



ITSV GmbH

Datenqualitätssicherung im DWH mit Artificial Intelligence Methoden

Mag. Thomas Schauer,
Dipl.-Ing. Vojkan Radak

Agenda

- Vorstellung IT-Services der Sozialversicherung GmbH
- Grundarchitektur der DWH-Landschaft der SV
- DWH Datenqualität
 - Herausforderungen
 - ... was bisher geschah ?
- DIGNA von dext.ai
 - Wie kann DIGNA helfen ?
 - DEMO
 - Installation („on premise“)
- DQS klassisch <> DIGNA

IT-Services der Sozialversicherung GmbH



Wir digitalisieren das österreichische Gesundheitswesen. **#herzderdigitalisierung**

**” 100%iges Tochterunternehmen der österreichischen Sozialversicherung
Full-Service IT-Provider**

Wir begleiten **9 Millionen ÖsterreicherInnen** ein Leben lang.

Als Management-GmbH zur Entwicklung der IT-Strategie, übergreifenden Steuerung und Koordinierung ist unser primärer Auftrag Kundenorientierung, Datensicherheit und Kosteneffizienz.

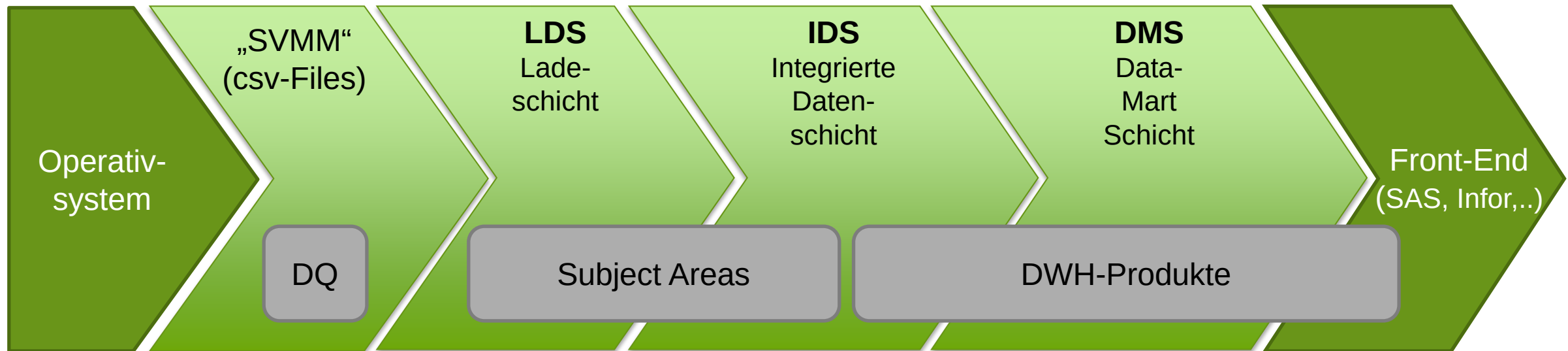
So konnten seit 2007 über 390 Mio. Euro IT-Kosten in der Sozialversicherung eingespart werden.

Meilensteine.

ITSV GmbH



Grundarchitektur der DWH-Landschaft der SV



DWH Datenqualität

- DQ ist ein „Thema“ ...
- jede/r redet darüber...
- üblich: eher negativ ...
- DQ wird in Unternehmen bemerkt, erst wenn es „zu spät“ ist...
- Maßnahmen zur „Verbesserung der DQ“ werden getroffen ...

um dann ziemlich oft nicht ernst
genommen zu werden !

DWH Datenqualität

- ITSV DWH „**Data Lake**“ in Zahlen
 - 40 Lieferanten der Quelldaten
 - 150.000 Lieferungen pro Jahr
 - 450 verschiedene Strukturen
 - 6.500 Felder
 - 50 Milliarden Datensätze
- Davon Abgeleitete **DWH/s** in Zahlen
 - 30 Abnehmer
 - 200 Strukturen
 - 200 Milliarden Datensätze

Wie kann Datenqualität in so komplexen Systemen kontrolliert und garantiert werden ?

