Lernzettel

Data Warehousing: Modellierung, ETL-Prozesse, OLAP, Star- und Snowflake-Schema

Universität: Technische Universität Berlin
Kurs/Modul: Informationssysteme und Datenanalyse

Erstellungsdatum: September 19, 2025

Lernzettel: Data Warehousing – Modellierung, ETL-Prozesse, OLAP, Star- und Snowflake-Schema

(1) Modellierung.
Data Warehouse ist eine zentrale analytische Umgebung, die operativen Transaktionen (OLTP) getrennt ist. Ziel: konsistente, historisierte Analysen und schnelle Abfragen über große Datenmengen. Zentrale Strukturen:

- Faktentabellen: Messgrößen/Zahlen (z. B. Umsatz, Stückzahl) mit Fremdschlüsseln zu Dimensionen.
- Dimensionstabellen: beschreibende Attribute (Zeit, Kunde, Produkt, Ort, Region usw.).
- Granularität: Feinste Detailtiefe der Faktentabelle; beeinflusst Speicherbedarf und Abfrageleistung.

(2) Dimensionale Modellierung – Star- vs. Snowflake-Schema.
Star-Schema

- Zentral: Faktentabelle, direkt mit allen Dimensionen verknüpft.
- Dimensionstabellen sind flach (denormalisiert).
- Vorteile: einfache Abfragen, gute Performance durch wenige Joins.
- Nachteile: Redundanzen, weniger Normalisierung, ggf. höherer Wartungsaufwand bei Änderungen.

Snowflake-Schema

- Dimensionstabellen sind normalisiert, Unterteilungen in Teil-Dimensionen möglich.
- Vorteile: geringerer Speicherbedarf, bessere Konsistenz, flexiblere Wartung.
- Nachteile: komplexere Abfragen, mehr Joins, potenziell längere Antwortzeiten.

(3) ETL-Prozesse (Extract, Transform, Load).
Zweck: Daten aus Quellsystemen extrahieren, transformieren (Qualität, Normierung, Historisierung) und in das Data Warehouse laden.

- Staging-Umgebung: temporäre Zwischenspeicherung vor dem Laden.
- Transformation: Formatierung, Harmonisierung, Berechnungen, Kennzahlenbildung.
- Qualitätsmanagement: Dublettenkontrolle, Fehlersuche, Validierung, Logging.
- Slowly Changing Dimensions (SCD): Typ 1, Typ 2, Typ 3.
- Scheduling und Monitoring: regelmäßig, robust, nachvollziehbar.

(4) OLAP und Analytik.
OLAP ermöglicht mehrdimensionale Analyse über Würfel (Cubes) und schnelle Aggregationen.

- Grundoperationen: Slice (Auswahl einer Dimension), Dice (Auswahl mehrerer Dimensionen),
- Drill-Down/Roll-Up (Hierarchieebenen wechseln), Pivot (Achsenneuausrichtung).
- Vorab-Aggregationen (materialisierte Würfel) verbessern die Reaktionszeit.

(5) Star-Schema – Details und Anwendungsfälle.

- Eignung: schnelle, einfache Analysepfade, klare Pfade zu Kennzahlen.
- Typische Kennzahlen (Measures): Umsatz, Kosten, Gewinn, Stückzahlen.
- Typische Dimensionen: Zeit (Tag, Monat, Quartal, Jahr), Produkt, Kunde, Ort, Vertriebskanal.

(6) Snowflake-Schema – Details und Abwägungen.

- Eignet sich, wenn Normalisierung Vorteile bei Konsistenz und Wartung bietet.
- Komplexere Abfragen durch zusätzliche Joins; oft Mischformen mit Star-Teildimensionen.

(7) Architektur und Datenfluss.

- Quellsysteme liefern Rohdaten (Transaktionssysteme).
- ETL-Server/Staging-Umgebungen bereinigen und transformieren.
- Data Warehouse als zentrale analytische Speicherung.
- Data Marts als themenspezifische Unterbereiche.
- Metadata-Management und Data Governance für Transparenz und Qualität.

(8) Data Warehousing vs Transaktionssysteme.

- Schreibrate: OLTP hoch, OLAP lese- bzw. aggregationslastig.
- Historisierung: Data Warehouse speichert historisierte Daten; OLTP typischerweise aktuelle Daten.
- Abfragearten: Transaktionsabfragen vs. analytische Abfragen (Aggregationen, Trends).
- Skalierung: Data-Warehouse-Architekturen oft speziell auf Analytik optimiert (z. B. Marts, MPP-Systeme).

(9) Typische Abfragen – Beispiele.
Beispielabfrage (Star-Schema) – Umsatz nach Monat und Region:

```
SELECT Monat, Region, SUM(Umsatz) AS Umsatzgesamt
FROM FaktUmsatz F
JOIN DimZeit Z ON F.kZeit = Z.kZeit
JOIN DimRegion R ON F.kRegion = R.kRegion
GROUP BY Monat, Region;
```

Beispielabfrage (Snowflake-Schema) – erfordert zusätzliche Joins zu Normalformen.

(10) Praktische Hinweise.

- Wähle Modellierung basierend auf Abfragepfaden, Aktualisierungsfrequenz und Skalierungsbedarf.
- Berücksichtige Historisierung, Data Quality und Governance bei ETL.
- Plane Metriken und Kennzahlen für betriebliche Entscheidungsunterstützung.