Projekt Data Warehouse

Bearbeiter: Benedikt Grothues, Julius Seiffert

Datum: 01.07.2016

Data Warehouse

Fakultät Informatik und Medien

Inhalt

Datenquellen

Für das Data Warehouse Projekt wurden folgende Datenquellen ausfindig gemacht.

Intern

- Tlc_green_trips_20<xx> - Fahrten der grünen Taxis in New York von 01.01.2009 bis 31.12.2018
- Tlc_yellow_trips_20<xx> - Fahrten der gelben Taxis in New York von 01.08.2013 bis 31.08.2018
- Taxi_zone_geom – Zonen, in denen Fahrer Passagiere aufnehmen, oder absetzen

Extern

- Wetterdaten: git – jfk_weather.csv – Wetterdaten des JFK Flughafen von 01.01.2009 bis 30.06.2016
- https://catalog.data.gov/ - Sehenswürdigkeiten in NewYork

Phase 1

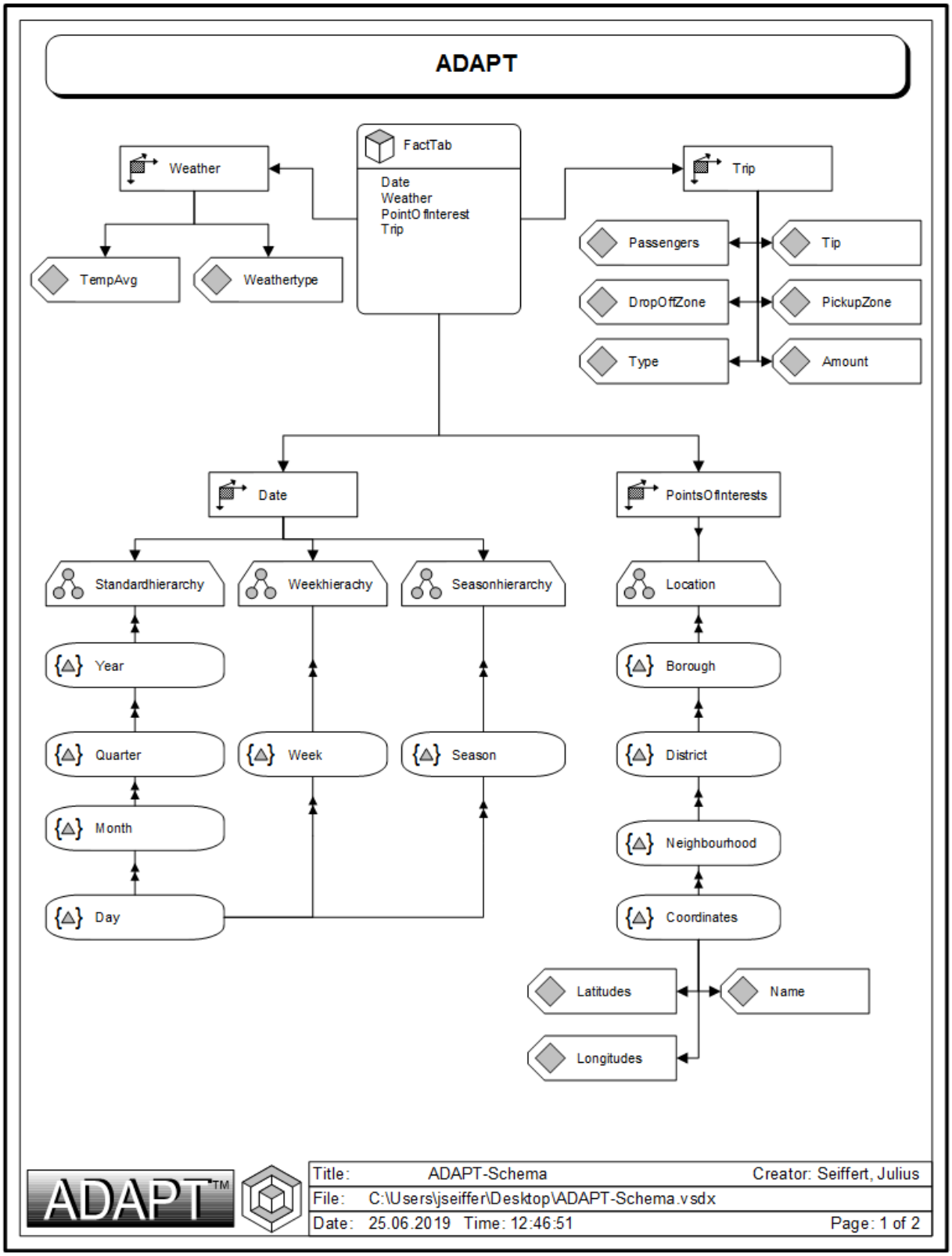
Dieses Datawarehouse soll in folgenden Fragestellungen unterstützen:

1. Welche Sehenswürdigkeiten werden am meisten zu welchem Wetter angefahren?
2. Bei welchem Wetter wird am meisten Taxi gefahren?
3. Bei welchen Sehenswürdigkeiten bekommt der Fahrer das höchste Trinkgeld?
4. Welche Sehenswürdigkeiten werden in welchen Jahreszeiten am häufigsten besucht.

Hierfür werden folgende Fakten und Maßzahlen benötigt.

- Datum (Fakt)
- Wettertyp (Fakt)
- Temperatur (Fakt)
- Sehenswürdigkeiten (Fakt)
- Summe der Fahrten - Maßzahl
- Summe der Passagiere - Maßzahl
- Trinkgeld pro Passagier bei Kreditkarte – Maßzahl

Adapt-Schema



Im ADAPT-Schema lassen sich die 4 Dimensionen und die Faktentabelle sehen. Bei den Dimensionen werden die Hierarchiestufen, sowie die Attribute angezeigt.

Bei der Dimension Date, gibt es als Besonderheit 3 Hierarchien:

- **Standardhierarchie:** Bei der Standardhierarchie wird das Datum nach Jahr, Quartal, Monat und Tag unterteilt.
- **Wochenhierarchie:** Bei der Wochenhierarchie wird einem Tag eine Wochennummer zugeordnet. Die Woche kann nicht in der Standardhierarchie untergebracht werden, das eine Wochennummer selbst zwei Jahre beinhalten kann.
- **Jahreszeithierarchie:** Bei dieser Hierarchie wird die Jahreszeit eingeführt. Diese kann nicht in den anderen Hierarchien untergebracht werden, das eine Jahreszeit Wochen und Jahre überschneiden kann

Bei der Dimension **PointsOfInterests** gibt es eine Hierarchie, in der die Koordinaten der Sehenswürdigkeiten einem *Borough*, einem *District* und einem *Neighbourhood* zugeordnet werden kann.

Die Dimension **Trip** hat keine Hierarchiestufen. Als Attribute werden hier die Passagieren einer Fahrt, das Trinkgeld, den Aufnahmeort, den Zielort, die Art des Taxis und den Betrag der Fahrt aufgelistet.

Die Dimension **Weather** hat keine Hierarchiestufen. Hier werden als Attribute die Durchschnittstemperatur und die Art des Wetter abgebildet.

Phase 2

In Phase 2 wurde die Basisdatenbank, sowie das MDM-Schema des Data Warehouse entworfen.

