Lection 11

Designing Data Warehouses

- To begin a data warehouse project, we need to find answers for questions such as:
  - Which user requirements are most important and which data should be considered first?
  - Which data should be considered first?
  - Should the project be scaled down into something more manageable?
  - Should the infrastructure for a scaled down project be capable of ultimately delivering a full-scale enterprise-wide data warehouse?

Designing Data Warehouses

- For many enterprises the way to avoid the complexities associated with designing a data warehouse is to start by building one or more data marts.

- Data marts allow designers to build something that is far simpler and achievable for a specific group of users.

Designing Data Warehouses

- Few designers are willing to commit to an enterprise-wide design that must meet all user requirements at one time.
- Despite the interim solution of building data marts, the goal remains the same: that is, the ultimate creation of a data warehouse that supports the requirements of the enterprise.

Designing Data Warehouses

- The requirements collection and analysis stage of a data warehouse project involves interviewing appropriate members of staff (such as marketing users, finance users, and sales users) to enable the identification of a prioritized set of requirements that the data warehouse must meet.

Designing Data Warehouses

- At the same time, interviews are conducted with members of staff responsible for operational systems to identify, which data sources can provide clean, valid, and consistent data that will remain supported over the next few years.

Designing Data Warehouses

- Interviews provide the necessary information for the top-down view (user requirements) and the bottom-up view (which data sources are available) of the data warehouse.

- The database component of a data warehouse is described using a technique called dimensionality modeling.

Data Warehouse Development Methodologies

- There are two main methodologies that incorporate the development of an enterprise data warehouse (EDW) and these are proposed by the two key players in the data warehouse arena.
  - Kimball's Business Dimensional Lifecycle (Kimball, 2008)
  - Inmon's Corporate Information Factory (CIF) methodology (Inmon, 2001).

Data Warehouse Development Methodologies

Methodology	Main Advantage	Main Disadvantage
Inmon's Corporate Information Factory	Potential to provide a consistent and comprehensive view of the enterprise data.	Large complex project that may fail to deliver value within an allotted time period or budget.
Kimball's Business Dimensional Lifecycle	Scaled down project means that ability to demonstrate value is more achievable within an allotted time period or budget.	As data marts can potentially be developed in sequence by different development teams using different systems; the ultimate goal of providing a consistent and comprehensive view of corporate data may never be easily achieved.

Kimballs' Business Dimensional Lifecycle

- Ralph Kimball is a key player in DW.
- About creation of an infrastructure capable of supporting all the information needs of an enterprise.
- Uses new methods and techniques in the development of an enterprise data warehouse (EDW).

Kimballs' Business Dimensional Lifecycle

- Starts by identifying the information requirements (referred to as analytical themes) and associated business processes of the enterprise.
- This activity results in the creation of a critical document called a Data Warehouse Bus Matrix.

Kimballs' Business Dimensional Lifecycle

- The matrix lists all of the key business processes of an enterprise together with an indication of how these processes are to be analysed.
- The matrix is used to facilitate the selection and development of the first database (data mart) to meet the information requirements of a particular department of the enterprise.

Kimballs' Business Dimensional Lifecycle

- This first data mart is critical in setting the scene for the later integration of other data marts as they come online.
- The integration of data marts ultimately leads to the development of an EDW.
- Uses dimensionality modeling to establish the data model (called star schema) for each data mart.

