ETL – Datenbereinigung - Harmonisierung

Datenbereinigung ist ein wichtiger Bestandteil des gesamten Datenmanagementprozesses und eine der Kernkomponenten der Datenaufbereitung,

mit der Datensätze für die Verwendung in Business-Intelligence- (BI) und Data-Science-Anwendungen vorbereitet werden.

Sie wird in der Regel von Datenqualitätsanalysten und -ingenieuren oder anderen Datenmanagementexperten durchgeführt.

Aber auch Data Scientists, BI-Analysten und Geschäftsanwender können Daten bereinigen

oder sich am Datenbereinigungsprozess für ihre eigenen Anwendungen beteiligen.

Data Cleansing versus Data Cleaning versus Data Scrubbing

Die Begriffe **Data Cleansing, Data Cleaning und Data Scrubbing** werden im Englischen oft synonym verwendet.

In den meisten Fällen werden sie als ein und dasselbe angesehen.

In einigen Fällen wird **Data Scrubbing** jedoch als ein Element der Datenbereinigung angesehen,

das speziell das Entfernen von doppelten, schlechten, nicht benötigten oder alten Daten aus Datensätzen beinhaltet.

Data Scrubbing hat auch eine andere Bedeutung im Zusammenhang mit der Datenspeicherung.

In diesem Zusammenhang handelt es sich um eine automatisierte Funktion,

die Festplattenlaufwerke und Speichersysteme überprüft, um sicherzustellen,

dass die darin enthaltenen Daten gelesen werden können und um fehlerhafte Sektoren oder Blöcke zu identifizieren.

Welche Arten von Datenfehlern werden durch Datenbereinigung behoben?

Datenbereinigung befasst sich mit einer Reihe von Fehlern und Problemen in Datensätzen, darunter ungenaue, ungültige, inkompatible und beschädigte Daten.

Einige dieser Probleme werden durch menschliches Versagen bei der Dateneingabe verursacht,

während andere aus der Verwendung unterschiedlicher Datenstrukturen, -formate und -terminologie in verschiedenen Systemen innerhalb einer Organisation resultieren.

Zu den Arten von Problemen, die üblicherweise im Rahmen von Datenbereinigungsprojekten behoben werden, gehören:

- **Tippfehler und ungültige oder fehlende Daten.**

Die Datenbereinigung korrigiert verschiedene strukturelle Fehler in Datensätzen.

Dazu gehören zum Beispiel Rechtschreibfehler und andere typografische Fehler, falsche numerische Einträge, Syntaxfehler und fehlende Werte, wie leere oder ungültige Felder, die Daten enthalten sollten.

- **Inkonsistente Daten.**

Namen, Adressen und andere Attribute sind oft von System zu System unterschiedlich formatiert.

So kann ein Datensatz beispielsweise die mittlere Initiale eines Kunden enthalten, ein anderer nicht.

Auch Datenelemente wie Begriffe und Bezeichnungen können variieren.

Datenbereinigung trägt dazu bei, dass die Daten konsistent sind, damit sie genau analysiert werden können.

- **Doppelte Daten.**

Bei der Datenbereinigung werden doppelte Daten in Datensätzen identifiziert und mit Deduplizierungsmaßnahmen entweder entfernt oder zusammengeführt.

Wenn zum Beispiel Daten aus zwei Systemen kombiniert werden, können doppelte Dateneinträge abgeglichen werden, um einzelne Datensätze zu erstellen.

- **Irrelevante Daten.**

Einige Daten, zum Beispiel Ausreißer oder veraltete Einträge, sind für Analyseanwendungen möglicherweise nicht relevant und könnten deren Ergebnisse verfälschen. Durch Datenbereinigung werden redundante Daten aus den Datensätzen entfernt,

wodurch Datenaufbereitung rationalisiert und die erforderliche Menge an Datenverarbeitungs- und Speicherressourcen reduziert wird.

Der Umfang der Datenbereinigung variiert je nach Datensatz und Analyseanforderungen.

Ein Datenwissenschaftler, der eine Betrugserkennungsanalyse von Kreditkartentransaktionsdaten durchführt,

möchte beispielsweise Ausreißerwerte beibehalten, da sie ein Anzeichen für betrügerische Käufe sein können.

Der Datenbereinigungsprozess umfasst jedoch in der Regel folgende Schritte:

1. **Inspektion und Profiling.**

Zunächst werden die Daten inspiziert und geprüft, um ihr Qualitätsniveau zu bewerten und Probleme zu identifizieren, die behoben werden müssen. Dieser Schritt umfasst in der Regel die Erstellung von Datenprofilen (Data Profiling), bei der die Beziehungen zwischen Datenelementen

dokumentiert, die Datenqualität geprüft und Statistiken über Datensätze erstellt werden, um Fehler, Diskrepanzen und andere Probleme zu erkennen

2.Bereinigung.

