Meilenstein 3 – ETL & DWH

BIC erhält einen neuen Kunden: die **ApoDeals GmbH**. Die ApoDeals GmbH ist ein Startup aus dem Aachener Raum, die eine Onlineplattform für den Verkauf von Medikamenten für Apotheken zur Verfügung stellt: https://www.apodeals.de. (Letzter Zugriff 1.3.2021 – Leider heute nicht mehr verfügbar ☹)

Geschäftsmodell: Apotheken können sich bei ApoDeals als Händler (vendor) registrieren und Artikel anbieten. ApoDeals stellt die Artikel im Internet Kunden vor. Ein Kunde stellt seinen Warenkorb zusammen und schickt eine Bestellung (order) ab. Dann bekommen alle Händler den Auftrag für ihre an der Bestellung beteiligten Artikel (order_assignment & order_assignmend_item). Die Händler verschicken die Waren. Sobald der Kunde die Ware an ApoDeals bezahlt hat und die Waren durch den Händler verschickt wurden, leitet ApoDeals das Geld an die Händler weiter. Für die Vermittlung sowie den Geldtransfer erhält ApoDeals eine Provision. ApoDeals möchte eine Auswertungsdatenbank (Data Warehouse) erhalten, um darüber später Abfragen durchführen zu können.

Sie erhalten den Datensatz ApoDeals.zip.

Aufgabe 1

Analyse. Analysieren Sie den Datensatz „ApoDeals.zip“. Zeichnen Sie ein ER-D nach Chen zu der Datenstruktur.

- Markieren Sie Entitätstypen durch Unterstreichung des Namens, die referenziert werden, aber nicht im Datensatz enthalten sind.
- Verzichten Sie bei dem ER-D auf die Darstellung von gewöhnlichen Attributen (wie Name, Reason, Amount, …).
- Zeichnen Sie Schlüssel- und Fremdschlüsselattribute ein.
- Beschreiben Sie kurz pro Tabelle, was diese für Daten beinhaltet.

Aufgabe 2

ETL-Prozess. Einübung.

a) Legen Sie eine Datenbank mit einer Tabelle an, die die Daten der ad_order.csv importieren kann. Importieren Sie die Daten der ad_order.csv mit KNIME in die erstellte Tabelle. Setzen Sie den Primärschlüssel. Lesen Sie die Datensätze der ad_order.csv in die Tabelle einmalig ein. Überführen Sie dabei
    - Umrechnung von Euro nach Cent: Fließkommawerte zu Ganzzahlen (Integer), wobei keine Informationen verloren gehen sollen
    - Datum- (und Uhrzeitwerte) zu TIMESTAMP
    - „NULL“ (Strings) zu NULL in der Datenbank.

b) Erweitern Sie den Workflow so, dass ein wiederholtes Einlesen der Datei die bestehenden Daten in der Datenbank aktualisiert.

Aufgabe 3

Die Geschäftsleitung der ApoDeals GmbH ist interessiert an der Beantwortung der folgenden Fragestellungen zu dem Datensatz ApoDeals.zip:

- Welcher Kunde hat welches Produkt zu welchem Zeitpunkt (Genauigkeit: Stunde) von welchem Händler in welcher Bestellung gekauft?
- Mit welchen und wie vielen Produkten wurde wie viel Umsatz erzielt?

Konzipieren Sie ein DWH nach dem Stern-Schema, um die Antworten auf diese Fragen zu liefern. Verwenden Sie bestehende Schlüssel aus den vorliegenden Daten für die Dimensionsids.

Erstellen Sie …

a) das ER-D sowie
b) das Datenbank-Schema in SQL in der Datenbank. Wählen Sie für jede Relation, die Sie erzeugen, maximal vier Attribute aus, die Sie für die Beantwortung der Fragestellung für relevant erachten.

Aufgabe 4

ETL-Prozess. Erstellen Sie einen ETL-Prozess mit KNIME, der die notwendigen Daten aus apodeals.zip **einmalig** in das Data Warehouse lädt.

Hinweis: Löschen Sie vor dem erneuten Laden der Daten die aktuellen Daten im DWH (manuell oder via KNIME). Wir betrachten zunächst den einfacheren Fall, wo Daten nicht weiter dem DWH hinzugefügt werden.