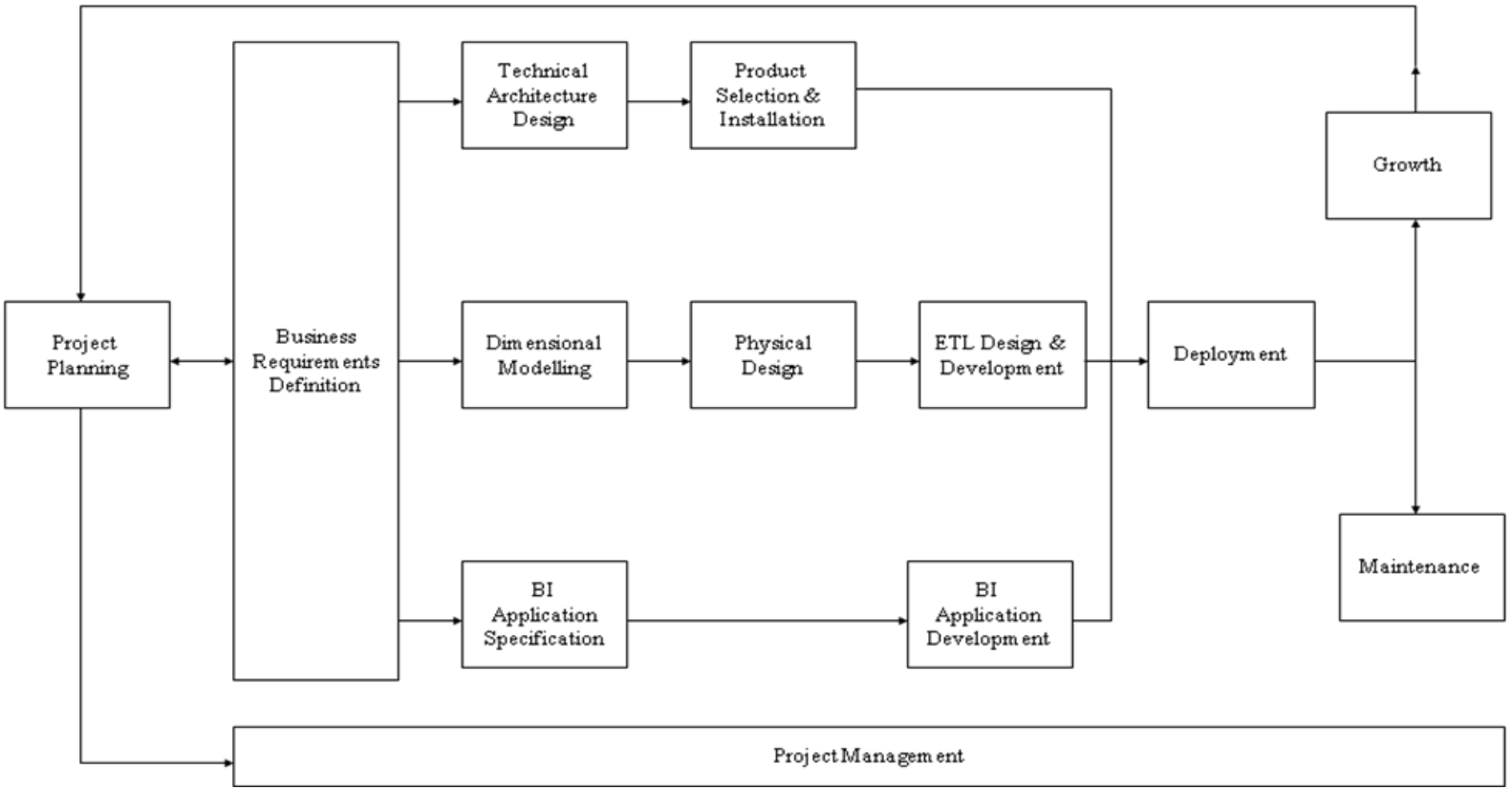
Kimballs' Business Dimensional Lifecycle

- Guiding principle is to meet the information requirements of the enterprise by building a single, integrated, easy-to-use, high-performance information infrastructure, which is delivered in meaningful increments of 6 to 12 month timeframes.
- Goal is to deliver the entire solution including the data warehouse, *ad hoc* query tools, reporting applications, advanced analytics and all the necessary training and support for the users.

Kimballs' Business Dimensional Lifecycle

- Has three tracks:
  - technology (top track),
  - data (middle track),
  - business intelligence (BI) applications (bottom track).
- Uses incremental and iterative approach that involves the development of data marts that are eventually integrated into an enterprise data warehouse (EDW).

Kimballs' Business Dimensional Lifecycle



Dimensionality modeling

- A logical design technique that aims to present the data in a standard, intuitive form that allows for high-performance access

- Every dimensional model (DM) is composed of one table with a composite primary key, called the fact table, and a set of smaller tables called dimension tables.

Dimensionality modeling

- Each dimension table has a simple (non-composite) primary key that corresponds exactly to one of the components of the composite key in the fact table.

