Konzept Datenqualität Zoo Pirmasens

10.06.2022, Version 6
Gruppe 3, Franziska Wolny

Inhaltsverzeichnis

1 Einleitung

Die Datenqualität beschreibt, wie gut die vorhandenen Datenbestände in der Lage sind, die tatsächliche Realität abzubilden, und wie gut die Datenbestände für die vorgesehenen Anwendungen geeignet sind.

Ein Ziel der Migration der Daten ins operative System ist die Performanz und Effizienz der Arbeitsprozesse zu sichern. Damit jeder Schritt dorthin richtig funktioniert, ist die Datenpflege ein notwendiger Schritt. Datenqualität ermöglicht die Vermeidung von veralteten Daten, Duplikaten, Redundanzen, Inkonsistenzen, Anomalien und weiteren Fehlern.

Bei einem ungepflegtem Datensatz könnten im vorliegenden Fall Fehler wie die Folgenden vorkommen:

- unvollständige Felder
- Formatfehler (Falsches Datumsformat (Monat-Tag-Jahr anstatt Tag-Monat-Jahr)

- mehrwertige Felder (Eingabe der Tiername, der Verwandtschaft und des Herkunftsort in einem Feld)
- uneinheitliche Bezeichnungen (Eingabe des Geschlechtes als „w“ oder „weiblich“)
- nicht plausible bzw. falsche Daten (das Gewicht des Penguins ist eigentlich das eines Elefanten)
- Transposition (umgekehrte Reihenfolge von Name und Nachname) bei fehlender Normalisierung
- Verletzung der referenziellen Integrität (Eintragung eines falschen Primär- oder Fremdschlüssels)

Eine schlechte Datenqualität hat weitreichende Konsequenzen auf den Geschäftsprozess. Einige Risiken sind:

- hohe Kosten durch Fehlerbehebung
- negative Auswirkungen auf Auswertungen und Entscheidungen durch falsche Datengrundlage
- die Echtheit und Ursprung der Daten sind nicht nachvollziehbar
- Redundanz und Duplikate führen zu unnötiger Akkumulation der Daten und Inkonsistenzen
- Daten sind nicht vertrauenswürdig, wenn sie sich wegen falscher und inkonsistenter Historisierung, fehlender Werte, oder Verletzung der referenziellen Integrität widersprechen

2 Ausgangssituation

Der Zoo Pirmasens besteht seit ca. 36 Jahren und wird derzeit modernisiert. Durch Vergrößerung und Umstrukturierung erfolgen Änderungen, die auf den Datenbestand einwirken. Die Dokumentation liegt bisher zum großen Teil in Papierform vor. Eine zentrale Datenbank soll nun alle relevanten Daten zentral und digital zusammenfassen und das operative Tagesgeschäft abbilden. Nur der derzeitige Stand und zukünftige Änderungen soll erfasst werden. Kunden und Ticketverkäufe sollen derzeit nicht erfasst werden. Der Fokus liegt auf den Tieren, ihrer Unterbringung, Fütterung und Pflege sowie den zuständigen Mitarbeitern. Der Zoo möchte die vorhandenen Daten selbst in die von uns erstellte Datenbank überführen.

3 Zielstellung

Es soll ein Datenmodell / eine Datenbank für die operative Abwicklung des Tagesgeschäfts erstellt werden. Alle Mitarbeiter sollen auf einen konsistenten, validen und aktuellen Datenbestand zugreifen. Der Datenbestand soll fortlaufend bearbeitet werden können und die Datenqualität auf ein Level von >97% gebracht werden. Zudem soll zur Unterstützung von BI ein Datawarehouse aufgesetzt werden.

4 Definition Datenqualität

Die Definition der **Datenqualität Q_{gesamt}** ist notwendig zur Überprüfung der Zielerreichung. Folgende Faktoren i können die Datenqualität beeinflussen:

i	Faktor	Maßnahmen (Beispiele)
1	Richtigkeit (korrekte Werte)	- Eingabefehler bei Datenübertragung vermeiden, z.B. 4-Augenprinzip (Gegencheck der Eingabewerte durch Kollegen) - automatische Rechtschreibkontrolle verwenden - Referenztabellen - im Nachgang Überprüfung auf Plausibilität z.B. durch Visualisierung der Verteilung
2	Redundanzfreiheit (keine Duplikate)	- Bei initialer Datenübertragung sicherstellen (z.B. durch Abhaken nach Abarbeitung von gelisteten Datenquellen) - Klare Verantwortlichkeiten definieren (wer trägt wann was ein)
3	Genauigkeit (z.B. Kommastellen)	- Relevant für numerische Werte, z.B. Gewicht, Größe - Große Verteilung über Tiere → Genauigkeit in % vom Absolutwert angeben, Festlegung: 1%, z.B. 10 g Genauigkeit bei 1000 g Gewicht
4	Vollständigkeit (keine fehlenden Werte)	- Soweit Daten vorhanden sind, eintragen, ggf. vervollständigen mit Hilfe der Kollegen der Fachabteilungen - fachliche Ansprechpartner definieren für Unklarheiten im weiteren Betrieb der DB
5	Konsistenz	- Sicherstellen durch Normalisierung, eindeutige und einzigartige Primärschlüssel und Verweise/Referenzen
6	Verständlichkeit	- Klare Spaltenbezeichnungen - Data Dictionary
7	Einheitlichkeit	- Alle Einträge einer Spalte haben die angegebene Einheit - Referenztabellen, Formatvorlagen (z.B. Datum) - festgelegte Datentypen und Feldlängen - interne Festlegung einheitlicher Bezeichnungen für z.B. Position, Befund usw. - Regelwerk für Festlegungen zur Schreibweise erstellen
8	Zuverlässigkeit (nachvollziehbare Entstehung der Daten)	- Dokumentation der initialen Befüllung der DB - Erstellung von klaren Anweisungen zur weiteren Eintragung von Daten Wer: Zugriffsrechte, Schulung Wann: wie oft neue Datensätze eintragen, Transfer ins DWH Was: welche Daten werden eingetragen Wie: Datentyp, Format, Eingabemaske, Scanner etc.
9	Aktualität (keine veralteten Werte)	- Klare Festlegungen, wann Daten einzutragen sind und ins DWH zu übertragen sind - möglichst kurzer Zeitverzug zwischen Änderungen und der entsprechenden Anpassung der Daten

In der Tabelle sind schon mögliche Maßnahmen benannt, weitere Überlegungen zu einigen der Faktoren folgen im nächsten Kapitel. Faktoren 1-8 müssen vorab festgelegt und schon bei der initialen Datenübertragung in die DB sichergestellt werden. Faktor 9 ist vor allem auch für eine langfristige Sicherstellung guter Datenqualität notwendig.

Alle Qualitätsfaktoren Q_i können prozentual bewertet werden (z.B. Anzahl aller korrekten Eintragungen im Verhältnis zur Anzahl aller Eintragungen für Faktor 1). Der Durchschnitt dieser Werte für alle Faktoren gilt als Maß für die Gesamtdatenqualität Q_{gesamt}:

$$Q_{gesamt} = \frac{1}{9}\sum_{i=1}^{9} Q_i * 100\% \qquad Q_i = \frac{Anzahl\ anforderungsgemäße\ Einträge}{Anzahl\ aller\ Einträge}$$

Q_{gesamt} kann bei Projektende (Ist-Stand des Stichtages ist komplett in der DB abgebildet) bewertet werden. Danach ist es sinnvoll, diese Kennzahl weiter zu tracken (**Monitoring**), zum Beispiel im Rahmen vom BI-Reporting einmal monatlich die Datenqualität zu berechnen. Hier ist eine verantwortliche Person und ein Intervall festzulegen. Falls die Datenqualität kontinuierlich unter dem Zielwert liegen sollte, müssen die hier beschriebenen Maßnahmen überprüft und ggf. neue definiert werden.

5 Sammlung, Bewertung und Sortierung der Ausgangsdaten

Nach Angaben des Kunden liegen die Ausgangsdaten in verschiedener Form analog vor. Es ist davon auszugehen, dass diese Papiere in Bezug auf die Datenqualität schwanken. Wir betrachten hier zudem ausschließlich Daten in Textform (Text bzw. Zahlen). Bilder und andere Medien werden vorerst nicht in die Datenbank aufgenommen. Folgende Schritte müssen zu Beginn unter Mithilfe der Personen durchgeführt werden, die die Daten ursprünglich aufgezeichnet haben. Eine Person sollte als Projektleiter die Planung und Leitung übernehmen und Verantwortlichkeiten, Meilensteine und Zeitplan erstellen.

- Zusammentragen ALLER verfügbaren relevanten Daten, um die im Data Dictionary aufgeführten Attribute mit Inhalt versehen zu können
- Verwerfen von veralteten Daten (z.B. ehemalige Mitarbeiter, nur aktuelle Daten sollen überführt werden)
- Verwerfen von doppelten Daten (z.B. Kopien von Bestellungen, Akten usw.)
- Fehlende Informationen nach Möglichkeit vorab bei den Verantwortlichen erfragen und digital festhalten
- Klärung von unklaren, unleserlichen oder kryptischen Inhalten zusammen mit dem Ersteller, Überführung in eine eindeutige Form (bestenfalls bereits digital)
- Vereinheitlichung von eventuell verschiedenen Nomenklaturen, Einheiten usw.
- Dokumentation des gesamten Prozesses
- Dokumentation von Auffälligkeiten bezüglich fehlender Daten, Formaten usw.

