ETL – Datenbereinigung - Harmonisierung

Datenbereinigung

Erklären sie die Datenbereinigung und 3 Varianten

Drei Varianten der Datenbereinigung:

1. ***Data Cleaning:***
    - ***Fokus:*** *Korrektur von Fehlern und Inkonsistenzen.*
    - ***Beispiel:*** *Anpassung falscher Werte und Standardisierung von Formaten, um die Genauigkeit der Daten sicherzustellen.*
2. ***Data Scrubbing:***
    - ***Fokus:*** *Entfernen oder Korrigieren unvollständiger oder falsch formatierter Daten.*
    - ***Beispiel:*** *Beseitigung von duplizierten Einträgen und Korrektur von fehlerhaften Datensätzen, die aufgrund von Eingabefehlern entstanden sind.*
3. ***Data Cleansing:***
    - ***Fokus:*** *Umfassender Prozess zur Sicherstellung der Datenqualität.*
    - ***Beispiel:*** *Erkennen und Entfernen von korrupten Daten, Überprüfung der Datenkonsistenz und Aktualisierung veralteter Informationen, um eine hohe Datenqualität zu gewährleisten.*

Welche Arten von Datenfehlern werden durch Datenbereinigung behoben?
Beschreiben Sie 4 „Arten“

Bei der Datenbereinigung werden folgende Arten von Datenfehlern behoben:

- ***Fehlende Werte (Missing Values):*** *Leere oder nicht erfasste Datenpunkte.*
- ***Duplikate (Duplicates):*** *Doppelte Einträge im Datensatz.*
- ***Inkonsistente Daten (Inconsistent Data):*** *Unterschiedliche Formate oder Schreibweisen für dieselbe Information.*
- ***Falsche Werte (Incorrect Data):*** *Fehlerhafte oder ungültige Datenpunkte.*
- ***Falsche oder irreführende Werte (Misleading Data):*** *Technisch korrekte, aber kontextuell falsche Daten.*
- ***Formatierungsfehler (Formatting Errors):*** *Falsche oder inkonsistente Datenformate.*
- ***Falsch zugeordnete Daten (Misfielded Data):*** *Daten in den falschen Feldern gespeichert.*
- ***Ungültige Daten (Invalid Data):*** *Daten, die den erwarteten Regeln nicht entsprechen.*
- ***Veraltete Daten (Outdated Data):*** *Nicht mehr aktuelle oder relevante Daten.*

Der Umfang der Datenbereinigung variiert je nach Datensatz und Analyseanforderungen. Wieso?????

Datenkomplexität: *Einfache Datensätze mit wenigen Variablen erfordern weniger Bereinigung als komplexe Datensätze mit vielen unterschiedlichen Datenquellen.*
Analyseanforderungen: *Die Genauigkeit und Detailtiefe der benötigten Analyse bestimmt, wie gründlich die Daten bereinigt werden müssen.*
Qualität der Rohdaten: *Datensätze mit bereits hoher Datenqualität benötigen weniger Bereinigungsaufwand im Vergleich zu stark fehlerhaften oder unvollständigen Daten.*
Datenquelle: *Daten aus zuverlässigen und standardisierten Quellen erfordern weniger Bereinigung als Daten aus heterogenen und unstrukturierten Quellen.*
Zeit- und Ressourcenverfügbarkeit: *Der verfügbare Zeitrahmen und die Ressourcen beeinflussen, wie umfassend die Datenbereinigung durchgeführt werden kann.*

Der Datenbereinigungsprozess umfasst jedoch in der Regel folgende Schritte:
Benennen Sie 4 Schritte und erklären Sie diese

Schritte im Datenbereinigungsprozess

1. ***Datenidentifikation und -prüfung:***
    - *Überprüfung der Datenquellen und ersten Qualitätskontrolle.*
2. ***Datenbereinigung:***
    - *Entfernen von Duplikaten, Korrigieren von Fehlern und Standardisieren von Formaten.*
3. ***Datenvalidierung:***
    - *Überprüfung der bereinigten Daten auf Richtigkeit und Konsistenz.*
4. ***Reporting und Dokumentation:***
    - *Erstellung von Berichten über die durchgeführten Bereinigungsmaßnahmen und die Datenqualität.*

Vorteile der Datenbereinigung
Erklären Sie 5 „Arten“

Vorteile der Datenbereinigung

1. ***Erhöhte Genauigkeit:***
    - *Datenbereinigung verbessert die Präzision der Daten, wodurch zuverlässigere Analysen und fundierte Entscheidungen möglich werden.*
2. ***Verbesserte Effizienz:***
    - *Durch das Entfernen unnötiger oder fehlerhafter Daten wird die Verarbeitungsgeschwindigkeit erhöht und die Effizienz der Datenanalyse gesteigert.*
3. ***Konsistenz der Daten:***
    - *Datenbereinigung stellt sicher, dass alle Daten einheitlich und im gleichen Format vorliegen, was die Vergleichbarkeit und Integration erleichtert.*
4. ***Reduzierung von Kosten:***

    - *Fehlerhafte Daten können zu falschen Entscheidungen und unnötigen Kosten führen; durch Datenbereinigung werden diese Risiken minimiert.*
5. ***Erfüllung von Compliance-Anforderungen:***
    - *Saubere und genaue Daten helfen Unternehmen, gesetzliche und regulatorische Anforderungen zu erfüllen und Compliance-Risiken zu reduzieren.*

Harmonisierung ist …

Was bedeutet Datenharmonisierung?

Letztlich müssen die Daten in der Fact-Tabelle und den Dim-Tabellen alle den gleichen Regeln und Gesetzen folgen, wie zum Beispiel Maßeinheiten zur Berechnung von Summen (Fact) oder die Bedeutung von Werten für die Aggregation (Dim). Harmonisierung dient dazu, dies zu gewährleisten. Datenharmonisierung ist ein Prozess, bei dem Daten aus verschiedenen Quellen in ein einheitliches Format gebracht werden, um Konsistenz und Vergleichbarkeit zu gewährleisten. Dies umfasst das Vereinheitlichen von Datenstrukturen, -definitionen und -standards, um eine nahtlose Integration und Analyse zu ermöglichen.

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

Der Hauptunterschied zwischen ETL und ELT liegt im Zeitpunkt der Transformation:

- ***ETL:** Transformation erfolgt vor dem Laden - einfacher, aber weniger flexibel.*
- ***ELT:** Transformation erfolgt nach dem Laden - flexibler, aber potenziell komplexer.*

Rahmenparameter für die Entscheidung zwischen ETL und ELT umfassen Datenmenge, Leistung, Datenqualität, Flexibilität, Systemressourcen und spezifische Geschäftsanforderungen.