Data Modelling for Analytical Environments

Linus Hjorth, Infotrek

INFOTREK

Data Modelling for Analytical Environments

BUSINESS VALUE

DELIVERY - ”IT”

…SO INFORMATION MODELLING AS WELL…

WHAT – WHY – WHERE – HOW

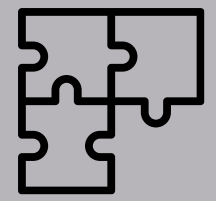
Two Worlds

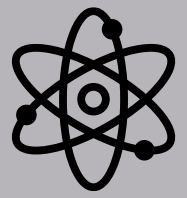
- Data quality
- Reliable figures
- Controlled delivery
- Agreed definitions
- GDPR

- Experimentation
- Flexibility
- Time to market
- ”Good enough”

Models – Types of (what)

- Information model
  - how things relate
- Logical data model
  - information rules, specific application





- Physical data model
  - specification of an application, or for development

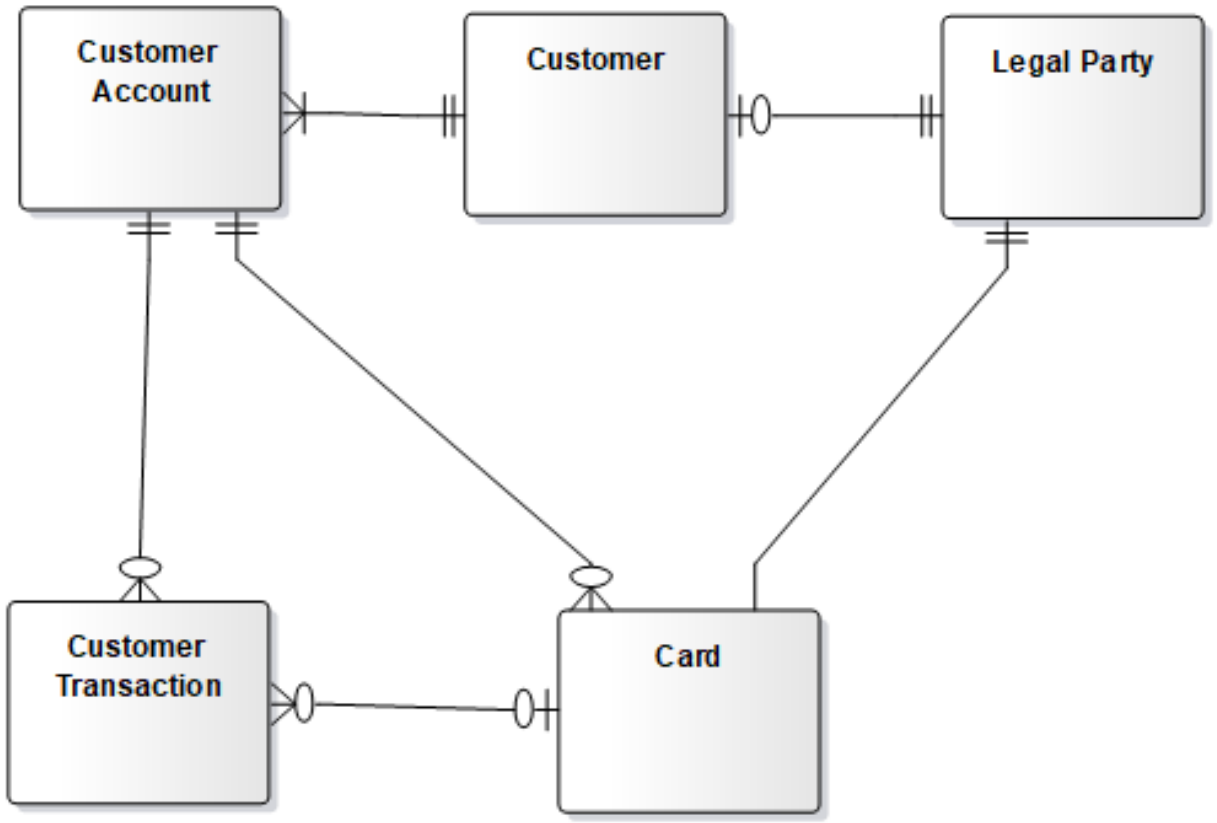
Icons made by Freepik from flaticon.com

INFOTREK

Information Models

- Different names…same thing?
  - Subject Model
  - Conceptual Model
  - Domain Model
  - Logical Business Model
- Describes how things are, or should be
- Not necessarily with system support

What about the use…

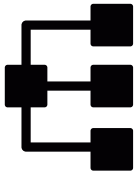


INFOTREK

Why create information models?

- Consensus
  - Different users
  - Data from disparate systems
  - Different applications

- Minimizing risk
  - Fewer misunderstandings
  - Help when prioritizing

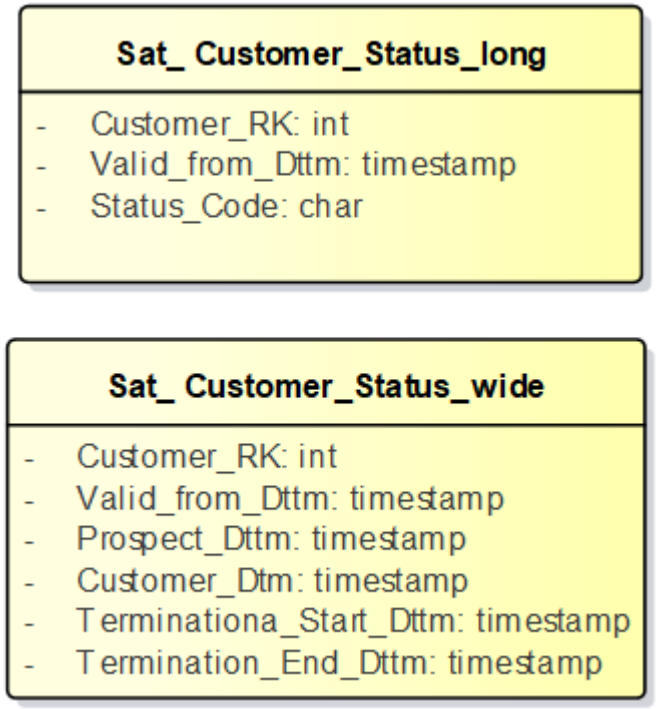


- Mappings towards
  - Rapport requirements
  - Source data

INFOTREK

Considerations

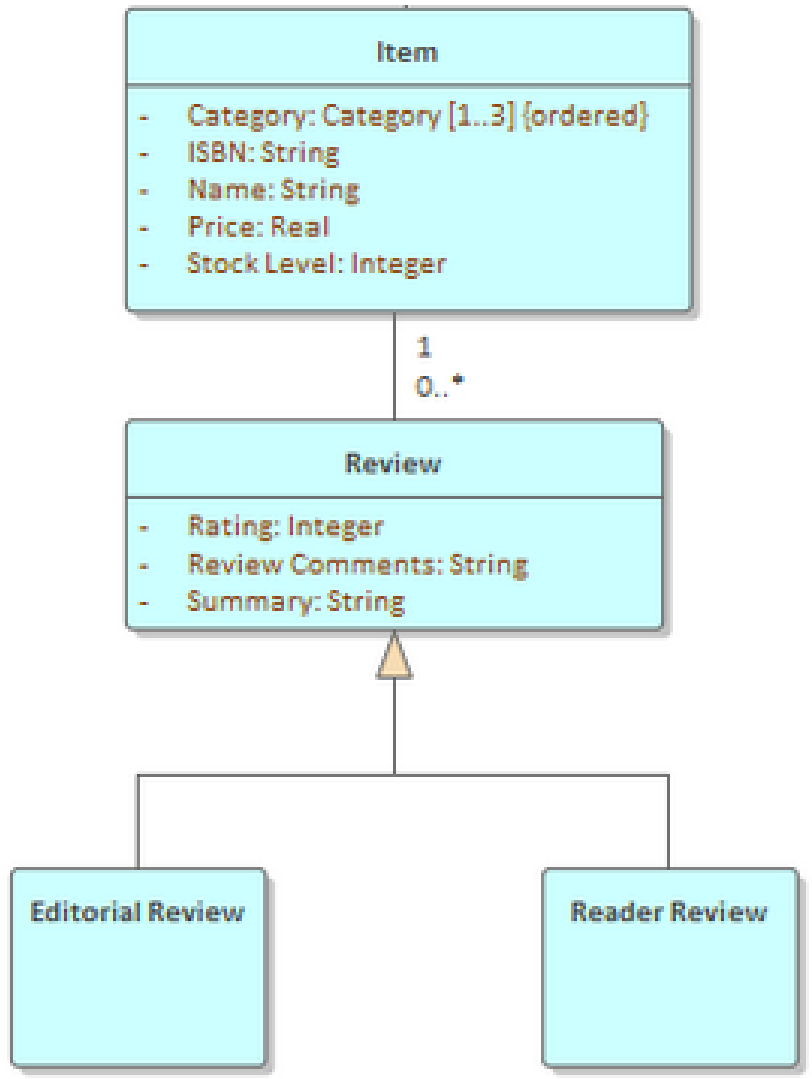
- Homonyms
- Synonyms
- Class vs instance
  - Problematic during verbal communication
  - ”Product”: model or unit?
- Aggregates
- Abstraction – using classifications
  - More efficient model
  - Business terms gets hidden from business stakeholders