Danach kann die Datenmenge bewertet werden, um die Dauer der Digitalisierung und die benötigte Anzahl Mitarbeiter bestimmen zu können, um in möglichst kurzer Zeit einen

arbeitsfähigen Zustand zu erreichen. So kann auch der Fortschritt dokumentiert werden (welcher Anteil ist bereits digitalisiert). Falls sich herausstellt, dass die Datenmenge zu groß ist, um von den eigenen Mitarbeitern bearbeitet zu werden, sollte über eine externe Beauftragung der Digitalisierung nachgedacht werden.

Die Daten sollten thematisch sortiert werden (bestenfalls nach Entitäten gemäß ERM bzw. Data Dictionary). Dies erleichtert den Überblick und die Aufgabenaufteilung beim Übertragungsprozess.

Zu diesem Zeitpunkt sollte auch eine Beratung eingeholt werden, inwiefern bei der Datendigitalisierung auf DSGVO-Konformität geachtet werden muss.

Es besteht die Möglichkeit, dass nach erfolgter Sammlung und Bewertung festgestellt wird, dass eine Vielzahl Daten in unzureichender Qualität vorliegt bzw. fehlt. Dann kann das gesetzte Ziel von 97% Datenqualität NICHT erreicht werden und das Projektziel muss neu definiert werden.

6 Übertragung der vorhandenen Daten in die erstellte Datenbank

Es ist keine Datenbank vorab vorhanden, sämtliche Tabellen und Spaltenüberschriften wurden neu definiert. Mittels der vorab gesichteten und sortierten Aufzeichnungen sollen die nach ERM erzeugten Tabellen in MS Access mit Daten gefüllt werden. Idealerweise könnte noch eine grafische Benutzeroberfläche erstellt werden, die die Eingabe erleichtert und Eingabefehler weiter reduziert. Dies war jedoch nicht Projektbestandteil. Für die Dateneingabe sollten IT-affine Mitarbeiter gezielt ausgewählt und **geschult** werden (z.B. durch die Weber AG). Diese könnten dann auch zukünftig als Ansprechpartner für das Hinzufügen und Ändern von Daten agieren.

Der Prozess des Überführens der Daten in die Datenbank sollte gut geplant, strukturiert und **dokumentiert** werden, um alle verfügbaren Daten zu erfassen, aber gleichzeitig **redundanzfrei** zu arbeiten, also Mehrfacheinträge zu vermeiden. Folgenden Informationen können z.B. hilfreich sein:

- Verantwortlicher Mitarbeiter
- Bearbeiteter Inhalt (z.B. bestimmte Tabelle bzw. Spalten oder bestimmter Zeitraum)
- Datenquelle (Notizbuch, Exceltabelle, Gedächtnisprotokoll, Personalakte etc.)
- Datum des Eintrags

Bei der Nutzung von Tools wie Schrifterkennungssoftware und Barcodescannern zum Einlesen der Daten muss nachträglich auf Fehler überprüft und diese Prüfung auch dokumentiert werden. Fehleranfällige Software oder Scanngeräte sollten ersetzt werden. Bei der Optimierung von Schrifterkennungsalgorithmen bzw. beim Erwerb geeigneter Software kann die Weber AG beratend zur Seite stehen.

Zudem sollte über den Verbleib der Originaldaten entschieden werden. Ein Aufbewahren im Lager für einen Übergangszeitrum bis zur Sicherstellung der Funktionsweise des neuen Systems empfiehlt sich. Die Datenbank sollte durch Backups gesichert werden (nicht Teil des Auftrags).

Im **Data Dictionary** finden sich alle vorhandenen Tabellen und Spalten mit den vorgeschriebenen **Datentypen** als Vorgabe für das Format der einzutragenden Daten. Diese Festlegungen sind beim Befüllen unbedingt einzuhalten. Die Spaltentitel sind so gewählt, dass sie möglichst selbsterklärend sind. Weitere Beschreibungen finden sich sonst im Data Dictionary.

Für Felder mit nur wenigen möglichen Werten sollten die vorhandenen **Referenztabellen** genutzt werden um Falscheingaben zu vermeiden, z.B.

