LECTION 10

The Evolution of Data Warehousing

- Since 1970s, organizations gained competitive advantage through systems that automate business processes to offer more efficient and cost-effective services to the customer.
- This resulted in accumulation of growing amounts of data in operational databases.

The Evolution of Data Warehousing

- Organizations now focus on ways to use operational data to support decision-making, as a means of gaining competitive advantage. However, operational systems were never designed to support such business activities.
- Businesses typically have numerous operational systems with overlapping and sometimes contradictory definitions.

The Evolution of Data Warehousing

- Organizations need to turn their archives of data into a source of knowledge, so that a single integrated / consolidated view of the organization's data is presented to the user.

- A data warehouse (DW) was deemed the solution to meet the requirements of a system capable of supporting decision-making, receiving data from multiple operational data sources.

Data Warehousing Concepts

- A subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process (Inmon, 1993).

Subject-oriented Data

- The warehouse is organized around the major subjects of the enterprise (e.g. customers, products, and sales) rather than the major application areas (e.g. customer invoicing, stock control, and product sales).

- This is reflected in the need to store decision-support data rather than application-oriented data.

Integrated Data

- The data warehouse integrates corporate application-oriented data from different source systems, which often includes data that is inconsistent.
- The integrated data source must be made consistent to present a unified view of the data to the users.

Time-variant Data

- Data in the warehouse is only accurate and valid at some point in time or over some time interval.
- Time-variance is also shown in the extended time that the data is held, the implicit or explicit association of time with all data, and the fact that the data represents a series of snapshots.

Non-volatile Data

- Data in the warehouse is not normally updated in real-time (RT) but is refreshed from operational systems on a regular basis. (However, emerging trend is towards RT or near RT DWs)

- New data is always added as a supplement to the database, rather than a replacement.

Benefits of Data Warehousing

- Potential high returns on investment
- Competitive advantage
- Increased productivity of corporate decision-makers

Comparison of OLTP Systems and Data Warehousing

CHARACTERISTIC	OLTP SYSTEMS	DATA WAREHOUSING SYSTEMS
Main purpose	Support operational processing	Support analytical processing
Data age	Current	Historic (but trend is toward also including current data)
Data latency	Real-time	Depends on length of cycle for data supplements to warehouse (but trend is toward real-time supplements)
Data granularity	Detailed data	Detailed data, lightly and highly summarized data
Data processing	Predictable pattern of data insertions, deletions, updates, and queries. High level of transaction throughput.	Less predictable pattern of data queries; medium to low level of transaction throughput
Reporting	Predictable, one-dimensional, relatively static fixed reporting	Unpredictable, multidimensional, dynamic reporting
Users	Serves large number of operational users	Serves lower number of managerial users (but trend is also toward supporting analytical requirements of operational users)

Data Warehouse Queries

- The types of queries that a data warehouse is expected to answer ranges from the relatively simple to the highly complex and is dependent on the type of end-user access tools used.
- End-user access tools include:
  - Traditional reporting and query
  - OLAP
  - Data mining

