Data Warehousing

- Outline
  - Issues associated with designing a data warehouse.
  - Technique for designing the database component of a data warehouse called dimensionality modeling.
  - How a dimensional model (DM) differs from an ER model.
  - A step-by-step methodology for designing a data warehouse.
  - Criteria for assessing degree of dimensionality provided by a data warehouse.

Decision Support (1)

- With OLTP (**On-Line Transaction Processing**):
  - Companies collect large amount of data for everyday operations
  - Examples:
    - Bank: client transactions
    - Supermarket: daily sales

- Above data could also be useful for management, planning and decision support:

  - Which of the financial products are the most successful?
  - How do the variations in sales of some supermarket products relate to various promotions of these and other products?

Decision Support (2)

- Decision support queries make traditional RDBMs (SQL) inadequate:
  - OR in WHERE clause poorly handled in many RDBMs
  - use of statistical functions (e.g. standard deviation) require embedded SQL
  - queries usually involved over time + require aggregation over time periods
  - expressing the queries often tedious even when queries (their type) are often related
  - optimisation difficult because related queries are not recognised

- Require comprehensive views of all aspects of the data $\Rightarrow$ **Data Warehouse**

Data Warehouse

- Provide access to data for complex analysis, knowledge discovery and decision making
  - Contain very large amount (GBs to TBs) of data from multiple sources augmented with summary information and covering a long time period
  - Store of integrated data from multiple sources, processed for storage in a multidimensional model

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

Differences between Data Warehouse and OLTP

DB transaction systems (OLTP)

- Holds current data
- Stored detailed data
- Data is dynamic
- Repetitive process
- High level of transaction throughput
- Predictable pattern of usage
- Transaction driven
- Application oriented
- Supports day-to-day decisions
- Serve large number of clerical/operational users

Data warehouse systems (OLAP)

- Holds historic data
- Stores detailed, lightly and highly summarised data
- Data is largely static
- Ad hoc, unstructured and heuristic processing
- Medium to low level of transaction throughput
- Unpredictable pattern of usage
- Analysis driven
- Subject oriented
- Supports strategic decisions
- Serves relatively lower number of managerial users

Designing Data Warehouses (1)

- To begin a data warehouse project, 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 project be scaled down into something more manageable?
  - Should infrastructure for a scaled down project be capable of ultimately delivering a full-scale enterprise-wide data warehouse?

Designing Data Warehouses (2)

- 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.
- 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, goal remains same: i.e. the ultimate creation of a data warehouse that supports the requirements of the enterprise.

Designing Data Warehouses (3)

- 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 identification of prioritized set of requirements that data warehouse must meet.
- 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 next few years.
- 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**.

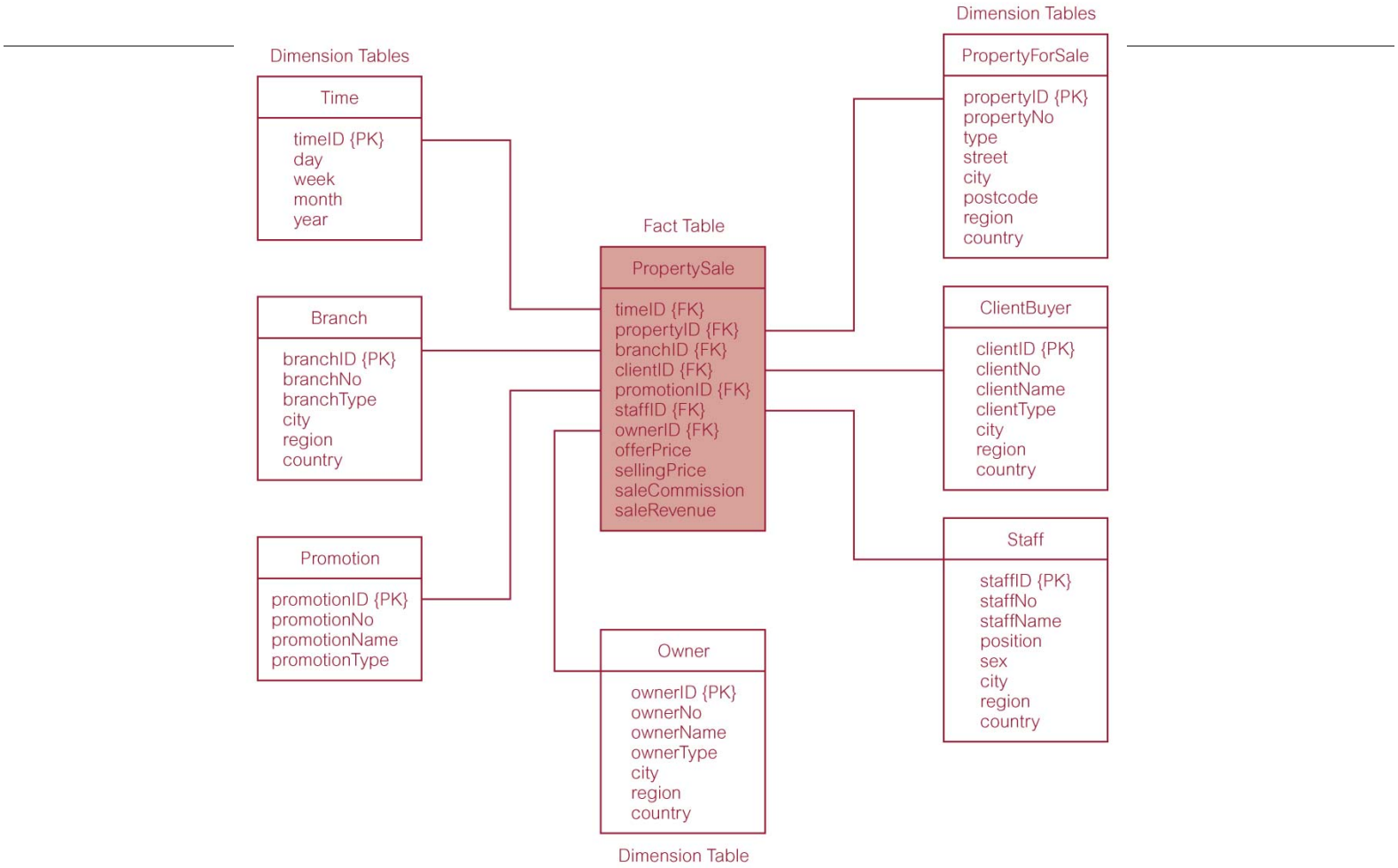
Dimensionality Modeling (1)

