

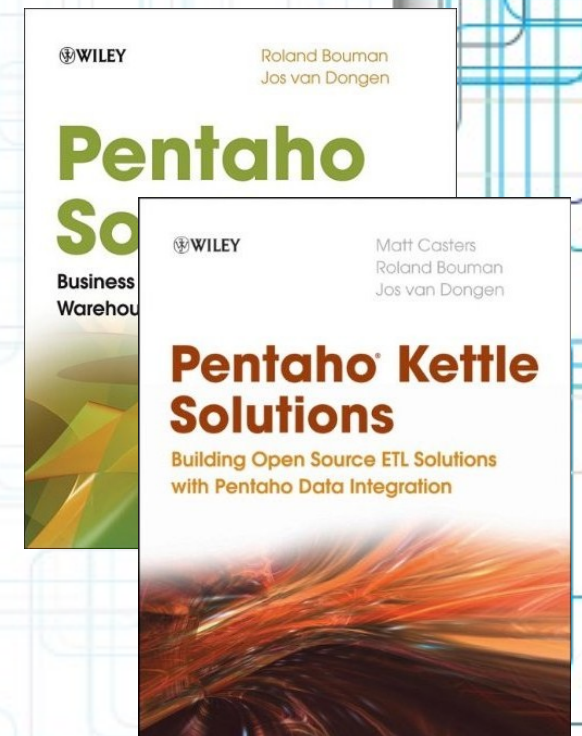
Trends in Data Warehouse Data Modeling:

Data Vault and Anchor Modeling



Thanks for Attending!

- Roland Bouman, Leiden the Netherlands
- MySQL AB, Sun, Strukton, Pentaho (1 nov)
- Web- and Business Intelligence Developer
- author:
 - Pentaho Solutions
 - Pentaho Kettle Solutions
- [Http://rpbouman.blogspot.com/](http://rpbouman.blogspot.com/)
- Twitter: @rolandbouman



Data Warehouse (DWH)

- Support Business Intelligence (BI)
 - Reporting
 - Analysis
 - Data mining
- General Requirements
 - Integrate disparate data sources
 - Maintain History
 - Calculate Derived data
 - Data delivery to BI applications

DWH Architectures

- Categories
 - Traditional
 - Hybrid
 - Modern
- Aspects
 - Modelling
 - Data logistics

DWH Architectures

- Traditional
 - Information Factory (Bill Inmon)
 - Enterprise Bus (Ralph Kimball)
- Hybrid
- Modern

DWH Architectures

- Traditional
- Hybrid
 - Hub-and-Spoke
- Modern

DWH Architectures

- Traditional
- Hybrid
- Modern
 - Data Vault (Dan Linstedt)
 - Anchor Modeling (Lars Rönnbäck)

Inmon DWH (Traditional): Corporate Information Factory

“A source of data that is subject oriented, integrated, nonvolatile and time variant for the purpose of management's decision processes.”