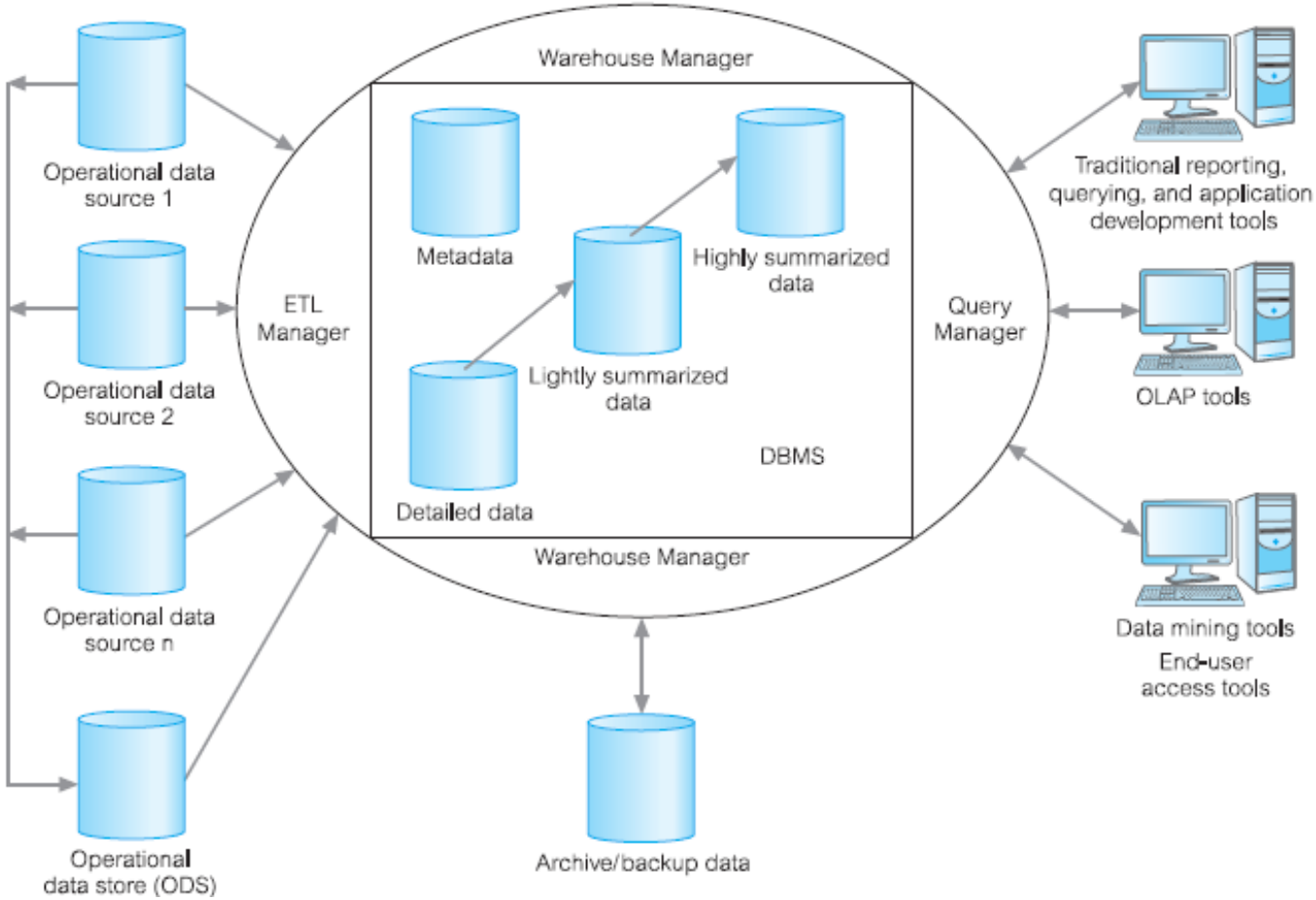
Data Warehouse Queries

- What was the total revenue for Scotland in the third quarter of 2001?
- What was the total revenue for property sales for each type of property in Great Britain in 2000?
- What are the three most popular areas in each city for the renting of property in 2001 and how does this compare with the figures for the previous two years?
- What is the monthly revenue for property sales at each branch office, compared with rolling 12-monthly prior figures?
- Which type of property sells for prices above the average selling price for properties in the main cities of Great Britain and how does this correlate to demographic data?
- What is the relationship between the total annual revenue generated by each branch office and the total number of sales staff assigned to each branch office?

Problems of Data Warehousing

- Underestimation of resources for data loading
- Hidden problems with source systems
- Required data not captured
- Increased end-user demands
- Data homogenization
- High demand for resources
- Data ownership
- High maintenance
- Long duration projects
- Complexity of integration

Example Data Warehouse Architecture



Operational Data Sources

- Main sources are online transaction processing (OLTP) databases.
- Also include sources such as personal databases and spreadsheets, Enterprise Resource Planning (ERP) files, and web usage log files.

Operational Data Store (ODS)

- Holds current and integrated operational data for analysis.
- Often structured and supplied with data in the same way as the data warehouse.
- May act as staging area for data to be moved into the warehouse.
- Often created when legacy operational systems are found to be incapable of achieving reporting requirements.

ETL Manager

- Data for an EDW must be extracted from one or more data sources, transformed into a form that is easy to analyze and consistent with data already in the warehouse, and then finally loaded into the DW.

ETL Manager

- Nowadays there are tools that automate the extraction, transformation, and loading (ETL) processes and also offer additional facilities such as data profiling, data quality control, and metadata management.

Warehouse Manager

- Performs all the operations associated with the management of the data in the warehouse such as:
  - Analysis of data to ensure consistency.
  - Transformation and merging of source data from temporary storage into data warehouse tables.
  - Creation of indexes and views on base tables.
  - Generation of denormalizations, (if necessary).
  - Generation of aggregations, (if necessary).
  - Backing-up and archiving data.

Warehouse Manager

- In some cases, also generates query profiles to determine which indexes and aggregations are appropriate.

- A query profile can be generated for each user, group of users, or the data warehouse and is based on information that describes the characteristics of the queries such as frequency, target table(s), and size of results set.

Query Manager

- Performs the operations associated with the management of user queries such as –
  - Directing queries to the appropriate tables and scheduling the execution of queries.
  - In some cases, the query manager also generates query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate.

Metadata

- Used for a variety of purposes and so the effective management of metadata is critical in achieving a fully integrated DW.
- Metadata (data about data) definitions are used by processes in the DW such as:
  - To map data sources to a common view of information within the warehouse.
  - To automate the production of summary tables.
  - To direct a query to the most appropriate data source.

End-User Access Tools

- Main purpose of DW is to support decision makers and this is achieved through the provision of a range of access tools including:
  - reporting and querying,
  - application and development,
  - OLAP,
  - data mining.

