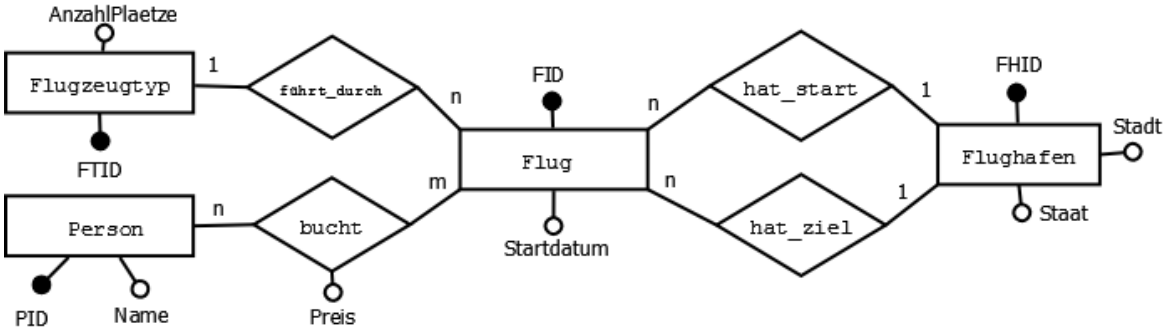
Aufgabenblatt 3 (10 Portfoliopunkte)

Ausgabe: 16. Juni 2015

Abgabe: Sonntag, 12. Juli 2015, 23:55 Uhr auf ISIS
Beachtet ggf. auch die Abgabemodalitäten auf ISIS

Aufgabe 1: SQL (3,5 Punkte)

Gegeben sei das folgende ERD und das dazugehörige Relationen Schema:



Relationales Schema:
Flugzeugtyp(FTID, AnzahlPlaetze)
Person(PID, Name)
Flughafen(FHID, Stadt, Staat)
Flug(FID, Startdatum, Flugzeugtyp → Flugzeugtyp, Startflughafen → Flughafen, Zielflughafen → Flughafen)
Bucht(PID → Person, FID → Flug, Preis)

Aufgabe:

Entwerfen Sie SQL-Anfragen (basierend auf dem obigen Schema), um die folgenden Fragen zu beantworten:

1. Was ist der meistgenutzte Flugzeugtyp (FTID, AnzahlFluege) für Flüge? (0,5 Punkte)
2. Wie lauten die Personen (PID, Name), die nach Barcelona ODER Rom geflogen sind? (0,25 Punkte)
3. Welche Person (PID, Name, Preis) hat für den Flug mit der Flugnummer (FID) '*10*' den billigsten Preis gezahlt? (0,5 Punkte)
4. Welche Personen (PID, Name) sind nie nach Spanien geflogen? (0,5 Punkte)
   Achtung: Personen, die nie einen Flug gebucht haben, sollen nicht in die Ergebnissrelation!
5. Welche Personen (PID, Name) sind nach Rom UND Athen geflogen? (0,25 Punkte)
6. Wie viele Flüge gab es im Jahr 2015 nach Athen und wie viele Personen sind dorthin geflogen? (0,5 Punkte)

7. Wie lauten die Namen aller Personen, die mindestens einmal in jedem Staat der Datenbank gelandet sind? (0,5 Punkte)
8. Welche war die beliebteste Stadt der Deutschen (Staat des Startflughafens = Deutschland) im April 2015? (0,5 Punkte)

Für diese Aufgabe ist eine .sql Datei abzugeben (**KEINE** .pdf-Datei!!!)

Aufgabe 2: OLAP und OLTP (1 Punkt)

Das Unternehmen Compuglobalhypermeganet ist ein internationaler Konzern, der sich auf den Verkauf von Pudding, Mehl, ISDN Kabeln und Katzenfutter spezialisiert hat. Das Unternehmen verfügt weltweit über mehr als 10.000 Filialen und ist in die vier Abteilungen „*Pudding*“, „*Mehl*“, „*ISDN Kabel*“, und „*Katzenfutter*“ unterteilt. Angespornt vom Erfolg der Konkurrenz möchte der Junior Vice President von Compuglobalhypermeganet nun ein Gutscheinsystem für seine Kunden einführen. Er hat dazu die folgenden vier Anforderungen definiert:

1. Kunden können Gutscheine mit definierten Guthaben (10€, 20€, 50€ oder 100€) in den Filialen und auf der Webseite erwerben. Gekaufte Gutscheine können in den Filialen und auf der Webseite eingelöst werden. Dem Kunden wird dazu der Wert des Gutscheins für den aktuellen Einkauf gutgeschrieben und der Gutschein wird als eingelöst markiert. Etwaiges Restguthaben verfällt und wird nicht ausgezahlt. (0, 25 Punkte)

2. Um den internen Wettbewerb zwischen den Abteilungen von Compuglobalhypermeganet anzuheizen, werden am Ende des Jahres die Filiale sowie die Abteilung mit den meisten und den wenigsten Gutscheineinlösungen in den einzelnen Monaten bestimmt. Zudem wird wöchentlich der Verkäufer mit den meisten Gutscheinverkäufen prämiert. (0,25 Punkte)