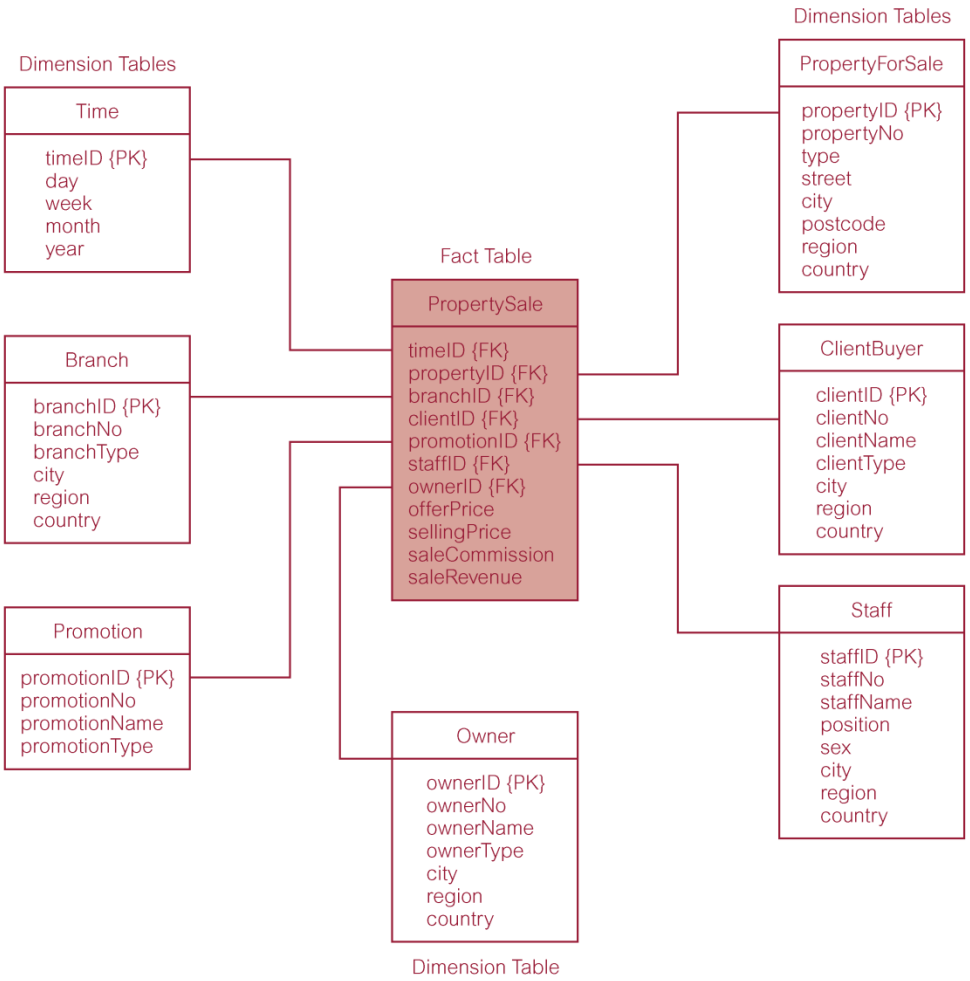
- Forms 'star-like' structure, which is called a star schema or star join.

Dimensionality modeling

- All natural keys are replaced with surrogate keys. Means that every join between fact and dimension tables is based on surrogate keys, not natural keys.

- Surrogate keys allows the data in the warehouse to have some independence from the data used and produced by the OLTP systems.

Star schema (dimensional model)



Dimensionality modeling

- Star schema is a logical structure that has a fact table (containing factual data) in the center, surrounded by denormalized dimension tables (containing reference data).

- Facts are generated by events that occurred in the past, and are unlikely to change, regardless of how they are analyzed.

Dimensionality modeling

- Bulk of data in data warehouse is in fact tables, which can be extremely large.
- Important to treat fact data as read-only reference data that will not change over time.
- Most useful fact tables contain one or more numerical measures, or 'facts' that occur for each record and are numeric and additive.

