Meilenstein 1 – Datenimport und DWH

Ziel ist das erste Einlesen von Daten im csv Format in eine relationale Datenbank, sowie das Bilden von Abfragen zu diesen Daten. Zudem konstruieren Sie ein erstes Data Warehouse.

Aufgabe 1

Installieren Sie lokal auf Ihrem Rechner eine PostgreSQL Datenbank Version 11.

Die Datenbank sowie eine Installationsanleitung finden Sie unter:

https://www.postgresql.org

Starten Sie die Datenbank und verwenden Sie ein Administrationswerkzeug Ihrer Wahl.

a) Erzeugen Sie eine Datenbank mit dem Namen „bi“ in UTF-8 Enkodierung.
b) Legen Sie einen Benutzer „bi“ an, der alle Rechte an der Datenbank besitzt. Recherchieren Sie online das hierfür notwendige Kommando. (Tipp: GRANT ALL PRIVILEGES …)

Den Benutzer „bi“ verwenden wir für alle folgenden Aufgaben.

Aufgabe 2

Daten einlesen. Sie erhalten den Datensatz shoes.csv eines Vergleichsportals, welcher Daten im Internet über Schuhverkäufe sammelt. Leider haben Sie neben den Daten keine weiteren Informationen erhalten.
Lesen Sie die Daten in die Datenbank „bi“ vollständig ein. Wählen Sie die Struktur so, dass Sie die Fragen aus Aufgabe 3) gut beantworten können.

Dokumentieren Sie Ihre Vorgehensweise.

Hinweis: Es empfiehlt sich die Fragen aus Aufgabe 3) vorher anzuschauen und die Struktur entsprechend zur Beantwortung der geforderten Fragen zu wählen.

Aufgabe 3

Abfragen. Beantworten Sie zu den eingelesenen Daten die folgenden Fragen, wobei Sie Dubletten nicht berücksichtigen müssen:

a) Welche Schuhmarke ist am häufigsten vertreten? Finden Sie die Top 5 absteigend sortiert.
b) Was ist das größte und geringste Gewicht der Schuhe?

c) Welche Zustände (condition) der Schuhe gibt es im Datensatz?

d) Wie viele Schuhe gibt es pro Zustand (condition)?

e) Welche Währungen kommen vor?

f) Wie häufig kommen die unterschiedlichen Währungen vor?

g) Welche Schuhmarke ist im Durchschnitt die teuerste bei der Währung USD? (Es soll ausschließlich die Währung USD betrachtet werden.)

h) Bei welchen Plattformen (wie Walmart) wurden die meisten Preise entdeckt? Zeigen Sie die Top 5.

Geben Sie …

1) die SQL Statements für die möglichen Vorbereitungen,

2) das von Ihnen verwendete SQL Statement für die eigentliche Abfrage sowie

3) die Ergebnisse des SQL Statements

ab.

Woche 2

Aufgabe 4

Erste Analyse. Was ist wahrscheinlich das Geschäftsmodell des Vergleichsportals? (Anders gefragt: Wie verdient das Vergleichsportal wahrscheinlich Geld?)

Aufgabe 5

Analyse: Fragestellungen. Das Management ist an der folgenden Fragestellung interessiert: „Bei welcher Schuhmarke gibt es die größte Preisspanne?“ Nennen Sie eine weitere Fragestellung, die das Management des Vergleichsportals zu den Daten wahrscheinlich stellt.

Aufgabe 6

Data-Warehouse. In Zukunft ist geplant, dass jede Stunde ein neuer Datensatz (shoes.csv) in diesem Umfang eintrifft. Ein Data Warehouse soll die Daten zentral speichern. Verwenden Sie aus dem erhaltenen Datensatz die folgenden Attribute, die Sie auf die Tabellen in einem Star Schema aufteilen:
id, brand, manufacturer, manufacturernumber, name, prices_amountmin, prices_amountmax, prices_currency, prices_condition, colors, categories, dateadded, dateupdated, imageurls, merchants_name, weight
Ignorieren Sie die übrigen Attribute.

a) Modellieren Sie in Form eines ER-Diagramms nach Chen ein Data Warehouse „dwh_shoes“ nach dem Star Schema, das die Informationen strukturiert speichert.

b) Erzeugen Sie die Relationen in Relationenschreibweise:
Relationsname (Attr1, Attr2, …). Markieren Sie Primärschlüssel durch Unterstreichung und Fremdschlüssel mit gestrichelter Unterstreichung. Verwenden Sie, wenn möglich, bestehende Schlüssel aus den vorliegenden Daten für die Dimensionids.

c) Wie können die in Aufgabe 5 gestellten Fragen des Managements durch das Data Warehouse beantwortet werden? Nennen Sie das SQL-Statement oder eine Beschreibung der Abfrage.
Sollte die Beantwortung der Frage nicht möglich sein, so beantworten Sie die Frage: Was müsste sich an Ihrem Entwurf verändern, damit die Frage des Managements beantwortet werden kann?

Aufgabe 7

Beschreiben Sie, wie das von Ihnen gewählte Modell die Eigenschaften nach Inmon erfüllt.

Aufgabe 8

Erzeugen Sie die Struktur des Data Warehouse des „dwh_shoes“ in SQL.

Aufgabe 9 (optional)

Befüllen Sie das entwickelte DWH mit den Daten der shoes.csv. Befüllen Sie die die Faktentabellen sowie eine von Ihnen gewählte Dimensionstabelle. Verwenden Sie eine Programmiersprache Ihrer Wahl.

Aufgabe

Abgabe. Laden Sie Ihre Ergebnisse als ZIP in ILIAS im Abgabeordner des Meilensteins hoch. Der Name des ZIPs soll sein:

<Nachname1>_<Nachname2>_<Meilensteinnummer>.zip

Das ZIP soll alle Ergebnisse des Meilensteins (Antworten zu Freitextaufgaben, Grafiken, Quelltexte, …) enthalten. Fügen Sie zusätzlich ein README ein, in der Sie Vor-, Nachname und Matrikelnummer der beteiligten Teammitglieder hinterlegen.