Data Warehousing Tools and technologies – ETL Processes

- *Extraction*
  - Targets one or more data sources and these sources typically include OLTP databases but can also include personal databases and spreadsheets, Enterprise Resource Planning (ERP) files, and web usage log files.
  - The data sources are normally internal but can also include external sources such as the systems used by suppliers and/or customers.

Data Warehousing Tools and technologies – ETL Processes

- *Transformation*
  - Applies a series of rules or functions to the extracted data, which determines how the data will be used for analysis and can involve transformations such as data summations, data encoding, data merging, data splitting, data calculations, and creation of surrogate keys.

Data Warehousing Tools and technologies – ETL Processes

- *Loading*
  - As data loads additional constraints defined in the database schema can be activated (such as uniqueness, referential integrity, and mandatory fields), which contribute to the overall data quality performance of the ETL process.

Data Warehousing Tools and technologies – ETL Tools

- Data profiling and quality control
  - Provides important information about the quantity and quality of the data coming from the source systems.

- Metadata management
  - Understanding a query result can require consideration of the data history i.e. What happened to the data during the ETL process? The answers are held in the metadata repository.

Data Warehouse DBMS Requirements

- Load performance
- Load processing
- Data quality management
- Query performance
- Terabyte scalability
- Mass user scalability
- Networked data warehouse
- Warehouse administration
- Integrated dimensional analysis
- Advanced query functionality

Data Mart

- A database that contains a subset of corporate data to support the analytical requirements of a particular business unit (such as the Sales department) or to support users who share the same requirements to analyse a particular business process (such as property sales).

Reasons for Creating a Data Mart

- To give users access to the data they need to analyze most often.
- To provide data in a form that matches the collective view of the data by a group of users in a department or business application area.
- To improve end-user response time due to the reduction in the volume of data to be accessed.
- To provide appropriately structured data as dictated by the requirements of the end-user access tools.

Reasons for Creating a Data Mart

- Building a data mart is simpler compared with establishing an enterprise-wide DW (EDW).
- The cost of implementing data marts is normally less than that required to establish a EDW.
- The future users of a data mart are more easily defined and targeted to obtain support for a data mart than an enterprise-wide data warehouse project.

Data Warehousing and Temporal Databases

- Data warehousing systems must manage relationships that exist between the historical data and the new data.
- Data warehouses are described as being temporal databases.

Data Warehousing and Temporal Databases

- Temporal data is data that changes over time.
- A temporal database contains time-varying historical data with the possible inclusion of current and future data and has the ability to manipulate this data.

Data Warehousing and Temporal Databases

- Shows historical property records with PK {propertyNo, year}

propertyNo	city	rent	year	ownerNo
PA14	Aberdeen	580	2011	CO46
PA14	Aberdeen	595	2012	CO46
PA14	Aberdeen	635	2013	CO46
PA14	Aberdeen	650	2014	CO46
PG21	Glasgow	578	2012	CO87
PG21	Glasgow	590	2013	CO87
PG21	Glasgow	600	2014	CO87

Data Warehousing and Temporal Databases

- Shows historical property records with PK {propertyNo, startDate, endDate}

propertyNo	city	rent	startDate	endDate	ownerNo
PA14	Aberdeen	580	01/01/2012	31/03/2012	CO46
PA14	Aberdeen	595	01/04/2012	31/04/2013	CO46
PA14	Aberdeen	600	01/05/2013	31/10/2013	CO46
PA14	Aberdeen	620	01/11/2013	31/03/2014	CO46
PG14	Aberdeen	635	01/04/2014	30/06/2014	CO46
PG14	Aberdeen	650	01/07/2014	31/12/2014	CO46
PG21	Glasgow	540	01/01/2012	30/02/2012	CO87
PA21	Glasgow	545	01/03/2011	30/04/2012	CO87
PA21	Glasgow	585	01/05/2012	31/10/2013	CO87
PA21	Glasgow	590	01/11/2013	31/03/2014	CO87
PG21	Glasgow	600	01/04/2014	31/12/2014	CO87

LECTION 10

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.