- Geschlecht
- Anrede
- Titel
- Einheiten
- Sozialer Status (Ehestand)
- Ort bzw. PLZ

Hier wurde der Access Nachschlageassistent mit den entsprechenden Referenztabellen verbunden, so dass nur die Optionen aus den Referenztabellen zur Verfügung stehen und Fehler vermieden werden. Weitere Attribute wurden mit **ja/nein-Auswahlfeldern** versehen, um eindeutige Werte sicherzustellen, z.B. (siehe auch Data Dictionary):

- Gehege: vergittert, frei laufend
- Teilweg/ Rundweg: barrierefrei
- Krankheit: Behandlung abgeschlossen, Meldepflicht

Falls weitere Spalten existieren, für die es nur wenige mögliche Einträge gibt, können diese über ein **Drop-Down-Menü** (Nachschlagsassistent) in Access befüllt werden. Dies vermeidet eventuelle fehlerhafte Einträge. Hierbei kann die Weber AG auch in der Phase der ersten Datenübertragung assistieren. Vorstellbar wäre dies z.B. für

- Nationalität der Mitarbeiter
- natürliche Lebensraum einer Tierart
- Krankheit
- Befund
- Medikation

Hier muss abgeschätzt werden, ob es Sinn macht, alle möglichen Fälle vorab einzutragen bzw. weitere Referenztabellen zu erstellen. Gegebenenfalls muss auch nachträglich die Liste erweitert werden. Vorteil ist, dass eine vorgeschriebene **konsistente** Bezeichnung von Lebensraum, Krankheit usw. eine spätere Auswertung erleichtert. Selbst wenn keine Referenztabellen erstellt werden, ist die interne Festlegung **einheitlicher Bezeichnungen** essenziell. Dies kann in einer entsprechenden Dokumentation geschehen oder bei allgemeinen Benennungsregeln auch ins Data Dictionary eingefügt werden.

Felder mit festgelegter Syntax (z.B. Telefonnummern, Email-Adressen) sollen einer Syntaxprüfung unterzogen werden.

Für manche Attribute empfiehlt sich die Definition von **Bedingungen und Grenzwerten**, um Fehler leichter identifizieren zu können. Diese können wie folgt aussehen:

- Geburtsdatum Mitarbeiter <01.01.2010, >01.01.1920
- Tier Sterbedatum > Geburtsdatum
- Lagerbestand, Meldebestand >0
- Rundweg, Teilweg Länge > 0 m, < 20.000 m
- Gewicht und Größe Tier abhängig von Tierart in bestimmtem Intervall
- usw.

Bei Verletzung der Regel muss eine Meldung erzeugt werden, die zur Überprüfung des entsprechenden Wertes durch die verantwortliche Person führt. Teilweise können diese Regeln auch im Data Dictionary festgehalten werden.

Zur Generierung eindeutiger und einzigartiger **Primärschlüssel** wird der AutoWert-Datentyp in Access verwendet.

Bei der Eingabe von Text sollte die **Rechtschreibprüfung** von MS Access verwendet werden, um Fehler zu vermeiden.

Numerische Werte wie z.B. die verfügbare Futtermenge beziehen sich auf bestimmte **Einheiten**, die in der entsprechenden Spalte der Tabelle angegeben werden. Dies sollte Unstimmigkeiten bei den Einheiten vorbeugen und Auswertungen **eindeutig** machen. Falls keine Einheit angegeben ist, steht die Einheit im Spaltentitel (z.B. Gewicht_gr, Groesse_cm). Die Daten müssen dann immer in dieser Einheit eingetragen werden.

Datumswerte sollten alle in einem einheitlichen Format normalisiert eingegeben werden wie im Data Dictionary vorgegeben, bestenfalls über eine Referenztabelle (spätestens im DataWarehouse).

Bei **fehlenden bzw. unbekannten Werten** sollen die entsprechenden Felder frei gelassen werden, falls auch eine Recherche im Fachbereich keine Antwort geliefert hat. Es sollen KEINE Werte wie z.B. 0, unbekannt, k.A. oder ähnliche Platzhalter eingegeben werden. Dies erschwert die Auswertung und muss später mühsam bereinigt werden.

Bei der Eingabe von Text sollte auf die **Verständlichkeit** geachtet werden, z.B. sind Bemerkungen ohne Abkürzungen, gut verständlich und eindeutig zu formulieren. Alles soll in deutscher **Sprache** eingegeben werden. Krankheiten, Tierarten, Gattungen etc. sollen ebenfalls mit deutschem Namen bezeichnet werden. Nur falls kein deutscher Name existiert soll der lateinische Name verwendet werden. Es sollte ebenfalls festgelegt werden, ob Umlaute und „ß“ verwendet werden, oder diese ersetzt werden durch international gebräuchlichere Zeichen (abhängig von den Anwendungsszenarien des Zoos).

