Start Your Roadmap to AI Success on a Solid Foundation

Dr. Thomas Petrik

Wien, Juni 2024

Dr. Thomas Petrik
Sphinx IT Consulting GmbH
Head of Technology Consulting
Management Board Member

High Performance Analytics

Database Architectures

Database Security

Attributes of a Solid Data Foundation

Data Quality Management

Data Quality Management

DQ-Rules

- technical Profiling
  - Where?
    - in the ETL-process
    - in the source
  - What?
    - techn. validation
      - distributions
      - formats
      - value ranges
    - compare to source
- content validation
  - base attributes and derived KPIs
  - consistency
  - target values & thresholds

DQ-Reporting

- active rules overview
- DQIs thresholds
- issues
- recomendations

Ad hoc Corrections

- Where?
  - in the source
  - in the DWH

  Regular corrections within the ETL process are transformations and not ad hoc corrections!
- How?
  - strictly audited
  - reproducable
  - documented (approval process)

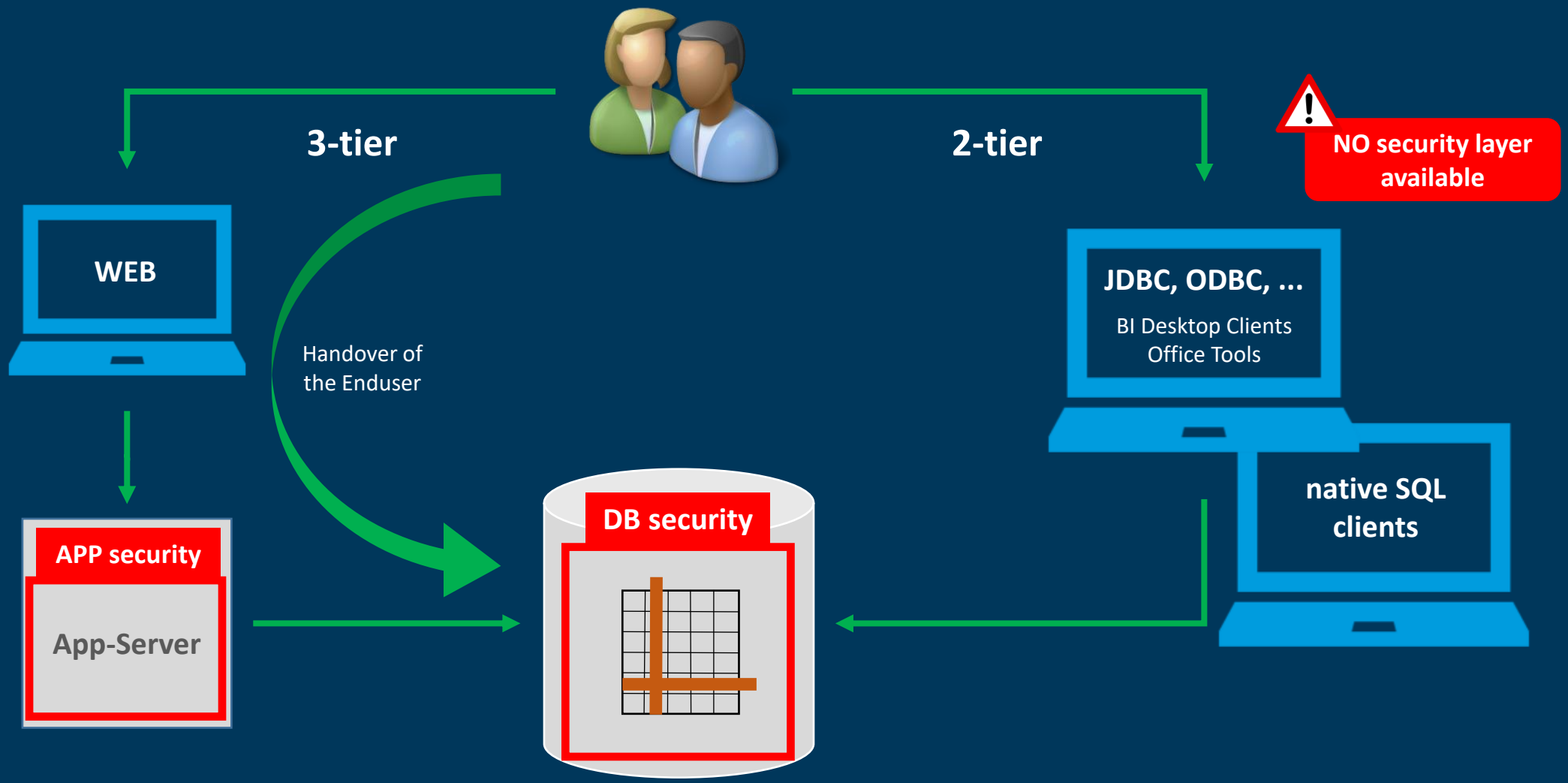
The reporducabilty of the reporting must not be compromised!.

