Konzept Data Quality

IT-System für den Zoo in Primasens

Kunde: Zoo in Primasens
Ansprechpartner: Herr Theo Teske

IT-Consulting - Wegmann AG

Dokumentversionen:

230418 V01: Ersterstellung des Dokuments nach einem Telefonat mit Herrn Teske und einem ersten Gespräch.
Verfasser: Uwe Dackermann

Inhaltsverzeichnis:

Abkürzungsverzeichnis:

DHW - Data Ware House
BI - Business Intelligence
ERM - Entity Relationship Modell
GUI – Graphical User Interface (Bsp.: Eingabemaske für den Anwender)

A) Einführung:

Ausgangssituation undAnforderungen:

Der Zoo in Primasens besteht seit 36 Jahren. Im Zuge einer umfassenden Modernisierung soll der Betrieb des Zoos „it-gestützt“ ablaufen.
Der Zoo hat zurzeit (17.04.2023) 60 Mitarbeiter, 6000 Tiere. Zur Beschaffung des Futters der Tiere greift der Zoo auf ca. 120 internationale Lieferanten zurück. Bei Krankheit der Tiere wird auf die Hilfe von ca. 50, ausschließlich externen, international tätigen Tierärzten zurückgegriffen. Für die Pflege des IT-Systems werden zukünftig zwei neue Stellen geschaffen.
Der Zoo möchte sich im weiteren Verlauf der Modernisierung zukünftig noch vergrößern. Es ist geplant neue Gebäude und neue Gehege zu bauen. Des Weiteren sollen die IT-Angebote Richtung Kunde ausgebaut werden bzgl. digitaler Zooführungen, Tierpatenschaften und einem Webshop.
Für die Sicherung der Datenqualität soll ein Konzept erstellt werden, das den Datenbestand fortlaufend bearbeitet und die Qualität über ein Level von 97% halten kann.

Die Datenerhebung soll nur durch die Mitarbeiter des Zoos erfolgen und nicht durch externe Quellen.

Ziel und Umfang des Projektes:
Das Projekt „IT-System für den Zoo Primasens“ umfasst folgende Punkte:

- Ein operatives Datenbank system soll die Geschäftsvorfälle, des Zoos abbilden können.
- Die Migratiion der Atldaten sollen in das operative Datenbanksystem eingeladen werden können.
- Es soll ein BI System entwickelt werden, mit dem Analysen, in einem DWH, zu ausgewählten Geschäftsprozessen durchgeführt werden können.

Einschränkungen und ausgeschlossen IT-Leistungen:
Die Gehaltsabrechnung der Mitarbeiter wird extern, von einem Personaldienstleister übernommen und ist nicht Inhalt dieses Projekts.
Die Erstellung des Frontend IT-Systems mit GUI-Masken und Scansytem wird in einem anderen Projekt bearbeitet und ist nicht Inhalt dieses Projekts.

B) Data Quality Konzept

1) operatives Datenbanksystem

a) Auflistung der abzubildenden Geschäftsfälle:

Mit dem operativen IT-System sollen folgende Geschäftsvorfälle abgedeckt werden können:

- Neues Tier anlegen / in diesem Zuge ggf. auch neue Gattung / Tierart anlegen
- Neuen Mitarbeiter anlegen
- Neuen Futterlieferant anlegen
- Pflege der Tiere
- Unterbringung der Tiere
- Geburt von Jungtieren
- Krankheit bei Tieren / Krankheitsverlauf
- Behandlung durch Ärzte
- Ärztevertretung im Krankheitsfall
- Ärztevertretung bei Urlaub
- Mitarbeiter Zuständigkeit für Tierart
- Mitarbeiter konkrete Zuordnung zu Tier
- Mitarbeitervertretung im Krankheitsfall
- Mitarbeitervertretung bei Urlaub
- Fütterung der Tiere
- Bestellung Futter
- Lagerung Futter, Bestandsverwaltung
- Erstellung Rundwege

b) Relevanz und Fokus auf Qualitätskriterien, mögliche Datenfehler und Vorbeugemassnahmen:

Zur generellen Verbesserung der Daten Qualität sind Schulungen der Mitarbeiter durchzuführen, zum Umgang mit dem IT-System und um ein Bewusstsein und Verständnis für die Relevanz der Data Quality Thematik zu schaffen.