Für solche und andere hier bereits benannte Festlegungen empfiehlt sich zusammenfassend das Anlegen eines **Format- und Stilregelwerks** mit verbindlichen Regeln zu Themen wie

- Verwendete Sprache
- Verwendung von Abkürzungen
- Sonderzeichen
- Art und Weise des Genderns
- Bezeichnungskonventionen für konkrete Attribute (ggf. auch im Data Dictionary)
- Art der Dezimaltrennung und 1000er-Trennung (Festlegung im OS)

- Datumsformat
- Ortsangaben in deutscher Sprache oder Landessprache (z.B. Warschau vs. Warszawa)
- usw.

Die Datenbank wurde **normalisiert** erstellt, um Redundanzen und Anomalien zu vermeiden und eine **Konsistenz** der Daten sicherzustellen. Die Attribute sind atomar (abgesehen von Freitext-Feldern) und die dritte Normalform wurde angestrebt (z.B. über die Nutzung einer Referenztabelle für Ort und Postleitzahl), so dass keine funktionalen Abhängigkeiten zwischen Attributen einer Tabelle vorhanden sind.

7 Laufender Betrieb: Erfassung der operativen Geschäftsabläufe

7.1 Aktualisierung der operativen Datenbank

Die Datenbank soll das operative Geschäft widerspiegeln und jederzeit möglichst aktuell sein. Im vorliegenden Fall des Zoobetriebs bietet sich für alle Entitäten eine ereignisgesteuerte Aktualisierung an. Das heißt, sobald Änderungen passieren, werden diese durch die verantwortlichen Mitarbeiter in der Datenbank aktualisiert. Ein Zeitfenster, innerhalb dessen dies ab Änderung passieren muss, sollte definiert werden, kann aber für verschiedene Entitäten unterschiedlich sein. Nachfolgend sind Vorschläge angegeben.

Mindestens folgende Werte sollten innerhalb einer Stunde nach Änderung aktualisiert werden:

- Futter (Lagerbestand)
- Mahlzeiten und benachbarte Entitäten (Menge)

Dies soll mit jeder Änderung des Futterlagerbestandes durch Lieferung, Fütterung oder Verwerfung von verdorbenen Lebensmitteln erfolgen, um jederzeit einen aktuellen Überblick über verfügbare Futtermengen und eine eventuelle Unterschreitung des Meldebestandes zu haben (Ausgabe Warnmeldung sobald Lagerbestand < Meldebestand). Dabei muss auf korrekte Datenerfassung geachtet werden (siehe im vorigen Kapitel genannte Richtlinien). Dies wird vereinfacht über eine noch zu erstellende Eingabemaske oder z.B. durch die Verwendung von Barcodescannern bei der Erfassung von Lieferungen oder verarbeitetem Futter.

Für andere Tabellen ist eine sofortige Aktualisierung weniger kritisch. Dies betrifft z.B.:

- Tiere und Tierarten
- Mitarbeiter
- Krankengeschichte
- Bestellungen
- Vertretungen (Mitarbeiter und Ärzte)

Hier reicht es, die Änderungen bis Ende des Geschäftstages vorzunehmen.

Manche Tabellen werden nach dem initialen Einfügen der aktuellen Datenmenge nur selten bis nie aktualisiert werden müssen, z.B.

- Lieferanten
- Gattung
- Rundwege
- Gebäude

Die anfänglich festgelegten Fristen zur Aktualisierung festgelegten sollten einige Wochen nach Inbetriebname überprüft werden. Zum jetzigen Zeitpunkt können eventuell noch nicht alle Erfordernisse des Geschäftsablaufs abgeschätzt werden kann. Nach der Testphase kann eine best practice formuliert, angewendet und regelmäßig (z.B. monatlich) überprüft werden.

In jedem Fall ist auch hier die Verantwortlichkeit zu klären, welcher Mitarbeiter bis wann für die Aktualisierung welcher Tabelle zuständig ist (z.B. Herr Mustermann, Tabelle Bestellungen, bei Änderungen im Status Aktualisierung bis spätestens täglich 16:00, Vertretung durch Frau Musterfrau).

7.2 Löschen von Daten

Zur Bewahrung der **referenziellen Integrität** muss darauf geachtet werden, dass beim Löschen von Datensätzen (z.B. Mitarbeiter, Tiere, etc.) die entsprechenden Einträge in Nachbartabellen mit Bezug auf diese Datensätze ebenfalls gelöscht werden.

