Prof. Dr.-Ing. Thomas Kudraß
HTWK Leipzig, F IM

Projekt zur Lehrveranstaltung Data Warehousing

New York City Taxi

Datenquellen

Interne Datenquellen zu Taxifahrten

Ausgangspunkt bilden die gesammelten Daten der New York City Taxi Cabs (kurz: NYC Taxi). Die Taxi und Limousine Commission (TLC) von New York City veröffentlicht eine Vielzahl der Daten von Taxis und Taxifahrten als Open Data in der Kategorie *Transportation* auf der OpenData Website der Stadt New York: https://opendata.cityofnewyork.us/ Die Daten selbst werden von Technologie-Providern dem TLC zur Verfügung gestellt, die durch das TPEP-Programm zertifiziert werden (TPEP = Taxicab Passenger Enhancement Program). Das TLC veröffentlicht die Daten der Taxifahrten unter:
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
Traditionell sind die Taxis in New York durch ihre unverwechselbare gelbe Farbe (Yellow Cab) zu erkennen.
Ein großer Teil aller Fahrten der Yellow Cabs werden auf Zuruf von der Straße entweder in Manhattan oder an einem der Flughäfen ausgelöst. Im Jahre 2012 wurden die apfel-grün lackierten Boro-Taxis eingeführt. Im Prinzip bieten sie die gleiche Leistung wie die gelben Taxis und dürfen im *Borough* ihrer Registrierung (daher der Name) Passagiere auf Zuruf mitnehmen oder vorbestellte Fahrten durchführen. Es ist ihnen nicht erlaubt, regelmäßig Fahrten zu den Flughäfen durchzuführen oder im Zentrum Manhattans „Pickups“ anzunehmen – als Grenzen gelten East 96th und West 110th Street. Es gibt fünf Boroughs in New York: The Bronx, Brooklyn, Manhattan, Queens und Staten Island. Die Tarife sind die gleichen wie für Yellow Cabs.
Außerdem gibt es in der Stadt die sogenannten For-Hire Vehicles (FHV) Diese sind dadurch gekennzeichnet, dass Fahrgäste gemeinsam auf einer von ihnen bestimmten Route transportiert werden. Somit unterscheiden sich die FHV-Taxis einerseits vom öffentlichen Personennahverkehr, funktionieren aber auch anders als Mietwagen oder Carsharing, wo Fahrzeuge vom Kunden selbst gesteuert werden.

Die Datensätze von gelben und grünen Taxis umfassen die Zeiten von Beginn und Ende der Fahrt sowie die Koordinaten von Beginn und Ende der Tour, die Entfernung, Fahrpreise und Gebühren, Bezahltyp sowie die Anzahl der Passagiere, die vom Fahrer erfasst wird.

Die Datensätze der FHV-Taxis umfassen Felder zur Erfassung der Lizenz-Nummer der Dispatcher-Basis, Datum und Zeit, wann ein Fahrgast ein- bzw. ausgestiegen ist (Pick-Up bzw. Drop-Off), ebenso die Zone von Beginn und Ende der Tour. Weiterhin wird in einem Flag vermerkt, ob die Tour Teil einer größeren gemeinsamen Fahrt mit anderen Fahrgästen war (shared ride chain), wie es typisch ist für große amerikanische FHV-Anbieter wie Uber oder Lyft Line. Die Taxizonen werden ebenfalls auf der Website der TLC dokumentiert. Die FHV-Datensätze stammen von den Dispatchern und werden von der TLC übernommen, die nicht für deren Datenqualität garantieren kann.

Die Daten der TLC sind auch zu finden auf der Google-Cloud-Plattform BigQuery:
https://cloud.google.com/bigquery/
Mit einem Gmail-Account kann man sich direkt in die GoogleCloud Plattform (GCP) einloggen. Dabei ist zu beachten, dass man Daten in geringem Umfang kostenlos herunterladen kann (keine Kosten unter 1 TB Download). Im Unterschied zu den Rohdaten der TLC kann man auf der GoogleCloud wesentlich komfortabler durch Abfragen bestimmte Rohdaten für den ETL-Prozess vorverarbeiten. Ebenso besteht die Möglichkeit, die Preview-Funktion zu nutzen, um den Umfang der herunterzuladenden Daten besser abzuschätzen.
Über die URL
https://bigquery.cloud.google.com/dataset/bigquery-public-data:new_york_taxi_trips?hl=de
kann man direkt auf alle Tabellen der TLC zurückgreifen (Anmeldung mit Google-Konto erforderlich).
Dabei sind folgende Tabellen verfügbar:

- tlc_fhv_trips_20<xx> - Fahrten der FHV-Taxis seit 2015
- tlc_green_trips_20<xx> - Fahrten der grünen Taxis seit 2013
- tlc_yellow_trips_20<xx> - Fahrten der gelben Taxis seit 2009
- taxi_zone_geom - Beschreibung der Taxizonen (ID, Name, Borough, geogr. Koordinaten)

Die ausführliche Beschreibung der Felder in den Datenquellen kann dem Anhang entnommen werden.

