Bachelorprojekt für 2015/2016
Fachgebiet Informationssysteme
Prof. Dr. Felix Naumann

idealo

Data Refinery: High-Performance-Daten-aufbereitung für den idealo-Preisvergleich

Für den idealo-Preisvergleich werden täglich eine Milliarde Datensätze mit Angebotsdaten aus verschiedenen Online-Shops zusammengeführt und verarbeitet. Diese müssen in ein Standardformat transformiert, gefiltert und katalogisiert werden um für die Endbenutzer geeignet zu sein. Die unterschiedlichen Eingangsformate und Katalog-Logiken, die Datenqualität und das schiere Volumen der Daten sind einige der spannendsten Herausforderungen hierbei.

Die aktuelle idealo-Architektur beruht auf einer Vielzahl von Prozessen, Datenbanken und Kommunikationskanälen, was einerseits zu hohem Betriebs- und Wartungsaufwand führt, andererseits den Durchsatz und die Verarbeitungsgeschwindigkeit beeinträchtigt.

Projektbeschreibung

Das Ziel des Bachelorprojektes ist es, eine innovative, ganzheitliche Architektur für die idealo-Angebotsverarbeitung zu konzipieren und prototypisch umzusetzen. Die Plattform soll auf Technologien verteilter Datenverarbeitung wie z. B. Hadoop, Spark und Flink basieren. Es kann dabei davon ausgegangen werden, dass die shop-spezifischen Angebotsdaten bereits von der Datenquelle empfangen und in ein einheitliches Rohdatenformat abgebildet wurden. Daher müssen beispielsweise das Herunterladen von CSV-Dateien und das Auslesen der relevanten Spalten nicht betrachtet werden.

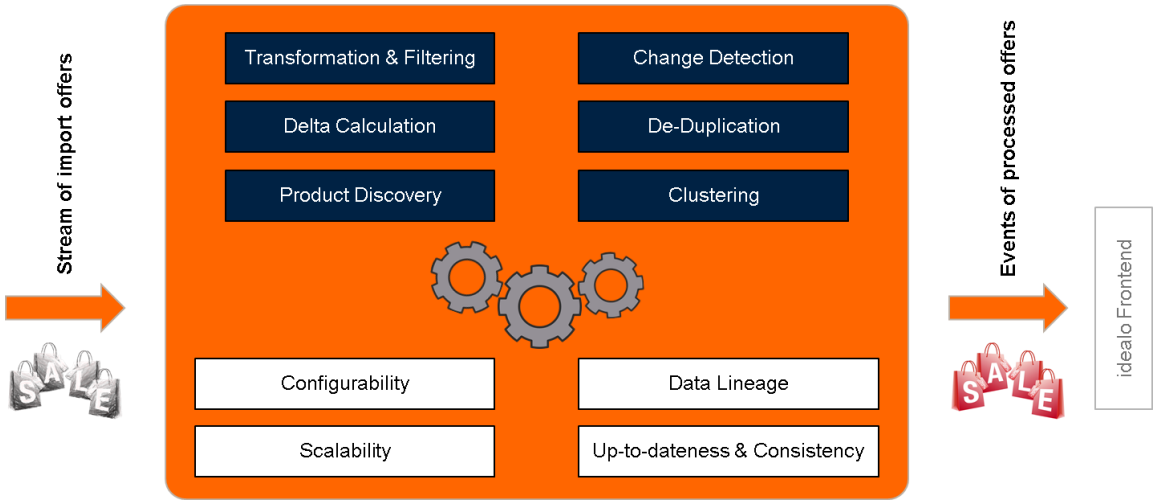
Im Projekt sollen die nachfolgend aufgeführten Anforderungen umgesetzt werden. Bis zu welcher Tiefe dies jeweils erfolgt, stimmen wir gemeinsam im Projektverlauf ab.

