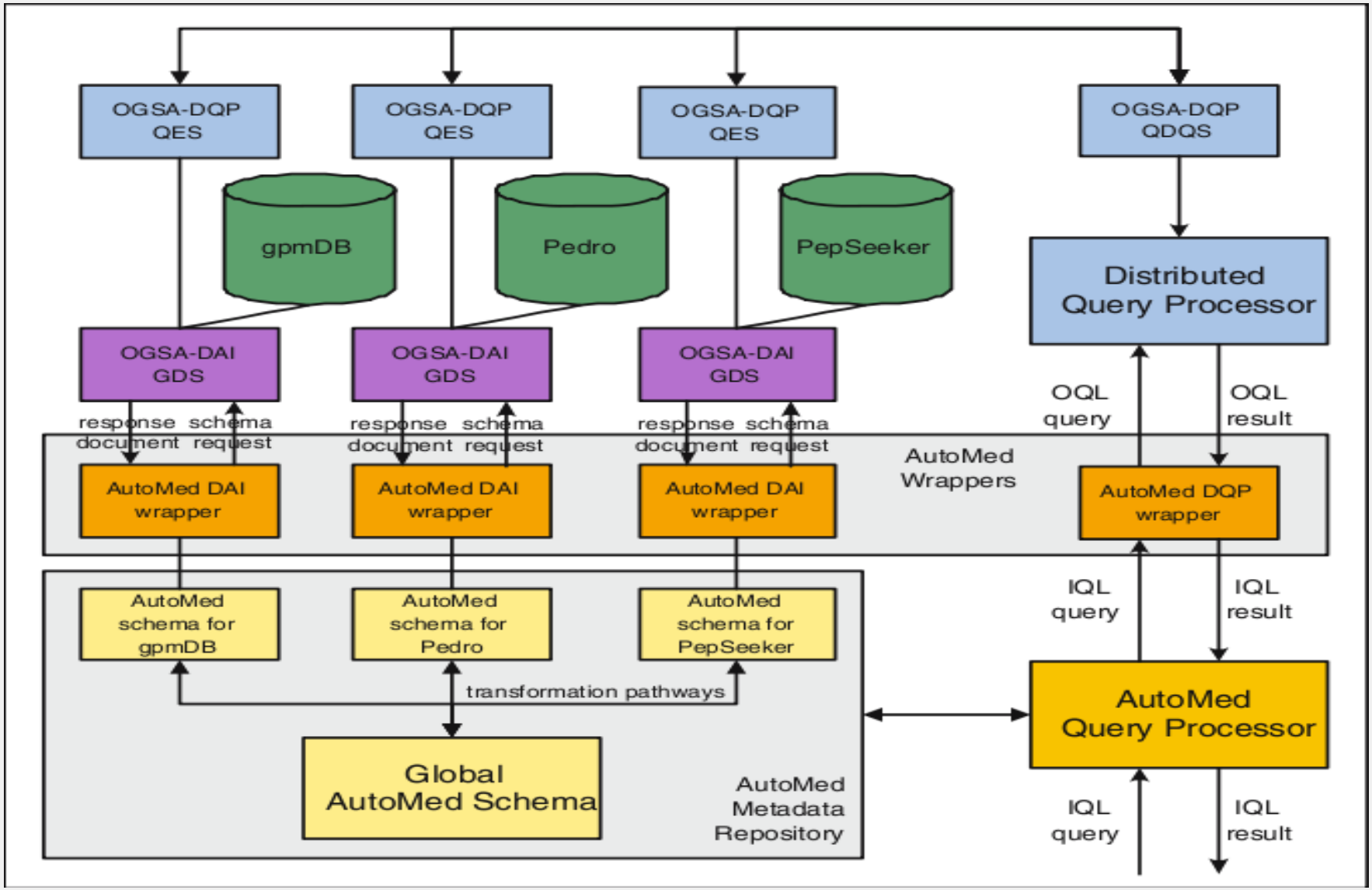
- Informatica On Demand (IOD):
  Webbasierte Konfiguration von Verbindungen
  - Source Connection, Target Connection
  - Filter Definition
  - Field Mappings, Transformation
  - Schedule



Zusammenfassung

- Integrationsproblematik in Grids und Clouds vorhanden
- Lösung in Grids: Standardisierung
  - Virtuelle Integration
  - Dienstbasierte Ansätze
  - OGF
- Lösung in Clouds: Eigenentwicklungen
  - Physische Integration
  - Migration in die Cloud
  - Proprietäre Ansätze

ISPIDER Architecture



[ZFB+ 06]

- [ZFB+ 06] Lucas Zamboulis, Hao Fan, Khalid Belhajjame, Jennifer A. Siepen, Andrew Jones, Nigel J. Martin, Alexandra Poulovassilis, Simon J. Hubbard, Suzanne M. Embury und Norman W. Paton. Data Access and Integration in the ISPIDER Proteomics Grid. In Ulf Leser, Felix Naumann und Barbara A. Eckman, Hrsg., DILS, Jgg. 4075 of Lecture Notes in Computer Science, Seiten 3–18. Springer, 2006.