Schwer ... 😊

DWH Datenqualität

■ Herausforderungen

- Datumsfelder mit vielen verschiedenen Formaten geliefert ✓
- Referential Integrity (PK, FK, ...) wenn nicht in Quellsystemen erledigt ✓
- Numeric Felder mit CHAR Inhalten ✓
- Unüblich andere Attribute, Datenformate, ... in Datenlieferungen
- Lieferungen verspart (z.B. 3/4 Tage, monatlich ...) an
- Neue, unbekannte Attribute, die nicht in den Datenlieferungen vorkommen ... Berichtsinkonsistenz
- Datenmengen (hunderte Strukturen, tausende Felder, ...)

■ Was bisher geschah? ... Wir haben es wirklich versucht ...

- Checks
 - Felder: Datum, Numeric, Längen, ..., PK, ..., FK, ... ✓
- Bei Produkten, die als „kritisch“ gezeichnet sind, obere/untere Grenzen der Datensatzanzahl bei gewählten Felder, definiert und auf „Ausreißer“ geprüft
- Verschiedene statistische Methoden: MAX, MIN, AVG, %, Quantile, K-Means, ... ausprobiert
- ...

DWH Datenqualität

DIGNA (dext.ai) !

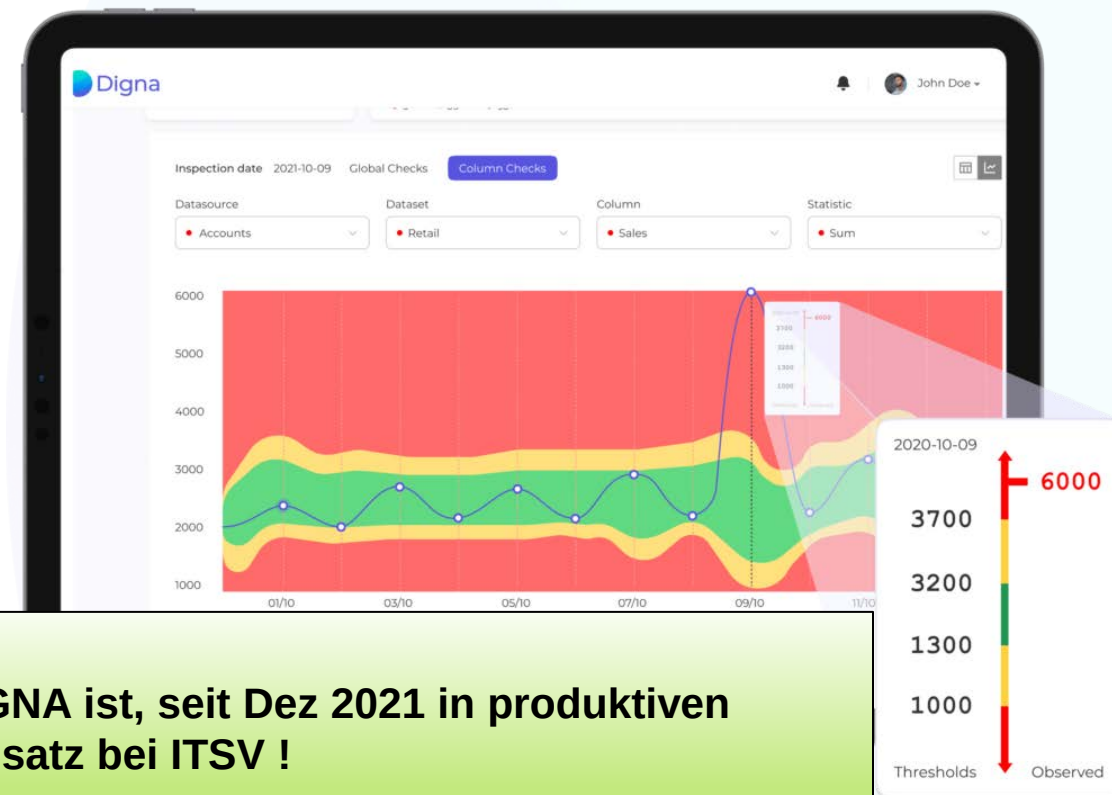
- Fragen ?...?...?
 - Wie **ändern** sich die „Grenzen“ ?
 - Welche Statistik, Funktion, Analytische Algorithmus ... folgen sie ?
 - Was ist „**normal**“, „üblich“ für eine Lieferung ?
 - Zyklen (Zeit)
 - Mengen (SUM , Anzahl)
 - Wie verwalten wir den Vielfalt der **unterschiedlichen** „Normalitäten“?
 - Haben wir alle Fehlerquellen und **(Un)Üblichkeiten** erwischt ?