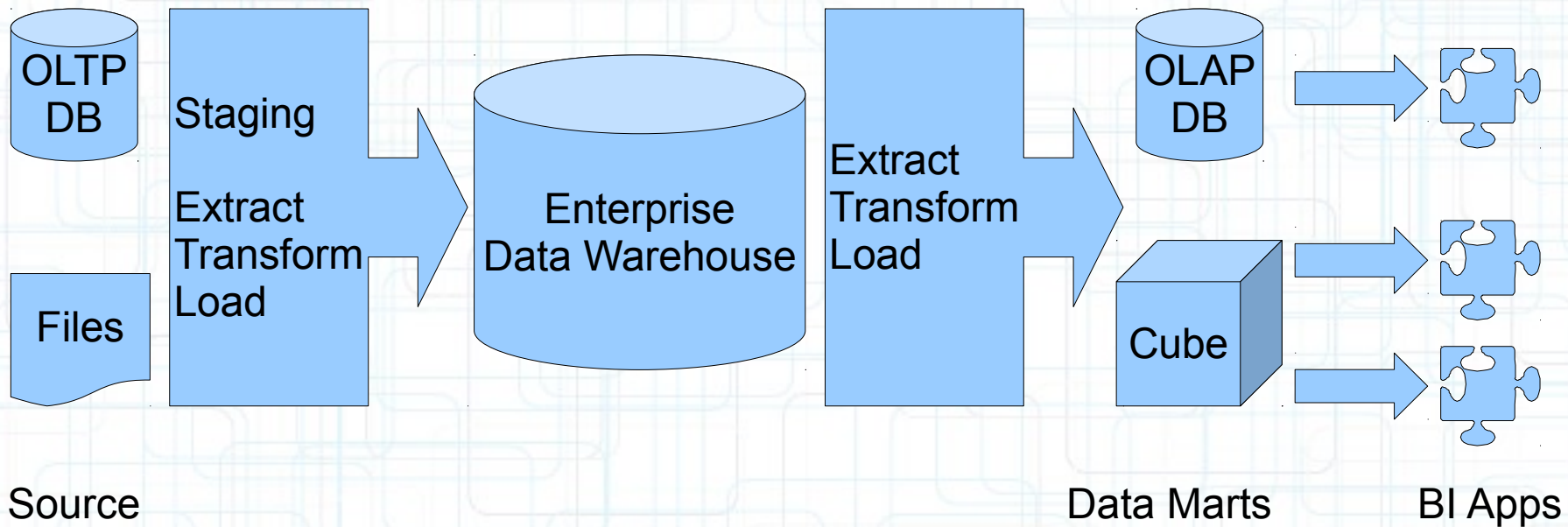
Bill Inmon (the Data Warehouse Toolkit)

• <http://www.inmoncif.com/home/>

Inmon DWH (Traditional): Corporate Information Factory

- Enterprise or Corporate DWH, DWH 2.0
- Focus on backroom data integration
 - Central information model
 - Single version of the truth
- Data delivery
 - Disposable data marts
- Bottom-up

Data logistics of the Corporate Information Factory



Data Modeling for the CIF Enterprise DWH

- Normalized, typically 3NF
- Organized in “subject areas”
 - Series of related tables
 - Example: Customer, Product, Transaction
 - Common key

Data Modeling for the CIF Enterprise DWH

- History
 - PK includes a date/timepart
- Contains both detail and aggregate data
 - Multiple levels of aggregation

Kimball DWH (Traditional): Dimensional Model and DWH Bus Architecture

“The data warehouse is the conglomeration of an organization's staging and presentation areas, where operational data is specifically structured for query and analysis performance and ease of use.”

Ralph Kimball (the Data Warehouse Toolkit)

• <http://www.kimballgroup.com/>

Kimball DWH (Traditional): DWH Bus Architecture

- Focus on data delivery
- Integration at the data mart level
- Top-down

Data Modeling for the DWH Bus Architecture

- Dimensional Modeling
 - Star schemas
- Organized in:
 - Fact tables
 - Dimension tables