INFOTREK

Logical Data Models

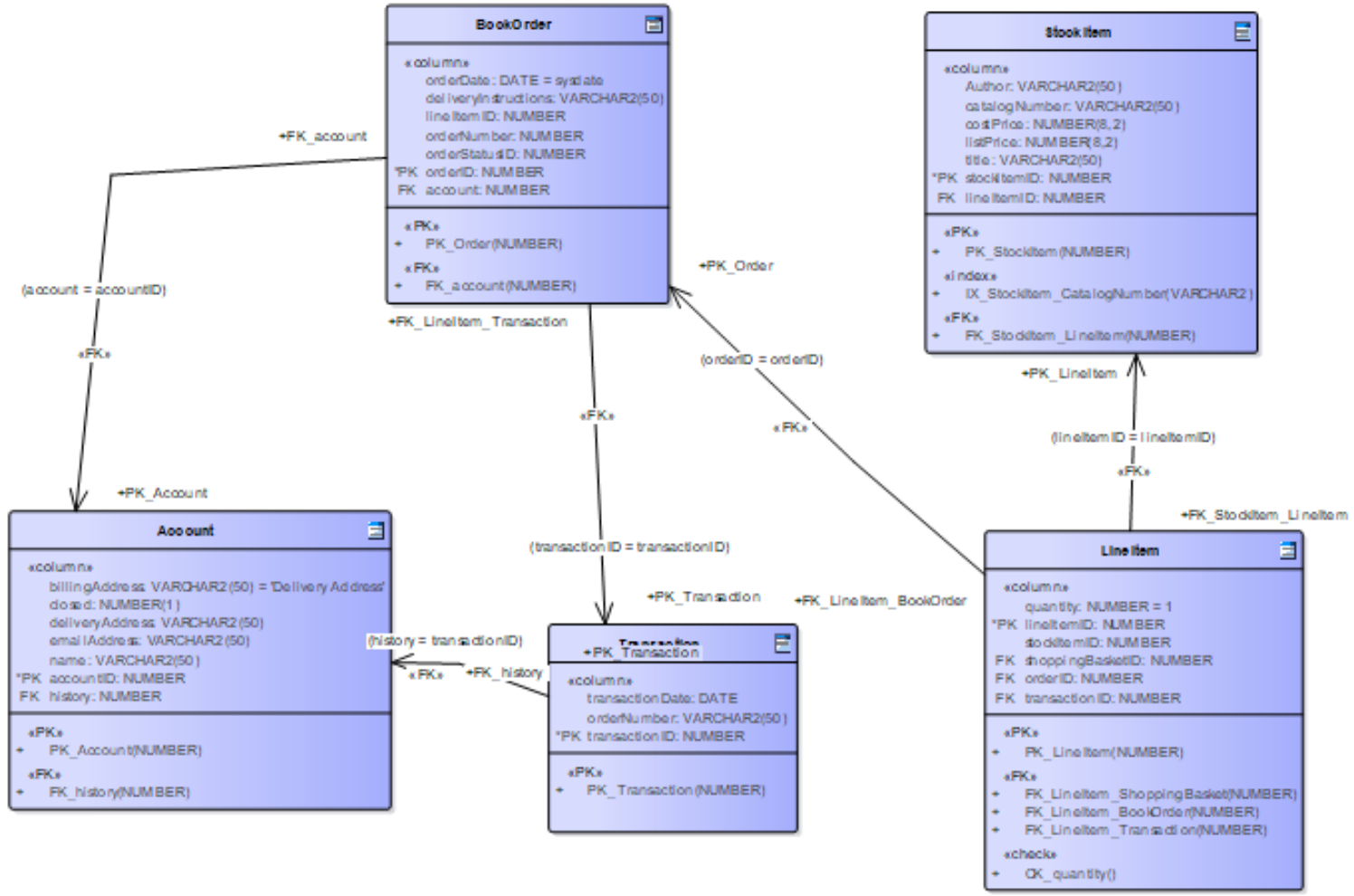
- Business or business like naming
- Logical data types (text – number – amount – no of - date)
- Describes rules for data
- Overlapping information modelling
- Data base independent
- Used as a requirement for development



INFOTREK

Physical Data Models

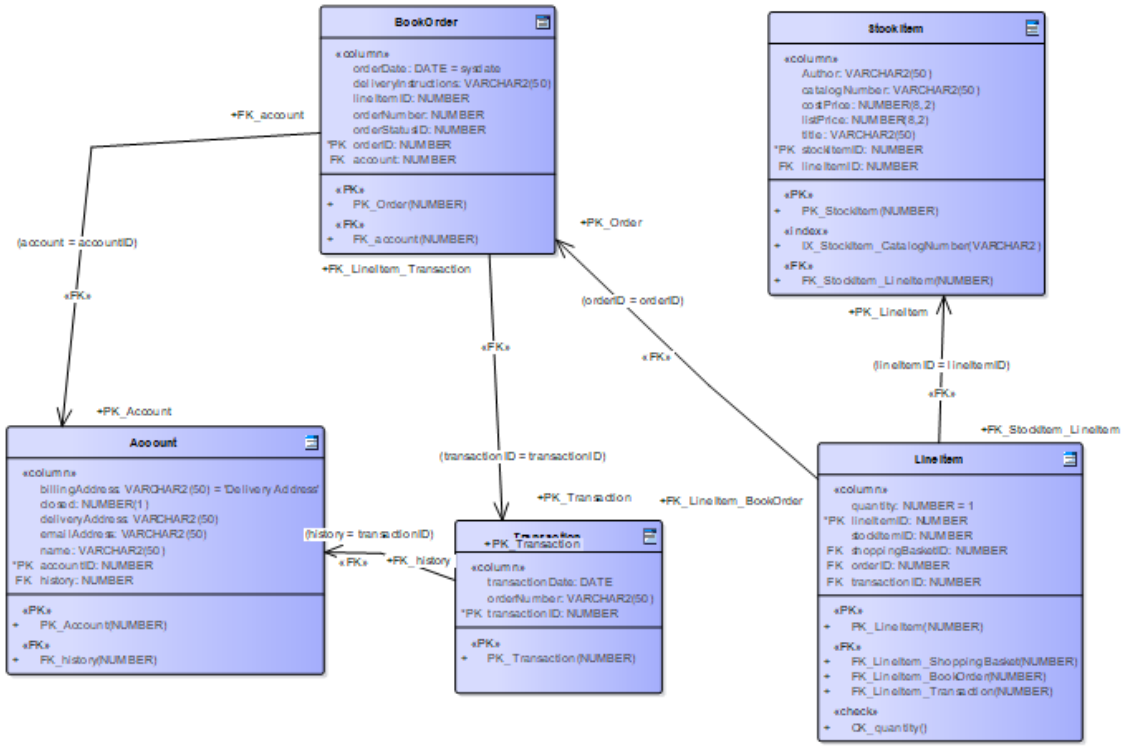
- Data base specific
- Table & column
- Physical data types
- Index, partitioning, schemas…



INFOTREK

Physical Data Models

- Can deviate from the logical model
  - Performance
  - Usability
  - De-normalization
  - Integrity managed by ETL
- Used for implementations