Änderungen müssen dokumentiert werden mit mindestens folgenden Informationen:

- Für Löschung zuständiger Mitarbeiter
- Zeitstempel
- Inhalt der Löschung

7.3 Regelmäßiges Monitoring der Datenqualität

Wie schon eingangs erwähnt, sollte die Datenqualität kontinuierlich überwacht werden. Hier müssen Kennzahlen, Grenzwerte, Maßnahmen, Prüfintervall und Verantwortlichkeiten festgelegt werden. Nachfolgend ist ein beispielhaftes Vorgehen skizziert.

Verantwortl. Mitarbeiter (ggf. für verschiedene Teilaspekte verschiedene Mitarbeiter)	Max Mustermann
Vertretung (ggf. mehrere)	Berta Beispiel
Prüfzeitpunkt	Jeden Ersten des Monats
Kenngrößen	Alle Q_i, Q_{gesamt}
Schwellwerte	Definieren für jeden Parameter, z.B. Aktualität $Q_9 > 0{,}95$
Maßnahmen bei Unterschreitung	Definieren für jeden Parameter, z.B. für Aktualität: - Ermittlung der kritischen Einflussgrößen mit größter Abweichung vom Ist-Wert (z.B. Attribut Lagermenge) - Kontrolle der Dokumentation der Änderung - Rücksprache mit für Änderung verantwortlichen Kollegen, werden Fristvorgaben eingehalten? - Kontrolle der Änderungsfristen, ggf. Anpassung - Kontrolle der vorhandenen Warnhinweise, ggf. Anpassung
Weiteres Vorgehen	ev. engmaschigere Kontrolle festlegen

Es sollten zudem Abläufe für folgende Prüfungen etabliert werden:

- Identifizierung fehlender Werte, Ursache (Übertragungsfehler, Daten fehlen, …)
- Identifizierung von Eingabefehlern
- Suchen nach Duplikaten
- Kontrolle der Datentypen
- Kontrolle des vorgegebenen Formates
- Kontrolle der Abhängigkeiten (Primär- und Fremdschlüssel)
- Kontrolle der inhaltlichen Korrektheit

Einige dieser Punkte können automatisiert werden (z.B. fehlende Werte, Datentyp, Format, Duplikate). Andere sind schwieriger zu lösen. Hier sollte der IT-Verantwortliche zusammen mit den Kollegen mit fachlichem Know-How Abläufe definieren. Dies betrifft vor allem die Kontrolle der inhaltlichen Korrektheit, hier verfügen wir nicht über den nötigen Einblick. Plausibilitätsprüfungen anhand von Grenzwerten (Kapitel 4) können hier unterstützen.

Zur **Nachvollziehbarkeit** sollten die Prüfergebnisse auch kontinuierlich dokumentiert werden, um z.B. die Ursache fehlender Werte festhalten zu können und nicht bei jeder Prüfung erneut zu fragen.

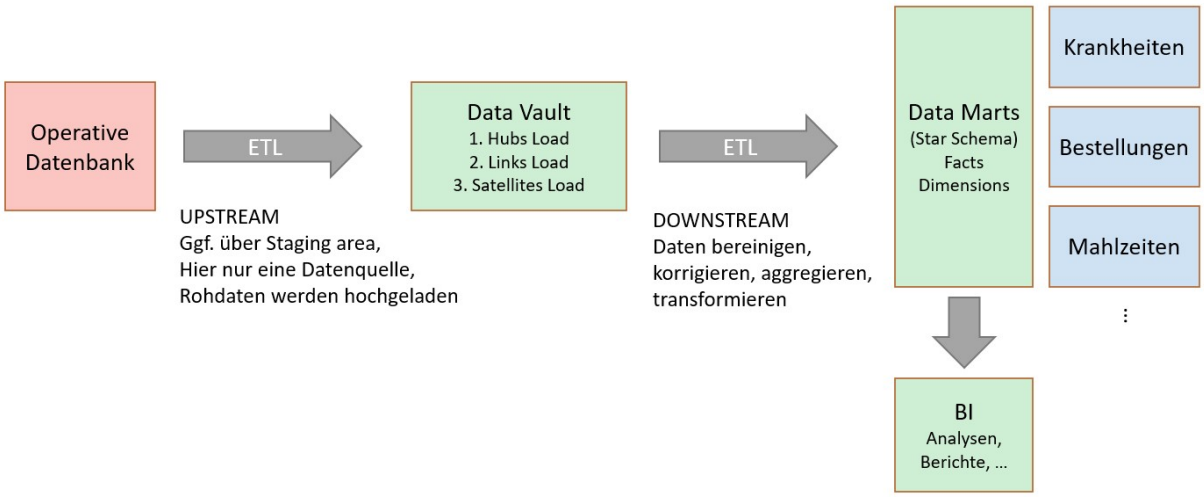
8 Qualitätsanforderung für Datawarehouse

Zur Erleichterung von Business Intelligence Anwendungen wurde ein Datawarehouse (DWH) aufgesetzt. Dies wurde als DataVault konzipiert, nähere Erklärungen dazu finden sich in der entsprechenden Dokumentation.

