

Data Vault

Pentaho BeNeDutch 2011, 24-11-2011

Kasper de Graaf / Edwin Weber

DIKW Academy

WWW.DIKW-ACADEMY.NL

FROM DATA TO INFORMATION TO KNOWLEDGE TO WISDOM

Architecture



WWW.DIKW-ACADEMY.NL

2

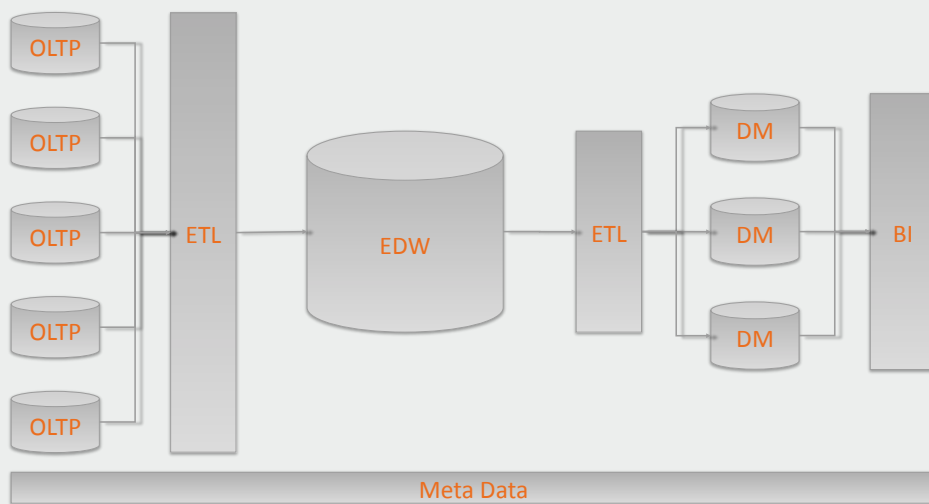
FROM DATA TO INFORMATION TO KNOWLEDGE TO WISDOM

Inmon, Kimball, History of Data Warehousing

THE GREAT DEBATE

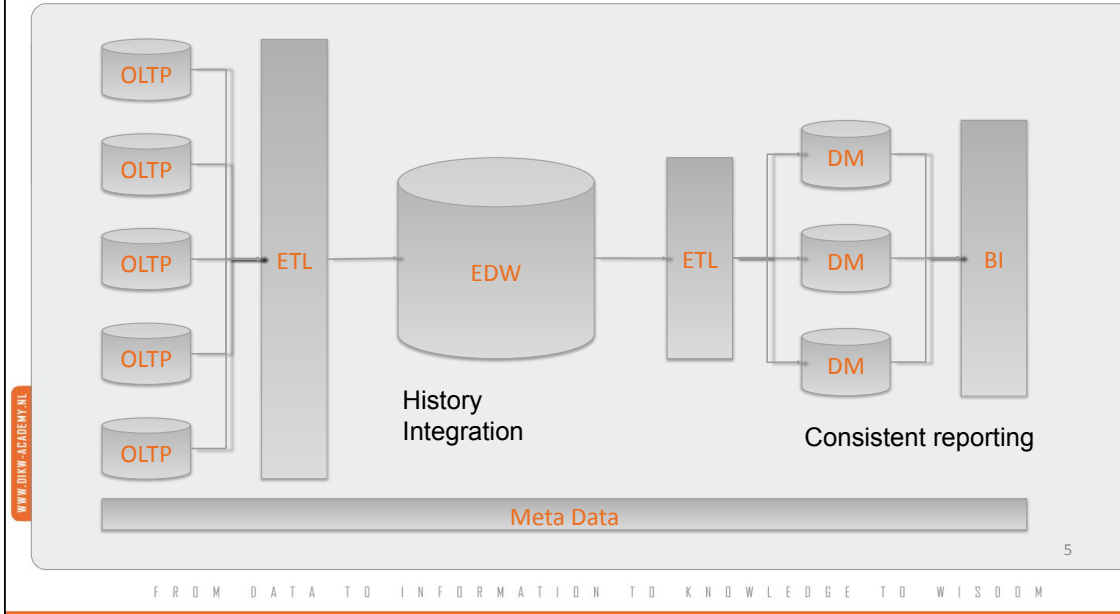
WWW.DIKW.ACADEMY.NL

Corporate Information Factory (CIF)