INFOTREK

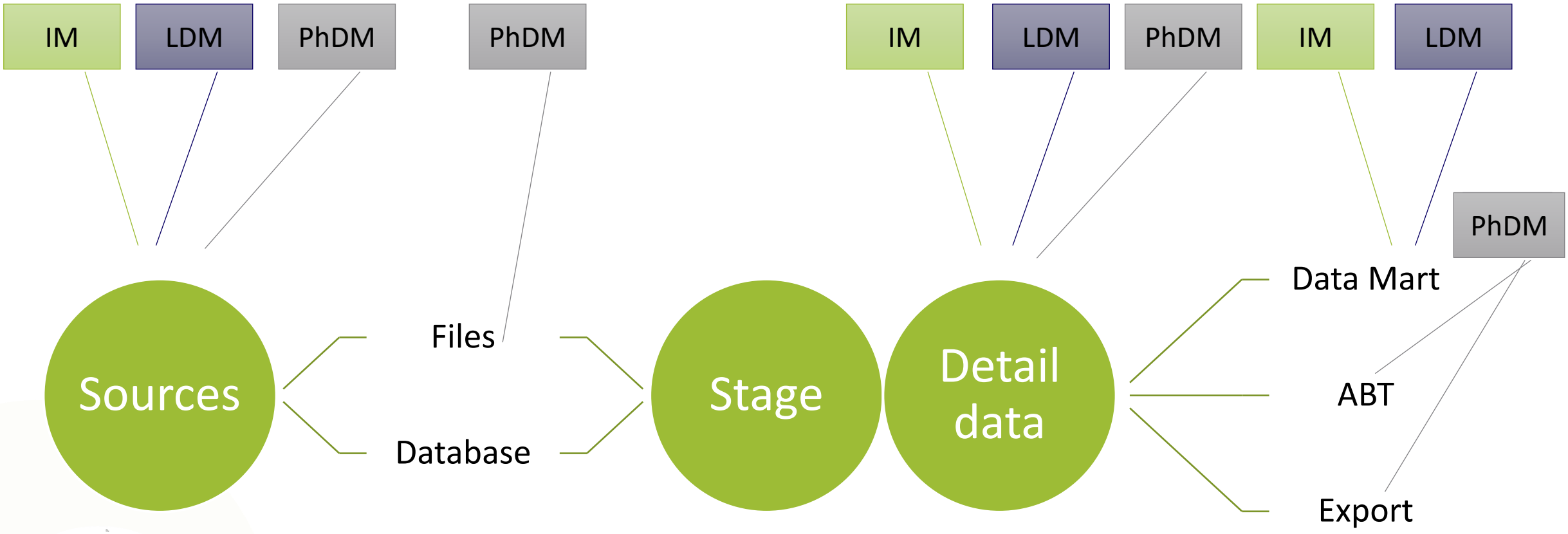
Summary Model Types

- If your concern is
  - Communication
    - use an information model
  - Data integrity
    - use a logical data model
  - Performance
    - use a physical data model

Ronald Damhof

What to use (where)?

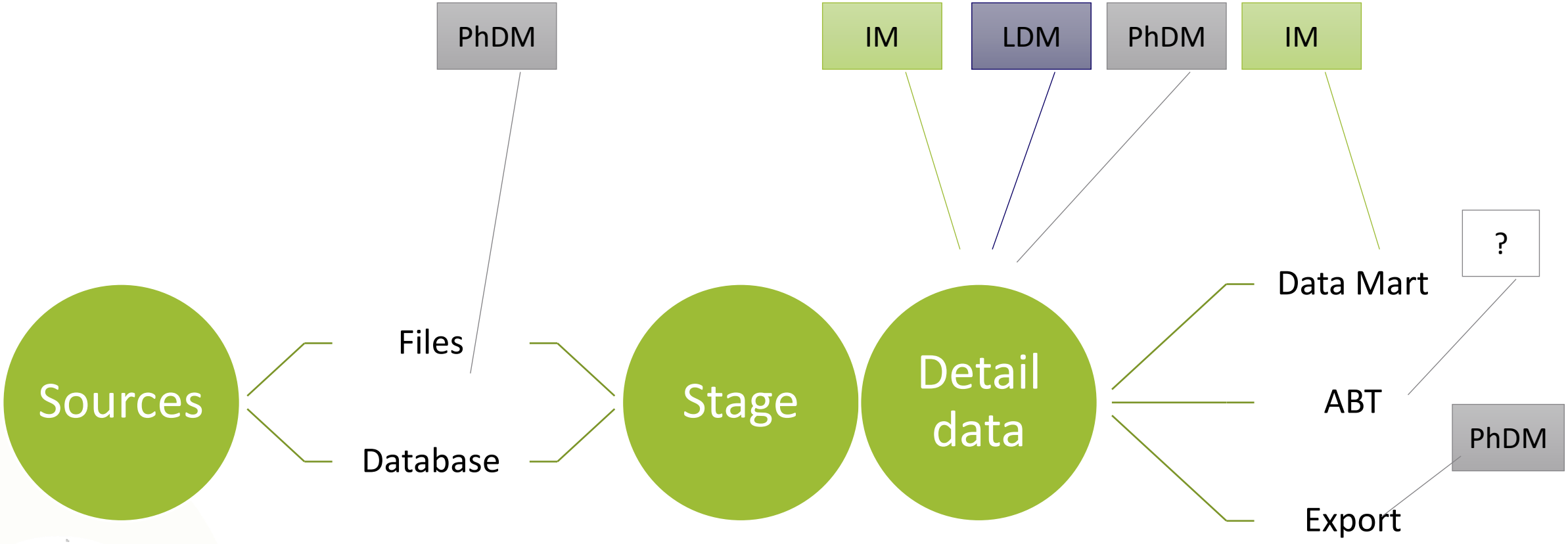
- Development



INFOTREK

What to use (where)?

- Maintenance



INFOTREK

Getting started (how)

- Interviews
- Workshops
- Documentation
- Business documents (top-down)
- Process models

- Never let it be a one person effort

INFOTREK

Getting started (how)

Examples of methodologies

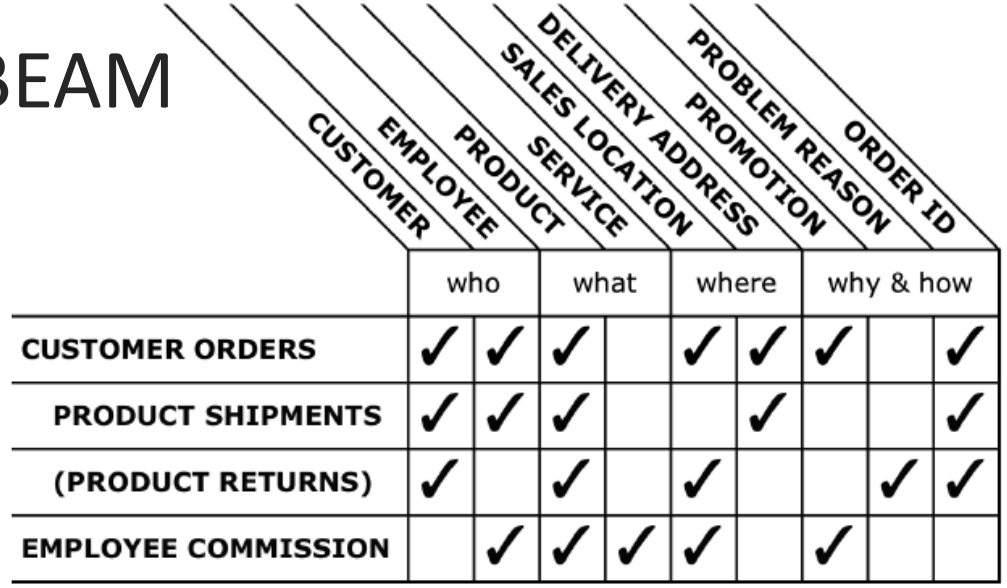
- SAS Analytics Life Cycle
- BEAM - Business Event Analysis & Modelling
- ELM – Ensemble Logical Model
- BIP: Business – Information - Prototype

Similar…

INFOTREK

Business Event Analysis & Modeling - BEAM

- Agile
- The ”7W’s”:
  - **W**ho
  - **W**hat
  - **W**hen
  - **W**here
  - **W**hy
  - ho**W** (did it happen)
  - ho**W** (many/much)
- Inspired from journalistic methodology



	CUSTOMER	EMPLOYEE	PRODUCT	SERVICE	SALES LOCATION	DELIVERY ADDRESS	PROMOTION	PROBLEM REASON	ORDER ID
	who		what		where		why & how		
CUSTOMER ORDERS	✓	✓	✓		✓	✓	✓		✓
PRODUCT SHIPMENTS	✓	✓	✓			✓			✓
(PRODUCT RETURNS)	✓		✓		✓			✓	✓
EMPLOYEE COMMISSION		✓	✓	✓	✓		✓		