Intelligent Data Inspection

Digna ist die **Data Quality Platform**, die Anomalien und Fehler in deinen Daten mit **Künstlicher Intelligenz** entdeckt.

Startup aus Wien mit langjähriger Expertise in den Bereichen Data Science und Computer Science.



DIGNA ist, seit Dez 2021 in produktiven Einsatz bei ITS V !

Gefördert von
mit über €200,000



 **Bundesministerium**
Digitalisierung und
Wirtschaftsstandort



DIGNA product by dext.ai

■ Wie kann DIGNA helfen ?

- Auf Grund der täglichen Analysen und ständigen **Modelanpassungen**, erkennt DIGNA „**Individualität**“ der Lieferungen und bestimmt automatisch was für eine Lieferung „normal“ und „üblich“ ist
- Eine Lieferung bekommt sowas wie eine Muster, ein „Fingerprint“ und alles was diesen nicht entspricht, wird als „**Anomalie**“ erkannt
- Untere und obere **Grenzen** bei allen gelieferten Inhalten werden **automatisch verwaltet/angepasst**
- **Lieferungsverspätungen** werden erkannt
- Erkennung der **neuen Inhalte** bei „slow changing“ Dimensionen
- Erkennung der neuen Inhalten bei „**very slow**“ Lieferungen („Quartalsdaten sind da“)

...und das alles, systematisch, für:
alle Strukturen (450)
alle Felder (6500)
alle Lieferungen (150 000 pro Jahr)

Lieferung: Grenzen tanzen aus der Reihe ...