Woche 2

Aufgabe 5

Abfragen. Da nun das Data Warehouse der ApoDeals GmbH erstellt und mit Daten befüllt ist, können wir die Daten über SQL abfragen. Erzeugen Sie SQL-Abfragen für die Fragen unter Aufgabe 3). Erzeugen Sie so viele SQL-Abfragen, wie Sie für die Beantwortung benötigen.
Geben Sie als Ergebnisse sowohl die SQL Abfragen als auch die SQL Ergebnisse ab.

Aufgabe 6

Schema. Überführen Sie das DWH „dwh_apodeals“ in ein eigenes Schema „ad_dwh“.

Aufgabe 7

Ableitung von Tabellen. Bei der Erstellung von Tabellen sollen bei den Attributen Redundanzen vermieden werden. Fügen Sie allen Tabellen des DWH dwh_apodeals die folgenden Spalten hinzu, wobei Sie eine Ableitung von Tabellen verwenden:

- insertdate (TIMESTAMP without timezone), Zeitpunkt des Anlegens des Datensatzes.
- insertsource (VARCHAR (200), NOT NULL), Name der Quelle, aus der die Daten stammen.

Setzen Sie bei Daten aus der Quelle „csv“ den insertsource entsprechend auf „csv“.

Hinweis: Zur Lösung müssen Sie die Struktur in der Datenbank löschen und mit Ableitungen neu anlegen.

Aufgabe 8

Wiederholen Sie jetzt, nach den Anpassungen in Aufgabe 6 und Aufgabe 7 den ETL-Prozess aus Aufgabe 4, wobei Sie notwendige Anpassungen an den Nodes in KNIME

a) für ein Schreiben in ein Schema und
b) das Setzen von insertdate und insertsource

durchführen.

Aufgabe 9

Recherche. Recherchieren Sie, …

a) wie in PostgreSQL mit Datumsfunktionen Abfragen durchgeführt werden können. Geben Sie mindestens fünf Beispiele, die Verwendungen dieser Datumsfunktionen zeigen.
b) wo der Unterschied zwischen einer VIEW und einer MATERIALIZED besteht. Wie wird eine MATERIALIZED VIEW erzeugt, aktualisiert und gelöscht?

Aufgabe 10

Views. Bei der Apodeals GmbH sollen nicht alle Mitarbeitenden auf alle Daten im DWH Zugriff erhalten. So sind die Daten des laufenden Monats eher unsicher, nicht abgeschlossen und müssen durch Neueintragungen häufig korrigiert werden. Auf die Daten des aktuellen Monats wird sehr häufig durch verschiedene Abteilungen zugegriffen.

Stellen Sie …

a) alle Daten aller abgeschlossenen Monate als eigene Materialized View bereit und

b) alle Daten des aktuellen Monats als eigene View bereit.

c) Der Kunde fragt sich, warum unterschiedliche Views verwendet werden. Argumentieren Sie und erklären Sie dem Kunden, warum sich einmal eine Materialized View und einmal eine View zur Lösung dieser Problemstellung anbietet.

Aufgabe 11

Berechtigung. Stellen Sie die Zugriffsberechtigung des DWH „dwh_apodeals“ für unterschiedliche Benutzer:innen ein. Es soll die folgenden Benutzer:innen geben:

- reporter_current: Fragt Daten des aktuellen Monats ab.
- reporter_all: Fragt Daten aller vorherigen Monate ab.
- updater: Befüllt das DWH.
- maintenance: Führt Backups und ggfls. Optimierungen durch.

a) Erstellen Sie eine Berechtigungsmatrix.
b) Legen Sie die notwendigen Benutzer:innen und Berechtigungen in der Datenbank an.
c) (optional) Schreiben Sie Statements für die Überprüfung, ob die Rechte korrekt gesetzt sind.

Aufgabe 12 (optional)

PostgreSQL als Data Warehouse. Als ein weiteres wertvolles Feature für ein Data Warehouse verwendet PostgreSQL intern Multiversion Concurrency Control (MVCC) für die Transaktionsverwaltung.

a) Recherchieren Sie, was MVCC bedeutet.

b) Wie funktioniert MVCC?

c) Welcher Vorteil ergibt sich dadurch für ein Data Warehouse?

Aufgabe

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

<TeamNr>_<Meilensteinnummer>.zip

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