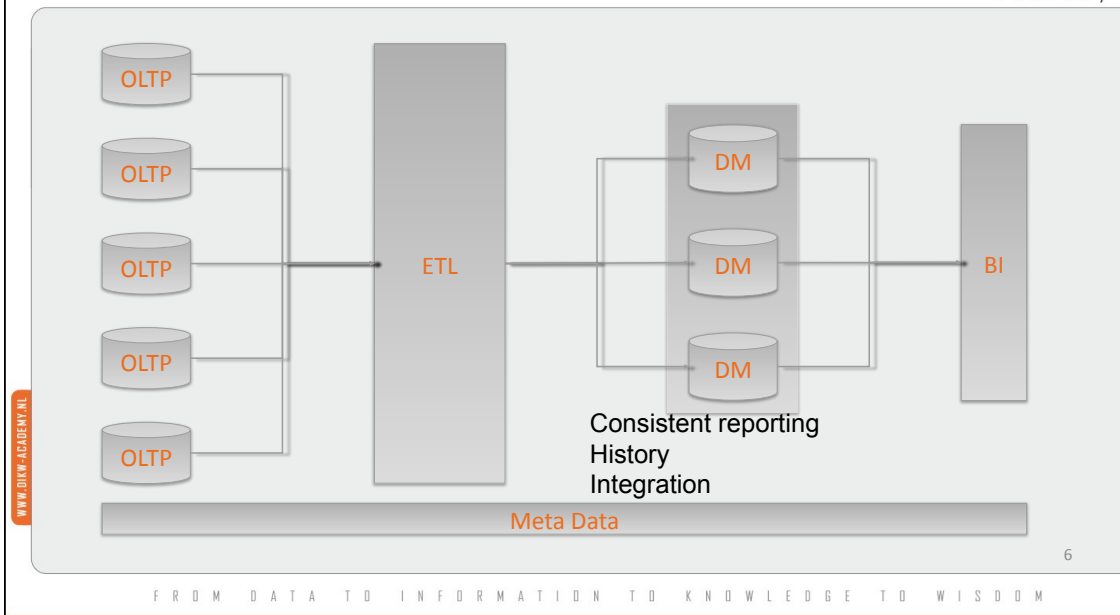
WWW.DIKW.ACADEMY.NL

CI: Inmon



FROM DATA TO INFORMATION TO KNOWLEDGE TO WISDOM

Bus Architecture: Kimball



FROM DATA TO INFORMATION TO KNOWLEDGE TO WISDOM

Inmon versus Kimball

Bill Inmon



- Father of DW'ing
- Corporate Information Factory
- Enterprise Data Warehouse

Ralph Kimball



- Father of Dimensional Modeling
- Bus Architecture
- Σ (Data Marts) = EDW

Inmon Characteristics

- 3NF Enterprise Data Warehouse
 - Tricky!
- Data Marts are 100% dependent on EDW
- Data Marts are disposable

Kimball Characteristics

- Lifecycle Toolkit
- No EDW
- DM + DM + DM = EDW
- Dimensional Everything!

BI & DW

MATURITY

Maturity of BI 1/2

- Organisations are currently struggling with their BI-initiatives in terms of:
 - Data Quality
 - Compliance
 - Traceability (enabling audits)
 - Scalability
 - Sustainability
- The current methodologies and techniques used for data warehousing were developed 15 years ago and never took auditability or compliance into account



Maturity of BI 2/2

- During these years the business has evolved in terms of the need for compliancy and traceability
- Many BI-solutions will FAIL any audit or certification against Compliance & Traceability requirements (whether this is ISO, SOX, BASELII, ...)