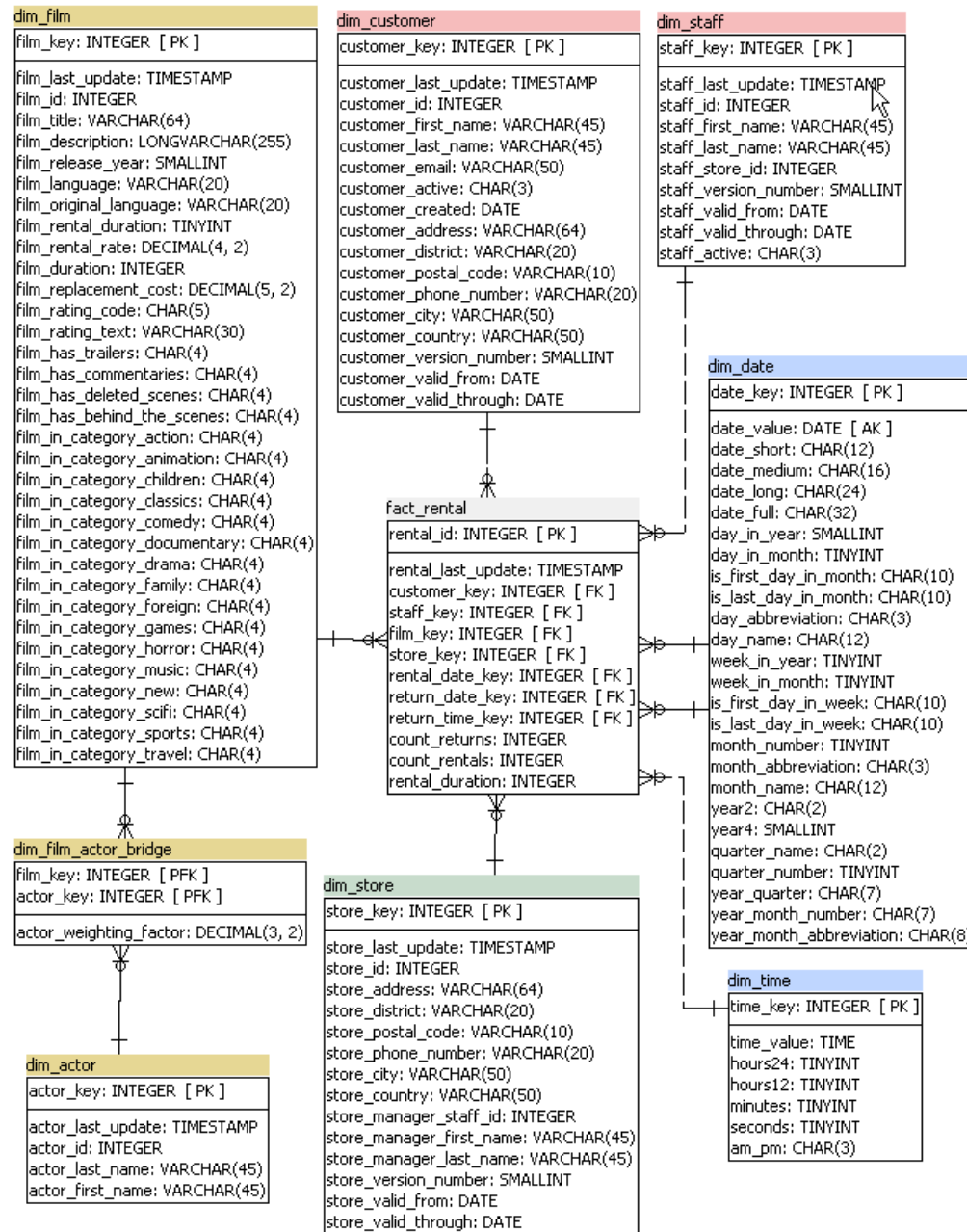
Data Modeling for the DWH Bus Architecture

- Fact tables
 - Highly normalized
 - Additive metrics
- Dimension tables
 - Highly denormalized
 - Descriptive labels
 - Shared across fact tables