Dimensionality modeling

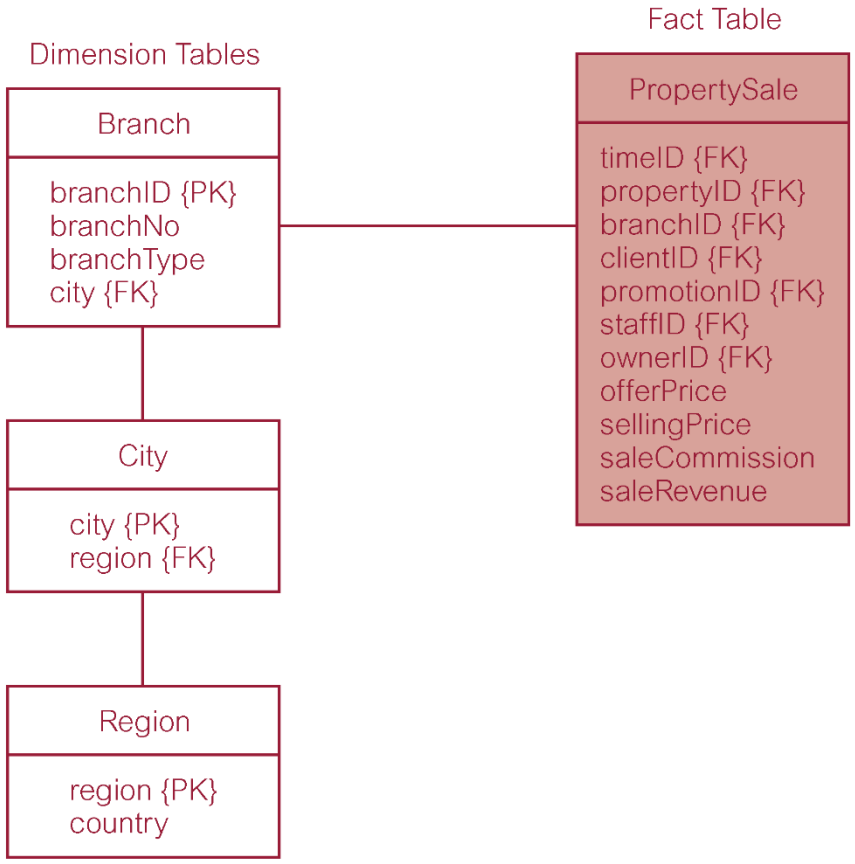
- Dimension tables usually contain descriptive textual information.
- Dimension attributes are used as the constraints in data warehouse queries.
- Star schemas can be used to speed up query performance by denormalizing reference information into a single dimension table.

Dimensionality modeling

- Snowflake schema is a variant of the star schema that has a fact table in the center, surrounded by normalized dimension tables .

- Starflake schema is a hybrid structure that contains a mixture of star (denormalized) and snowflake (normalized) dimension tables.

Property sales with normalized version of Branch dimension table



Dimensionality modeling

- Predictable and standard form of the underlying dimensional model offers important advantages:
  - Efficiency
  - Ability to handle changing requirements
  - Extensibility
  - Ability to model common business situations
  - Predictable query processing

Comparison of DM and ER models

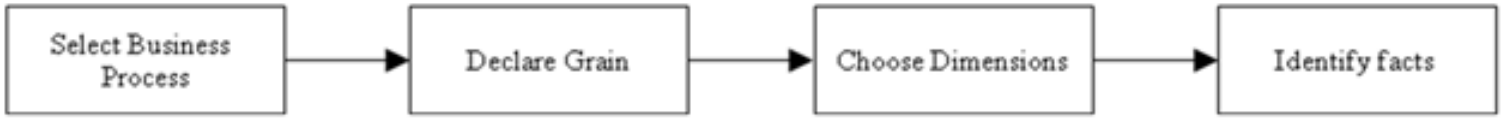
- A single ER model normally decomposes into multiple DMs.
- Multiple DMs are then associated through 'shared' dimension tables.

Dimensional Modeling Stage of Kimball's Business Dimensional Lifecycle

- Begins by defining a high-level dimension model (DM), which progressively gains more detail and this is achieved using a two-phased approach.
- The first phase is the creation of the high-level DM and the second phase involves adding detail to the model through the identification of dimensional attributes for the model.