CUSTOMER ORDERS [DE]

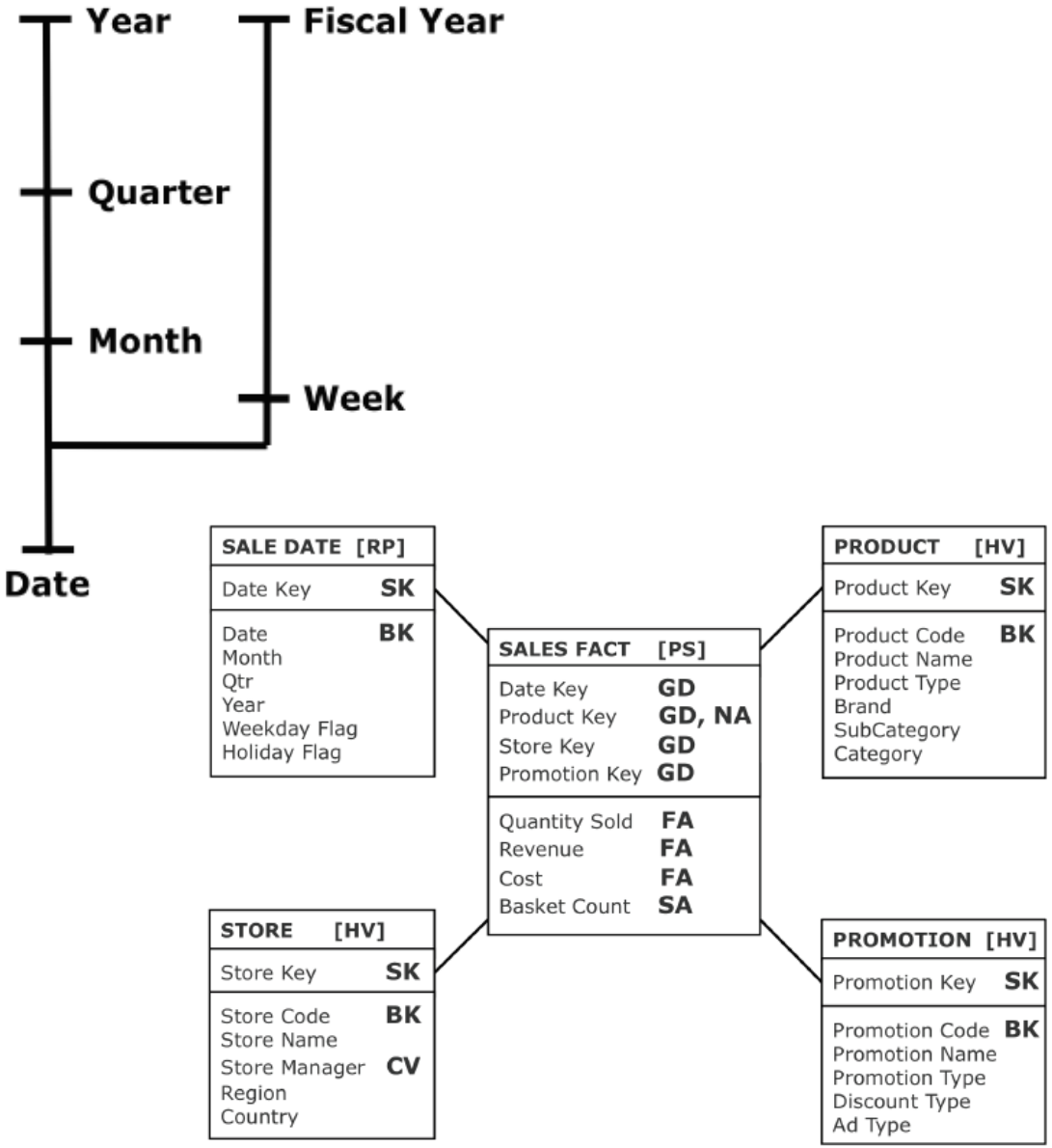
CUSTOMER	**orders** PRODUCT	on ORDER DATE	QUANTITY	for REVENUE	with DISCOUNT	using ORDER ID
[who]	[what] MD, GD	[when] MD	[Retail Units]	[$, £, €]	[$, £, €, %]	[how] GD
Elvis Priestley	iPip Blue Suede	18-May-2011	1	$249	0	ORD1234
Vespa Lynd	POMBook Air	29-Jun-2011	1	£1,400	10%	ORD007
Elvis Priestley	iPip Blue Suede	18-May-2011	1	$249	0	ORD4321
Phillip Swallow	iPOM Pro	14-Oct-2011	1	£2,500	£150	ORD0001
Walmart	iPip G1	10 Years Ago	750	$200,000	$10,000	ORD0012
US Senate	iPOM + Printer	Yesterday	100	$150,000	$20,000	ORD5466
US Senate	iPip Touch	Yesterday	100	$25,000	$1,000	ORD5466

Figures from the book *”Agile Data Warehouse Design – Collaborative Dimensional Modeling, from Whiteboard to Star Schema”*

INFOTREK

Business Event Analysis & Modeling - BEAM

- Focus on dimensional modelling
- Fact: ho**W** (many/much)
- Dimensions: the rest
- *”Agile Data Warehouse Design – Collaborative Dimensional Modeling, from Whiteboard to Star Schema*”: Lawrence Corr och Jim Stagnitto

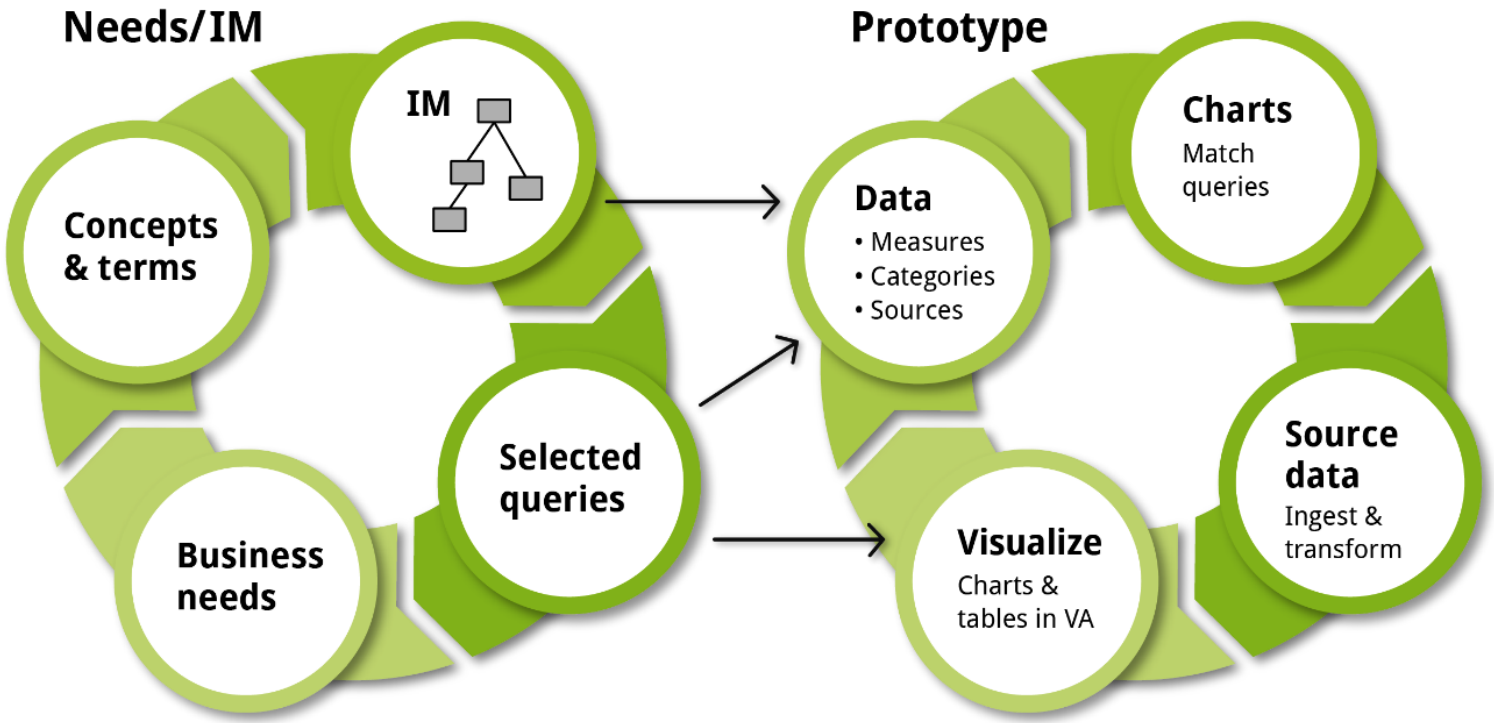


Figures from the book *”Agile Data Warehouse Design – Collaborative Dimensional Modeling, from Whiteboard to Star Schema”*

