sphinx
THE DIGITAL SPARK

Data Management heute und morgen:
Ein pragmatischer Streifzug durch den Dschungel der Begrifflichkeiten

Dr. Thomas Petrik

Wien, Juni 2024

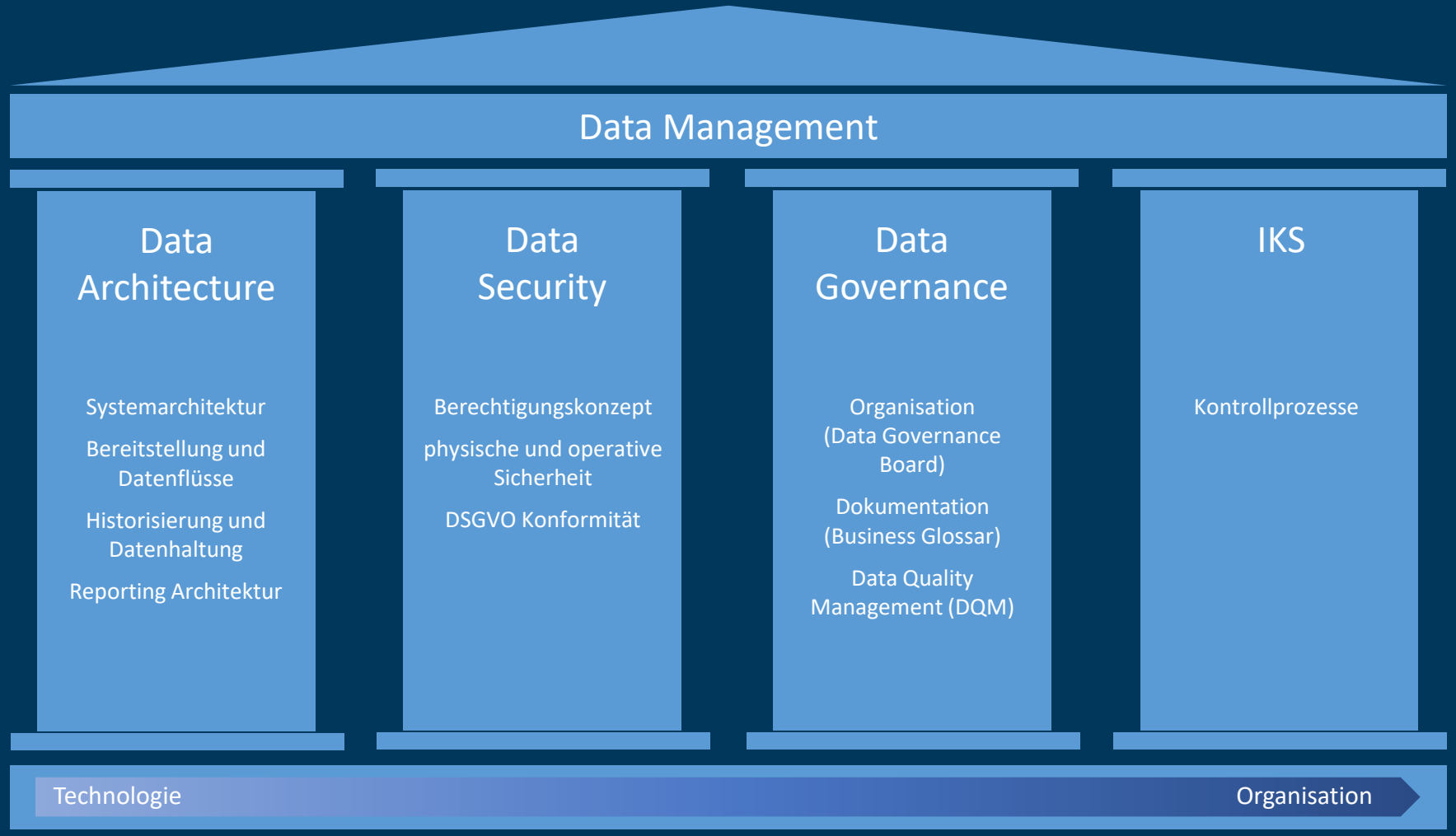
Dr. Thomas Petrik
Sphinx IT Consulting GmbH
Head of Technology Consulting