Die ***Vollständigkeit*** der Datensätze ist wird als wichtig eingestuft.
Fehler können hier gerade bei der Eingabe der Datensätze entstehen. Fehlende Daten können in allen Geschäftsfällen auftreten.
Um fehlende Werte in den Datensätzen zu vermeiden sind in der Datenbank Bedingungen gesetzt, die unbedingt eine Befüllung der Felder verlangen (NOT NULL).
Des Weiteren werden in den GUI-Masken ebenso vollständige Dateneingaben verlagt, sofern dies möglich ist.

Die ***Eindeutigkeit*** der Daten, dass alle Daten zweifelsfrei interpretiert werden können, wird als wichtig eingesuft und betrifft alle Datensätze. Um die Daten in der Datenbank eindeutig identifizieren zu können ist ein umfassendes Data Dictionary erstellt, das jede Enittät mit ihren Attribuben beschreibt.

Die ***Korrektheit*** der Daten, ob die Daten mit der Realiät übereinstimmen und die **Aktualität**, der Daten, ob die Daten dem aktuellen Zustand der Realität entsprechen ist, ist besonders wichtig, bei der Unterbringung der Tiere und der Fütterung der Tiere, da das Risiko bei Fehlern sehr große Gefahren mit sich bringt. Hier wird im Frontendsystem ein Scansystem eingerichtet, mit dem die Mitarbeiter sowohl die Tiere und die Gehege eindeutig über ihre IDs

identifizieren und umbuchen müssen. So kann gewährleistet werden, das die Fehlerquote in dem Bereich sehr gering ist.
Bei der Datenerhebung der anderen Geschäftsvorfälle werden über GUI-Masken des Frontendsystems werden folgende Massnahmen getroffen, um Fehleingaben zu unterbinden:

- Soweit möglich werden Referenztabellen verwendet, um Schreibfehler zu Vermeiden.
- Eine Abrage nach dem Datenformat und unerlaubte Zeichen oder Ziffern, bei der Eingabe in die einzelnen Datenfelder verhindert, dass die Daten aus Versehen in ein falsches Feld geschrieben werden.

Die **Genauigkeit** der Daten, also die Exaktheit der numerischen Werte, ist in allen Geschäftsvorgängen nicht von besonderer Relevanz. Eine Präzision der Fließkommazahlen mit zwei Stellen nach dem Komma ist ausreichend.

Die **Konsistenz** der Daten, also dass sich Daten nicht widersprechen ist in allen Geschäftsvorgängen wichtig. In der Datenbank sind referentielle Integritätsbedinungen (mit Fremdschlüsseln und automatisch erstellten Primärschlüsseln) implementiert, so dass hier die Datenwiederspruchfrei verwaltet werden.
Bei der Datenerhebung der werden Referenztabellen, so weit möglich verwendet, so dass über die GUI-Masken des Frontendsystems, nur auf eingeschränkte vorgegebene Werte ausgewählt werden können. Dies bewirkt eine einheitliche Schreibweise der Datenwerte.

Die **Redundanzfreiheit,** dass Dupletten von Datensätzen existieren, ist in allen Geschäftsvorfällen wichtig. Vorbeugemassnahmen können hier nur duch Mitarbeiterschulungen getroffen werden.

Die **Relevanz** der Daten, ob die Datensätze dem Informationsbedürfnis auch gerecht werden, ist bei der Konzeption in Absprache mit dem Geschäftsführer Herrn Teske erfolgt und muss bei weiterem Handlungsbedarf in weiteren Modifizierzungen der Datenbank erfolgen.

Die **Einheitlichkeit** der Daten, die einheitliche Strukturierung der Daten, ist dadurch gegeben, da die Daten nur in einem Datenbank System verwaltet werden und externe Datenquellen nicht zugelassen sind.