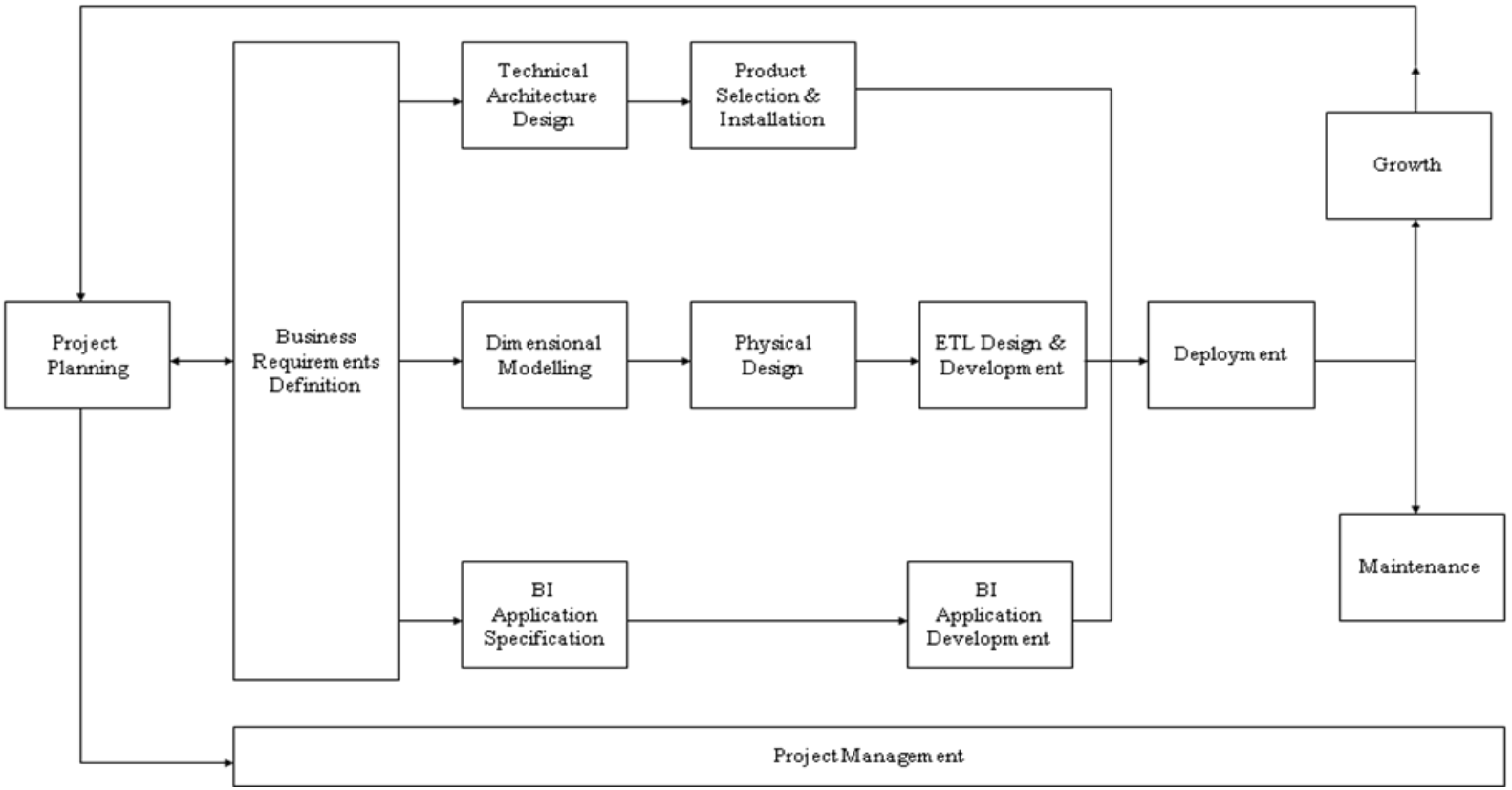
Kimballs' Business Dimensional Lifecycle

- 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.

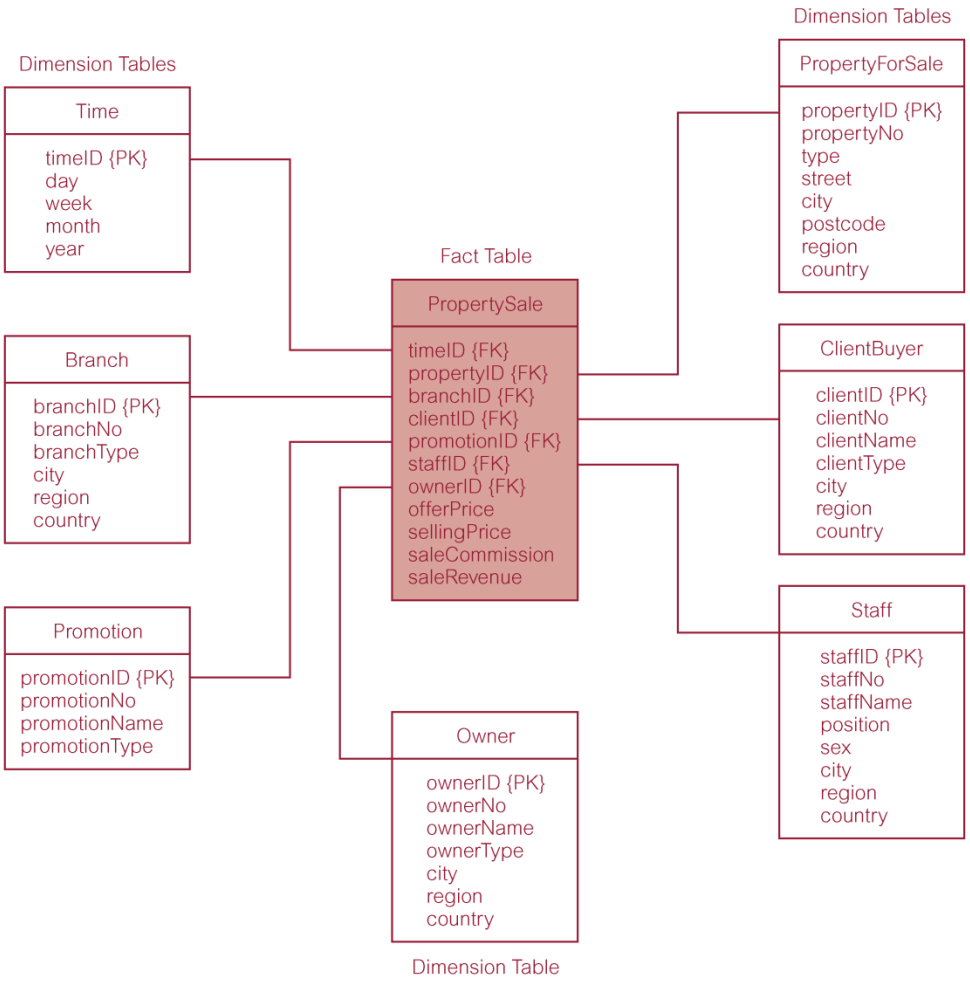
- Forms 'star-like' structure, which is called a star schema or star join.