3. Für jeden zehnten gekauften und jeden fünften eingelösten Gutschein wird dem Kunden die Teilnahme an einem Gewinnspiel angeboten. Hierfür muss der Kunde einige statische Informationen preisgeben (Postleitzahl, Alter, Einkommen, Familienstand, Anzahl der Kinder, Blutgruppe, Erbkrankheiten), die vom System gespeichert werden. (0,25 Punkte)

4. Basierend auf den gesammelten statistischen Informationen über Kunden des Gutscheinsystems soll analysiert werden, wie für welche Kundenkreise das Gutscheinsystem idealerweise beworben werden soll. (0,25 Punkte)

Beschreiben Sie für diese Anforderungen jeweils, ob die benötigten Anfragen eher in das Gebiet OLTP oder OLAP fallen. Erläutern Sie ihre Entscheidung kurz (max. 1 Satz pro Anforderung).

Aufgabe 3: Mehrdimensionale Modellierung (1 Punkt)

Entwerfen Sie ein Stern-Schema (kein Schneeflocken-Schema) das für die analytischen Probleme aus Aufgabe 2 verwendet werden kann. Dabei soll die Faktentabelle *Transaktionen* sowohl Verkäufe als auch Einlösungen von Gutscheinen beinhalten. Beachten Sie, dass alle wichtigen Attribute und Dimensionen für die Aufgaben 2 und 4 in Ihrem Schema vorkommen, und dass Primär- sowie Fremdschlüssel klar gekennzeichnet sind.

Hinweis: Um zwischen Gutscheinverkäufen und –Einlösungen in der Faktentabelle unterscheiden zu können, können optionale Dimensionen verwendet werden. Zum Beispiel könnte die Dimension *Einkauf* nur für jene Transaktionen referenziert werden, bei denen ein Gutschein eingelöst wurde. Die Unterscheidung zwischen Einlösung und Verkauf kann dann z.B. am fehlenden (=NULL) Fremdschlüssel für *Einkauf* festgemacht werden.

Aufgabe 4: Data Warehousing Anfragen (1 Punkt)

Entwerfen Sie SQL Anfragen (basierend auf Ihrem Schema für Aufgabe 3), um die folgenden Fragen zu beantworten:

1. Wie viele Gutscheine wurden insgesamt verkauft? (0,25 Punkte)

2. Wieviel hat Compuglobalmegahypernet an den Gutscheinen verdient, d.h. wieviel Geld haben Kunden insgesamt verschenkt, indem sie Restguthaben von Gutscheinen verfallen lassen haben (Wert von Gutschein > Kosten für Einkauf). (0,25 Punkte)

3. Geben sie für jeden Verkäufer (Name) an, welchen Gesamtwert an Gutscheinen er in der letzten Woche verkauft hat. Sortieren sie das Ergebnis absteigend nach dem Gesamtwert. Gehen Sie davon aus, dass eine Funktion TODAY() existiert, die den aktuellen Tag zurückgibt. (0,25 Punkte)

4. Geben Sie die insgesamt eingelösten Gutscheine für alle Kombinationen aus Abteilung, PLZ des Kunden und Jahr zurück (Dicing nach Abteilung, Kunden PLZ sowie Jahr, Slicing nach eingelösten Gutscheinen). (0,25 Punkte)

Aufgabe 5: Klassifikation von Daten (1 Punkt)

Betrachten sie das folgende Datenschema einer Kreditagentur.

ID	Alter	Nationalität	Verheiratet?	Anz. Kinder	Beruf	Einkommen	Credit Score
0	46	Deutsch	Ja	2	Verkäufer	Niedrig	126,4
1	19	Deutsch	Nein	1	NULL	Keines	78,1
2	22	China	Nein	0	Ingenieur	Sehr Hoch	254,1
3	34	Argentinien	Nein	1	Lehrer	Hoch	105
…	…	…	…	…	…	…	…

Klassifizieren Sie zunächst die acht Attribute danach, ob es sich eine um *quantitative* oder *qualitative* Variable handelt. Klassifizieren Sie anschließend die quantitativen Attribute in *kontinuierlich* und *diskret*, sowie die qualitativen Attribute in *nominal*, *ordinal* und *binär*. Begründen sie ihre Entscheidungen kurz (max. 1. Satz pro Attribut).

Aufgabe 6: Zentrale Statistiken in SQL (1 Punkte)

In der Vorlesung wurden die sogenannten Zentralwerte der Statistik (Durchschnitt, Median und Modus), besprochen. Entwerfen Sie SQL Anfragen, die für das Attribut *Alter* aus Aufgabe 5 diese Werte bestimmen. Sie können davon ausgehen, dass a) *Alter* keine NULL Werte enthält b) dass die Tabelle eine ungerade Anzahl an Zeilen hat und c) dass der Modus eindeutig ist.

Achtung: Es dürfen nur die Standard Aggregationsfunktionen aus dem SQL-99 Standard verwendet werden. Insbesondere bedeutet dies, dass Funktionen wie MEDIAN() oder MODE() die einige Datenbanksysteme anbieten nicht verwendet werden dürfen.