Business' requirements for DW/BI

- Integrated data
- Historically correct (temporal)
- Performance
- One version of the truth
- Compliant and traceable
- IT-agility
- High quality
- Understanding the data



13

FROM DATA TO INFORMATION TO KNOWLEDGE TO WISDOM

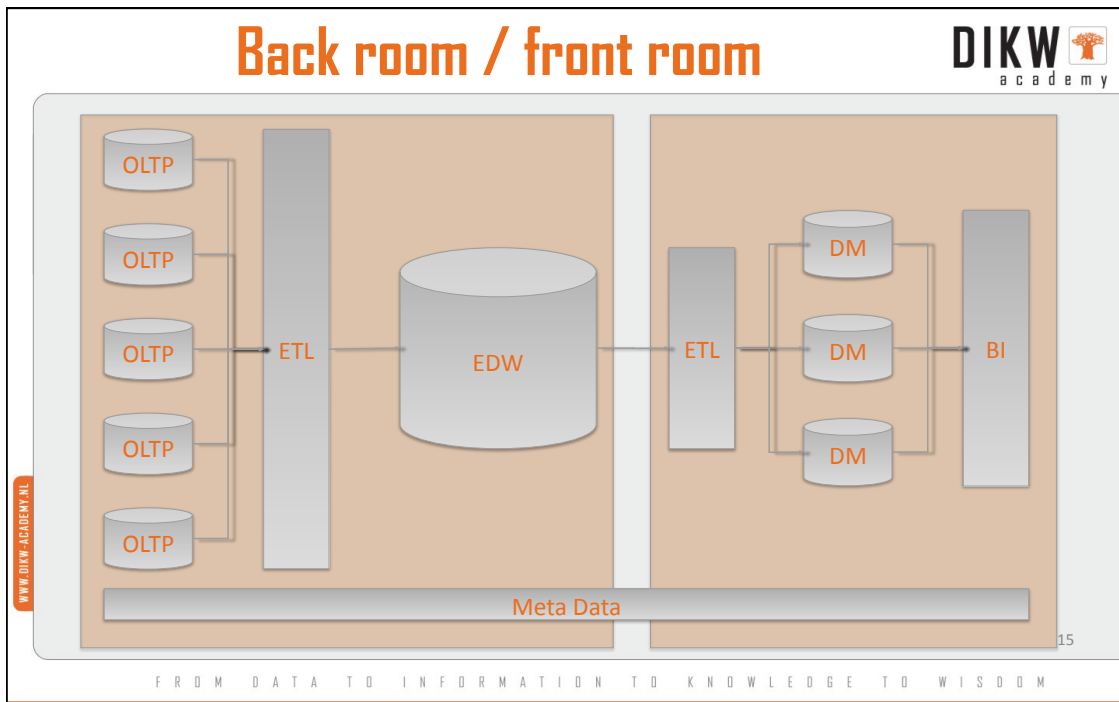
One version of the truth?

- Truth is subjective (who's truth?)
 - Truth is subject to change (when's truth?)
 - Temporal truth?
 - Can historical data be tested on truth?
 - Truth is hard to define
 - What if two source systems disagree with each other?
- Due to the nature of truth we need the ability to redefine *yesterday's, today's and tomorrow's truth*. This implies that we need to store the *facts!*

14

FROM DATA TO INFORMATION TO KNOWLEDGE TO WISDOM

Back room / front room



FROM DATA TO INFORMATION TO KNOWLEDGE TO WISDOM

Back / front room responsibilities

- Back room
 - Integration
 - Capture history
 - Maintain lineage
 - “Yesterday’s and today’s **facts** as stated in the source systems”
- Front room
 - Consistent reporting
 - Subject orientation
 - According to the business rules as defined by the business
 - “Today’s version of the truth”

16

FROM DATA TO INFORMATION TO KNOWLEDGE TO WISDOM

The architecture

DATA VAULT

F R O M D A T A T O I N F O R M A T I O N T O K N O W L E D G E T O W I S D O M

What is a Data Vault

Real name: Common Foundational Integration Model Architecture
Author: Dan Linstedt

The Data Vault is a detail oriented, historical tracking and uniquely linked set of normalized tables that support one or more functional areas of business,

It is a hybrid approach encompassing the best of breed between 3rd normal form (3nf) and star schema. The design is flexible, scalable, consistent and adaptable to the needs of the enterprise. It is a data model that is architected specifically to meet the needs of today's Enterprise Data Warehouse.