Projektziele

- **Schnittstellenimplementierung:** Anbindung an die bereits in einheitlicher Struktur vorliegenden Rohdaten
- **Transformation und Filterung**: Globale und shop-spezifische Normalisierungen, Ersetzungen, Erkennung ungültiger Werte, etc.
- **Änderungserkennung**: Frühzeitiger Abbruch der Verarbeitungskette, wenn sich für nachfolgende Verarbeitungsschritte wesentliche Daten nicht geändert haben
- **Delta-Berechnung**: Erkennung zu löschender Angebote aus aktualisierten Angebotslisten
- **Deduplizierung:** Entfernung von Produktduplikaten global und innerhalb der Lieferung einzelner Shops anhand shop-spezifischer Kriterien
- **Produktzuordnung:** Zuordnung von Shop-Angeboten zu idealo-Kategorien sowie Produkten aus dem idealo-Katalog
- **Produktvorschläge:** Ermittlung geeigneter Angebotsgruppen zur Erweiterung des Produktkatalogs basierend auf erwarteter Angebotsanzahl, Popularität und Marktrelevanz
- **Wartung:** Konfigurierbarkeit globaler und shop-spezifischer Regeln für alle o. g. Schritte
- **Data Lineage**: Nachvollziehbarkeit der Verarbeitungsschritte und -fehler je Angebot
- **Horizontale Skalierbarkeit** für kleine (Testdaten) und sehr große Angebotsmengen, idealerweise mit automatischer Anpassung an Lastspitzen

- **Datenaktualität und -konsistenz** durch eine event-getriebene Verarbeitungskette mit minimalen Durchlaufzeiten von der Datenanlieferung bis zum Endnutzer

Weiterhin sollen die üblichen Anforderungen in der Softwareentwicklung beachtet werden: Wiederverwendung (ggf. existierender idealo-Module zur Angebotsanreicherung), Modularität (Erweiterbarkeit), Einfachheit, Effizienz, Robustheit (Fehlertoleranz), Transparenz (Monitoring) und Integrierbarkeit mit anderen Systemkomponenten (Importer, Content-Tools, Frontend).



Als zu untersuchende Zielplattform für die Implementierung sollen vor allem streaming-orientierte In-Memory-Frameworks berücksichtigt werden, wie z. B. **Apache Spark** oder **Apache Flink**. Hinsichtlich der Anwendungsarchitektur ist ein möglicher Lösungsansatz CQRS (Command-Query-Responsibility-Separation), bei dem Angebots-Rohdaten als Command-Stream ein verteiltes Objektmodell aktualisieren und die daraus resultierenden Änderungen als Event an Query-Datenbanken (z. B. eine Suchmaschine) für die Endnutzer weitergeleitet werden.

Um eine gemeinsame und flexible Konkretisierung der zu erreichenden Ziele zu ermöglichen, ist ein **agiles Vorgehen** erwünscht.

Projektpartner

idealo ist Deutschlands größter Online-Preisvergleich. Seit seiner Gründung im Jahr 2000 hat sich das Unternehmen vom Drei-Mann-Startup zum Marktführer unter den Produktpreisvergleichen entwickelt. In den Büros in Berlin-Kreuzberg arbeiten über 600 Menschen täglich daran, Verbraucher umfassend, aktuell und genau über Online-Angebote und -Händler zu beraten.

idealo arbeitet eng mit tausenden Onlineshops zusammen. Egal ob Global Player wie Amazon oder spezialisierter Anbieter mit kleinem Produktsortiment: Bei idealo hat jeder Händler die Chance, mit dem günstigsten Angebot ganz oben zu stehen.

idealo ist der erste und einzige Produktpreisvergleich mit dem Siegel des TÜV Saarland für geprüfte Vergleichsportale. Das Unternehmen legt Wert auf Nachhaltigkeit und kooperiert u. a. mit Verbraucherschutz-Vertretern wie der Stiftung Warentest, HTV-Life und Der Blaue Engel.

Das Projekt für 4-8 Studenten beginnt am 1. Oktober 2015 und wird durch Prof. Dr. Felix Naumann, Sebastian Kruse und Thorsten Papenbrock betreut. Fragen können gerne an felix.naumann@hpi.de gerichtet werden.