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

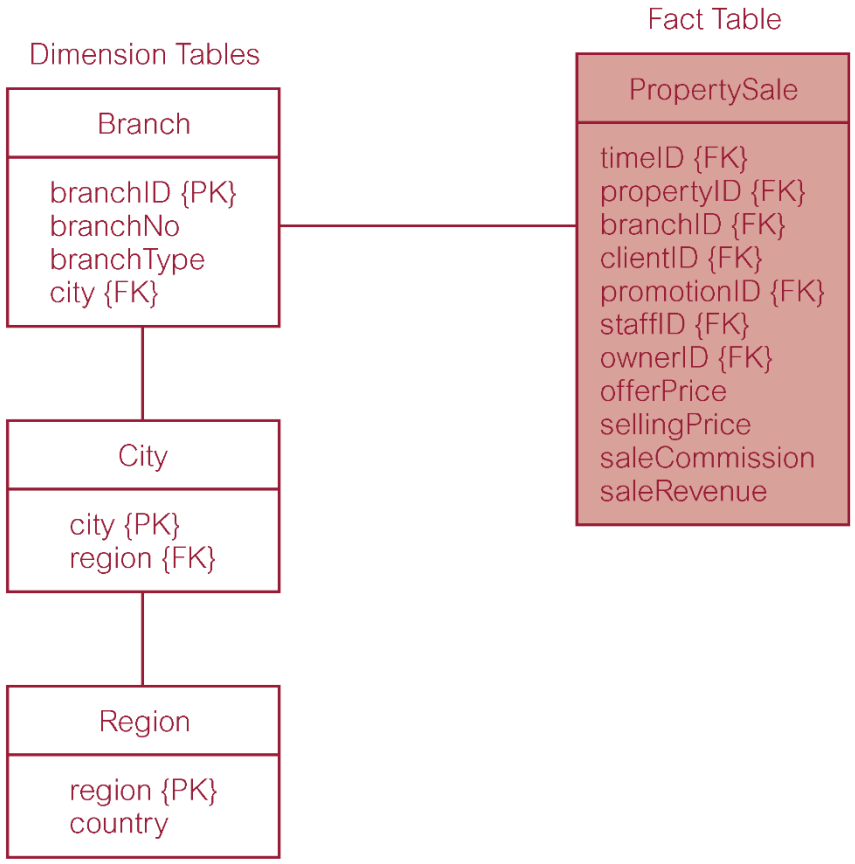
- Dimension tables usually contain descriptive textual information.
- Dimension attributes are used as the constraints in data warehouse queries.
- Star schemas can be used to speed up query performance by denormalizing reference information into a single dimension table.

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

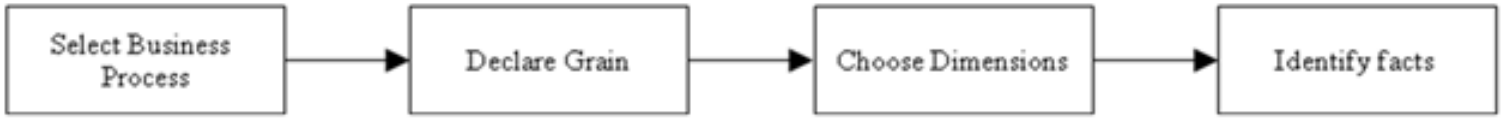
- A single ER model normally decomposes into multiple DMs.
- Multiple DMs are then associated through 'shared' dimension tables.