Basisdatenbank

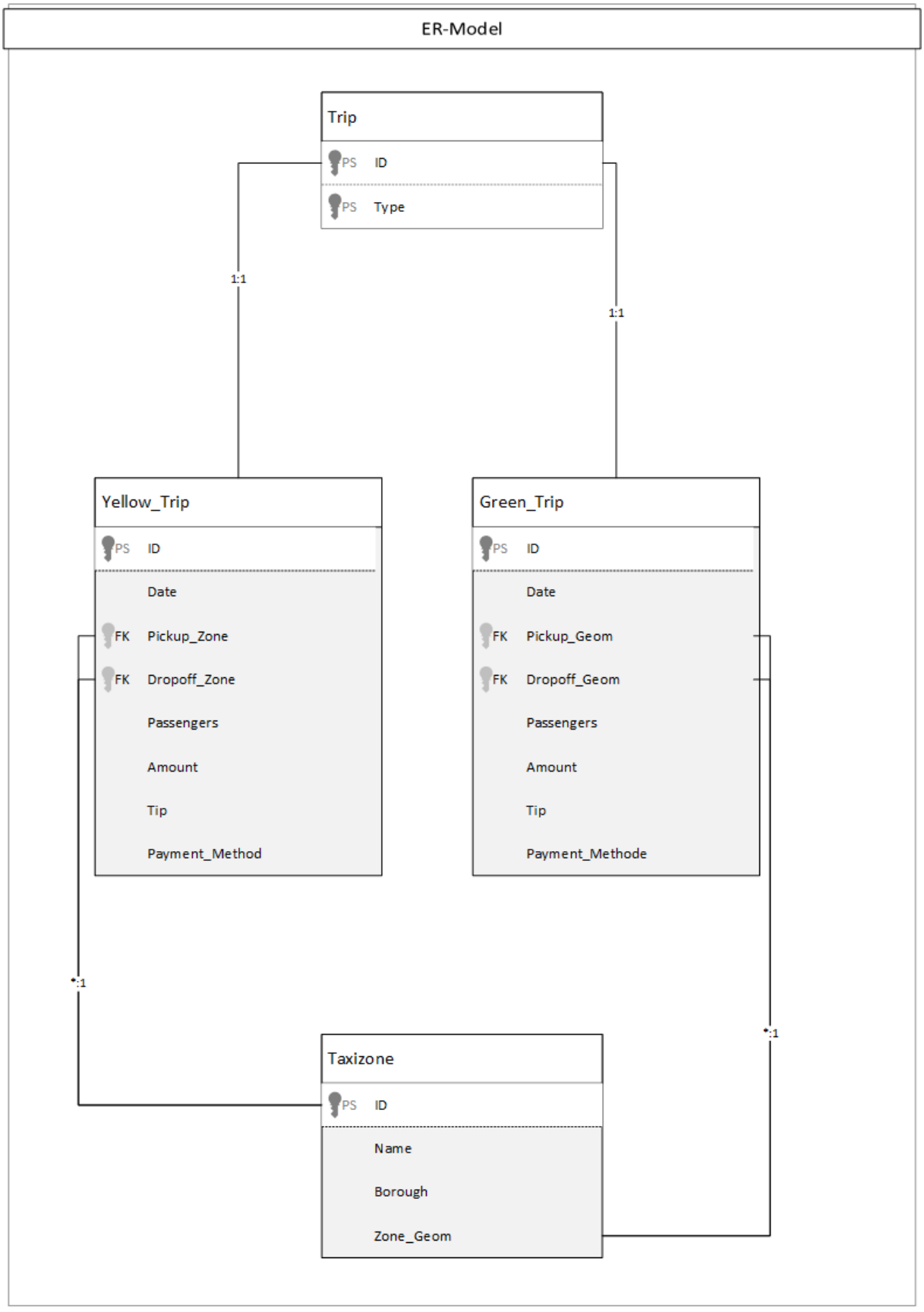


Abbildung 1: ER-Modell

In der Basisdatenbank wird in der Tabelle **Trip** auf die Tabellen **Yellow_Trip** und **Green_Trip**. Je nach Tabelle bekommt **Trip** ein Attribut *Type*.

Bei den Fahrten der gelben Taxis wird der die Zonen ID des Aufnahme- und Zielortes mit angegeben. Hier kann direkt auf die Tabelle **Taxizone** referenziert werden.

Bei den Fahrten der grünen Taxis werden leider nur die Koordinaten des Aufnahme- und des Zielortes angegeben. Hier können die Koordinaten dem Attribut *Zone_Geom* zugeordnet werden, welches einen größeren geometrischen Bereich abdeckt.

Mehrdimensionales Datenbankschema

Als mehrdimensionales Datenbankschema wird das Starschema benutzt. Hierbei gibt es eine Faktentabelle und mehrere Dimensionstabellen, die allerdings nicht weiter unterteilt werden.

