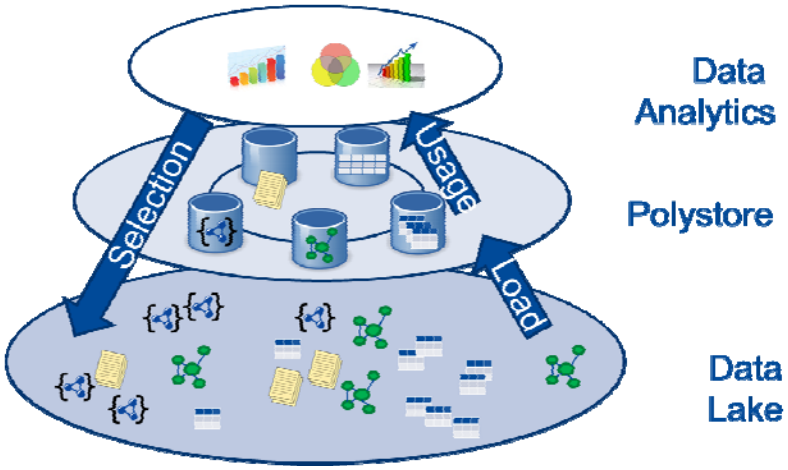
Aufbau eines Polystores für Data Analytics

Data Analytics Auswertungen setzen auf verschiedenen, meist heterogenen Daten auf. Diese sind häufig auf unstrukturierte Weise in verschiedenen Systemen vorhanden und müssen für konkrete Auswertungen (Data Analytics) zusammengeführt werden. Diese Sammlung verschiedener Daten wird auch als Data Lake bezeichnet. Für Auswertezwecke benötigt man die Daten in einer strukturierteren Variante, dazu muss das Laden „on demand“ in einen Polystore erfolgen. Der Gesamtprozess wird als Data Preproccesing bezeichnet. Die Übersicht über diese Zweistufigkeit des Data Preprocessings und die sich anschließende Analyse der Daten ist in der nachfolgenden Abbildung zu finden.



Thema der Arbeit:

Im Rahmen dieser Masterarbeit sollen Daten in einem Data Lake strukturiert werden. Dazu sollen die verfügbaren Datenquellen in einem einheitlichen Katalog (Data Dictionary) dargestellt werden.

Es soll zunächst ein einheitliches Data Dictionary für die Verwaltung und das Auffinden von Daten in verschiedenen DBMS konzipiert werden. Zielsetzung ist eine Orientierung an den Data Dictionaries von relationalen Datenbanken und deren Erweiterung zur Speicherung von Objekten (JSON-Dokumenten) und Graphen. Anschließend sollen Programme erstellt werden, die das DD für vorhandene Datenquellen füllen. Diese Abschlussarbeit ist auf der Ebene Data Lake einzuordnen.

Arbeitsschritte:

- Einarbeitung in die Themenbereiche Polystores, heterogene Datenbanken, Data Dictionaries
- Konzeption eines Data Dictionaries (DD) für Data Lakes, Aufbau des Data Dictionary für rel. DBMS, Objekte (JSON) und Graphen, (es besteht die Anforderung der Erweiterbarkeit des Data Dictionary um zusätzliche Feature descriptions)

- Erstellung eines solchen DD innerhalb einer relationalen DB (mysql)
- Implementierung eines Programms zur Ableitung der Metadaten und Aufnahme in das DD: Entwicklung für relationale Datenbanken (am Bsp. mysql), JSON-DB (am Beispiel mongoDB) und einer GraphDB (am Beispiel Neo4J)
- Evaluation mit Testdatenbanken für alle drei DBMS

Literatur:

- Data Lakes, Polystores
    - Michael Stonebraker: The Case for Polystores, https://wp.sigmod.org/?p=1629
    - Fatemeh Nargesian, ErkangZhu, Renée J. Miller, Ken Q. Pu, Patricia C. Arocena: Data Lake Management: Challenges and Opportunities, Tutorial, VLDB 2019
    - Jeyhun Karimov, Tilmann Rabl, and Volker Markl: PolyBench: The First Benchmark for Polystores
    - Publikationen von Poly18 und Poly19@VLDB
- Data Dictionaries (DD) von rel. DBMS
    - https://www.ibm.com/support/knowledgecenter/SSFGJ4_7.6.0/com.ibm.mbs.doc/configur/r_data_dictionary_tables.html
- Föderierte Datenbanken
    - Sheth/Larson: Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases, 1990,

      http://static.cs.brown.edu/courses/csci2270/papers/federated.pdf