High Performance Analytics

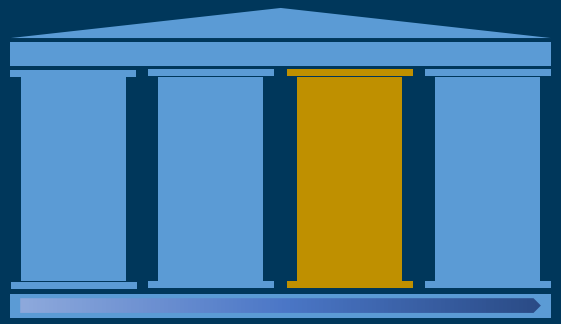
Database Architectures

Database Security

Die Säulen des Data Management


Data Management
Data Architecture
Systemarchitektur
Bereitstellung und Datenflüsse
Historisierung und Datenhaltung
Reporting Architektur
Data Security
Berechtigungskonzept
physische und operative Sicherheit
DSGVO Konformität
Data Governance
Organisation (Data Governance Board)
Dokumentation (Business Glossar)
Data Quality Management (DQM)
IKS
Kontrollprozesse
Technologie
Organisation

Die Data Governance Säule



Data Governance: Organisation

Data Governance: Rollen

Data Steward (DS)

- operative Tätigkeit
  - Datenpflege
  - Datenqualität
  - Dokumentation
- Expertise
  - fachlich
  - technisch
    - Datenflüsse
    - Datenhaltung

Data Owner (DO)

- Hoheit über zugewiesene Attribute
  - in den Quellsystemen
  - in den nachgelagerten Systemen
- primäre Auskunftsperson
  - zu fachlichen Fragen
    - Berechnungsmethoden
    - Querbeziehungen zu anderen Attributen
  - zu technischen Fragen
    - Formate, Wertebereiche, etc.
- DQ-Verantwortlicher
  - DQ-Reporting
  - DQ-Regelwerk
- Dokumentations-Verantwortlicher
- Einbindung im Change-Prozess
  - Review & Freigabe
- Freigaberolle im Berechtigungsprozess
  - Definition der Security Attribute
  - Kooperation mit Datenschutz

Data Governance Office

Data Governance Officer (DGO)

- Leitung des Data Governance Boards
  - Geschäftsordnung
  - Bericht an das Management Board
- Koordination von DG-Projekten
- Ernennung durch das Management Board

Data Governance Board (DGB)

- besteht aus DGO + DO
- Ziele
  - Förderung der Kollaboration durch Schaffung der organisatorischen Voraussetzungen für Data Sharing.
  - **Schaffung von Transparenz durch Aufbau und Pflege eines Business Glossars.**
  - Verbesserung von Datenschutz und Compliance durch Einpflegen von Datenklassifizierungen.
  - Regelung von Verantwortlichkeiten, insbesondere die Zuordnung von Data Ownern.
  - Steigerung der Datenqualität: Überprüfung bzw. Plausibilisierung der Konsistenz, Integrität, Aktualität sowie der zeitgerechten Bereitstellung der Daten.
- Schulungen
- Abstimmung mit Security, BCM, IT Governance, IKS

Business Glossar

Basisinformationen:
Name des Begriffs, Definition,
Synonyme, Wertebereiche, Formate