Dimensional Modeling Stage of Kimball's Business Dimensional Lifecycle

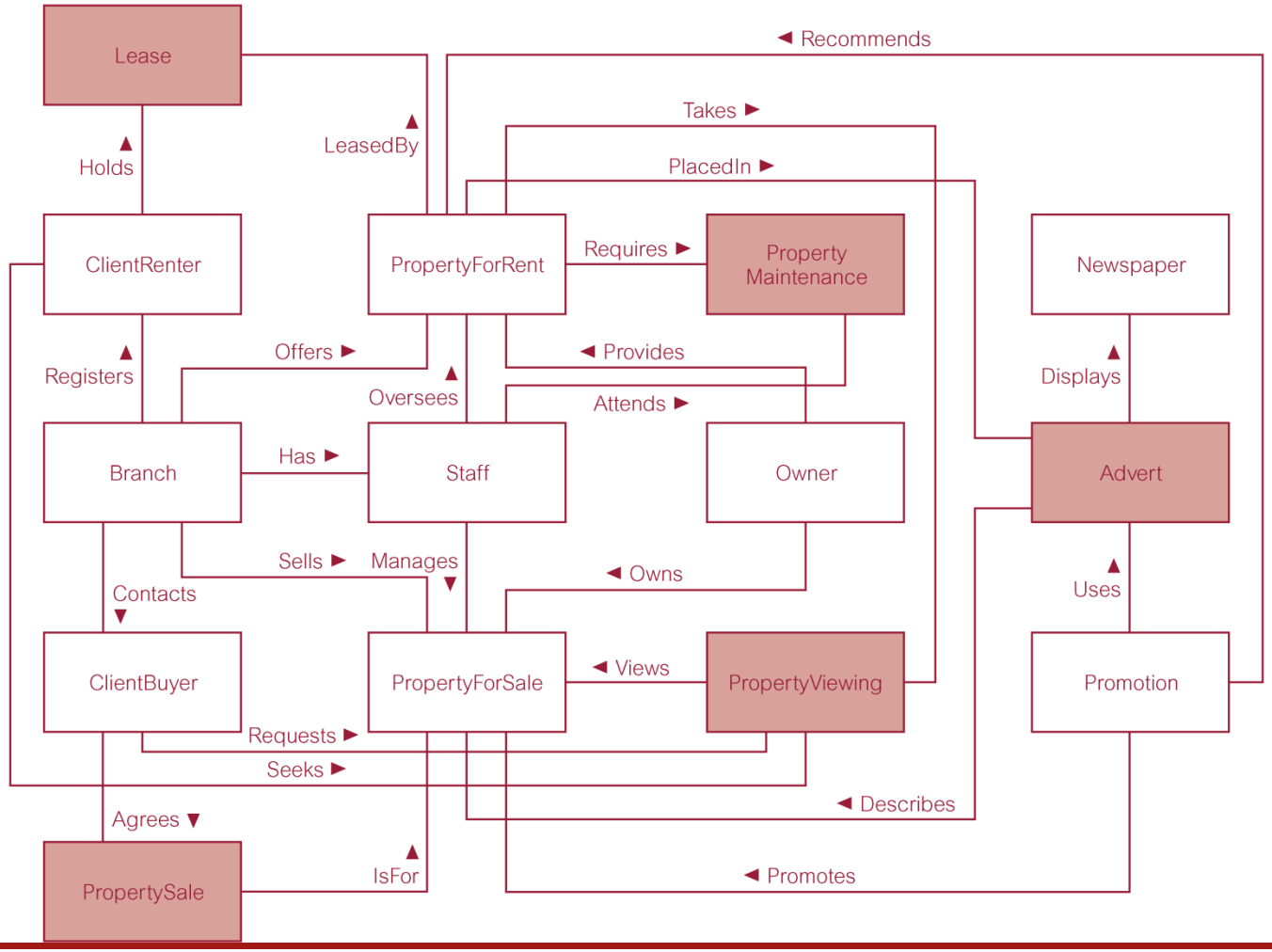
- Phase 1 involves the creation of a high-level dimensional model (DM) using a four-step process.