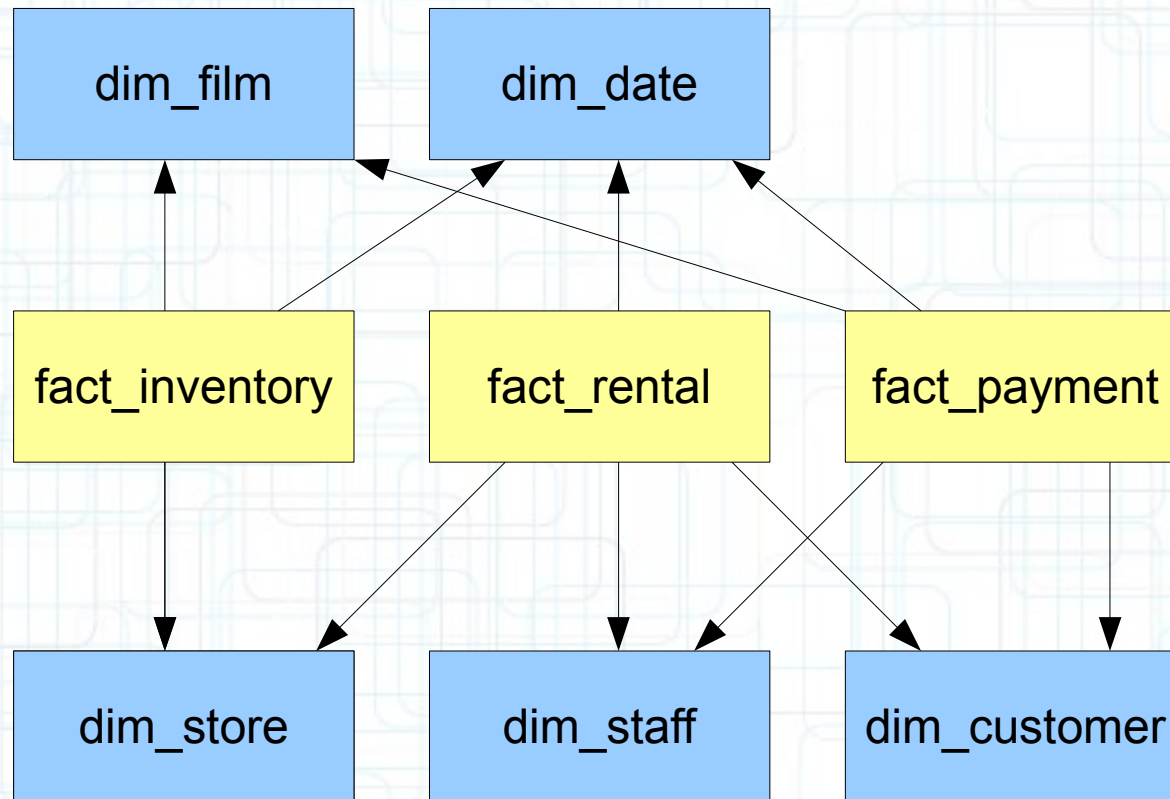
Data Modeling for the DWH Bus Architecture

- History
 - Slowly changing dimensions (versioning)
 - Fact links to Date and/or Time dimensions
- Detailed, not aggregated

Sakila Rental Star Schema



Sakila DWH Bus Architecture



Problems with traditional DWH architectures

- General Problems
 - Lack of flexibility and resilience to change
 - Loading (ETL) Complexity
- Problems with Inmon
 - Centralization requires upfront investment
 - Single version of whose truth, when?
- Problems with Kimball
 - Dimensional Model anomalies

Dimensional Modeling Anomalies

- Snowflaking (dimension normalization)
 - Monster dimensions
 - Outriggers
 - Ex: Customer Demographics
- Hierarchical data
 - Bridge table (closure table)
 - Ex: Employee/Boss,
- Multi-valued dimensions
 - Bridge table
 - Ex: Account/Customer bridge table

Hybrid DWH: Hub-and-Spoke

- Inmon back-end (hub)
- Kimball front-end (satellites)

Modern: Data Vault

“The Data Vault is a detail oriented, historical tracking and uniquely linked set of normalized tables that supports 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.”

Dan Linstedt (Data Vault Overview)

- <http://danlinstedt.com/>

Data Vault

- Focus on
 - Data Integration
 - Traceability and Auditability
 - Resilience to change
- Single version of the facts
 - Rather than single version of the truth
- All of the data, all of the time
 - No upfront cleansing and conforming
- Bottom-up

Data Vault Modelling

- Hubs
- Links
- Satellites

Data Vault Modelling: Hubs

- Hubs Model Entities
- Contains business keys
 - PK in absence of surrogate key
- Metadata:
 - Record source
 - Load date/time
- Optional surrogate key
 - Used as PK if present
- No foreign keys!

