# DB & DWH Leistungsnachweis Aufgaben

Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

## Einführung

Beim Leistungsnachweis im Modul DB & DWH geht es darum, Teile des vermittelten Stoffes anhand eines konkreten Beispiels anzuwenden. Ein Leistungsnachweis ist keine Prüfung. Das Resultat wird lediglich mit „bestanden/nicht bestanden“ bewertet. Ein nicht-bestandener Leistungsnachweis kann maximal einmal wiederholt werden.

Im Modul DB & DWH stehen zwei verschiedene Aufgaben zur Verfügung. Es kann frei entschieden werden welche der beiden Aufgaben gelöst werden soll (späteste Festlegung hierzu muss bis am Montag, 5.11.2018, erfolgen).

## Aufgabe Variante 1 (Teamarbeit, 2er-Teams, Aufwand ≈ 2 Personentage, 3er-Teams mit Zusatzaufgabe ≈ 3 Personentage)

In dieser Aufgabe geht es darum, ein kleines data warehouse (eigentlich einen kleinen data mart) aufzubauen. Dabei müssen sowohl ETL-Probleme gelöst als auch Modellierungsaufgaben und konkrete Auswertungen umgesetzt werden.

### Tools

Zur Umsetzung müssen folgende Tools verwendet werden: Pentaho Data Integration (ETL), MySQL (DWH), Excel oder PowerBI (Auswertungen).