18

F R O M D A T A T O I N F O R M A T I O N T O K N O W L E D G E T O W I S D O M

Inmon, Kimball, Linstedt

- 3NF (inmon)
 - Originally build for On-line Transaction Processing (OLTP). It was **adapted** to meet the needs of data warehousing.
- Star Schema (Kimball)
 - Originally architected to solve subject-oriented problems. It was **adapted** to meet the needs of data warehousing
- Data Vault is a hybrid, best of breed solution
 - The Data Vault is **architected and designed** to meet the needs of data warehousing. It is **not an adaption.**

19

FROM DATA TO INFORMATION TO KNOWLEDGE TO WISDOM

HUBS, LINKS & SATELLITES

FROM DATA TO INFORMATION TO KNOWLEDGE TO WISDOM

Hub

- A hub is a list of unique business keys
- *Every* hub has the following attributes (all required, no additions allowed):
 - Primary Key
 - Business Key
 - Load DTS
 - Record Source

Sample hub_customer

hub_customer_id	customer_id	load_dts	record_source
1	1	2009-10-29 20:38:52	sakila-db.customer
2	2	2009-10-29 20:38:52	sakila-db.customer
3	3	2009-10-29 20:38:52	sakila-db.customer
4	4	2009-10-29 20:38:52	sakila-db.customer
5	5	2009-10-29 20:38:52	sakila-db.customer
6	6	2009-10-29 20:38:52	sakila-db.customer
7	7	2009-10-29 20:38:52	sakila-db.customer

Column	DataType
hub_customer_id	INT
customer_id	INT
load_dts	TIMESTAMP
record_source	VARCHAR(100)

Indexes

- PRIMARY
- BK

Hub – attributes 1

- Primary Key
 - Surrogate, system generated, unique identifier
 - Used internally
- Business Key
 - Unique identifiable business element
 - Used in the source systems
 - Known to the business
 - Does not change

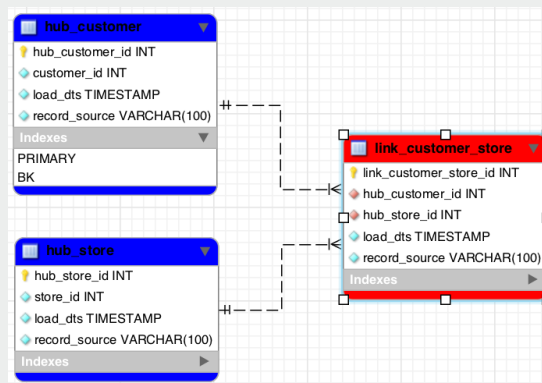
Hub – attributes 2

- Load DTS
 - The timestamp the record was inserted(!) into the Data Vault (the first time the EDW saw the business key)
 - System generated
- Record Source
 - Defines the origin of the record (e.g. Source system, table, etc.)
 - Lowest grain possible
 - Be elaborate!

Link

- A link is an intersection of business keys (hubs)
- Attributes:
 - Primary Key
 - {Hub Surrogate Keys 1..n}
 - Load DTS
 - Record Source
- The link's business key is a composite unique index on the n Hub surrogate Keys