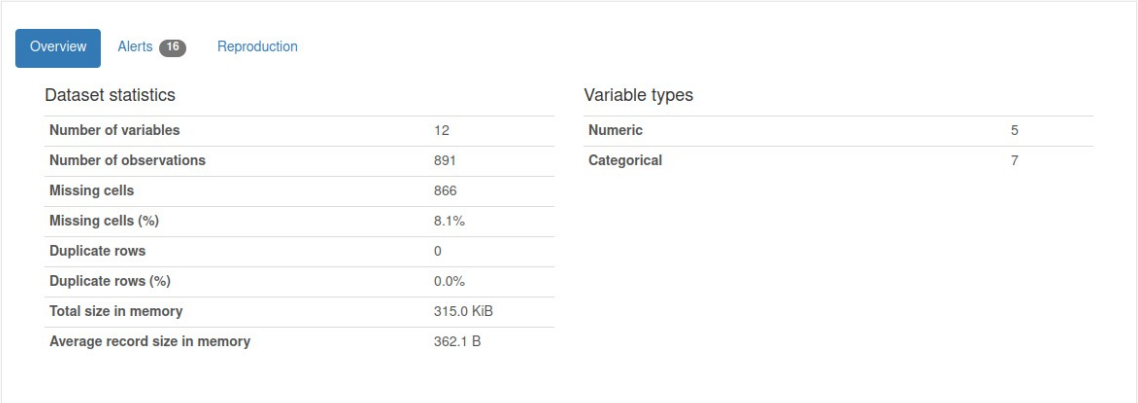
Die **Zuverlässigkeit** der Daten, die Nachvollzihbarketi der Entstehung der Daten, betrifft alle Geschäftsvorfälle und ist ein Thema mit geringerer Relevanz, da keine externen Datenquellen zugelassen sind. Dennoch ist in der Datenbank bei den meisten Entitäten ein Datumsfeld für die Erstellung und der Modifikation der Daten enthalten. Bei den Bestellungen des Tierfutters ist ebnso die ID des Mitarbeiteres hinterlegt.

Die ***Verständlichkeit*** der Daten ist ein wichtiges Thema in allen Bereichen und wird durch eine treffende Bezeichnung der Spaltennamen in der Datenbank umgesetzt, sowie dem Data Dictionary.

c) Erkennung und Behebung von Datenfehlern:

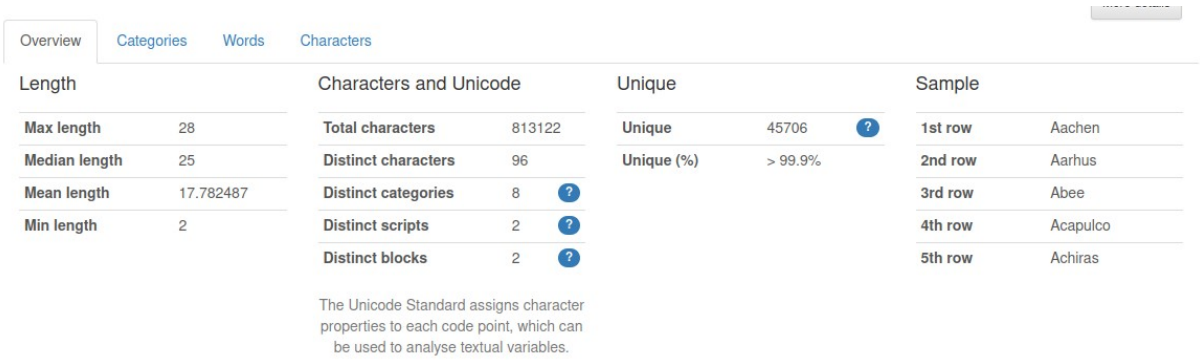
Die Identifikation und Erkennung von möglichen Datenfehlern wird täglich über ein Profiling im operativen Datenbank System durch geführt. Mit hilfe des python scripts db-profiler werden die Daten des operativ Systems analysiert und ein Bericht im .html Format erzeugt. Dem Bericht können Fehler wie Duplikate, fehlende Werte entnommen werden (Siehe Bild1 Bsp: .html Report 1)

Bild1 Bsp: .html Report 1



Des Weiteren sind statistische Auswertungen in dem Bericht enthalten, die genutzt werden können um Ausreißer in Daten zu identifizieren, die auf weitere Fehler hindeuten können, wie Schreibfehler bei der Eingabe oder Formatfehler (Siehe Bild2 Bsp: .html Report 2)

Bild2 Bsp: .html Report2



Eine Mögliche Korrektur nach der Erkennung von Datenfehlern kann manuelle oder teilweise automatisiert erfolgen. Dies ist jedoch noch nicht mit Herrn Teske besprochen worden und muss Fehlerart und Datenfeld spezifisch erfolgen. So könnten beispielsweise bei fehlenden Werten Default werte eingesetzt werden oder bei der Verwendung von ungewollten Sonderzeichen, diese script gesteuert entfernt werden.

Bei den Beständen in den Futterlagern ist regelmäßig eine Inventur durchzuführen, und anschließend der Abgleich mit den Beständen im operativ System erfolgen, um die Aktualität der Daten zu erhöhen. Die zeitlichen Abstände zwischen den Inventurdurchführungen, sind dabei mit dem Anspruch an die Aktualität abzustimmen.