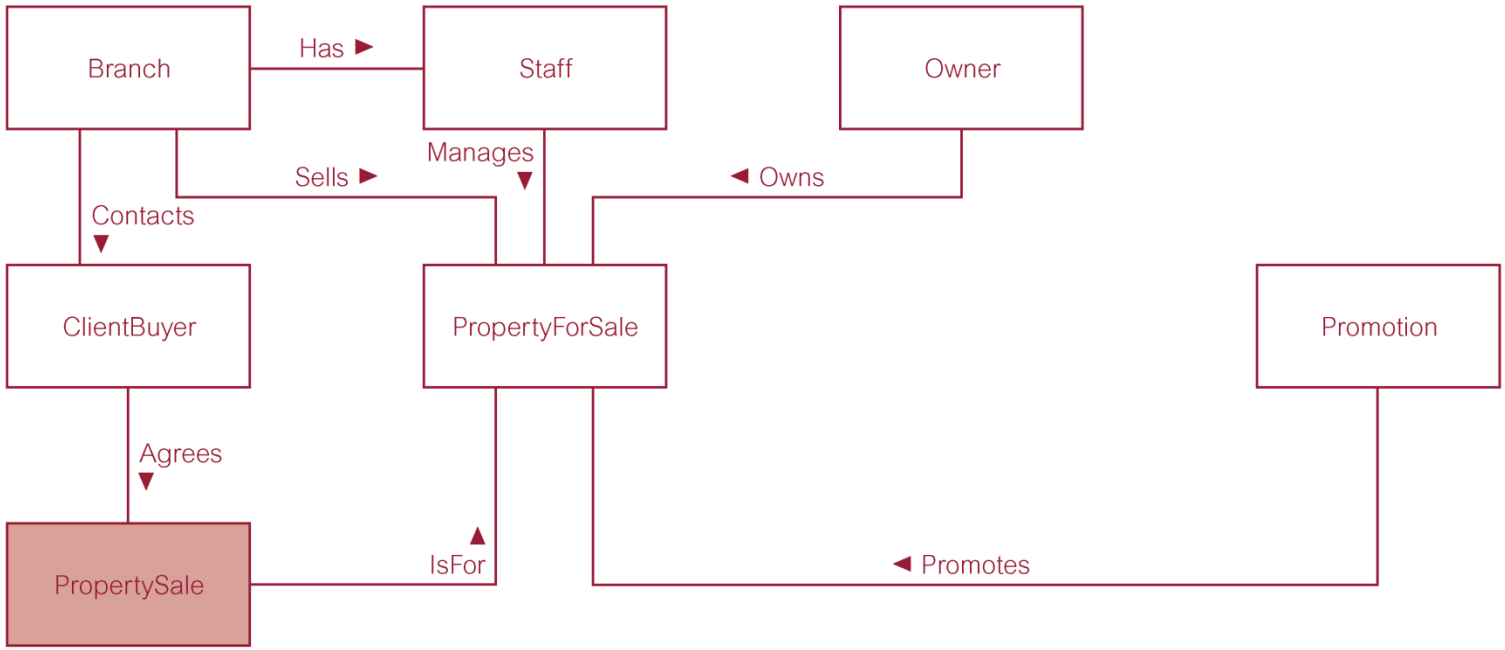
Step 1: Select business process

- The process (function) refers to the subject matter of a particular data mart.
- First data mart built should be the one that is most likely to be delivered on time, within budget, and to answer the most commercially important business questions.

