ETL – Datenbereinigung & Harmonisierung

Was ist Datenbereinigung?

Datenbereinigung (Data Cleansing) bezeichnet den Prozess, bei dem Daten in einem Datensatz auf Fehler, Inkonsistenzen, Duplikate oder Unvollständigkeiten überprüft und korrigiert werden. Ziel ist es, die Qualität, Genauigkeit und Konsistenz der Daten zu verbessern.

3 Varianten der Datenbereinigung:

1. **Manuelle Bereinigung:** Einzelne Fehler werden durch Menschen in Tools wie Excel oder direkt in Datenbanken bereinigt.
2. **Regelbasierte automatische Bereinigung:** Regeln (z. B. "keine NULLs erlaubt in Spalte X") werden auf große Datensätze angewendet.
3. **Machine Learning-basierte Bereinigung:** Intelligente Algorithmen identifizieren Anomalien oder fehlerhafte Muster in großen Datenmengen.

Arten von Datenfehlern (4 Beispiele):

1. **Rechtschreibfehler & Tippfehler** (z. B. "Müncchen" statt "München")
2. **Duplikate** (gleiche Kunden mehrfach vorhanden)
3. **Formatinkonsistenzen** (Datum als "01.01.2024" vs. "2024-01-01")
4. **Unvollständige Daten** (fehlende Adressen, Telefonnummern, IDs)

Warum variiert der Umfang der Bereinigung?

Weil:

- **Unterschiedliche Datenquellen** (z. B. Webformulare vs. ERP-Systeme) unterschiedlich fehleranfällig sind
- **Anwendungsfall-spezifisch:** Eine Marketinganalyse toleriert ggf. fehlende Telefonnummern, während ein CRM-System diese zwingend benötigt
- **Unterschiedliche Qualitätsstandards** je nach Branche oder Projekt

4 typische Schritte im Datenbereinigungsprozess:

1. **Datenprofilerstellung:** Erste Analyse der Datenqualität (z. B. NULL-Werte, Häufigkeiten, Ausreißer)
2. **Fehleridentifikation:** Lokalisieren fehlerhafter Daten (z. B. leere Felder, doppelte IDs)
3. **Korrektur & Standardisierung:** Behebung der Fehler und Vereinheitlichung (z. B. Schreibweise, Formate)
4. **Validierung & Dokumentation:** Prüfen, ob die Änderungen korrekt waren, und dokumentieren der Datenqualität

5 Vorteile der Datenbereinigung:

1. **Höhere Datenqualität** führt zu besseren Entscheidungen
2. **Effizientere Datenanalysen** mit weniger Nachbearbeitung
3. **Kundenbindung durch korrekte Kommunikation** (richtige Namen/Adressen)
4. **Kosteneinsparung** durch Vermeidung falscher Datenverarbeitung
5. **Bessere Compliance** durch Einhaltung gesetzlicher Datenanforderungen

Harmonisierung ist ...

Ein Spezialfall der Datenbereinigung, bei dem **heterogene Daten aus verschiedenen Quellen vereinheitlicht** werden – z. B. unterschiedliche Produktnamen für denselben Artikel.

Was bedeutet Datenharmonisierung?

Datenharmonisierung ist der Prozess, verschiedene Datenquellen in ein gemeinsames Format oder eine konsistente Struktur zu überführen, damit sie sinnvoll zusammengeführt und analysiert werden können.

Unterschied: ETL vs. ELT

Aspekt	ETL	ELT
Transformation sort	Vor dem Laden (in Staging-Umgebung)	Nach dem Laden (im DWH selbst)

Geeignet für	Traditionelle Systeme, On-Prem	Cloud-basierte Systeme (BigQuery, Snowflake)
Geschwindigkeit	Langsamer bei großen Datenmengen	Schneller durch skalierbare DWH-Leistung
Kontrolle	Hohe Kontrolle über Transformationen	Weniger Kontrolle, mehr im DWH selbst

ETL-Testing-Konzept

Data Warehouse Testing ist ...

Ein Testprozess zur Sicherstellung, dass ein Data Warehouse korrekte, konsistente und analysierbare Daten liefert.

Was ist ETL?

ETL steht für **Extract, Transform, Load** – ein Prozess zur Datenintegration aus unterschiedlichen Quellen.

Was ist ETL-Testing?

ETL-Testing ist das Überprüfen, ob die Daten **korrekt extrahiert, transformiert und geladen** wurden. Dabei wird sichergestellt, dass keine Daten verloren gehen oder verfälscht werden.

ETL-Testing-Prozess (5 Schritte):

1. Anforderungen & Mapping-Verständnis
2. Testplanung & Datenvorbereitung
3. Testdurchführung (z. B. Row Count, Hash Check, Validierung)
4. Fehleranalyse & Eskalation
5. Regressionstest / Abnahme

Test Life Cycle (Testlebenszyklus):

- Testplanung
- Testdesign (Testfälle schreiben)
- Testausführung (manuell oder automatisiert)
- Fehlerprotokollierung
- Testreporting / Abschlussbericht

Mögliche Testtypen:

- Row Count Vergleich
- Datenvalidierung (z. B. Formate, Wertebereiche)
- Duplikattests
- Null-/Leereintragsprüfung
- Geschäftsregelvalidierung
- Performance-/Lasttests

Wie erstellt man einen Testcase?

1. Beschreibung des Ziels (Was wird geprüft?)
2. Eingabedaten/Quellen definieren
3. Erwartetes Ergebnis festhalten
4. SQL-Statement oder Testskript formulieren
5. Testergebnis dokumentieren (Pass/Fail)