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

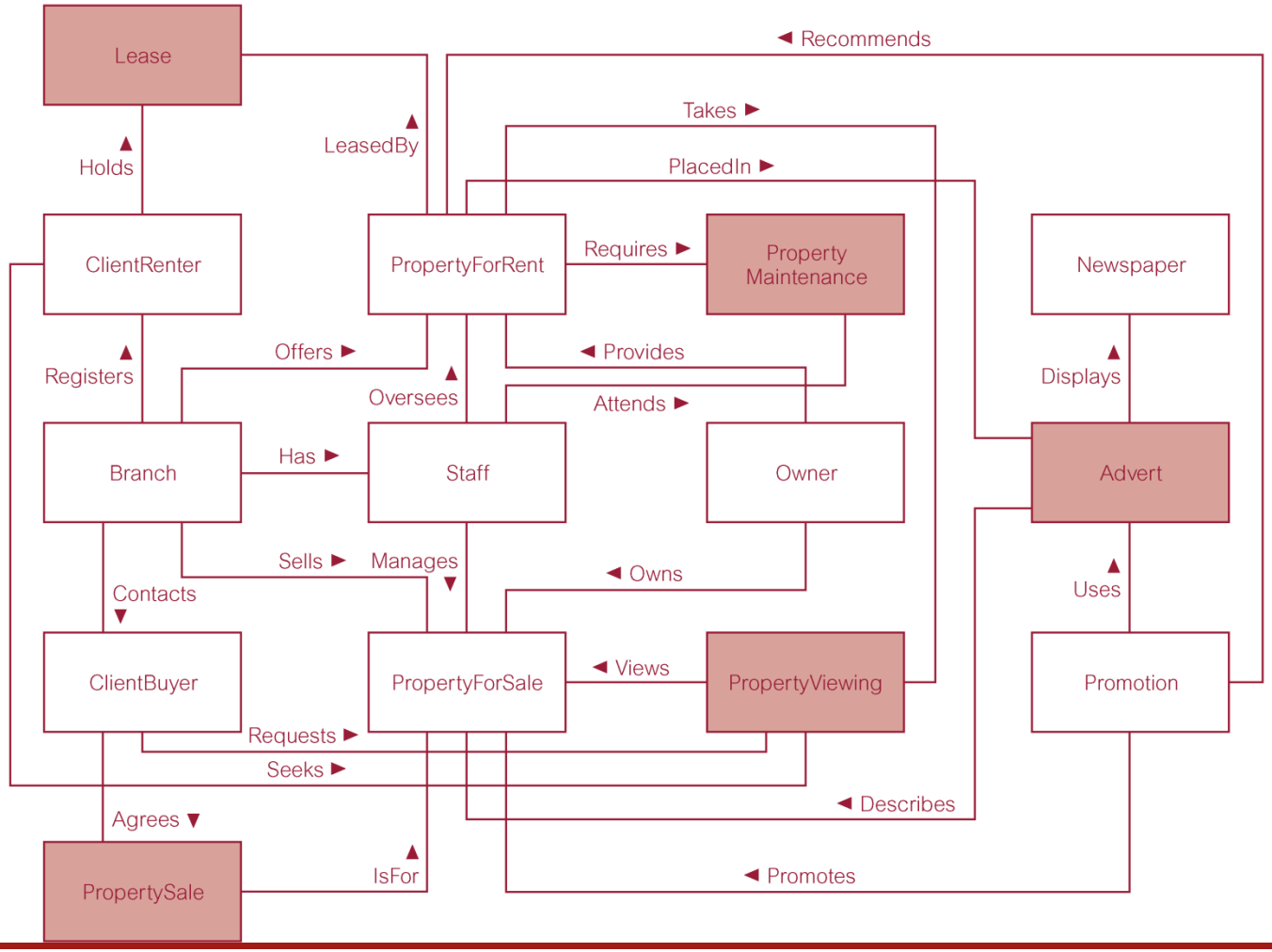
- Phase 1 involves the creation of a high-level dimensional model (DM) using a four-step process.