Verantwortlichkeit:
Data Owner

Kontextinformationen:
Verlinkung auf verwandte Begriffe,
Relevanz im Unternehmen

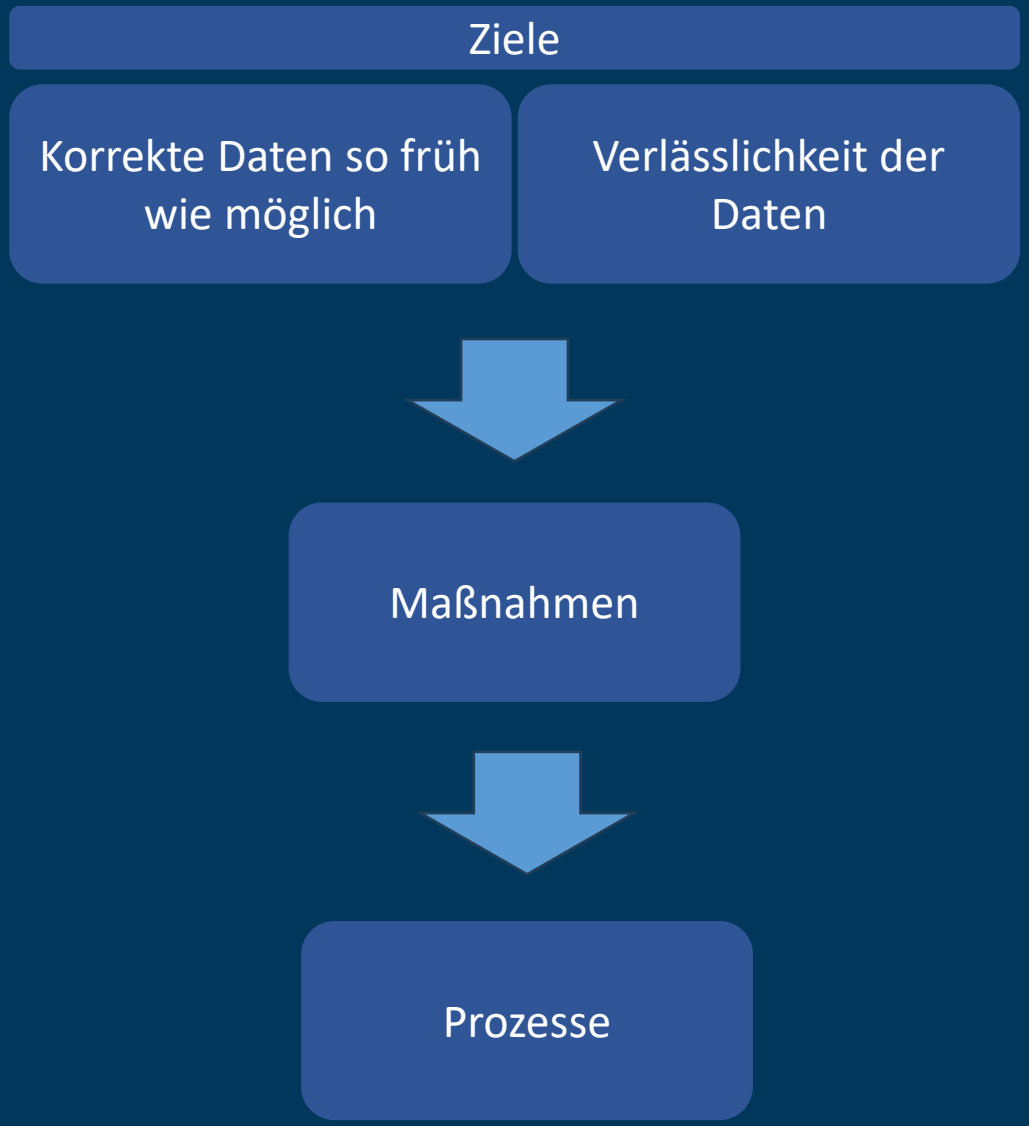
Datenklassifizierung:
z.B. öffentlich, vertraulich, streng vertraulich

DSGVO-Relevanz:
Verweis auf Datenschutzrichtlinien,
Aufbewahrungsfristen, zulässige
Anonymisierungsverfahren