Sample link customer_store



Link - characteristics

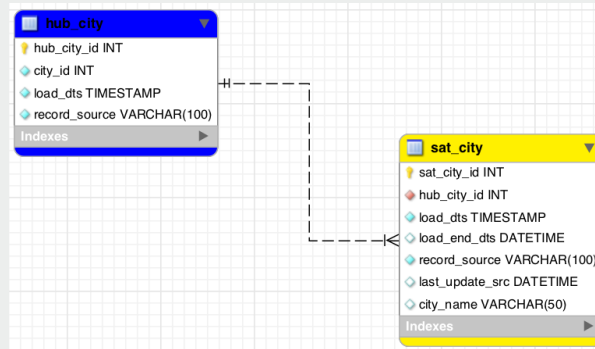
- A link's business key is a Composite Unique Index
- A Link Load Data Represents the first time the EDW saw the data
- A link is based on an identifiable business element relationship, otherwise known as a foreign key, also known as a business event or transaction between business keys.
- It should not change over time. It is established as a fact that occurred at a specific point in time and will remain that way forever.
- All attributes are mandatory

Satellite

- A satellite contains detailed information (attributes) of a single hub or link, including historical changes
- Attributes:
 - Primary Key
 - Foreign Key (to hub or link sat describes)
 - Load DTS
 - Load end DTS
 - Record Source
 - {attributes 1..n}

Sample sat_city

Sat_city id	Hub_city id	load_dts	Load_end_dts	record_source	Last_update_src	City_name
1	1	2009-10-29 20:38:52		sakila-db.city	2009-01-01	Amsterdam
2	2	2009-10-29 20:38:52	2009-10-29 20:44:16	sakila-db.city	2009-01-01	Paris
3	2	2009-10-29 20:44:16		sakila-db.city	2009-11-30	Parijs



Satellite - characteristics

- A satellite has only 1 foreign key
- Parent table is always a Hub or a Link
- The Business Key is composite: foreign key + Load DTS
- Avoid outer joins; one row for every row in Hub
- Hubs or Links can have more than one Satellite
 - Satellites are defined by type of data, rate of change or source system.

Colors of data modeling

* adapted from Hans Hultgren, Genesee Academy

31

The colors of data modeling

- Any data model consists of the following constructs:

– Business Keys



– Associations

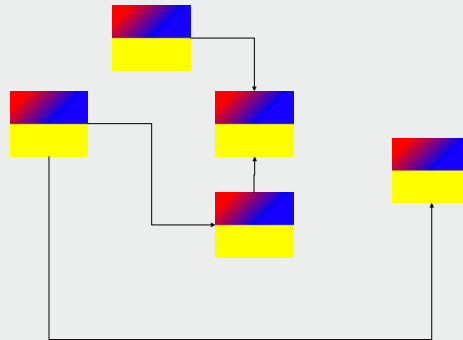


– Details

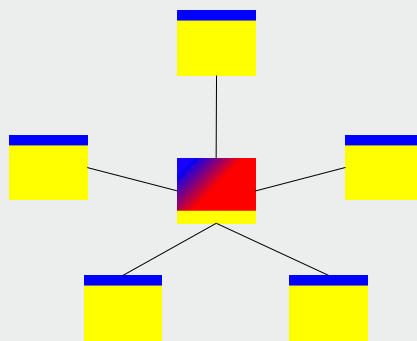


32

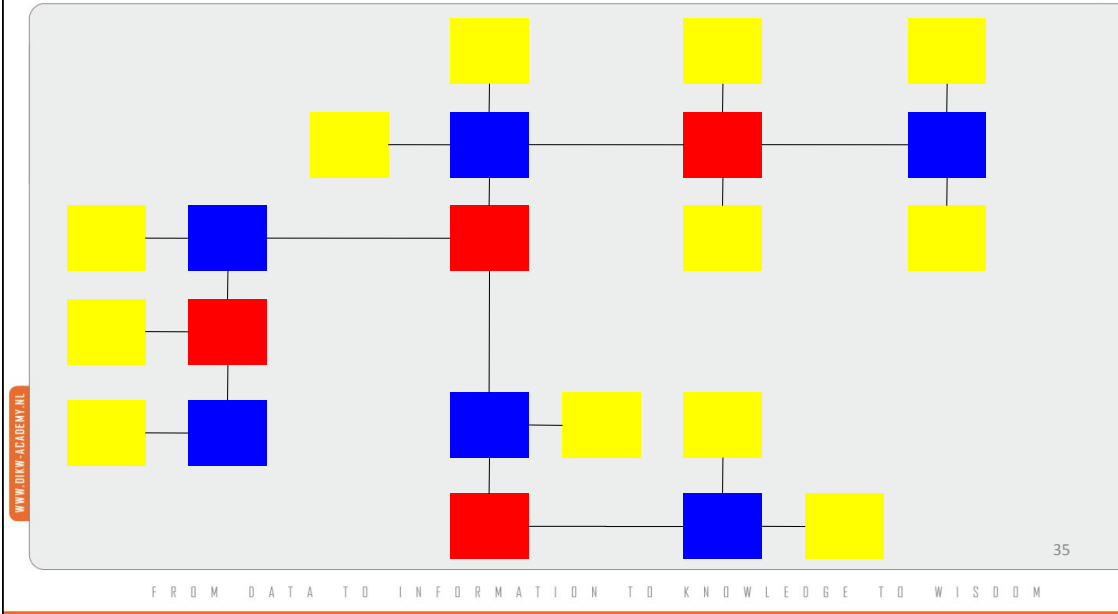
Colors of 3NF



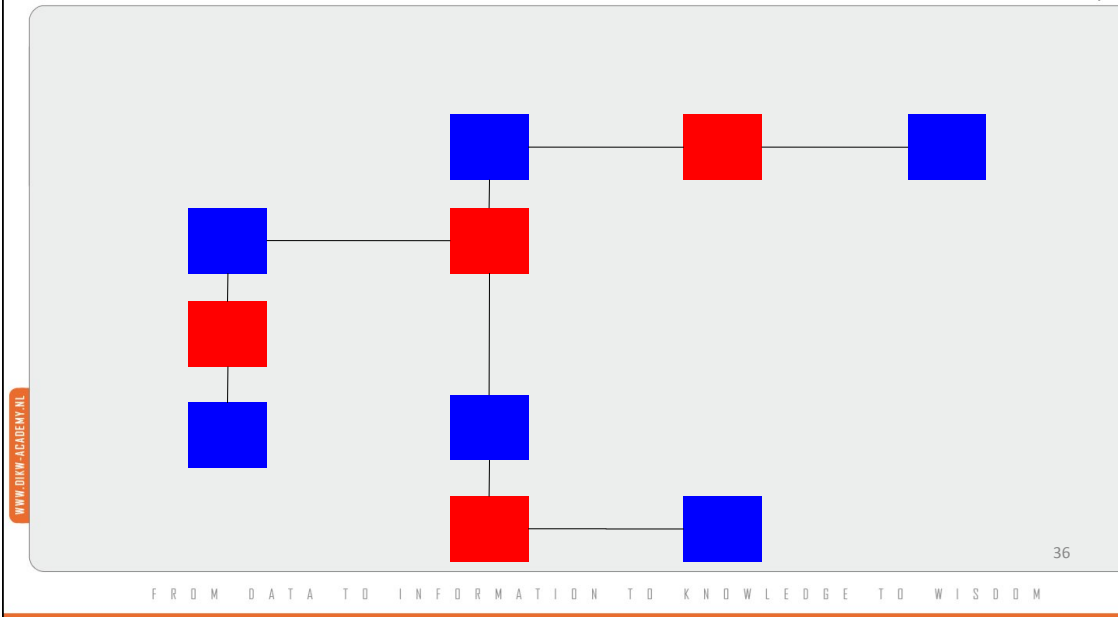
Colors of the star schema



Colors of the data vault



Colors of the data vault – the 'skeleton'



Why Data Vault

- Predictable framework
- 100% of the data is loaded 100% of the time
- Traceable / lineage: Complete history trace
- Extensible: no more <alter table>

- Very well suited to house large data volumes (due to large degree of decoupling and bulk loading)
- Very well suited to generation, hence: DV Automation conference (Damhof, Breur, DIKW Academy)
- Our solution using PDI ... Edwin Weber

37

Thank you

38