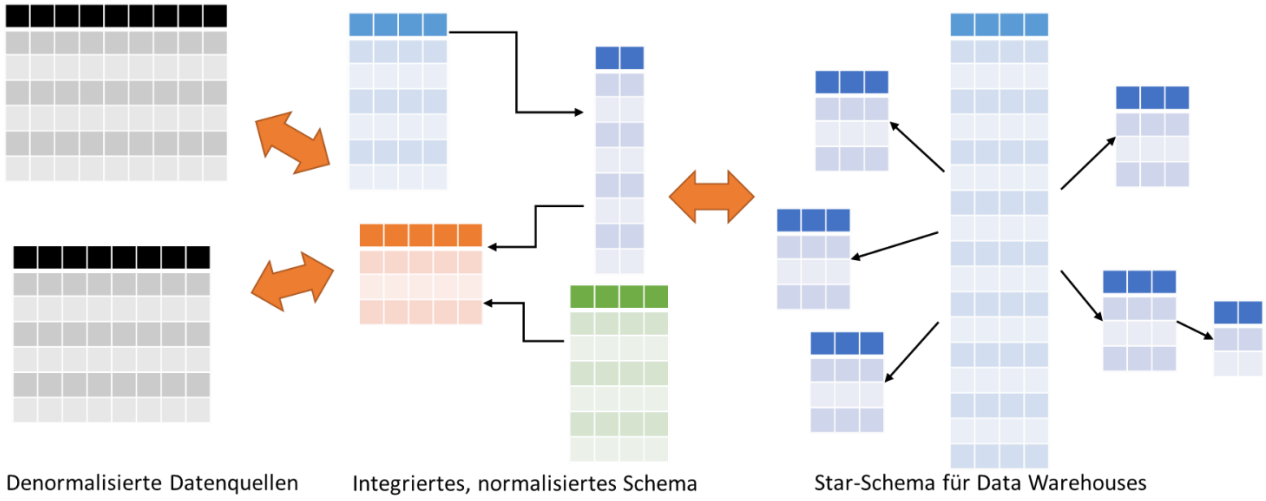
BCNF*: Automatische Schemaanalyse und Datentransformation für Data Warehouses

Bachelorprojekt für 2021/2022
Fachgebiet Informationssysteme, Prof. Dr. Felix Naumann

Jährlich transportiert SBB Cargo AG rund 29,8 Millionen Tonnen Güter netto im Wagenladungs-, Ganzzugs- und im kombinierten Verkehr innerhalb der Schweiz – dies entspricht knapp 10 000 Lastwagenfahrten pro Tag. Die täglich anfallenden Daten aus den verschiedenen Geschäftsbereichen werden in einem Data Warehouse konsolidiert und analysiert. Im Zuge des weiteren Ausbaus des bestehenden **Data Warehouse** werden kontinuierlich weitere Datenquellen integriert und zur Analyse in BI-Tools aufbereitet. Diese Daten sind sehr vielfältig mit historischen **unterschiedlichen, dahinterliegenden Datenmodellen** und Schemata. Auch der Grad der Normalisierung ist heterogen – teils sind die Daten nur in völlig denormalisierter Form erhältlich. Die typische Grössenordnung solcher Schemata befindet sich im Bereich 10-15 Tabellen, im Einzelfall aber auch zwischen 80 und 100 Tabellen. Die dazugehörigen Datenmengen umfassen bis zu 100 GB.

Um eine strukturierte Ablage mit wenigen Redundanzen zu gewährleisten, und um zugleich eine Vielfalt an analytischen Anwendungen zu ermöglichen, sollen Verfahren entwickelt werden, die automatisiert **verschiedene Schemavarianten** erzeugen und nahtlose Datentransformation zwischen diesen Varianten ermöglichen. So können mit der gleichen Datenmenge verschiedene Anwendungsfälle unterstützt werden, insbesondere die Redundanz-freie Speicherung in normalisierten Schemata und die **hoch-performante Analyse** in Business-Intelligence-Tools auf Sternschemata.



Projektbeschreibung

Zur Unterstützung der Data Engineers und Data Scientists wollen wir einen skalierbaren und toolgestützten Prozess entwickeln, welcher bestehende konventionelle Ansätze zum **Schemamanagement** algorithmisch beschleunigt und eine fachliche Beurteilung erleichtert. In diesem Projekt sollen die vielfältigen Versprechen

der Schemamanagement-Forschung fachlich erweitert und in ein praktisch nützliches Werkzeug überführt werden. Das Projekt soll die folgenden Schritte ermöglichen:

- Semi-automatisiertes Erzeugen der jeweiligen Boyce-Codd- oder anderer Normalformen aus denormalisierten Tabellen, basierend auf **Verfahren aus der Forschung** oder mit neuen Ansätzen
- **Schemaintegration** und -inferenz über mehrere normalisierte Inputtabellen durch
    - Reduktion der so erzeugten Dimensionen und Fakten durch **Schema Matching** auf Tabellen- und Attributebene, basierend auf Verfahren aus der Forschung oder mit neuen Ansätzen
    - Einsatz und Erweiterung von Schlüssel/Fremdschlüssel-Erkennungsverfahren
- Semi-automatisiertes Erzeugen von **Starschemata** durch ein definiertes Regelwerk unter Zuhilfenahme eines Object-Relational-Mapping-Frameworks
- Bewertung der erzeugten Schemata und Datenbanken anhand von Anfrage-Workloads und **Optimierung der Schemata** für diese Workload

Das Bachelorprojekt ermöglicht das Entwickeln einer entsprechenden Softwareanwendung auf Basis realer Daten in einer modernen durchgehenden Multi-Cloud-Plattform.

Projektpartner

Die SBB ist mit mehr als einem Viertel Anteil an der gesamten Güterverkehrsleistung das führende Unternehmen im Schweizer Güterverkehr. Knapp 16 Prozent entfallen auf SBB Cargo AG, die als Puls der Schweizer Wirtschaft die grossen Wirtschaftsräume verbindet.

Das Projekt beginnt im Oktober 2021 und wird durch Prof. Dr. Felix Naumann, sein Team und Partner bei SBB betreut. Fragen können gerne an felix.naumann@hpi.de gerichtet werden.