INFOTREK

BIP

- ❑ Business Needs
- ❑ Information Model
- ❑ Prototype

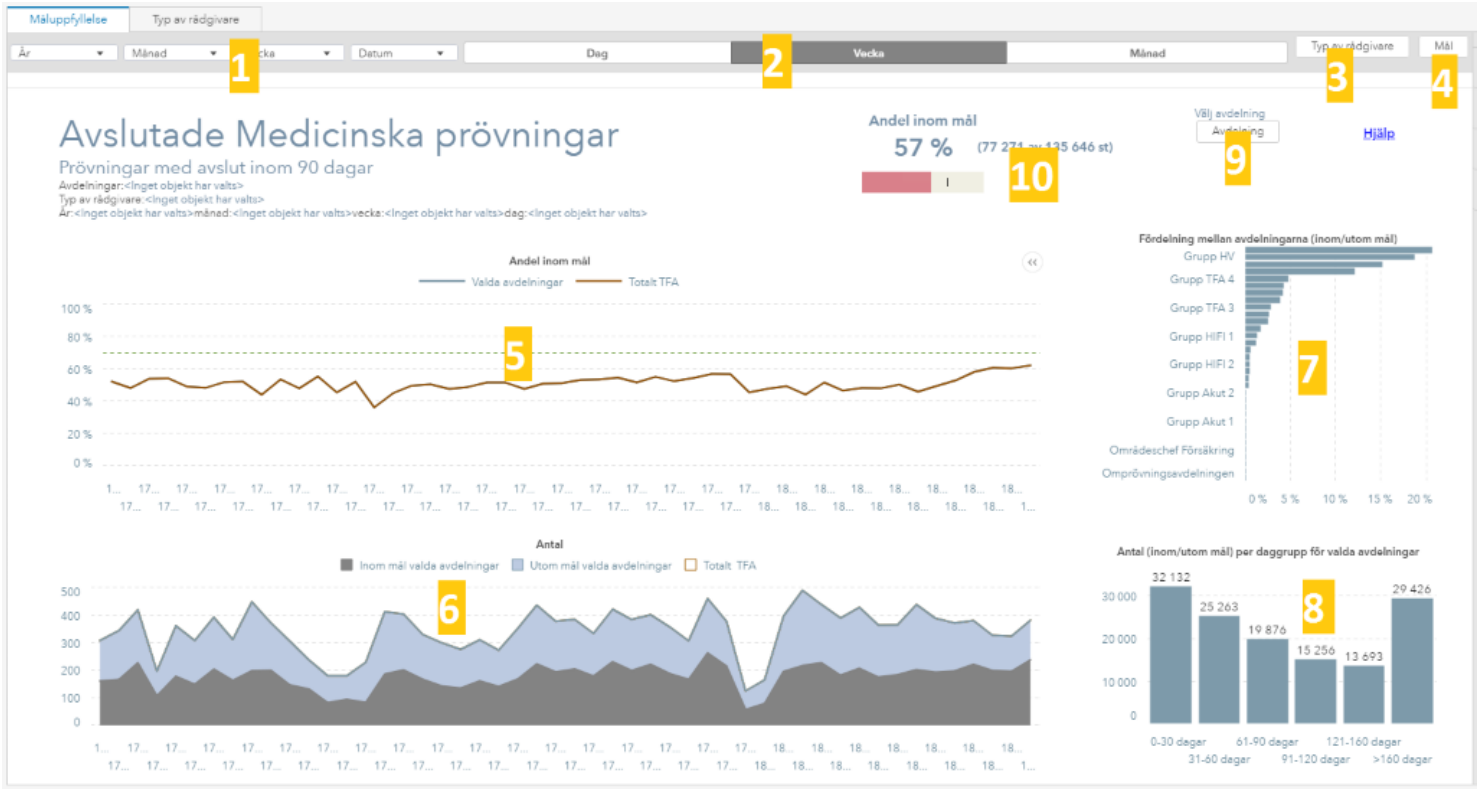


- Interviews and workshops – to gather **business** requirements
- create **information model**
- Map prioritized requirements to source data
- Build **prototype**

INFOTREK

BIP - Business Needs - Information Model - Prototype

- Inspired by BEAM
- Whole cycle should fit into a sprint
- Quick verification of requirements and feasability
- Created by Infotrek



INFOTREK

Data Modelling Paradigms

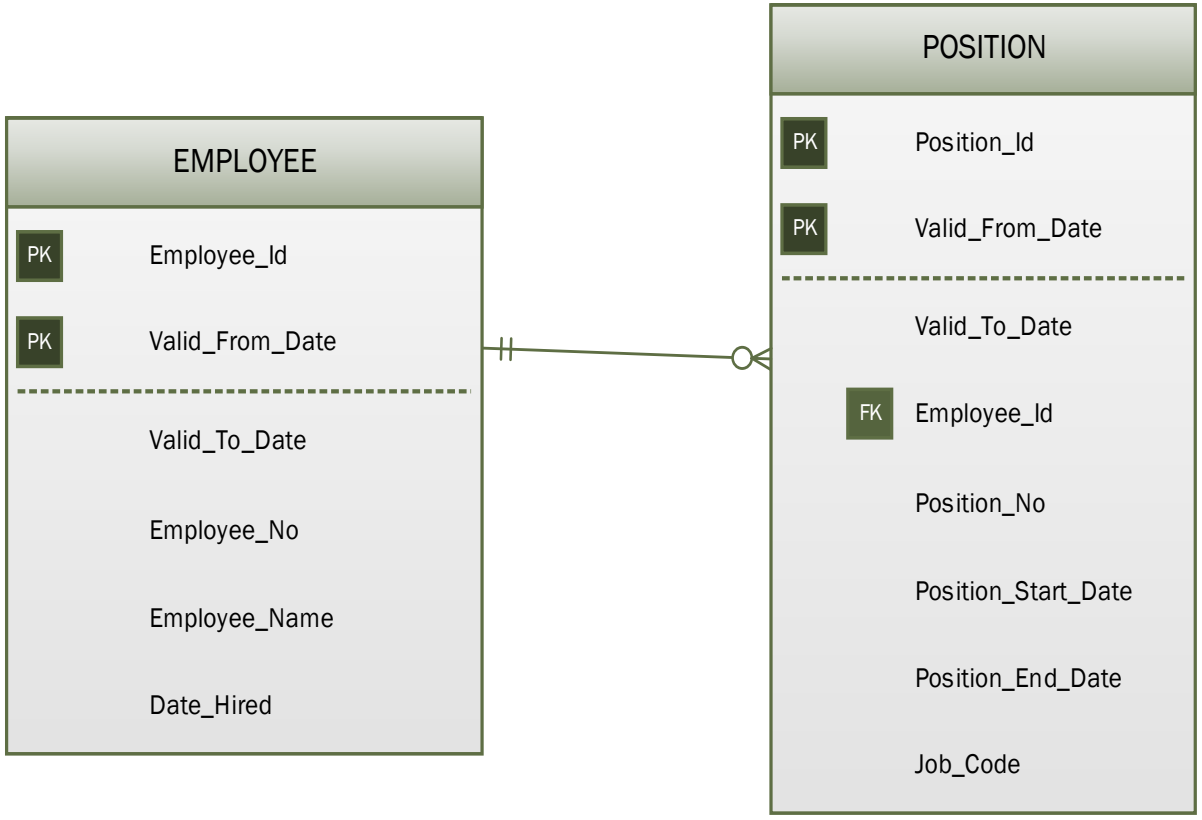
…within analytics/BI

- Third normal form - 3NF
- Dimensional modelling
- Data Vault
- Data prepared for analysis
- ”Big Data”