Das Datawarehouse ist generell besonders anfällig für Qualitätsprobleme, da sich die Fehler der Quelldaten dort akkumulieren. Fehlerhafte ETL-Prozesse (extract transform load), können die Daten weiter verschlechtern. Meist fällt dies dann erst am Ende der Prozesskette, bei der Auswertung in den Fachabteilungen auf, beziehungsweise können kostenintensive Folgefehler entstehen, wenn Fehler nicht bemerkt werden. Daher ist auf eine hohe Datenqualität der Quelldaten, Quellsysteme und Ladeprozesse zu achten. Ersteres versuchen wir mit den Maßnahmen in den vorigen Kapiteln zu gewährleisten.

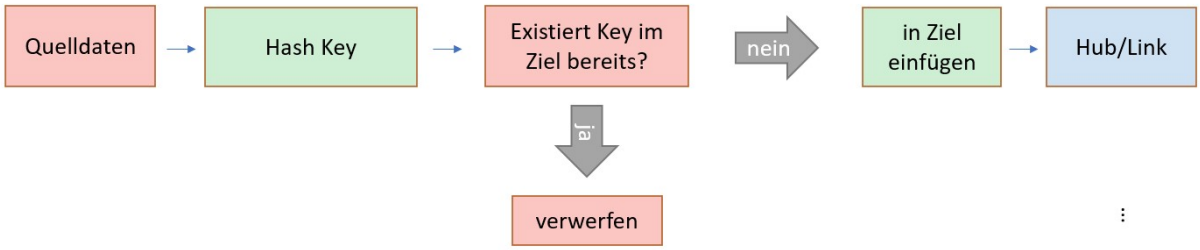
8.1 DataVault, ETL-Prozess und DataMarts

Beim DataVault gibt es zwei Load-Prozesse (ETL), einmal upstream beim Transfer der Rohdaten in den Vault und dann downstream zu den speziellen DataMarts für die entsprechenden Fachabteilungen. Schematisch sieht das wie folgt aus:



Die Ladeprozesse erfolgen automatisiert über geeignete Anwendungen, die sicherstellen, dass der Prozess pausiert, abgebrochen und fortgesetzt werden kann ohne Daten und Konsistenz zu verlieren. Hubs, Links und Satelliten werden standardisiert auf gleiche Art und Weise geladen.

Da die Datenqualität schon bei der Eingabe in die operative Datenbank kontrolliert wird, erspart man sich hier einige Bereinigungsschritte. Upstream werden Rohdaten hochgeladen, es werden HashKeys erzeugt. Diese sorgen auch für die Eindeutigkeit und Redundanzfreiheit der Daten im DataVault und beschleunigen Abfragen:



Es gibt in unserem Fall nur eine Datenquelle (operative Datenbank), was Fehler beim Zusammenführen der Daten verhindert. Es sind Checkpoints integriert, die z.B. prüfen, ob alle erforderlichen Daten geladen wurden (Datum), alle eindeutigen Primärschlüssel vorhanden sind, die Datentypen wie definiert vorliegen und keine Duplikate vorhanden sind. Fehlende Daten werden als NaN behandelt. Falls die Prüfung fehlschlägt, wird eine entsprechende Warnmeldung für die verantwortliche Person (IT) generiert.

Die Richtlinien aus Kapitel 4 und Maßnahmen aus Kapitel 7.3 sollten auch für den DataVault implementiert werden um Fehler zu vermeiden und ggf. frühzeitig zu erkennen (z.B. monatliches Monitoring).

Downstream können die Daten je nach Anwendungsszenario transformiert werden, z.B. kann eine Aggregation je nach Bedarf, z.B. über Zeitintervalle, erfolgen. Es wurden zunächst folgende Use Cases abgebildet:

Faktentabelle	**Dimensionen**	**Use Case**
Krankheit	Arzt Tier Zeit	Auswertung verschiedener Krankheiten für verschiedene Tiere, z.B. - Häufigkeit aller Krankheiten über Tierarten - Zeitverlauf verschiedener Krankheiten - Sind manche Tiere einer Tierart häufiger krank - Ist der Behandlungserfolg mit manchen Medikamenten besser als mit anderen
Bestellung	Futter Futterart Lieferant Zeit	Abbildung der Bestellhistorie, z.B. - Wie oft muss ein bestimmtes Futter nachbestellt werden - Lohnen sich dafür ggf. größere Bestellmengen mit Verhandlungsbasis für Rabatte