Externe Datenquellen

Um die Aussagekraft der mit einem Data Warehouse zu gewinnenden Informationen noch zu erhöhen, sollte mindestens zwei weitere externe Datenquellen hinzugezogen werden. Diese sollten vorzugsweise der zeitlichen oder der geographischen Dimension zugehörig sein.

In der Dimension Zeit wäre die Kombination mit historischen Wetterdaten von New York aus dem Untersuchungszeitraum interessant. Historische Wetterdaten sind leider zumeist kostenpflichtig, eine mögliche Quelle für tägliche Wetterdaten ist in der Datei jfk_weather.csv zu finden unter: https://github.com/geekman1/nyc_taxi
Eine weitere interessante Kombination wäre die Verknüpfung mit Feiertagen oder anderen großen Ereignissen in New York, die im gleichen Zeitraum wie die betrachteten Taxifahrten stattgefunden haben (z.B. Sportereignisse, Paraden).
Mögliche Quellen hierfür:
http://www.feiertagskalender.ch/
https://en.wikipedia.org/ Category: Annual events in New York City

In der geographischen Dimension wäre es sinnvoll, die in den Datenquellen gespeicherten Geokoordinaten der Taxifahrten mittels *Reverse Geocoding* anzureichern. Hierbei ist zum Verständnis die Verwaltungsgliederung von New York zu betrachten. Die Stadt New York besteht aus 5 Boroughs (Stadtbezirke), die in 59 Community Districts gegliedert sind. Die Community Districts wiederum bestehen aus verschiedenen Neighboorhoods. Die Neighborhoods sind allerdings keine fest definierten Verwaltungseinheiten, sondern

Bezeichnungen, die sich im Laufe der Zeit für einzelne Stadtteile herausgebildet haben (z.B. historische Ortschaften). Eine Übersicht zur Verwaltungsgliederung von New York befindet sich hier: https://de.wikipedia.org/wiki/Verwaltungsgliederung_von_New_York_City

Reverse Geocoding könnte hierbei aus den Koordinaten die Postadresse von Start und Ziel ermitteln oder (somit bereits verdichtet) eine Zuordnung zu Borough und Neighborhood. Weiterhin wäre es noch denkbar, die Nähe bestimmter Sehenswürdigkeiten (Landmarks) wie z.B. Theater, Stadien, mit in die Untersuchung einzubeziehen.

Hierbei könnte entweder die Adresse von Start und Ziel ermittelt werden oder in verdichteter Form eine Zuordnung der Koordinaten zu Borough/Neighborhood.
https://github.com/geekman1/nyc_taxi/blob/master/find_pickup_dropoff_nbhds.py
Die Boroughs und ihre Neighborhoods sind ebenfalls dort zu finden, in einem großen JSON File.
https://raw.githubusercontent.com/geekman1/nyc_taxi/master/nyc_neighborhoods.json

Projekt-Aufgabe

Mit dem zu bearbeitenden Projekt ist der komplette Ablauf eines kleinen Data Warehouse-Projekts zu vollziehen

Phase 1: Analyse

1. Datenverarbeitungsanforderungen
   Es sind mögliche Auswertungen (aus Business-Sicht) und deren Anforderungen (z.B. Häufigkeit, Aktualität, Genauigkeit) zu beschreiben. Dabei sind bestimmte Maßzahlen zu definieren, die auf den Fakten beruhen (z.B. Anzahl transportierter Passagiere).
2. Konzeptuelle Modellierung
   Zur konzeptuellen Modellierung des Data Warehouse sind geeignete Dimensionen, Fakten bzw. Kennzahlen eines multidimensionalen Modells zu spezifizieren. Hierfür ist die ADAPT-Technik zu nutzen [1] [2]. Grundsätzlich sind Auswertungen in einer zeitlichen oder geographischen Dimension vorzunehmen. Die Kennzahlen können die Anzahl der Passagiere, die erzielten Einnahmen oder die gefahrenen Distanzen betreffen.

Das Ergebnis der Analyse ist schriftlich zu dokumentieren.

Phase 2: Entwurf des Data Warehouse

1. Relationaler Entwurf der Basisdatenbank (BDB)
   Diese Datenbank enthält die gesamten Daten des Data Warehouse auf feingranularer Ebene als Basis der Analysen. Die BDB ist nach den Regeln des Datenbankentwurfs konzipiert – nicht nach dem Analysezweck. Sie ist somit redundanzfrei gespeichert (relational 3NF) und bereinigt (Operational Data Store).
2. Entwurf eines MDM-Schemas
   Es ist ein Schema für einen oder mehrere Data Cubes auf der Basis des multidimensionalen Datenmodells (MDM) zu entwerfen, das auf relationalen Strukturen umzusetzen ist (ROLAP). Das MDM-Schema ist zugleich das Zielschema für die aus der Basis-Datenbank zu ladenden Daten in die Cubes.