Data Lineage

Data Quality

Data Quality Management



DQM Maßnahmen

DQ-Regeln

- technisches Profiling
  - Wo?
    - im ETL-Prozess
    - am Quellsystem selbst
  - Was?
    - techn. Plausibilisierung
      - Verteilungen
      - Formate
      - Wertebereiche
    - Abgleich mit Quelldaten
- fachliche Plausibilisierung
  - Basisattribute und abgeleitete KPIs
  - Konsistenz
  - Sollwerte & Thresholds

DQ-Reporting

- Überblick über aktive Regeln
- DQIs und Schwellwerte
- Issues
- Empfehlungen

Ad hoc Korrekturen

- Wo?
  - an der Quelle
  - im DWH

  Permanente Modifikationen im ETL-Prozess zur Verbesserung der DQ sind keine ad hoc Korrekturen!
- Wie?
  - strikt auditiert
  - nachvollziehbar
  - dokumentiert (Freigabeprozess)

Die Reproduzierbarkeit des Reportings darf dadurch nicht beeinflusst werden.

Exkurs:
Historisierung

Historisierung – wozu?

100% reproduzierbares Reporting

fachlich

An welchem Buchungstag war der Satz gültig?

Buchungstagslogik

Business Date (BDate)

technisch

Wann kam der Satz ins DWH?

systimestamp

Technical Date (TDate)

Stammdaten & Eventdaten

Stammdaten

- andauernde Gültigkeit
- z.B. Kontodaten, Kundendaten

KundenNr	Name	BDate
123	Müller	2023-11-21
123	Müller	2023-11-22
123	Mayer	2023-11-23

KundenNr	Name	BDate_From	BDate_To
123	Müller	2023-11-21	2023-11-22
123	Mayer	2023-11-23	9999-12-31

fachlich:
Erweiterung auf einen Gültigkeitsbereich

technisch:
Deduplizierung

Eventdaten

- nur zu einem bestimmten Zeitpunkt gültig
- z.B. Transaktionsdaten, Umsätze
- BDate ergibt sich unmittelbar aus dem Transaktionsdatum

RE_Nr	Name	RE_Datum	BDate
100	Popcorn	2023-11-21 17:21	2023-11-21
100	Nachos	2023-11-21 17:21	2023-11-21
…	…	…	…
187	Cola	2023-11-23 20:11	2023-11-23

Das SCD-Dilemma

- SCD1 - "Slowly Changing Dimension" Type 1
  - keine Historisierung, nur Updates
  - **sollte in einem modernen DWH nicht mehr existieren!**
- SCD2 – "Slowly Changing Dimension" Type 2
  - von/bis-Logik (Effective Date, Business Date)
  - ursprünglich gedacht für Historisierung von Dimensionen, die sich nur langsam verändern
  - Einsparung von Speicherplatz

Beispiel: Kontosaldo
ändert sich fast täglich

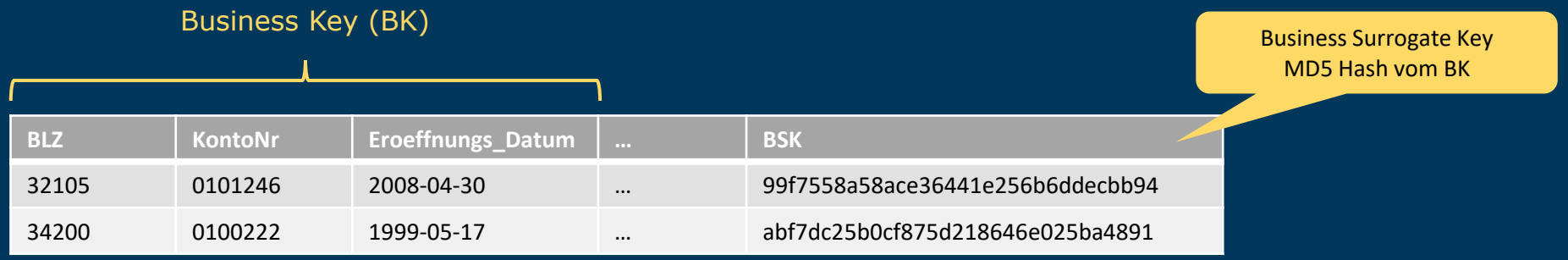
von/bis-Logik oder als Event-Zustand?