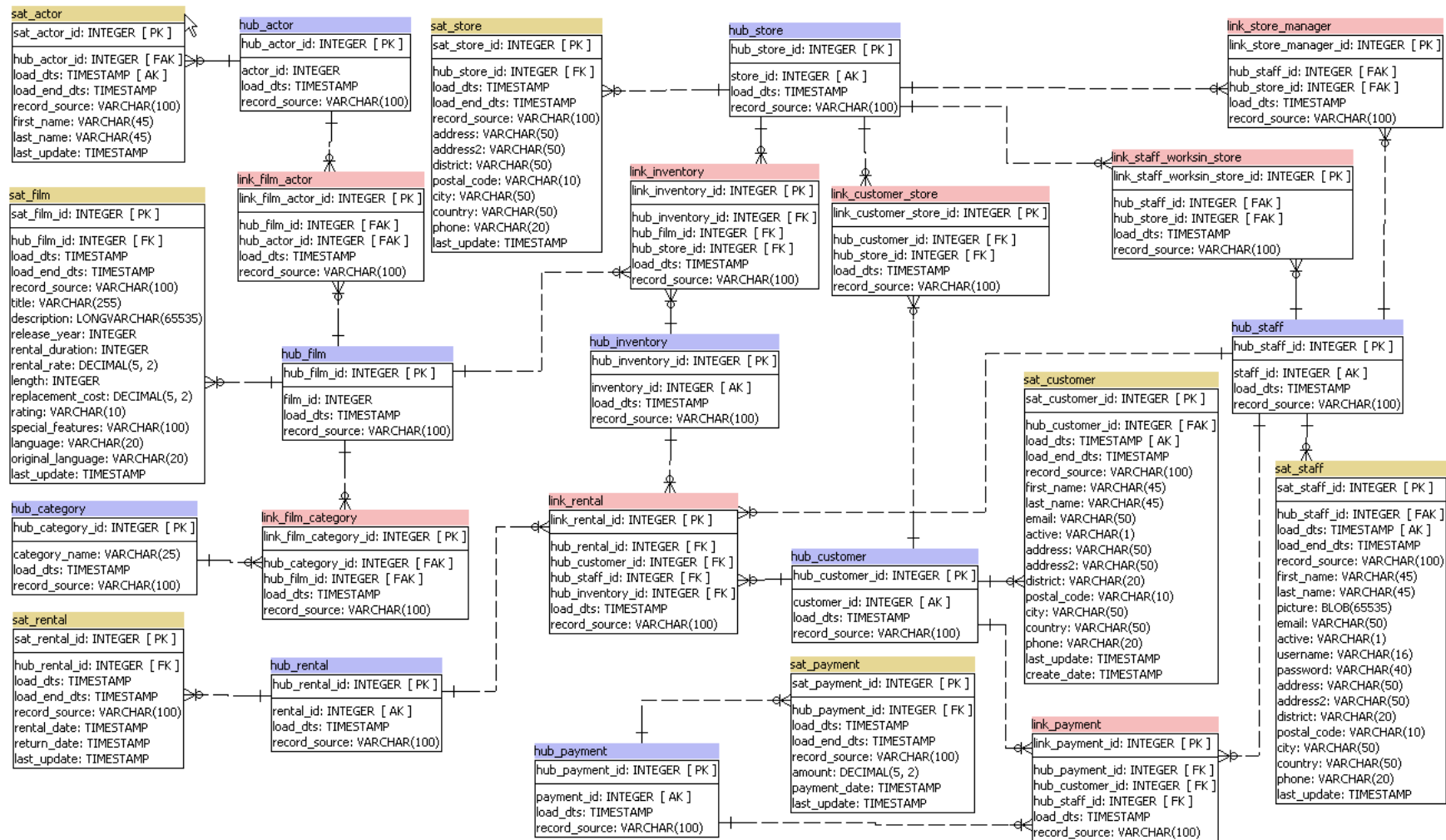
Data Vault Modelling: Links

- Links model relationships
 - Intersection table (M:n relationship)
- Foreign keys to related hubs or links
 - Form natural key (business key) of the link
- Metadata:
 - Record source
 - Load date/time
- Optional surrogate key