- Logical design technique that aims to present the data in a standard, intuitive form that allows for high-performance access
- Uses the concepts of ER modeling with some important restrictions.
- 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.
- 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 (2)

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

Star Schema for Property Sales of DreamHome



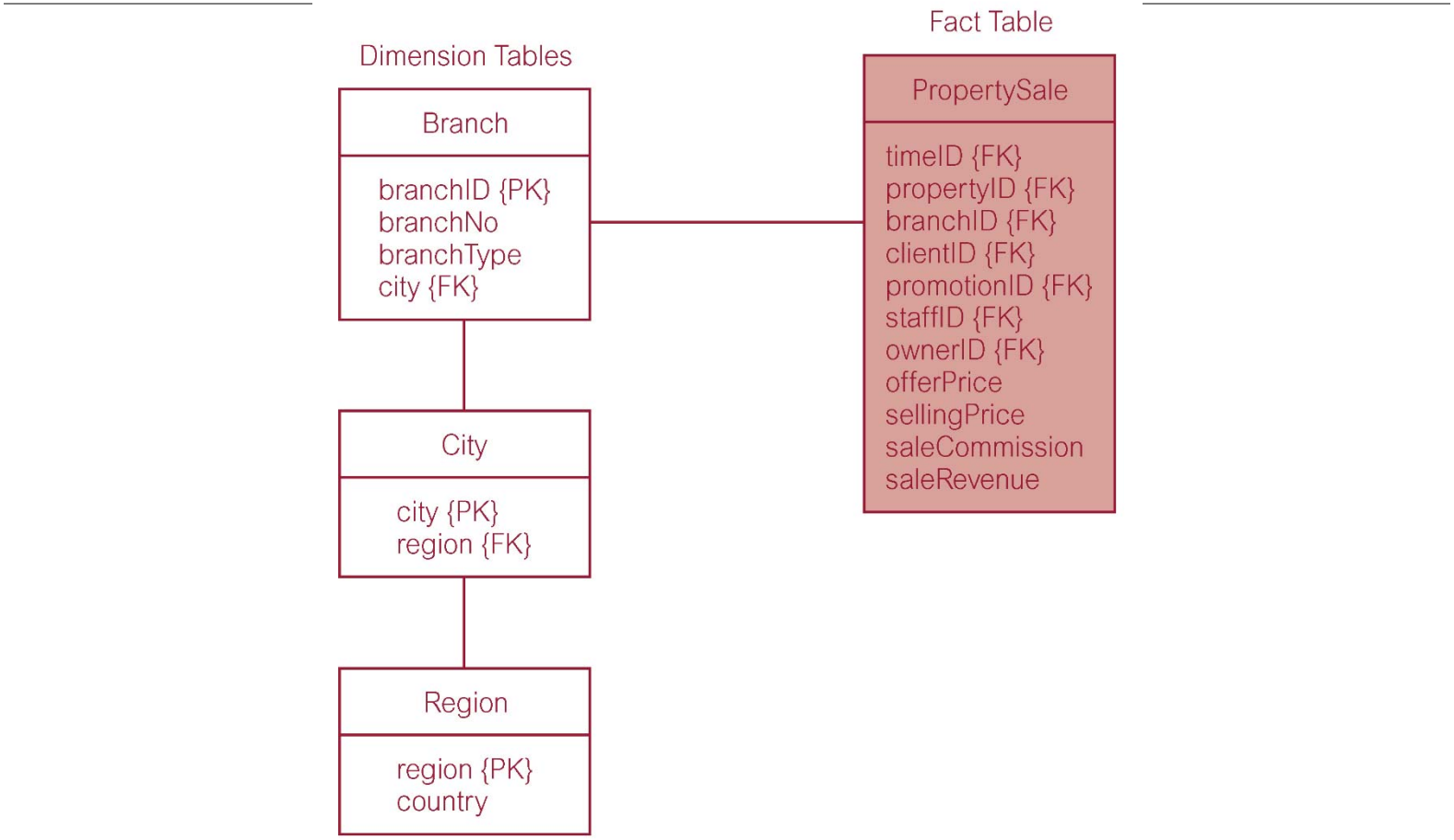
Dimensionality Modeling (3)

- Star schema is a logical structure that has a fact table containing factual data in the center, surrounded by dimension tables containing reference data, which can be denormalised.
- Facts are generated by events that occurred in the past, and are unlikely to change, regardless of how they are analysed.
- 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 (4)

- 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 denormalising reference information into a single dimension table.
- Snowflake schema is a variant of the star schema where dimension tables do not contain denormalised data.
- Starflake schema is a hybrid structure that contains a mixture of star (denormalised) and snowflake (normalised) schemas. Allows dimensions to be present in both forms to cater for different query requirements.

Property Sales with Normalised Version of Branch Dimension Table



Dimensionality Modeling (5)

- 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.