Meilenstein 2 – ETL

Ziel ist das Kennenlernen und Einüben von ETL-Prozessen, mit denen Daten aus unterschiedlichen Quellen in ein Data Warehouse geladen werden können.

Aufgabe 1

Recherche. Vergleichen Sie ETL-Software. Finden Sie 5 Vergleichskriterien und vergleichen Sie 3 Produkte anhand dieser Kriterien. Verlassen Sie sich bei Ihrer Recherche auf Onlinequellen und schauen Sie sich die Werkzeuge nicht selbst an. Dokumentieren Sie die verwendeten Onlinequellen.

Aufgabe 2

Installation. Installieren und starten Sie KNIME.

Aufgabe 3

Lesen Sie in KNIME die sales1.csv[1] ein und beantworten Sie mit Hilfe von KNIME die folgenden Fragen zu dem Datensatz:

a) Wie viele Datensätze haben die order_priority = „H“ (High)
b) Wie viele Einheiten (units_sold) wurden in Asien (region) in Summe verkauft?
c) Wie viele Einheiten (units_sold) wurden in Asien und Europa (region) in Summe verkauft?
d) Wie viele Einheiten (units_sold) wurden im Jahr 2013 Online (sales_channel) verkauft?
e) Wie viele Bestellungen gab es in 2013 (order_date), die erst in 2014 (ship_date) verschickt wurden?
f) Überprüfen Sie, ob die Spalte „total_profit“ einen Berechnungsfehler enthält.

Exportieren Sie für die Abgabe den Workflow in KNIME (ohne Verarbeitungsdaten).

Aufgabe 4

Finden Sie die Funktionsweise der folgenden Komponenten in KNIME heraus, und geben Sie jeweils ein Beispiel inklusive Beispieldaten für:

- Table Creator[2]
- Sorter
- Rule-based Row Filter
- Concatenate

[1] Originäre Quelle: https://data.world/bobmajor/sales; Inhalte verändert: bereinigt und reduziert.
[2] Der Table Creator ist besonders geeignet um Daten für Beispiele zu konstruieren

Aufgabe 5

Lesen Sie die bereits bekannte shoes.csv in KNIME ein und beantworten Sie mit KNIME die folgenden Fragen, wobei Sie doppelte Einträge eliminieren (doppelte Einträge erkennen Sie an der gleichen id):

- F1: Welche Schuhfarbe kommt am häufigsten vor? Ignorieren Sie unbekannte Farben.
- F2: Wie häufig kommt die Schuhfarbe vor?

Exportieren Sie für die Abgabe den Workflow in KNIME (ohne Verarbeitungsdaten).

Woche 2

Aufgabe 6

Sie erhalten neben der shoes.csv die shoes_sizes.csv, die auch Schuhgrößen und Versandkosten zu den einzelnen Einträgen enthält. Beantworten Sie die folgenden Fragen mit KNIME:

- F3: Bei welcher Marke gibt es am häufigsten bedingungslos freien Versand?
- Aus welchen Quellen stammen die meisten Preise? Nennen Sie die Top 5 und geben Sie nur den Domain- und Top-Level-Domainname aus. (z.B. ebay.com)
- (optional) Normieren Sie die Schuhgrößen auf UK (7,7.5,…), wobei Sie eine online verfügbare Schuhgrößentabelle verwenden. Welche Schuhgröße kommt am häufigsten vor?

Aufgabe 7

Schreiben Sie Ihre Ergebnisse unter Verwendung von KNIME aus Aufgabe 5 und 6 in eine CSV und Excel Datei mit den Namen output.csv und output.xls. Die Datei output.csv soll das folgende Format besitzen:

```
question;answer
F1;<your answer>
F2;<your answer>
F3;<your answer>
```

Nehmen Sie das Präfix bei den Fragen „Fx“ als Eintrag in der ersten Spalte. Setzen Sie in die zweite Spalte den Ergebniswert ein.

Aufgabe 8

Wählen einen zuvor von Ihnen umgesetzten Workflow aus und stellen Sie diesen auf Streaming um, wobei mindestens 4 Nodes am Streaming beteiligt sein sollen.

Aufgabe 9

Startet Sie KNIME von der Kommandozeile. Übergeben Sie als Variable den Namen der zu schreibenden CSV Datei, der vorher konstant „output.csv“ war.

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.