Hinweis: Für einige der Aufgaben müssen Sie gezielt bestimmte Werte aus einem SQL Ergebnis zurückgeben. Hierzu können sie das Konstrukt LIMIT x OFFSET y verwenden, welches die nächsten X Zeilen, beginnend ab Zeile y+1, aus dem Ergebnis zurückgibt. Zum Beispiel würde

```
SELECT * FROM test LIMIT 2 OFFSET 4
```

die fünfte und sechste Zeile der Tabelle *test* zurückgeben. LIMIT x OFFSET y kann für beliebige Anfragen verwendet werden. Für die Werte x und y können Unteranfragen verwendet werden. Zum Beispiel würde

```
SELECT x FROM test LIMIT 1 OFFSET (SELECT count(*) - 2 FROM test)
```

den vorletzten Wert der Tabelle *test* zurückgeben.

Aufgabe 7: Explorative Datenanalyse (1,5 Punkte)

In der bereitgestellten DB2 Datenbank finden sie die Tabelle *housing* mit dem folgenden Schema:

```
CREATE TABLE housing(
crim DECIMAL(8,5), zn DECIMAL(8,5), indus DECIMAL(8,5), chas SMALLINT,
nox DECIMAL(8,5), rm DECIMAL(8,5), age DECIMAL(8,5), dis DECIMAL(8,5),
rad DECIMAL(8,5), tax DECIMAL(8,5), ptratio DECIMAL(8,5),
b DECIMAL(8,5), lstat DECIMAL(8,5), medv DECIMAL(8,5));
```

Die Tabelle enthält statistische Informationen über die Vororte von Boston im Jahr 1978. Detaillierte Informationen über die Bedeutung der einzelnen Attribute können auf der folgenden Webseite gefunden werden: https://archive.ics.uci.edu/ml/datasets/Housing

In dieser Aufgabe werden wir einige explorative Analysen der Daten in der Tabelle *housing* vornehmen.

1. Die Variable NOX gibt die gemessene Konzentration an Stickoxiden in pptm (Anteil pro 10 Millionen Volumenanteilen) in der Luft für den jeweiligen Vorort an. Diese Variable kann daher als Proxy für die Luftqualität in den Vororten verwendet werden. Geben Sie bitte die Basisstatistiken (Minimum, Maximum, Mittelwert, Median, Standardabweichung) für NOX an. (0,25 Punkte)

2. Der von der EPA (amerikanische Umweltbehörde) festgelegte Grenzwert für die Konzentration von NOX liegt bei 0.53 pptm. Liegt die Konzentration von Stickoxiden über diesem Grenzwert können langfristig gesundheitliche Probleme auftreten. Für eine Studie möchten wir nun bestimmen ob NOX auch einen verhaltensstörenden Einfluss hat. Berechnen Sie hierzu bitte jeweils die durchschnittliche Kriminalitätsrate (CRIM) für Vororte, die den Grenzwert über- bzw. unterschreiten. (0,125 Punkte)

3. Die letzte Analyse zeigt deutlich, dass die Kriminalitätsrate in Vororten mit einer überkritischen Konzentration an Stickoxiden ca. um den Faktor 50 ~~25~~ höher ist als in Vororten mit unkritischen Konzentrationen. Ist dies ein Beleg für die These das hohe NOX Konzentrationen verhaltensändern wirken können? Warum? Warum nicht? (0,125 Punkte)

4. Die Variable MEDV gibt den mittleren Wert von Häusern für den jeweiligen Vorort in Tausend Dollar an. Um einen guten Überblick über die Verteilung von MEDV zu bekommen, plotten Sie bitte ein Histogramm von MEDV. Was können Sie aus diesem Histogramm ablesen? (0,25 Punkte)

5. Wir betrachten nun ob wir aus den Daten Zusammenhänge zwischen anderen Variablen und dem mittleren Wert der Häuser auslesen können. Berechnen sie dazu den Korrelationskoeffizienten (Pearson) zwischen MEDV und allen anderen Variablen. (0,125 Punkte)

6. MEDV zeigt eine starke negative Korrelation zu LSTAT, welche den jeweiligen Bevölkerungsanteil einkommensschwacher Personen im Vorort angibt. Was genau bedeutet diese Korrelation? (1 Satz) Visualisieren Sie diesen Zusammenhang mit einem Scatterplot. (0,125 Punkte)

7. Betrachten Sie nun bitte den Zusammenhang zwischen LSTAT, CRIM & NOX (insbesondere für NOX über/unter dem Grenzwert). Welche Zusammenhänge können Sie erkennen? Inwiefern belegt/wiederlegt dies ihre These aus Aufgabenteil 2 & 3? Belegen Sie ihre Erkenntnisse mit entsprechenden Datenanalysen und/oder Plots. (0,5 Punkte)

Hinweis: Diese Aufgabe lässt sich am Einfachsten bearbeiten, indem Sie die Daten aus DB2 exportieren, und anschließend in ein Datenanalysetool ihrer Wahl (z.B. Excel, R, Octave, NumPy) importieren (z.B. mittels CSV Import), um die gewünschten Analysen durchzuführen.