Beispiel: Kontosaldo
ändert sich fast täglich

Speicherung für einen Gültigkeitszeitraum (von/bis-Logik) trotz hoher Änderungsrate
Der Saldo gilt ununterbrochen bis zur nächsten Änderung.

fachliche Entscheidung

Der Business (Surrogate) Key



BLZ	KontoNr	Eroeffnungs_Datum	...	BSK
32105	0101246	2008-04-30	...	99f7558a58ace36441e256b6ddecbb94
34200	0100222	1999-05-17	...	abf7dc25b0cf875d218646e025ba4891

- BSK ist die Basis für die Historisierung
  - es wird die Veränderung von Attributen für den selben BSK betrachtet
- BK ergibt sich meist aus dem PK der Quelle
  - nicht zwingend
  - oft kein PK vorhanden
  - PK der Quelle enthält manchmal zeitabhängige Information
    - vor allem in bereits historisierten Systemen

Technische Historisierung

Quelle

KundenNr	Name	BDate
123	Müller	2023-11-15

KundenNr	Name	BDate
123	Müller	2023-11-16

keine Änderung
Satz ignoriert

KundenNr	Name	BDate
123	Mayer	2023-11-17

Korrekturbuchung

KundenNr	Name	BDate
123	Meier	2023-11-17

DWH

KundenNr	Name	BDate_From	BDate_To	TDate_From	TDate_To
123	Müller	2023-11-15	9999-12-31	2023-11-15 23:15:41	9999-12-31

KundenNr	Name	BDate_From	BDate_To	TDate_From	TDate_To
123	Müller	2023-11-15	9999-12-31	2023-11-15 23:15:41	**2023-11-17 23:16:00**
123	**Müller**	**2023-11-15**	**2023-11-16**	**2023-11-17 23:16:00**	**9999-12-31**
123	**Mayer**	**2023-11-17**	**9999-12-31**	**2023-11-17 23:16:00**	**9999-12-31**

KundenNr	Name	BDate_From	BDate_To	TDate_From	TDate_To
123	Müller	2023-11-15	9999-12-31	2023-11-15 23:15:41	2023-11-17 23:16:00
123	Müller	2023-11-15	2023-11-16	2023-11-17 23:16:00	9999-12-31
123	Mayer	2023-11-17	9999-12-31	2023-11-17 23:16:00	**2023-11-18 08:00:00**
123	**Meier**	**2023-11-17**	**9999-12-31**	**2023-11-18 08:00:00**	**9999-12-31**

Technische Historisierung

Abfrage des Zustands vom Buchungstag 16.11., der am 17.11. um 8:30 gültig war:

```
SELECT *
  FROM kunde
 WHERE      '2023-11-16' BETWEEN bdate_from AND bdate_to
       AND '2023-11-17 08:30:00' BETWEEN tdate_from and tdate_to;
```

KundenNr	Name	BDate_From	BDate_To	TDate_From	Tdate_To
123	**Müller**	**2023-11-15**	**9999-12-31**	**2023-11-15 23:15:41**	**2023-11-17 23:16:00**
123	Müller	2023-11-15	2023-11-16	2023-11-17 23:16:00	9999-12-31
123	Mayer	2023-11-17	9999-12-31	2023-11-17 23:16:00	2023-11-18 08:00:00
123	Meier	2023-11-17	9999-12-31	2023-11-18 08:00:00	9999-12-31

100% reproduzierbar

sphinx
MASTERING THE GAP

Dr. Thomas Petrik

E thomas.petrik@sphinx.at
M +43 664 155 8304
T +43 1 599 31- 0

Sphinx IT Consulting GmbH
Aspernbrückengasse 2
1020 Wien

www.sphinx.at