INFOTREK

Third Normal Form - 3NF

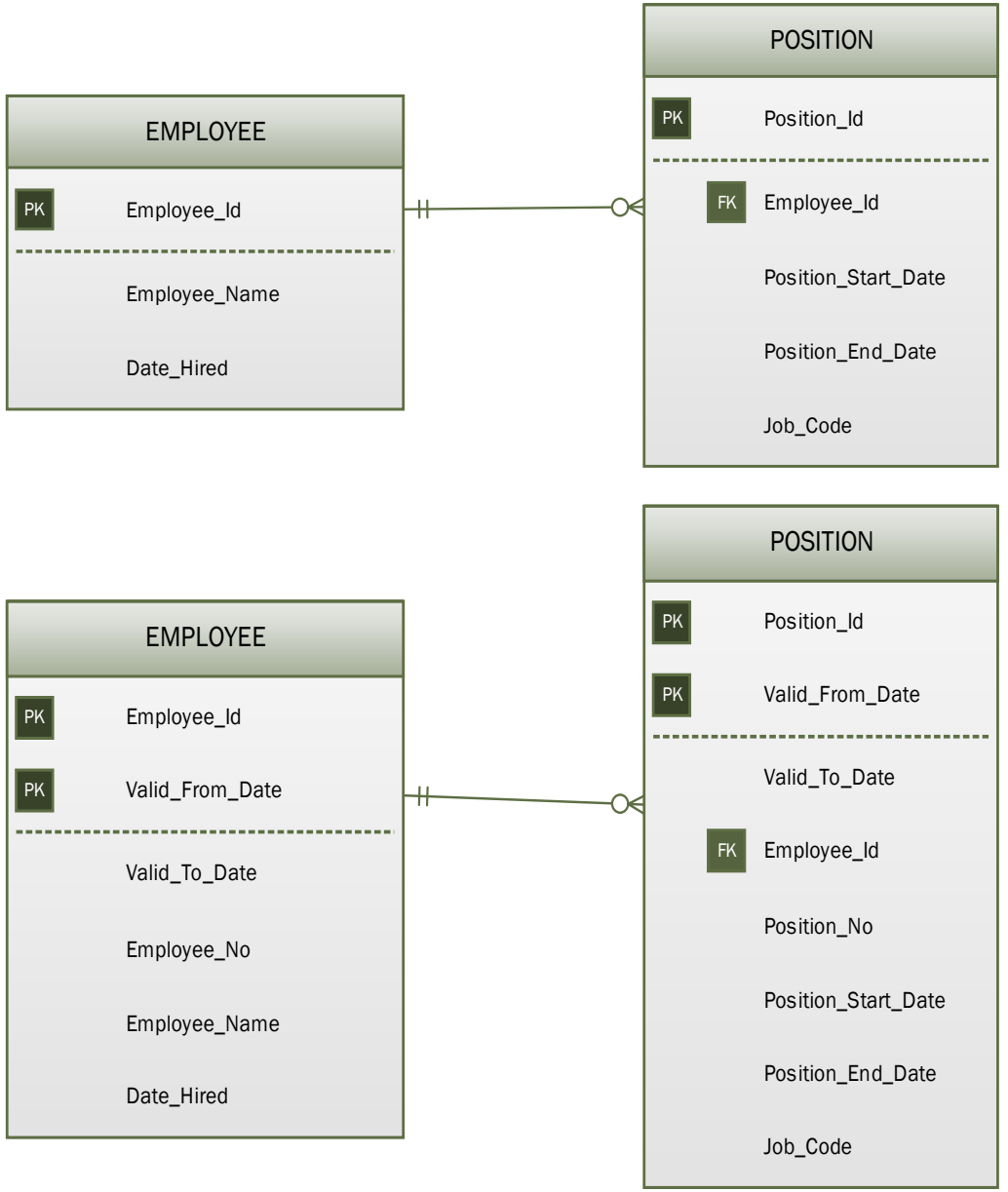
- Foundation for modelling in relational data bases since dawn of time
- Brought to the BI/DW domain by Bill Inmon
- Main use – the detail data layer (atomic)
- Addition: Data versioning (new record in case of changed values)



INFOTREK

Third Normal Form - 3NF

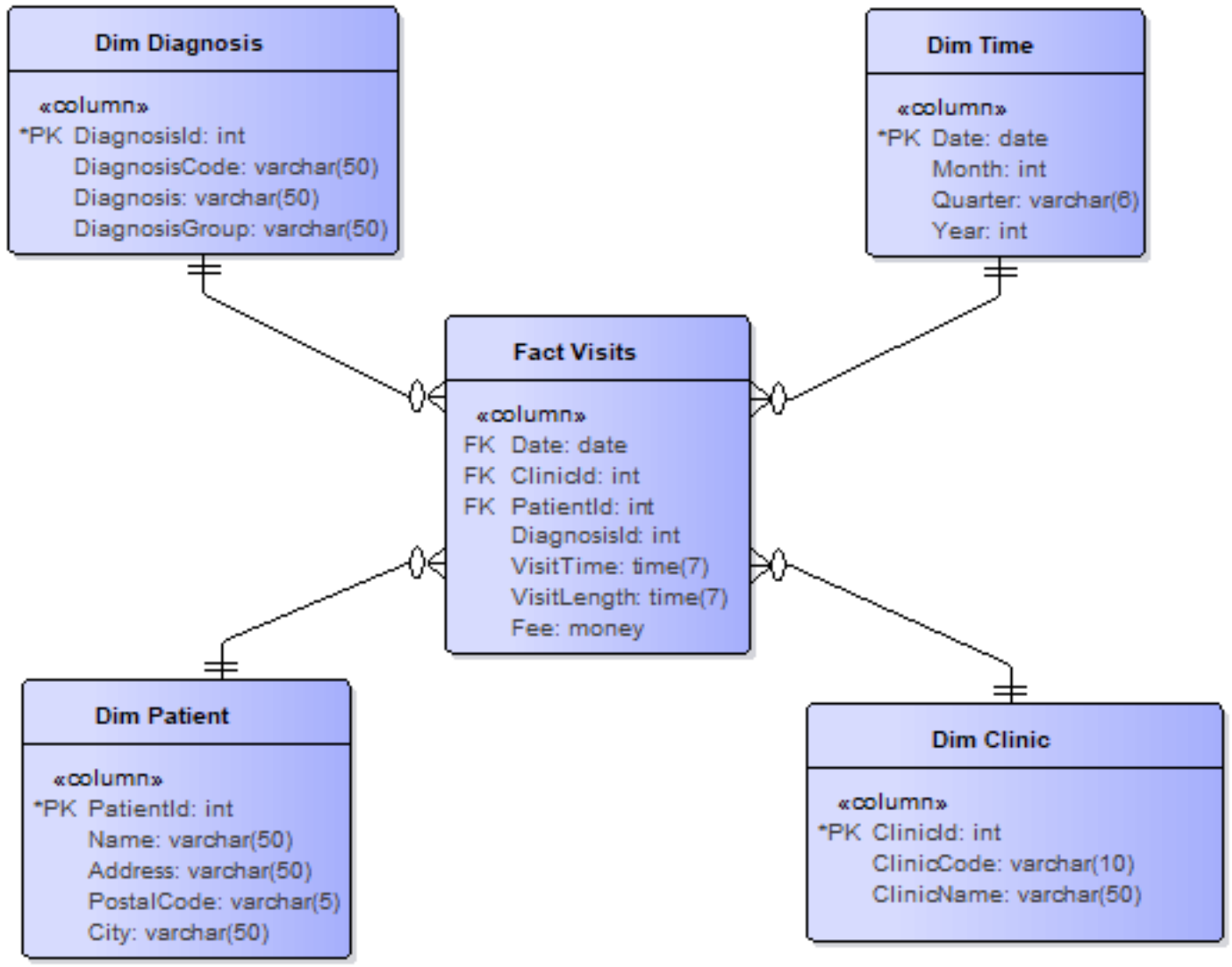
- Hard to implement as a physical model if you wish to comply with relational algebra
  - …and the data modelling tool might protest
- Pros:
  - wide spread knowledge
  - optimized for storage
- Cons
  - Inflexible
  - not optimized for query