Data Vault Modelling: Satellites

- Satellites model a group of attributes
- Foreign key to a Hub or Link
- Metadata:
 - Record source
 - Load date/time

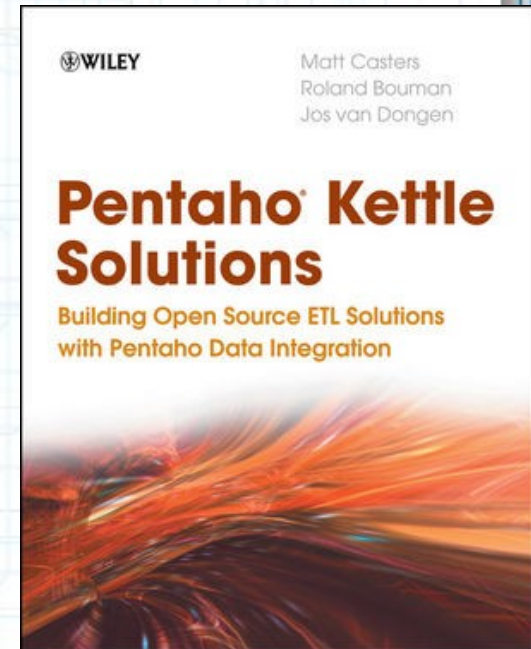
Sakila Data Vault Example



Data Vault tools and Example

- Kettle Data Vault Example

- Sakila Data Vault
- Chapter 19
- Kasper van de Graaf
- <http://www.dikw-academy.nl>



- Quipu

- Data Vault Generator
- Kettle templates
- Johannes van den Bosch
- <http://www.datawarehousemanagement.org/>



Modern: Anchor model

“Anchor Modeling is an agile information modeling technique that offers non-destructive extensibility mechanisms enabling robust and flexible management of changes. A key benefit of Anchor Modeling is that changes in a data warehouse environment only require extensions, not modifications.”

Lars Rönnbäck (Agile Information Modeling in Evolving Data Environments)

• <http://www.anchor modeling.com/>

Anchor Modelling

- Focus on
 - Resilience to change
 - Agility
 - Extensibility
 - History tracking
- Bottom-up

Anchor Modelling

- 6NF (Date, Darwen, Lorentzos)
- Table features no non-trivial join dependencies at all
- Translation: A 6NF table cannot be decomposed losslessly
- Translation
- Temporal Data

Anchor Modelling Constructs

- Anchors
- Attributes
- Ties
- Knots

Anchor Modelling: Anchors

- Entities are modeled as Anchors
- Relationships may be modeled as Anchors
 - m:n relationships having properties
- Only a surrogate key

Anchor Modelling: Ties

- Ties model relationships
 - 1:n relationships
 - m:n relationships without properties
- Static vs Historized
 - History tracked using date/time
- May be Knotted
 - Knot holds set of association types
- Two or more “anchor roles”
 - Relationships may be broken into several ties having only mandatory anchors