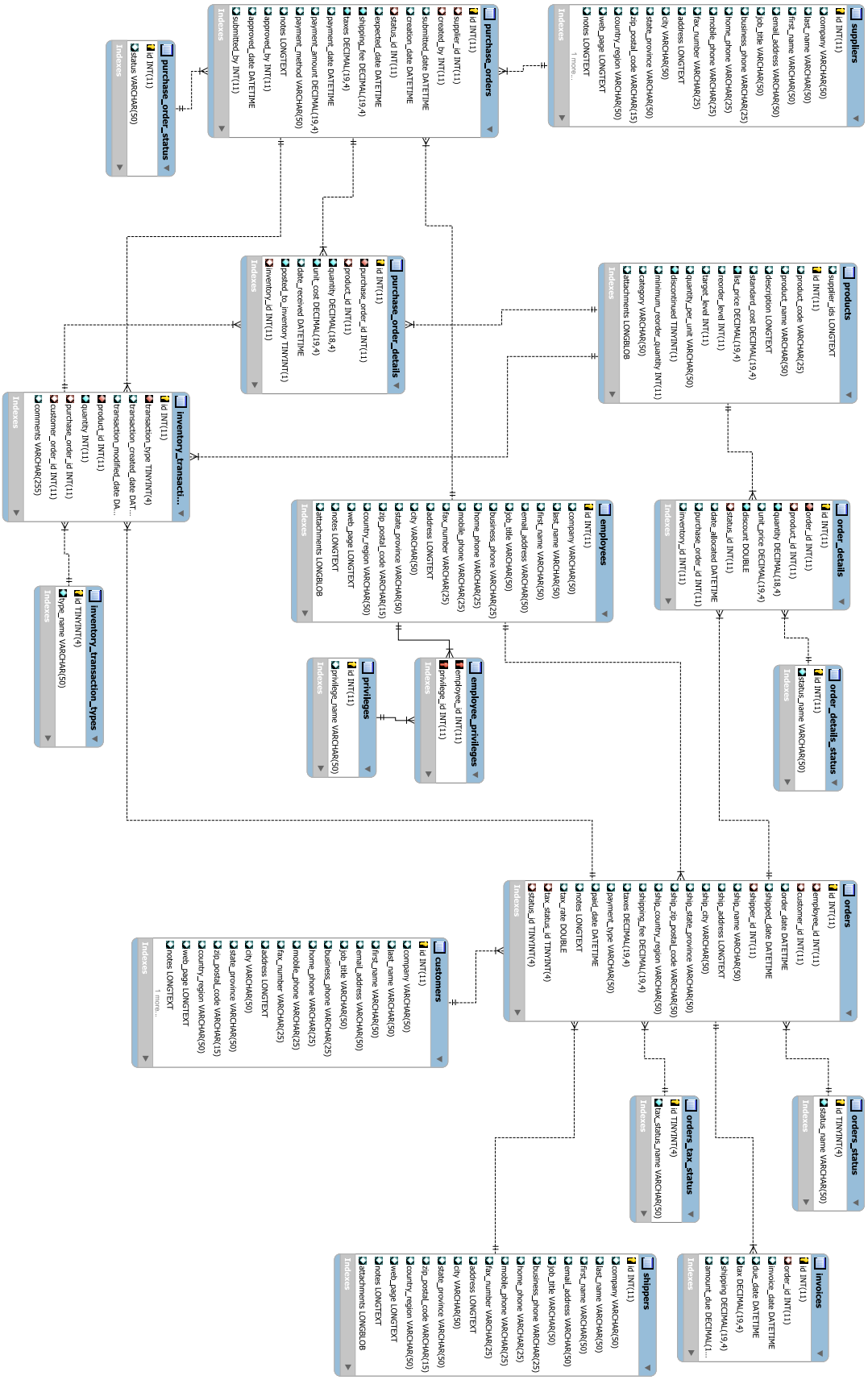
### Ausgangslage

Das Unternehmen „Northwind“ betreibt ein ERP-System (OLTP). Diesem liegt eine relationale Datenbank zugrunde. Aus verschiedenen Gründen ist es jedoch nicht gestattet, für analytische Zwecke direkt auf dieser Datenbank Abfragen zu tätigen. Es soll daher ein eigenständiges DWH aufgebaut werden. Die Daten stammen zwar aus dieser ERP-Datenbank, müssen aber von dort zuerst in .csv-Dateien exportiert werden, die dann via ETL-Prozesse ins DWH gelangen.

### Ausgangsdaten (zu finden auf OLAT, Datei: DB & DWH-Leistungsnachweis Daten.zip)

| | |
|---|---|
| Northwind-Database.sql | Skript, um die ERP-Datenbank zu erzeugen. |
| Northwind-Daten.sql | Skript, um die ERP-Datenbank zu füllen. |
| Northwind-ERM.pdf | ER-Diagramm der ERP-Datenbank (anderer Dialekt als der im Modul besprochene). Siehe nächste Seite. |

Schema der ERP-Datenbank:



**Teilaufgaben 2er-Teams**

- Aufsetzen der ERP-Datenbank (d.h. obige Skripte laufen lassen). Studium des Schemas und vertraut werden mit den Datenbankinhalten.
- Überlegen, welche Auswertungen mit dem DWH ermöglicht werden sollen und festlegen, welche Daten dazu benötigt werden. Das DWH soll Auskunft über das Kaufverhalten der Kunden liefern, es werden also minimal Kunden-, Produkt- und Bestellungsdaten benötigt. Je nach Ihren Auswertungswünschen aber auch noch weitere.
- Exportieren (manuell) der erforderlichen Tabellen in einzelne .csv-Dateien. Dabei können Attribute, die später nicht benötigt werden, bereits weggelassen werden.
- Implementieren einer „staging area".
- Implementieren eines ETL-Prozesse mit PDI, d.h. einlesen der .csv-Dateien in diese staging-area. Durchführen der notwendigen Transformationen (innerhalb oder ausserhalb der staging-area).
- Beheben der folgenden Datenqualitätsprobleme im ETL-Prozess:
  1. Aufgrund eines technischen Problems wurden alle Discounts in der Tabelle `order_details` negativ erfasst. Diese Werte sollen im ETL-Prozess in positive Werte umgewandelt werden.
  2. In der Tabelle `products` wurden einige Produkte irrtümlicherweise doppelt erfasst. Finden Sie heraus, inwiefern sich die Einträge der identischen Produkte unterscheiden. Eliminieren Sie dann alle Duplikate, so dass es pro Produkt genau einen Eintrag in der Zieltabelle gibt (und zwar den, mit der jeweils kleineren id).
- Erstellen eines zu den geplanten Abfragen passenden „multidimensionalen" Schemas. **Dazu muss der in der Vorlesung verwendete ERM-Dialekt benutzt werden.**
- Füllen des implementierten Schemas im ETL-Prozess mit Daten.
- Ausführen von mindestens **drei verschiedenen** selbst definierten Analysen und präsentieren der Ergebnisse.