INFOTREK

Dimensional Modelling

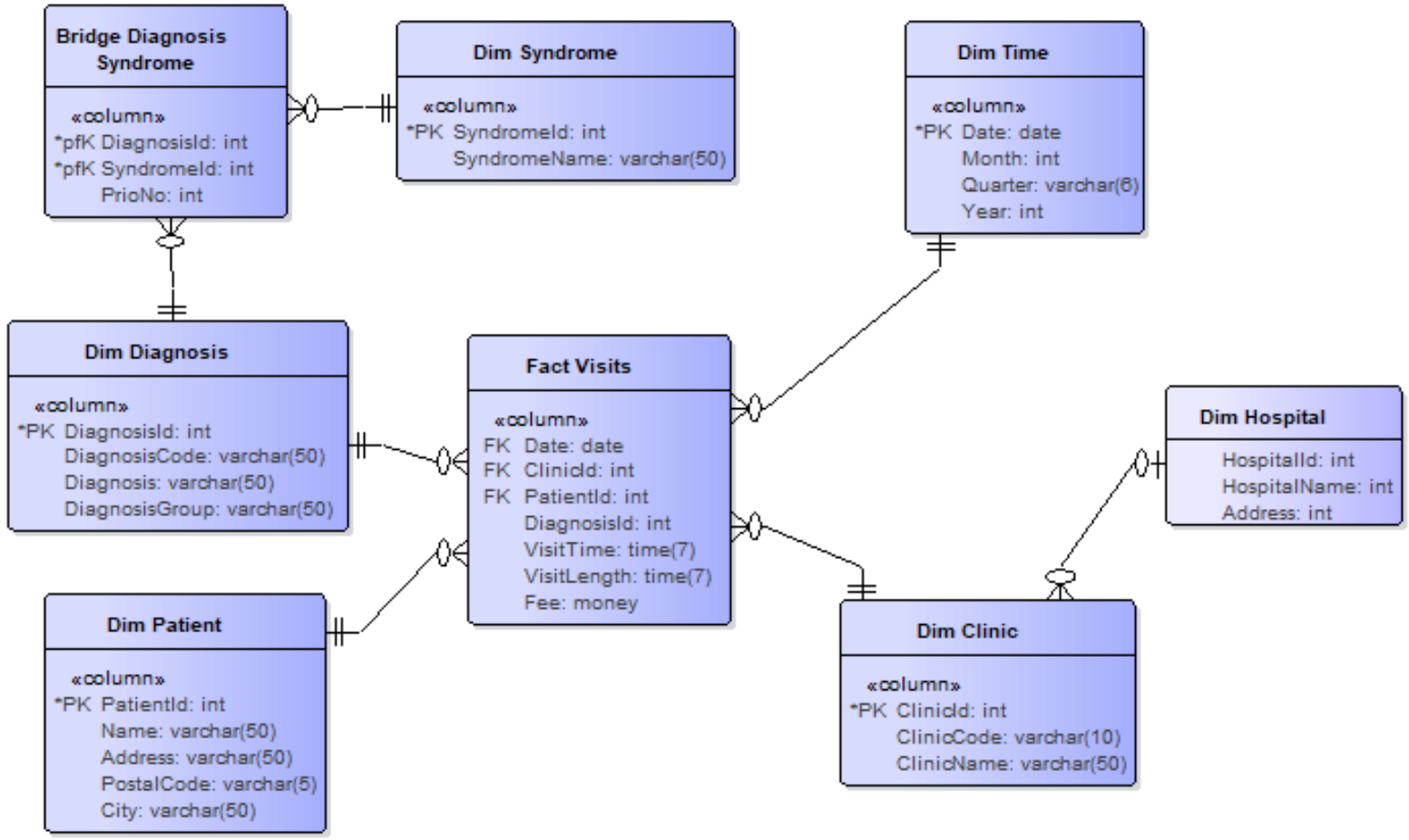
- Made famous by Ralph Kimball and Margy Ross
- Physical implemented as Star Schemas
- Works best with predictable query patterns
- Relatively easy to understand
- Main use: data marts



INFOTREK

Dimensional Modelling

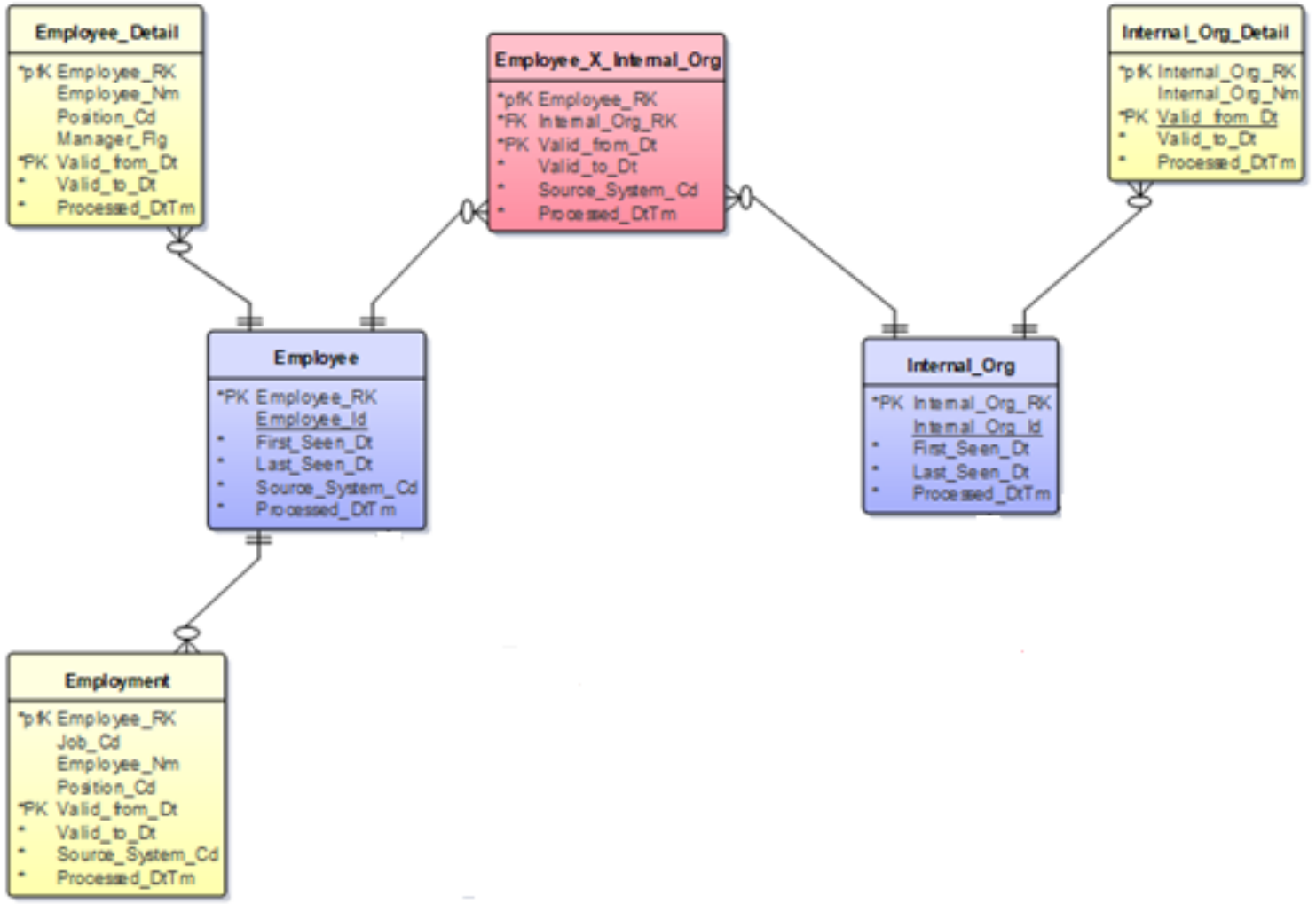
- Logical model ”Snowflake” – hierarchies normalized
- Not optimal for certain data structures
  - N-occurences
  - M-M relationsships
- If used in a detail data store: bridge-tables and other constructs – as complicated as 3NF