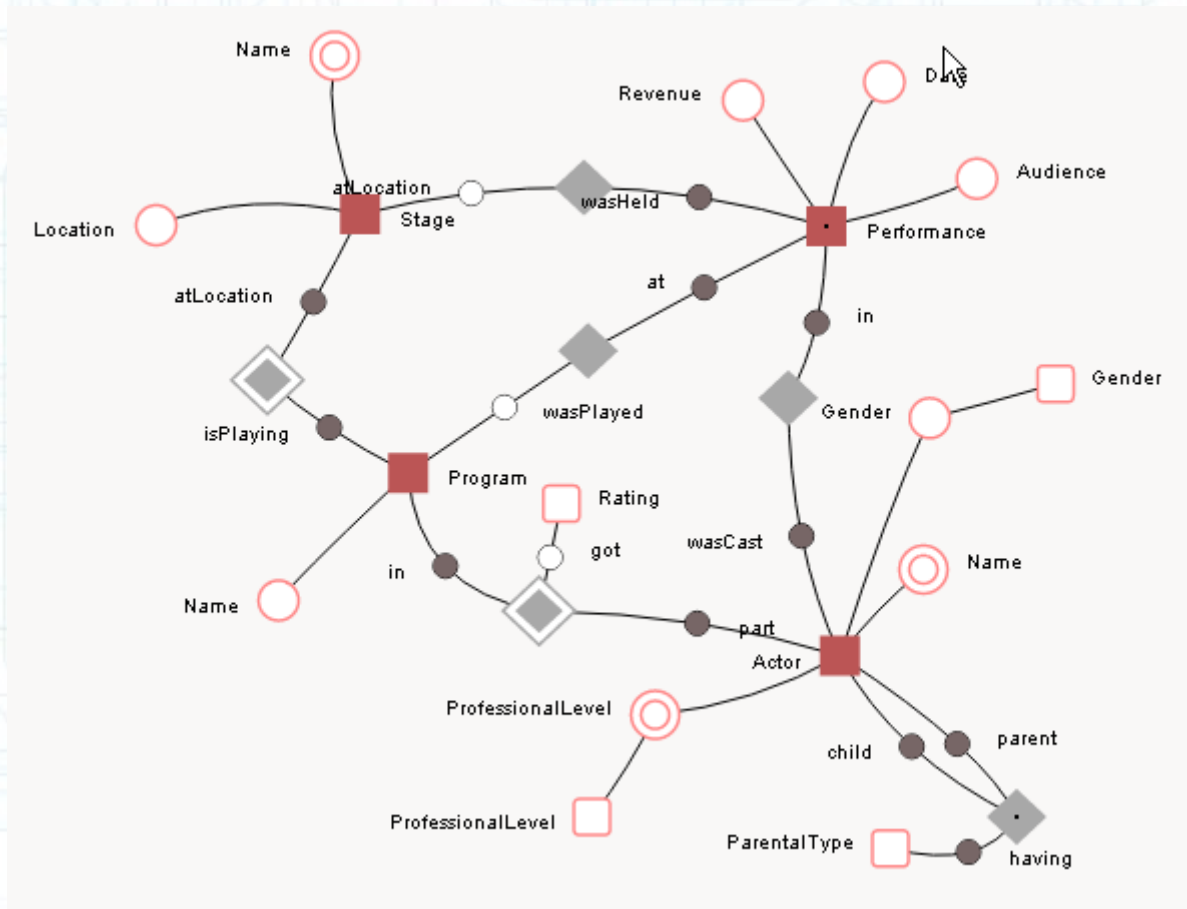
Anchor Modelling: Attributes

- Models properties of an Anchor
- Static vs Historized
 - History tracked using date/time
- May or not be Knotted
 - Knot holds set of valid attribute values

Anchor Modelling: Knots

- Reference table
 - Fairly small set of distinct values
- Dictionary lookup to qualify
 - Attributes
 - Ties
- “Knotted” Attributes and Ties

Anchor Model Diagram



- anchor
- Static attribute
- ◆ Static tie
- knot
- ◎ Historized attribute
- ◊ Historized tie

<http://www.anchor modeling.com/modeler/latest/>

Aknowledgements

- Kasper de Graaf
 - Twitter: @kdgraaf
 - <http://www.dikw-academy.nl>
- Jos van Dongen
 - Twitter: @josvandongen
 - <http://www.tholis.com/>