Hierfür müssen mindestens folgende Qualitätschecks beim Downstream ETL erfolgen:

- IDs eindeutig (Arzt, Tier, Futter, Lieferant)
- Korrekte Zuweisung über Schlüssel: Futter → Futterart, Tier → Tierart
- Korrekte, eindeutige Schreibweise gemäß Richtlinien (z.B. Namen)
- Einheitliches Zeitformat
- Check der numerischen Werte auf Plausibilität und stichprobenartige Gegenrechnung

8.2 Aktualisierung und Historisierung

Zur Auswertung der Daten im Rahmen von BI sollte ein sinnvolles Intervall bestimmt werden, in dem die Daten in den DataVault hinzugefügt werden. Für die Praxis im Zoo können verschiedene Intervalle sinnvoll sein, so dass veränderte Daten z.B. bei Tieren, Mitarbeitern, Krankheiten und Futterbestände bedarfsgerecht aktualisiert werden:

Uploadfrequenz für verschiedene Entitäten:

Täglich	**Wöchentlich**	**Monatlich**
Futter	Tiere	Ärzte
Mahlzeiten	Tierart	Lieferanten
Bestellung	Lieferanten	Rundwege
Krankengeschichte		Gehege
		Herkunftzoos
		Mitarbeiter
		…

Auch hier sollten die anfänglich festgelegten Frequenzen des Uploads regelmäßig auf ihre Zweckmäßigkeit für BI überprüft werden. In der Anfangsphase sollte nach Rücksprache mit BI und IT häufiger nachjustiert werden, um den Erfordernissen gerecht zu werden.

Der DataVault ermöglich über das Load Date eine vollständige Historisierung der Daten (es erfolgt KEINE Löschung von Daten im DWH).

9 Zugriffsrechte

Bestimmte Bereiche des Datensystems sollten nur für bestimmte Personengruppen zugänglich sein, um einen Eintrag von Fehlern durch ungeübte Personen zu vermeiden. Eine Fachabteilung sollte immer nur Zugriff auf die jeweils benötigten DataMarts haben. Der DataVault mit ETL-Prozessen sollte dem IT-Beauftragten vorbehalten sein. Die Eingabe in der operativen Datenbank sollten wie schon erwähnt nur entsprechend geschulte Mitarbeiter unter Beachtung der oben genannten Regeln und Intervalle übernehmen.

10 Zusammenfassung

Zusammenfassend müssen initial folgende Schritte mit den entsprechenden Qualitätsmaßnahmen durchgeführt werden:

- Festlegung von verantwortlichen Personen (IT und Projektleiter für Datensammlung und -übertrag) und Planung des Ablaufs durch diese Personen
- Einarbeitung dieser Personen in Datenbankstruktur und ERM zur Feststellung benötigter Daten
- Sammlung der vorhandenen Aufzeichnungen und Informationen
- Bewertung auf Relevanz, Duplikate, ggf. Korrektur
- Beschaffung fehlender Informationen, Zwischenbewertung zur Vollständigkeit der Daten und Erreichbarkeit des Qualitätsziels
- Planung des Übertrags durch einen Verantwortlichen Mitarbeiter (Festlegung Anzahl der Mitarbeiter, Aufgabenverteilung, Meilensteintermine, Dokumentation)
- Erstellung eines internen Formatregelwerks
- Schulung der involvierten Mitarbeiter im Umgang mit den hier angegebenen und selbst definierten Konventionen bei der Dateneingabe
- Durchführung der Übertragung in die Datenbank
- Abschließende Bewertung der Datenbank, Test, ev. Korrekturen
- Testlauf und Bestimmung der Datenqualität bzw. Zielerreichung

Im weiteren Betrieb sollten folgende Schritte befolgt werden:

- Festlegung von verantwortlichen Personen, Zeiträume/Fristen und Dokumentationsrichtlinien für die Aktualisierung der entsprechenden Tabellen
- Schulung der involvierten Mitarbeiter im Umgang mit den hier angegebenen und selbst definierten Konventionen bei der Dateneingabe
- Festlegung von Monitoringintervallen und -vorgehensweisen im Hinblick auf die Datenqualität

In Bezug auf das Datawarehouse sollte folgendes beachtet werden:

- Festlegung von verantwortlichen Personen, Zeiträume/Fristen und Dokumentation für die ETL-Prozesse
- Klärung der Zugriffsrechte
- Festlegung von Monitoringintervallen und -vorgehensweisen im Hinblick auf die Datenqualität