INFOTREK

Data Vault & ensembles

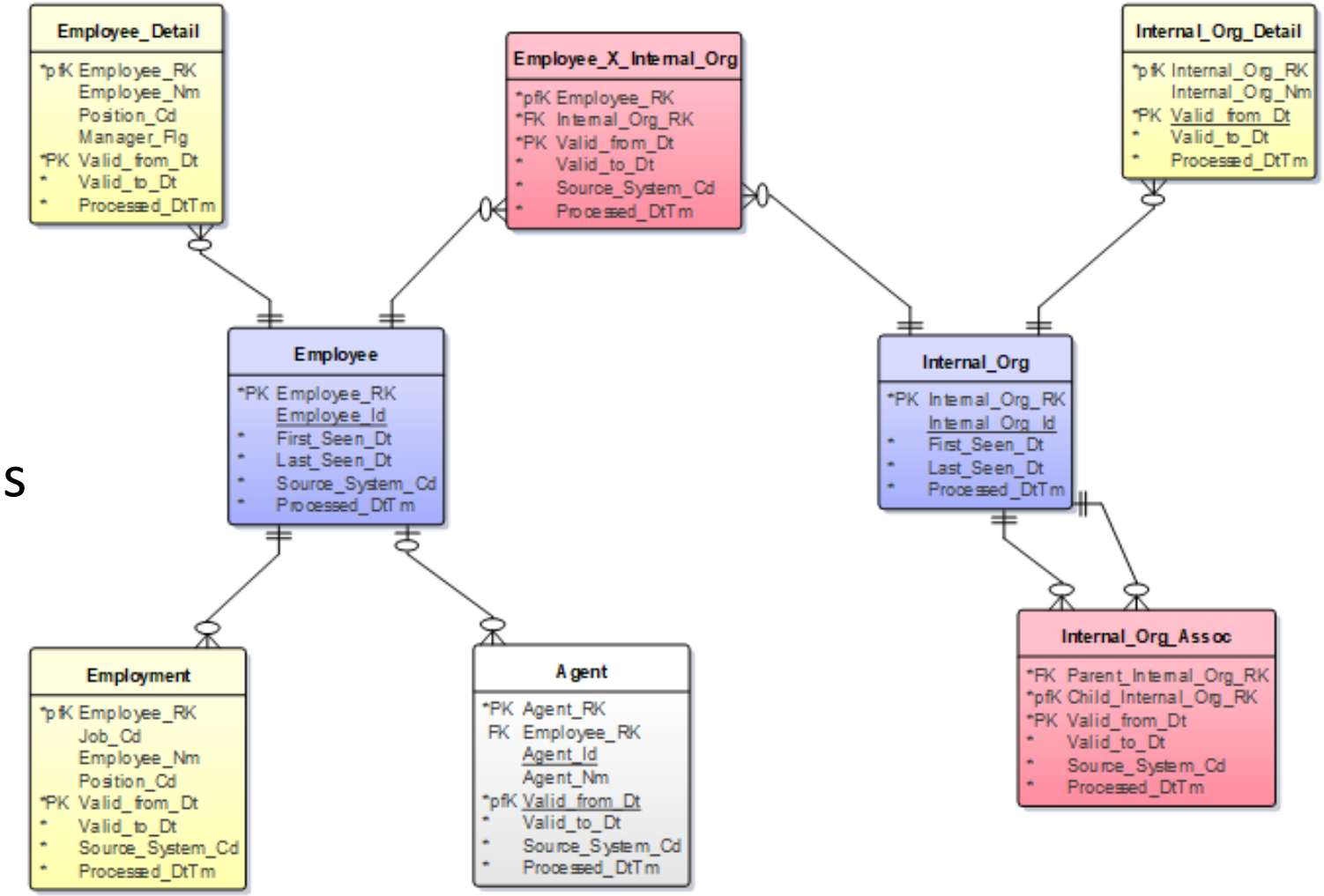
- De-facto standard for data warehouse
- Segmentation of data into three categories
  - Core Business Concept (unique identifier)
    - **Hub (blue)**
  - Relationships between business concepts
    - **Link (red)**
  - Descriptions of business concepts
    - **Satellite (yellow)**



INFOTREK

Data Vault & ensembles

- Considered flexible
  - Relations always Many-to-Many
  - Separation of keys, relations & attributes
  - Separation of attribute entities – as you like it
- Do’s and don'ts, still a craft
- Invented by Dan Linstedt



INFOTREK

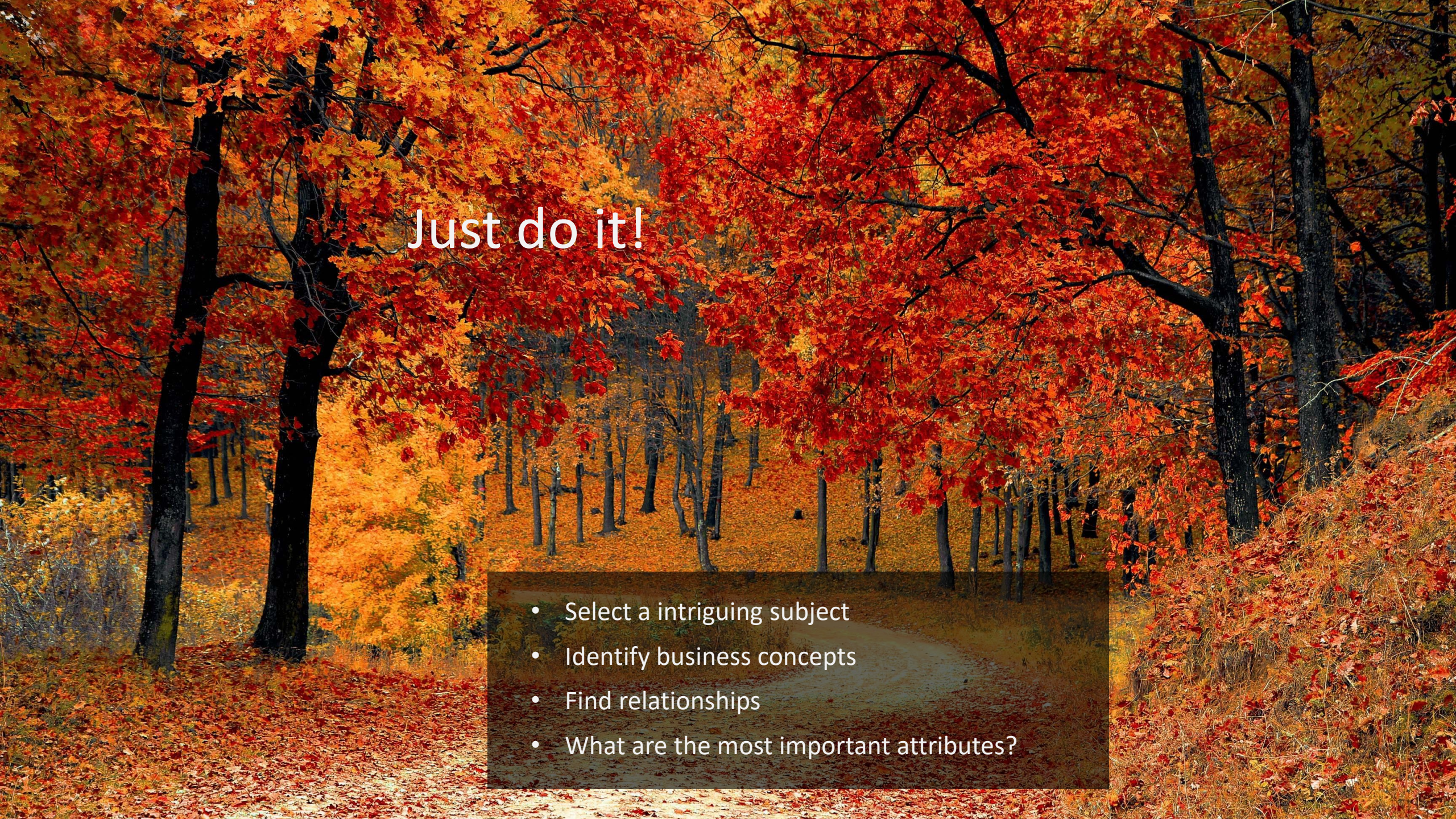
Modelling for Analysis – ”Analytical Base Table”

- Different analysis – different data layout
- Common requirements:
  - Continues values -> grouping using intervals, ranking
  - One row per individual – transpose (long to wide)
  - History/trends -> across multiple variables (Amount_Jan, Amount_Feb…)
  - Flags – numerical (0/1 instead of J/N)

Weight	Wheelbase	Length	rank_MPG_Highway	rank_Horsepower	rank_Wheelbase
4451	106	189	7	2	5
2778	101	172	1	5	8
3230	105	183	2	5	6
3575	108	186	3	2	4
3880	115	197	7	3	1
3893	115	197	7	3	1

And then what?

- Should everyone do everything?
  - Maturity
    - Competence
    - Maintenance & development processes
    - Organization
  - Phase
    - Established or evolving?
  - Complexity


Just do it!
• Select a intriguing subject
• Identify business concepts
• Find relationships
• What are the most important attributes?

Thank you!

Contact Information

linus.hjorth@infotrek.se

linkedin.com/in/linushjorth

@LinusHjorth

Acknowledgements:
Ylva Andersson – Statistics & Data Science
Siavoush Mohammadi – Structure