Levels & Goals of Database Security

Basic Principals

1. Security is always bound to the human user
2. Security is tool-agnostic
3. Audit shows the human user on each and every access path
4. Human users cannot delegate their access rights without permission
5. Security must not prevent users from collaboration in various projects

Access Paths – The Big Picture



DB-Security: 3 Levels

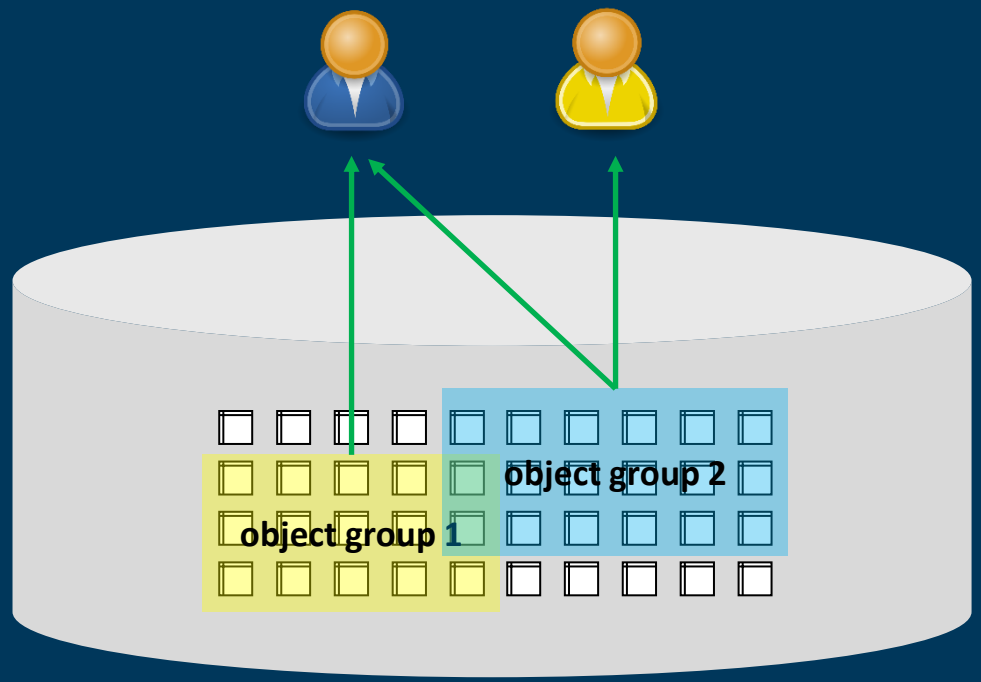
Object Level Security
OLS

Row Level Security
RLS

Column Level Security
CLS

OLS – Define Object Groups

- object groups contain tables and views
- object groups may be overlapping
- object groups may be related to roles 1:1
- object groups are associated with users