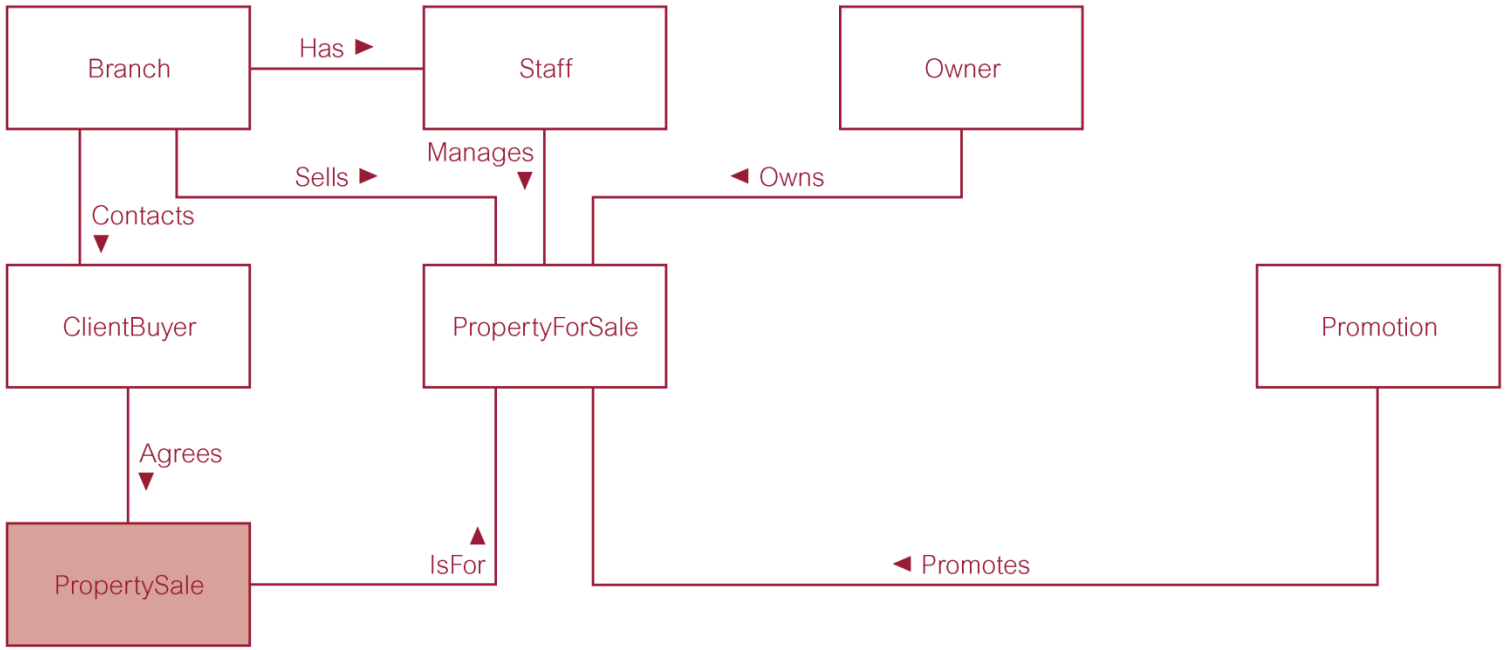
Step 1: Select business process

- 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



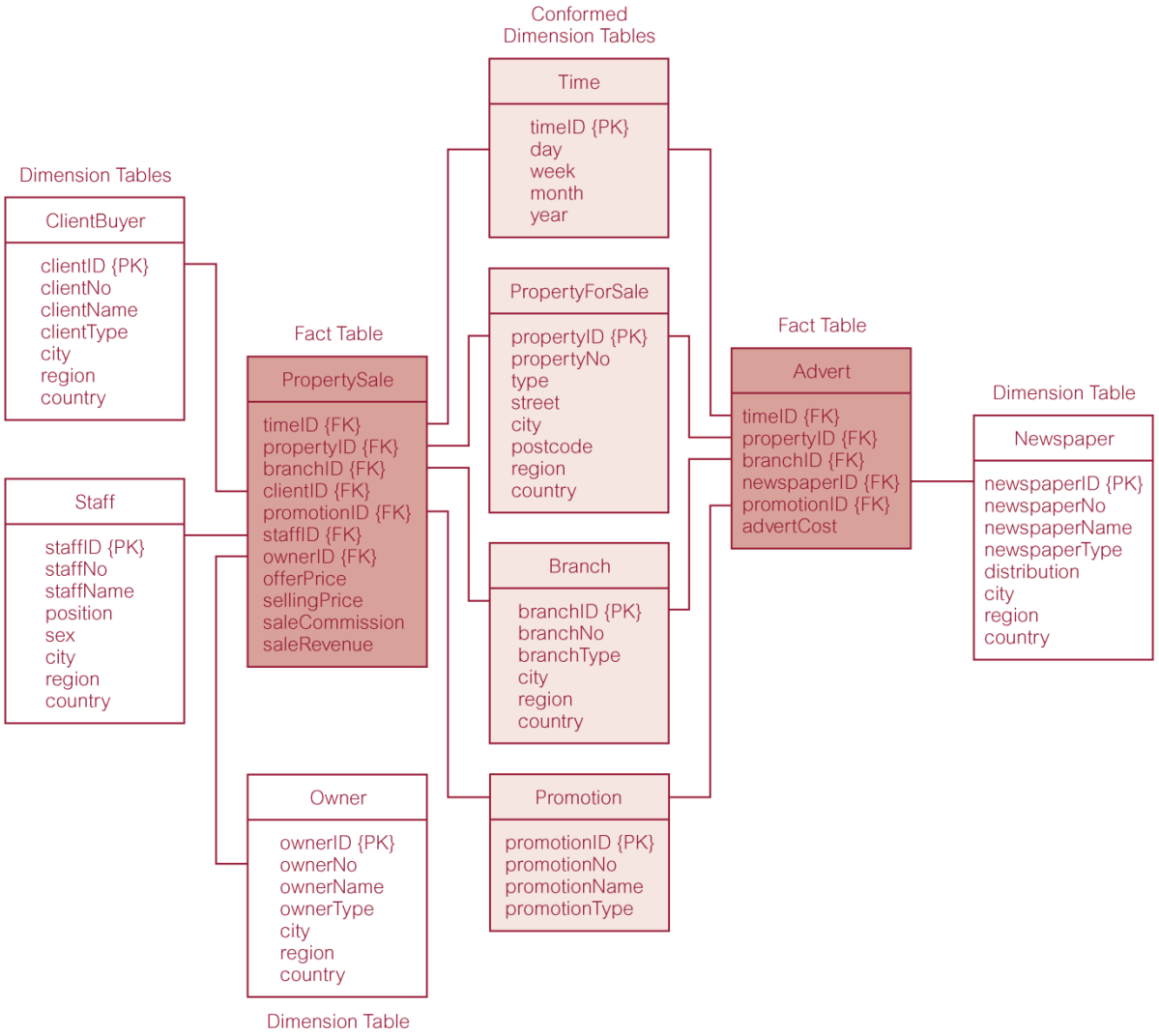
Step 2: Declare grain

- 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

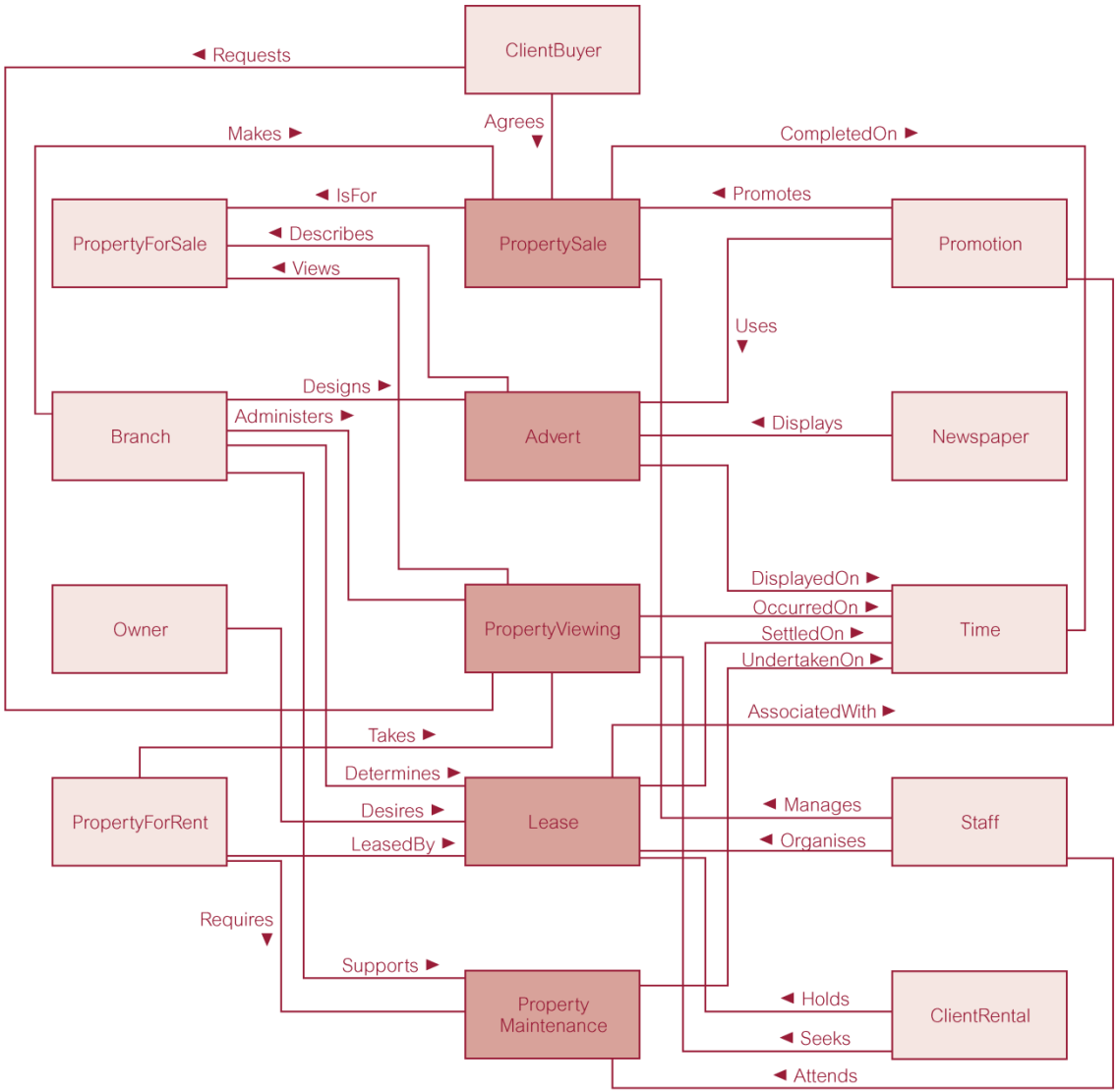
- There are three basic types of slowly changing dimensions:
  - Type 1, where a changed dimension attribute is overwritten
  - Type 2, where a changed dimension attribute causes a new dimension record to be created
  - Type 3, where a changed dimension attribute causes an alternate attribute to be created 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.