ER model



ER model of property sales business process



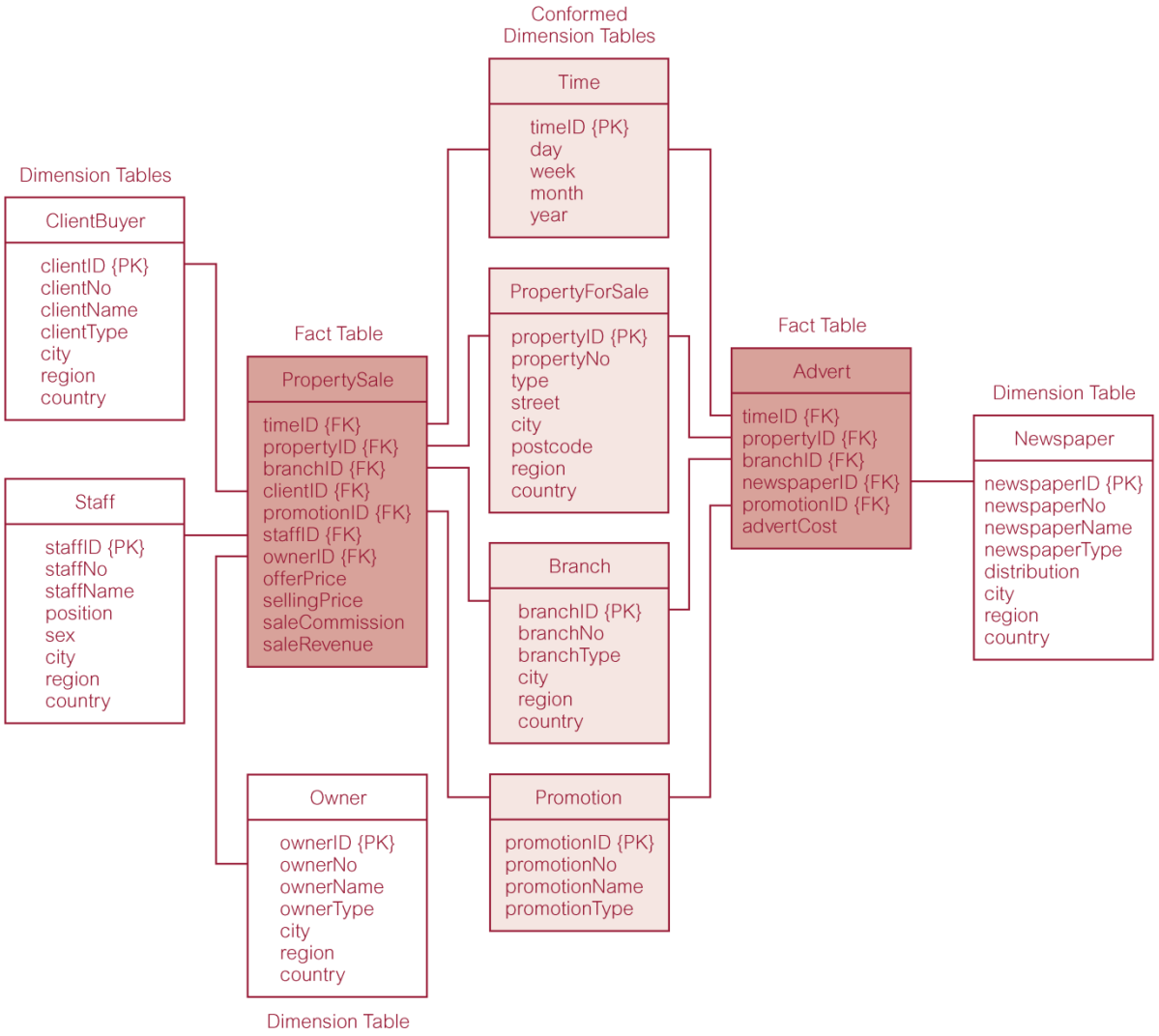
Step 2: Declare grain

- Decide what a record of the fact table is to represents.
- Identify dimensions of the fact table. The grain decision for the fact table also determines the grain of each dimension table.
- Also include time as a core dimension, which is always present in star schemas.

Step 3: Choose dimensions

- Dimensions set the context for asking questions about the facts in the fact table.
- If any dimension occurs in two data marts, they must be exactly the same dimension, or one must be a mathematical subset of the other.
- A dimension used in more than one data mart is referred to as being conformed.

Star schemas for property sales and property advertising



Step 4: Identify facts

- The grain of the fact table determines which facts can be used in the data mart.
- Facts should be numeric and additive.
- Unusable facts include:
  - non-numeric facts
  - non-additive facts
  - fact at different granularity from other facts in table

Step 4: Identify facts

- Once the facts have been selected each should be re-examined to determine whether there are opportunities to use pre-calculations.

Dimensional Modeling Stage of Kimball's Business Dimensional Lifecycle

- Phase 2 involves the rounding out of the dimensional tables.
- Text descriptions are added to the dimension tables and be as intuitive and understandable to the users as possible.
- Usefulness of a data mart is determined by the scope and nature of the attributes of the dimension tables.

Additional design issues

- Duration measures how far back in time the fact table goes.

- Very large fact tables raise at least two very significant data warehouse design issues.
  - Often difficult to source increasing old data.
  - It is mandatory that the old versions of the important dimensions be used, not the most current versions. Known as the 'Slowly Changing Dimension' problem.

Additional design issues

- Slowly changing dimension problem means that the proper description of the old dimension data must be used with the old fact data.

- Often, a generalized key must be assigned to important dimensions in order to distinguish multiple snapshots of dimensions over a period of time.