Welche Analysen Ihnen interessant erscheinen, ist Ihnen überlassen. Ebenso die Auswertungs- und Darstellungsweise. Experimentieren Sie ruhig ein wenig.

**Zusatzaufgabe für 3er-Teams**

Migration der ERP-Datenbank in eine Datenbank, die dem in der Vorlesung behandelten ERM-Dialekt entspricht. Lieferobjekte: Vollständiges ERM-Schema, SQL-Skripte zur Erzeugung dieser neuen Datenbank, SQL-Skripte zum Befüllen dieser neuen Datenbank (d.h. alle notwendigen Skripte zum Erzeugen der neuen Datenbank und migrieren der Daten aus der alten Datenbank). Achtung: Es gibt zusätzliche Tabellen für die Beziehungstypen, die auch gefüllt werden müssen!

| | |
|---|---|
| Erwartete Resultate: | **Bericht** im Umfang von **mindestens zehn Seiten A4** (inkl. Screenshots, Beispiele etc.).<br>**Entwicklungsobjekte**: Skripte, PDI-Transformationen, Installationshinweise, … (alles was nötig ist um die Lösung nachvollziehen und „laufen lassen“ zu können). |
| 3er-Teams zusätzlich: | Vollständiges ERM der migrierten ERP-Datenbank, SQL-Skripte zum Erzeugen und Füllen dieser migrierten Datenbank. |
| Abgabetermin: | Sonntag, 18.11.2018, 1800. Verspätete/unvollständige Abgaben gelten als nicht eingereicht und müssen wiederholt werden. |
| Abgabeart: | **E-Mail** an aebd@zhaw.ch |
| Abgabeformat: | **.pdf** (Bericht)<br>**.zip** (Entwicklungsobjekte) |

## Aufgabe Variante 2 (Einzelarbeit, Aufwand ≈ 1 Personentag)

Analysieren Sie einige OLTP-Anwendungen in Ihrem beruflichen Umfeld und wählen Sie dann eine konkrete Anwendung aus, bei der Sie Datenqualitätsprobleme vermuten oder bereits wissen, dass es solche gibt (zur Not kann die Problemstellung auch aus ihrem privaten Umfeld stammen).

Identifizieren Sie in diesem System (es können auch mehrere sein) konkrete Daten, die Qualitätsprobleme verursachen (könnten).

Beschreiben Sie diese Probleme, am Besten anhand von konkreten Beispielen (ggf. anonymisiert).

Beschreiben Sie dann, welche konkreten «data-profiling»-Verfahren Sie anwenden würden, um den oben beschriebenen Problemen auf die Spur zu kommen.

Überlegen Sie anschliessend, wie die fehlerhaften Daten im Rahmen eines sogenannten «data cleaning»-Prozesses, bereinigt werden könnten.

Diskutieren Sie organisatorische & technische Massnahmen, die geeignet wären, die Probleme dauerhaft zu lösen.

| | |
|---|---|
| Erwartetes Resultat: | **Bericht** im Umfang von **mindestens zehn Seiten A4** (inkl. Abbildungen, Beispiele etc.). |
| Abgabetermin: | Sonntag, 18.11.2018, 1800. Verspätete/unvollständige Abgaben gelten als nicht eingereicht und müssen wiederholt werden. |
| Abgabeart: | **E-Mail** an aebd@zhaw.ch |
| Abgabeformat: | **.pdf** |