2) Altdatenmigration

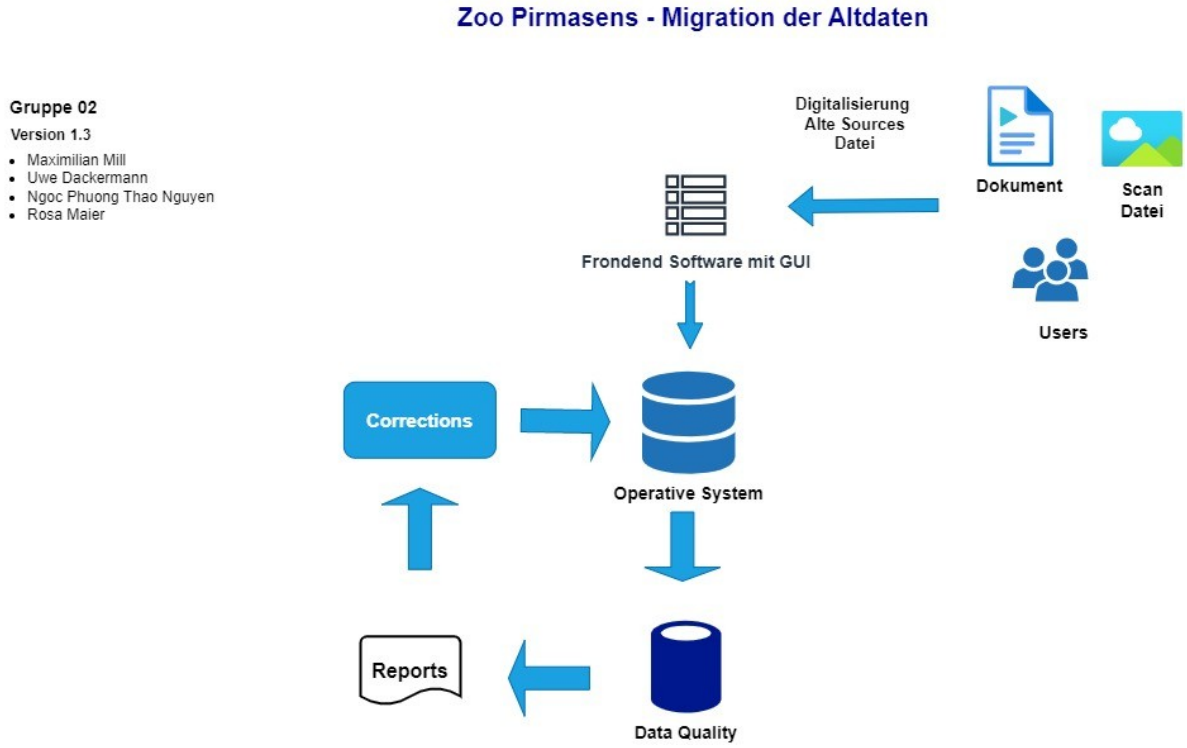
a) Auflistung der Daten für die Integration

- 60 Datensätze von den Mitarbeitern
- 6000 Datensätze von Tieren
- 50 Datensätze von Tierärzten
- 120 Datensätze von Lieferanten

b) Konzept für die Integration der Altdaten

Für die Integration der Altdaten ist das Frontend IT-System zu verwenden, damit die gleichen Fehlervermeidungsmechanismen greifen können, die schon in Abschnitt 1) b) genannt wurden.
Nachdem die Daten in der Datenbank sind wird das in Abschnitt 1) c) verwendete Profiling script verwendet um unentdeckte Fehler zu erkennen und im Report zu dokumentieren (siehe Bild3: Ablaufdiagramm Altdatenmigration).
Anschließend muss dann noch entschieden werden, wie und ob die Fehler korrigiert werden.