Üblich: Grenzen

Anomalie. Wert nicht in Grenzen

Lieferung – Muster



Übliche Muster: „Zickzack“

Anomalie zu: „Zickzack“
Muster

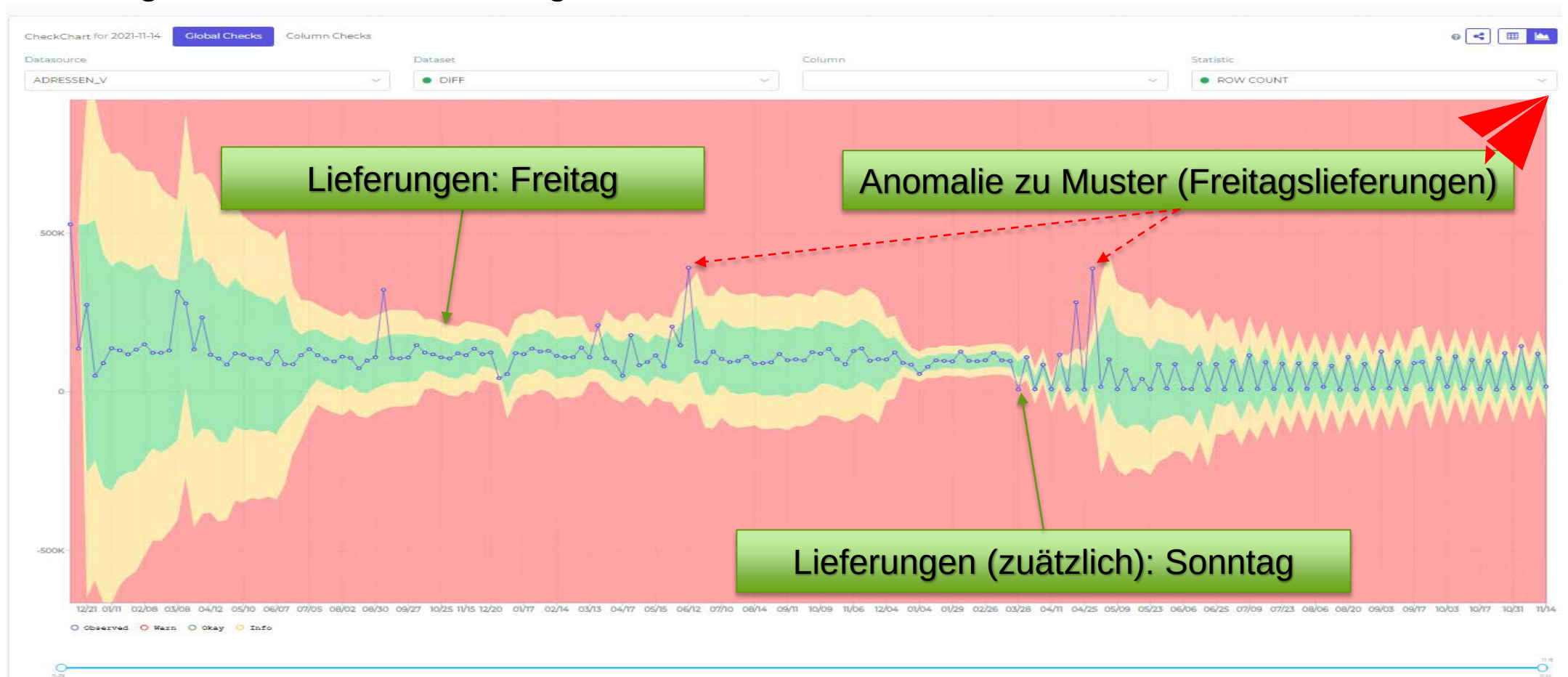
Lieferungsverspätungen



Üblich: 1 - 3 Tage

Nicht üblich: 5, 6, 11, ... Tage

Lieferung – neue Mustererkennung



Installation („on premise“)

- DIGNA Software Installation: ~ **1 Stunde 15 Minuten**
- ITSV Vorarbeit **pro Projekt** (derzeit ~ 6 Projekte produktiv)
 - **0.5 Tage Data Profiling** (Kategorisierung der Felder: „numeric“, „categorical“, „other“)
 - **0.5 Tage Meta Data Definition** in DIGNA
- DIGNA **Inspection/Daten** Sammlung
 - **20 - 30 Min täglich** bei täglichen Zuwachs von cca 50 Millionen Datensätze (in allen 6 Projekten)

ITSV DIGNA – Fakten

- Seit Dezember 2021 in **produktiven** Einsatz
- Entdeckte **Anomalien** bisher:
 - Drei verspätete Lieferungen
 - Fehler an Lieferanten gemeldet, die Export Routinen korrigiert
 - Job Scheduler restartet
 - Neue Lieferung von einer bisher unbekannten Quelle („Code: Art der ...“)
 - Wird in nächsten Tagen in ITS V Prozesse integriert
- **Ersparnis:**
 - ~ **einige Tage** (DWH Korrekturen, Reloads, Berichtskorrekturen, ..., erspart)
- ... und das ist erst der Anfang ! 😊



DQS klassisch

PKs, FKs, ... können geprüft werden

Feldtypen Check: numeric, date, char, int, ...

Für **bekannte, schon passierte** Fehler,
können Warnungen ausgegeben werden

Auf **bestimmte**, bekannte oder **vermutete**
Fehlerquellen kann Aufmerksamkeit geweckt
werden

DQS DIGNA

Lernt **dynamisch** was „**üblich**“ ist und leitet
davon die „**Anomalien**“ ab

Dadurch können globale Analysen für **alle**
Daten gemacht werden.

Auf bisher **unbekannte** „**Anomalien**“ kann
Aufmerksamkeit geweckt werden

**DIGNA spart Zeit und
hältet uns wach !**





**Herzlichen Dank für Ihre
Aufmerksamkeit!**