Additional design issues

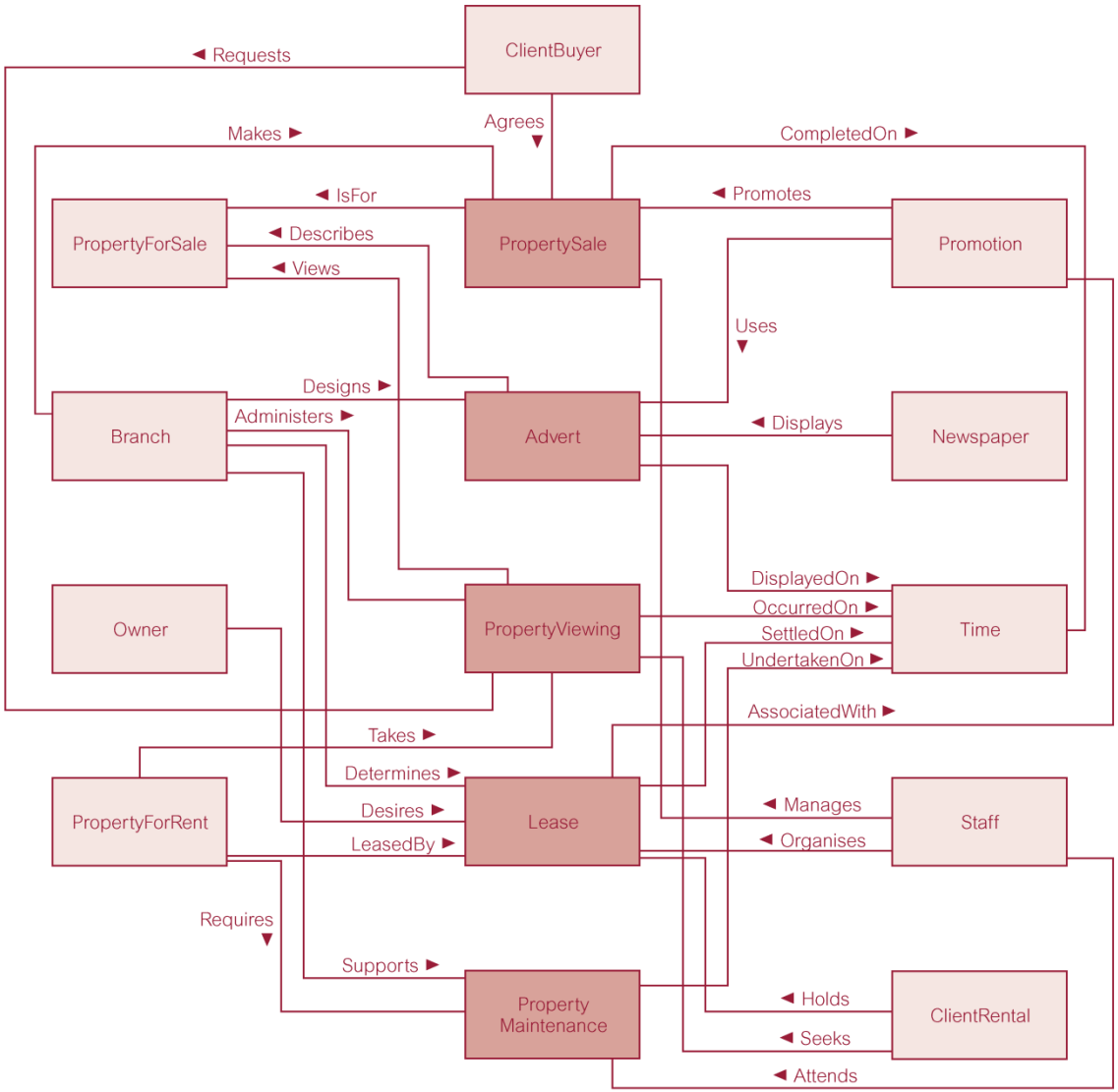
- There are three basic types of slowly changing dimensions:
  - Type 1, where a changed dimension attribute is overwritten
  - Type 2, where a changed dimension attribute causes a new dimension record to be created
  - Type 3, where a changed dimension attribute causes an alternate attribute to be created so that both the old and new values of the attribute are simultaneously accessible in the same dimension record

Kimball's Business Dimensional lifecycle

- Lifecycle produces a data mart that supports the requirements of a particular business process and allows the easy integration with other related data marts to form the enterprise-wide data warehouse.

- A dimensional model, which contains more than one fact table sharing one or more conformed dimension tables, is referred to as a fact constellation.

Dimensional model (fact constellation) for the *DreamHome* data warehouse



Data Warehouse Development Issues

- Selection development methodology.
- Identification of key decision-makers to be supported their analytical requirements.
- Identification of data sources and assess the quality of the data.
- Selection of the ETL tool.
- Identification of strategy for meta-data be management.

Data Warehouse Development Issues

- Establishment of characteristics of the data e.g. granularity, latency, duration and data lineage.
- Establish storage capacity requirements for the database.
- Establishment of the data refresh requirements.
- Identification of analytical tools.
- Establishing an appropriate architecture for the DW/DM environment .
- Deal with the organisational, cultural and political issues associated with data ownership.