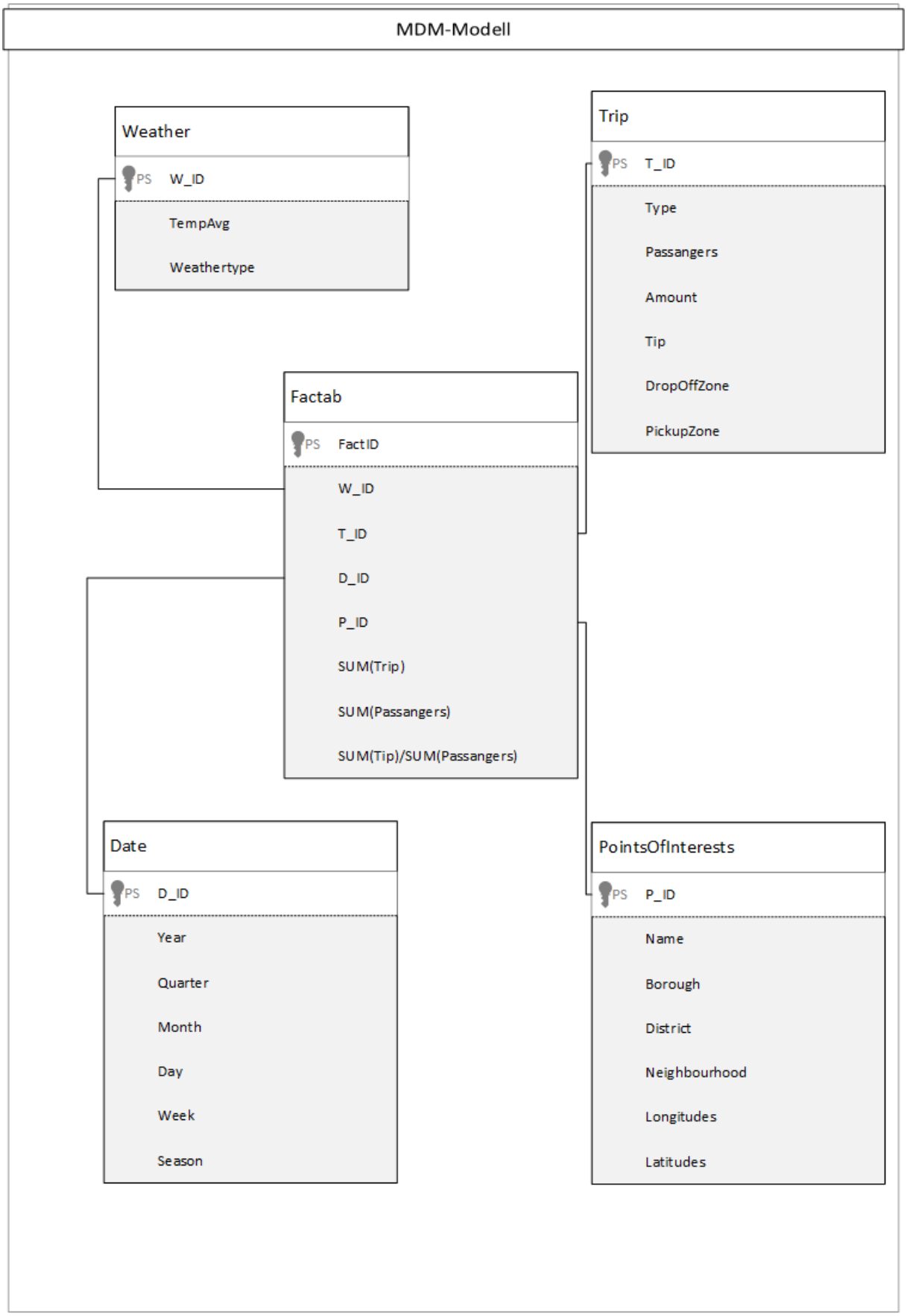


Abbildung 2: MDM-Schema

Die Tabelle **Facttab** enthält die Fremdschlüssel zu den anderen Dimensionen, sowie die Maßzahlen.

Die Dimensionen enthalten alle in der eigenen Dimensionstabelle Ihre Hierarchiestufen. Hier werden keine extra Tabellen angelegt, um über Tabellen auf die verschiedenen Hierarchiestufen zuzugreifen, wie es beim Snowflake-Schema der Fall gewesen wäre.