Dies ist das Herzstück des Bereinigungsprozesses, bei dem Datenfehler korrigiert und inkonsistente, doppelte und redundante Daten bereinigt werden.

3.Verifizierung.

Nach Abschluss der Bereinigung sollte die Person oder das Team, die beziehungsweise das die Arbeit durchgeführt hat, die Daten erneut überprüfen, um sicherzustellen, dass sie sauber sind und den internen Regeln und Standards für die Datenqualität entsprechen.

4.Reporting.

Die Ergebnisse der Datenbereinigung sollten anschließend an die IT- und Geschäftsleitung gemeldet werden, um Trends und Fortschritte bei der Datenqualität aufzuzeigen. Der Bericht könnte die Anzahl der gefundenen und behobenen Probleme sowie aktualisierte Metriken zum Qualitätsniveau der Daten enthalten.

Die bereinigten Daten können dann in die verbleibenden Phasen der Datenaufbereitung überführt werden, beginnend mit der Datenstrukturierung und Datentransformation, um sie für die Analysezwecke vorzubereiten.

Vorteile der Datenbereinigung

Eine gut durchgeführte Datenbereinigung bietet folgende Vorteile für Unternehmen und Datenmanagement:

- **Verbesserte Entscheidungsfindung.**

Mit genaueren Daten können Analyseanwendungen bessere Ergebnisse erzielen. Dadurch können Unternehmen fundiertere Entscheidungen zu Geschäftsstrategien und -abläufen sowie zu Themen wie Patientenversorgung und staatliche Programme treffen.

- **Effektiveres Marketing und Vertrieb.**

Kundendaten sind oft falsch, inkonsistent oder veraltet.

Die Bereinigung der Daten in Kundenbeziehungsmanagement- und Vertriebssystemen trägt dazu bei, die Effektivität von Marketingkampagnen und Vertriebsbemühungen zu verbessern.

- **Bessere operative Leistung.**

Saubere, qualitativ hochwertige Daten helfen Unternehmen, Fehlbestände, Lieferengpässe und andere Geschäftsprobleme zu vermeiden,

die zu höheren Kosten, geringeren Einnahmen und schlechten Kundenbeziehungen führen können.

- **Verstärkte Nutzung von Daten.**

Daten sind zu einem wichtigen Unternehmenswert geworden, aber sie können keinen Geschäftswert generieren, wenn sie nicht genutzt werden.

Indem die Datenbereinigung die Vertrauenswürdigkeit der Daten erhöht, trägt sie dazu bei,

Manager und Mitarbeiter davon zu überzeugen, sich bei ihrer Arbeit auf die Daten zu verlassen.

- **Geringere Datenkosten.**

Datenbereinigung verhindert, dass sich Datenfehler und Probleme in Systemen und Analyseanwendungen weiter ausbreiten.

Langfristig spart dies Zeit und Geld, da IT- und Datenmanagementteams nicht immer wieder die gleichen Fehler in Datensätzen beheben müssen.

Harmonisierung ist ...

Was bedeutet Datenharmonisierung? Daten sind harmonisiert, wenn alle Unternehmensinformationen (Daten) von allen Mitarbeitern in allen Standorten und in allen genutzten Systemen und Prozessen identisch verstanden, interpretiert und genutzt

Datenharmonisierung ist die Methode, unterschiedliche Datenfelder, Formate, Dimensionen und Spalten in einem zusammengesetzten Datensatz zu vereinen . Damit eine Organisation erfolgreich sein kann, benötigen Benutzer demokratischen Zugriff auf saubere, qualitativ hochwertige Daten und konforme Dimensionen – die Datenkomplexität wird ausgeblendet und Datenformate werden vereinbart.

Unterschied ETL and ELT

- Siehe ETL-vs.-ELT-Which-One-Is-Right-for-Your-Organization.pdf

  Lesen Sie sich das pdf durch und versuchen Sie die Unterschiede/Vorteile und Nachteile zu erkennen

ETL-testing

Sie sind als externer Consultant (Data Engineer) im Integration Team aufgenommen worden.

Ihre erste Aufgabe ist es DWH-Testing und da im speziellen ein Konzept für ETL-Testing zu erstellen .

Ihre Ausarbeitung soll folgende Punkt beinhalten

- Data Warehouse Testing ist .....
- Was ist ETL?
- Was ist ETL-Testing
- ETL-Testing -Prozess .... (5 Schritte)
- Test Life Cycle
- Mögliche Test Typen
- Wie kann ein Testcase erstellt werden