Example: Row Level Security

INSTITUTE	DEPARTMENT	…	AMOUNT
103	DEP2	…	300
196	DEP2	…	400
104	DEP2	…	10000

INSTITUTE	DEPARTMENT	…	AMOUNT
196	DEP2	…	400

select * from table;

department=DEP2

institute=196 and department=DEP2

INSTITUTE	DEPARTMENT	…	AMOUNT
196	DEP1	…	5000
103	DEP2	…	300
196	DEP2	…	400
104	DEP2	…	10000

Example: Column Level Security

INSTITUTE	DEPARTMENT	…	AMOUNT
196	DEP1	…	(null)
103	DEP2	…	300
196	DEP2	…	(null)
104	DEP2	…	10000

INSTITUTE	DEPARTMENT	…	AMOUNT
196	DEP1	…	(null)
103	DEP2	…	(null)
196	DEP2	…	(null)
104	DEP2	…	(null)

select * from table;

sensitive columns only for INSTITUTE 103,104

no sensitive columns at all

INSTITUTE	DEPARTMENT	…	AMOUNT
196	DEP1	…	5000
103	DEP2	…	300
196	DEP2	…	400
104	DEP2	…	10000

Historisation

Historization – what for?

100% reproducable Reporting

business point of view

On which business date was the row valid?

Business Day Logic

Business Date (BDate)

technical point of view

When did the row arrive in the DWH?

systimestamp

Technical Date (TDate)

Master Data & Event Data

Master Data

- valid for a period of time
- e.g. account data, customer data

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

business-related: extend valifity to a period of time

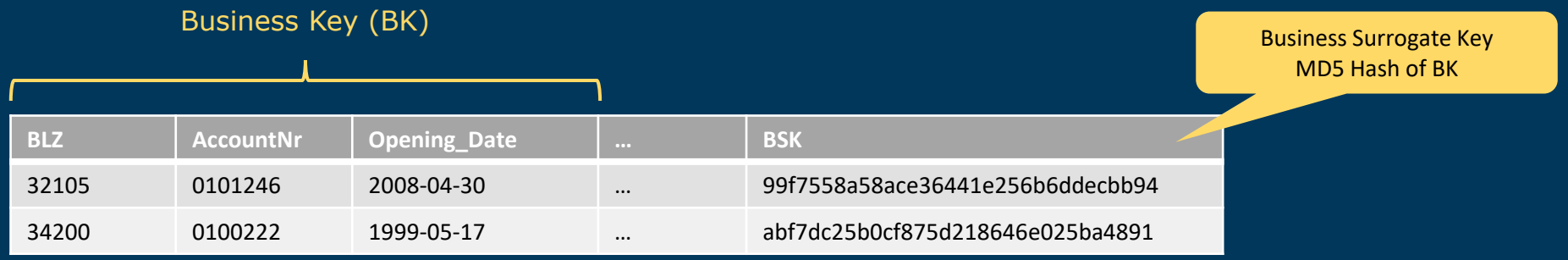
technical: deduplication

Event Data

- valid only at a specific point in time
- z.B. transactions, purchases
- BDate results from the transaction date

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

The Business (Surrogate) Key



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

- BSK is the basis for historisation
  - monitor the changes over time for the same business key
- BK mostly results from the primary key (PK) in the source
  - but: not always!
  - sometimes no PK is defined
  - PK in the source may contain time dependend information
    - especially if the source is already somehow historised itself

Technical Historisation

Source

CustNr	Nam	BDate
123	Müller	2023-11-15

CustNr	Nam	BDate
123	Müller	2023-11-16

no change
row ignored!

CustNr	Nam	BDate
123	Mayer	2023-11-17

Correction

CustNr	Nam	BDate
123	Meier	2023-11-17

DWH

CustNr	Nam	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

CustNr	Nam	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**

CustNr	Nam	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**

Technical Historisation

Ask for the status as of the booking date 16.11., which was valid at 8:30 am on 17.11.

```
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;
```

CustNr	Nam	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% reproducable

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