Bild3: Ablaufdiagramm Altdatenmigration



3) DWH für Analysen

a) Auflistung der Geschäftsprozesse, die im DWH analysiert werden sollen:

Mit dem DWH sollen folgende Dimensionen bedient werden können:

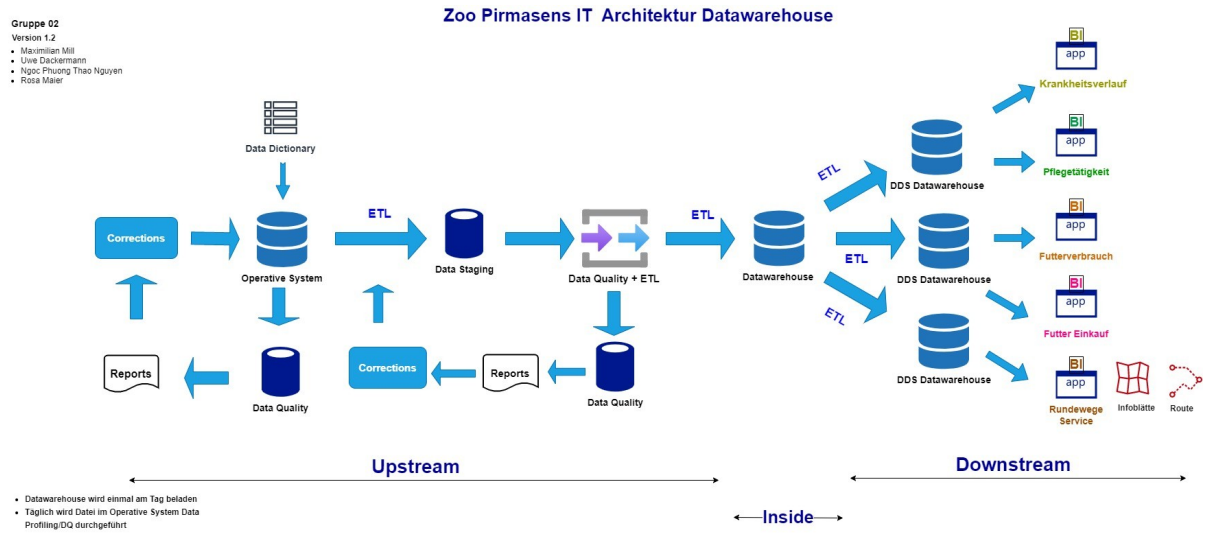
- Krankenverlauf
- Pflege der Tiere
- Futterlieferant
- Futterlagerhalterung
- Bewertung Strecken

b) Fehlervorbeugung, Fehlererkennung und Fehlerbehebung der Datensätze für den ETL Prozess:

Die möglichen Fehler bei dem ETL Prozess sind für den Zoo, dadurch reduziert dass es nur eine Datenquelle für das DWH gibt, nämlich das operativ System. Hier werden in dem in Abschitt 1) beschriebenen Fehlervorbeugungs-, Fehlererkennungs- und Fehlerkorrekturmechanismen angewandt.
So beschränken sich die Fehlerquellen auf den Extraktionsvorgang und den Beladevorgang. Nach dem Extraktionsvorgang werden die Daten einem Data Quality Prozess ähnlich dem des operativ Systems durchlaufen. Hierbei wird wieder ein Python gesteuertes Profing durchgeführt und ein Repot erstellt, in dem diverse Fehler erkannt werden und gegebenenfalls die Korrektur der Daten durchgeführt. Dieser Prozess wird solange durchlaufen, bis die gewünschte Daten Qulität erreicht ist.

Anschließend wird durch ein automatisches Belade script gestartet, das die Datensätze in das DWH lädt (siehe Bild4: IT-Architektur DWH).

Bild4: IT-Architektur DWH



In der Testphase werden die bereits genannten Profiling und Reporting Tools auf die Daten im DWH anwendet, um den Beladevorgang des DWH abzustimmen, um hier mögliche Fehler zu vermeiden.

C) Offene Punkte und Verbesserungsmöglichkeiten

In den Anforderungen wurde zu der Datenqualität gesamtes Level > 97% genannt, das erreicht werden soll. Dieses Level muss noch deutlicher in Zusammenarbeit mit dem Kunden definiert werden. Das tägliche Reporting durch das Profiling script bietet heirbei die Möglichkeit, dieses Level zu quantifizieren.

Bei der Analyse der Routen für den Zoo soll ein Besucherfeedback mit eingehen, bei dem die Struktur und Umfang noch nicht bestimmt sind.

Eine Verbesserung der Datenqualität wurde auch der erweiterte Einsatz von Referenztabellen bringen, um Eingabefehler zu reduzieren. So existiert aktuell noch keine Referenztabelle für Postleitzahlen im Ausland. Weitere Referenztabellen können noch für die Routenerstellung verwendet werden (Bsp. Farbe der Route).