Das Ergebnis des Entwurfs ist in Form von Diagrammen, versehen mit notwendigen Erläuterungen zu bestimmten Entwurfsentscheidungen, zu dokumentieren.

Phase 3: Implementierung des Data Warehouse

1. Implementierung der Basisdatenbank (BDB)
   Das Schema der Basisdatenbank ist mittels SQL-Anweisungen zu implementieren. Dabei sind die Gegebenheiten der Datenquellen zu berücksichtigen.

2. ETL-Prozess
   Die Datenquellen soll so aufbereitet werden, dass sie im MS SQL Server für OLAP-Anwendungen genutzt werden können. Dazu müssen die Rohdaten extrahiert, und ggf. gefiltert und vorverdichtet werden, um den Umfang der weiterzuverarbeitenden Daten möglichst frühzeitig zu reduzieren. Diese Daten können z.B. als CSV-Datei in die Basisdatenbank geladen werden. Dabei können alternativ die Original-Daten der TLC verwendet werden oder die von der GoogleCloud bereitgestellten Tabellen der TLC. Darüber hinaus sollten auch andere externe Datenquellen (außerhalb der TLC) in die BDB geladen werden, um sie mit den TLC-Daten zu verknüpfen.
3. Implementierung der Data Cubes
   Auf Grundlage der Basisdatenbank sind die Data Cubes zu implementieren, mit den jeweils konzipierten Auswertungsdimensionen entsprechend des entworfenen MDM-Schemas. Dieses sollte nicht manuell, sondern mit Tool-Unterstützung (MS Analysis Services) erfolgen.

Das Schema der Basisdatenbank ist in Form eines SQL-Skripts zu dokumentieren.
Hinweis: Die Projektdatenbank ist auf 20 GB beschränkt, so dass bei der Auswahl der Beispieldaten das Datenvolumen a-priori begrenzt werden sollte.

Phase 4: Auswertungen und Visualisierung

1. Berichte und Diagramme für Datenbank-Auswertungen.
   Es sind mindestens vier ausgewählte Auswertungen zu implementieren. Darin enthalten sollten sein sollte die zeitliche und die räumliche Dimension, angereichert mit Informationen aus den externen Datenquellen. Durch Dimensionshierarchien sollte die Möglichkeit zu verfeinerten Analysen bestehe, wo es sich anbietet. Die Auswertung sollte graphisch erfolgen unter Nutzung eines Reporting Tools (z.B. PowerPivot, Analysis Services).
2. Optimierung der Data Cubes
   Zur Verbesserung der Performance sollten mögliche Optimierungen des DWH Schemas betrachtet werden. Hierzu zählen z.B. Denormalisierung einzelner Tabellen, Dimensions-reduktion, Verkleinerung von Cubes, Partitionierung, Indexierung, Materialisierte Sichten.
3. Evaluation der Ergebnisse
   Ausgehend von der Informationsbedarfsanalyse sind die Ergebnisse der Auswertungen hinsichtlich ihrer Aussagekraft und Steuerungsfähigkeit für die Stadt New York bzw. die New York Taxi Commission (TLC) zu bewerten.

Die Ergebnisse der Phasen 3 und 4 werden in einem Abschlusskolloqium präsentiert. Vorzubereiten sind ein paar Präsentationsfolien sowie die zugehörige Live-Demo.

Termine

Phase 1 und 2 (schriftlich bzw. elektronisch): 01.07.2019
Phase 3 und 4 (mündlich): 16.09.2019 (Kolloq-Termine individuell in KW 38)

Als Datenbank-Plattform kann genutzt werden:

- Microsoft SQL Server 2014 https://www.microsoft.com/de-de/download/details.aspx?id=42299

Für die konzeptuelle Modellierung des MDM-Schemas wird das Visio Stencil für ADAPT empfohlen [1].
Für die Datenauswertung und Visualisierung steht das Excel Add-In PowerPivot bereit, das mit MS SQL Server zusammenarbeiten kann [3] [4], für die mehrdimensionale Analyse können die SQL Server Analysis Services [5] genutzt werden.

Das Projekt ist in Zweiergruppen durchzuführen.

Weitere Referenzen:

[1] ADAPT™: http://www.symcorp.com/tech_expertise_design.html

[2] Symmetry Corporation: Getting Started with ADAPT™ . OLAP Database Design, White Paper: http://www.symcorp.com/downloads/ADAPT_white_paper.pdf

[3] A. Ferrari, M. Russo: Microsoft Excel 2013: Building Data Models with PowerPivot. Microsoft Press, 2013.

[4] S. Giessen: PowerPivot: Einstieg in die Arbeit mit PowerPivot für Microsoft Excel 2013 (Taschenbuch). Create Space Independent Publishing, 2015.

[5] Microsoft: Mehrdimensionale Modellierung, SQL : https://docs.microsoft.com/de-de/sql/analysis-services/multidimensional-modeling-adventure-works-tutorial?view=sql-analysis-services-2017