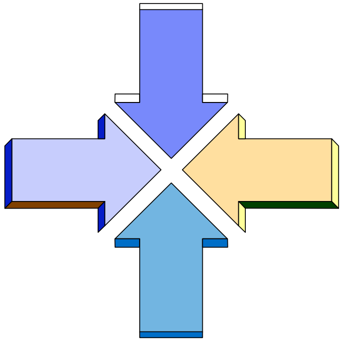
Dimensional Data Modeling Introduction

Agenda

- Basic Terminology
- Dimensional Model Schemas
- Types of Dimensions
- Types of Facts
- Dimensional Modeling Process

Dimensional data modeling

Dimensional Data Modeling techniques organize the content of the data warehouse. It structures the data according to the way users ask business questions.

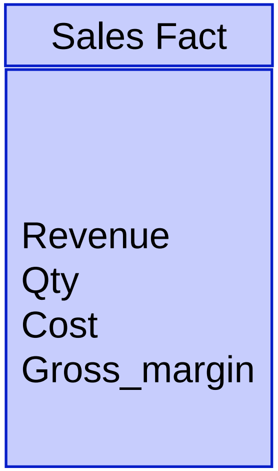


Dimensional Data Models

- Dimensional Data Models
  - Developed top-down
  - Depicts a business process through its relevant facts and dimensions
  - Groups data into categories of business measure and characteristics
  - More suitable for analytical applications where the focus is querying large sets of data

DDM: basic terminology fact table

- Fact Table



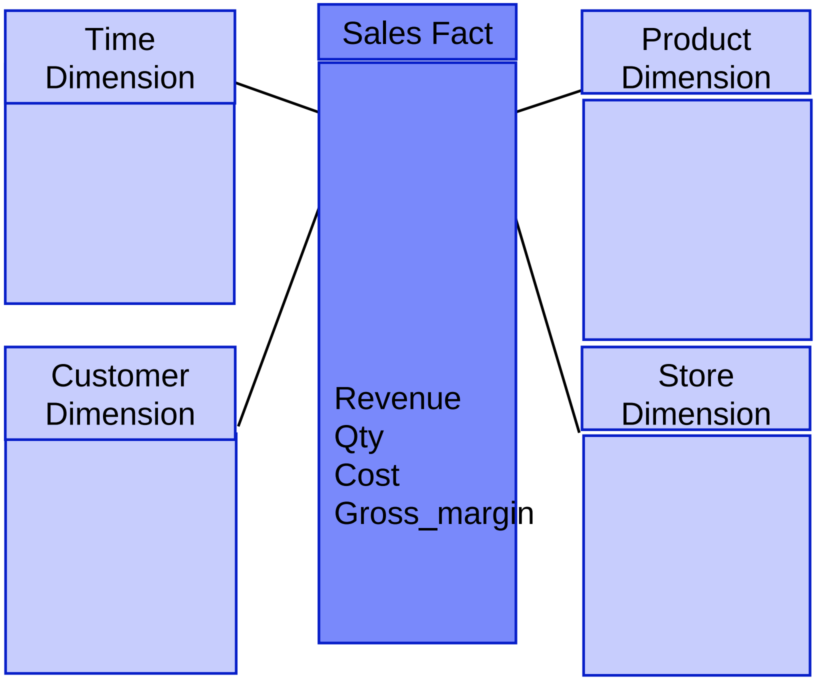
Sales Fact

Revenue
Qty
Cost
Gross_margin

- Definition
  - The performance measures of the business
  - Usually stores numerical and additive measures
  - The “what I want to know”
- Characteristics
  - Basis for analysis
  - Continuously valued
  - Can be derived or calculated
  - Column headers in query results
- Examples
  - Revenue
  - Quantity
  - Cost
  - Gross Margin

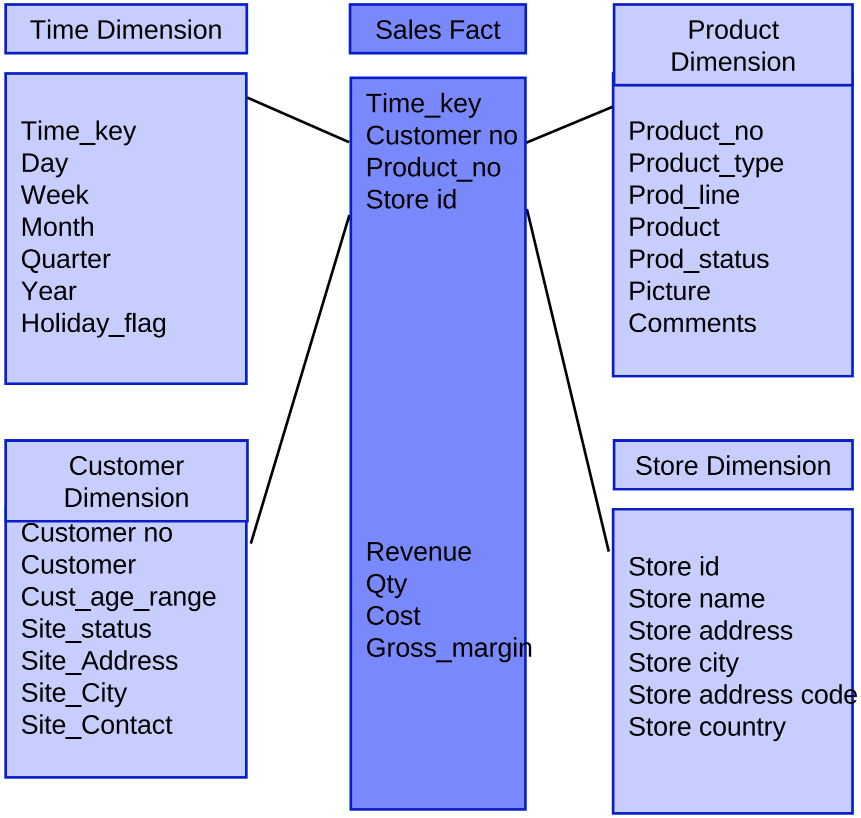
DDM: basic terminology dimension tables

- Dimension Tables



- Definition
  - Descriptions of the business;
  - The “which, who, how, where, or when that describes or explains the fact.”
- Characteristics
  - Constant
  - Enables “slicing and dicing” the facts by different variables
- Examples
  - Time
  - Customer
  - Product
  - Store

DDM: basic terminology attributes



Time Dimension

Time_key
Day
Week
Month
Quarter
Year
Holiday_flag

Sales Fact

Time_key
Customer no
Product_no
Store id

Revenue
Qty
Cost
Gross_margin

Product Dimension

Product_no
Product_type
Prod_line
Product
Prod_status
Picture
Comments

Customer Dimension

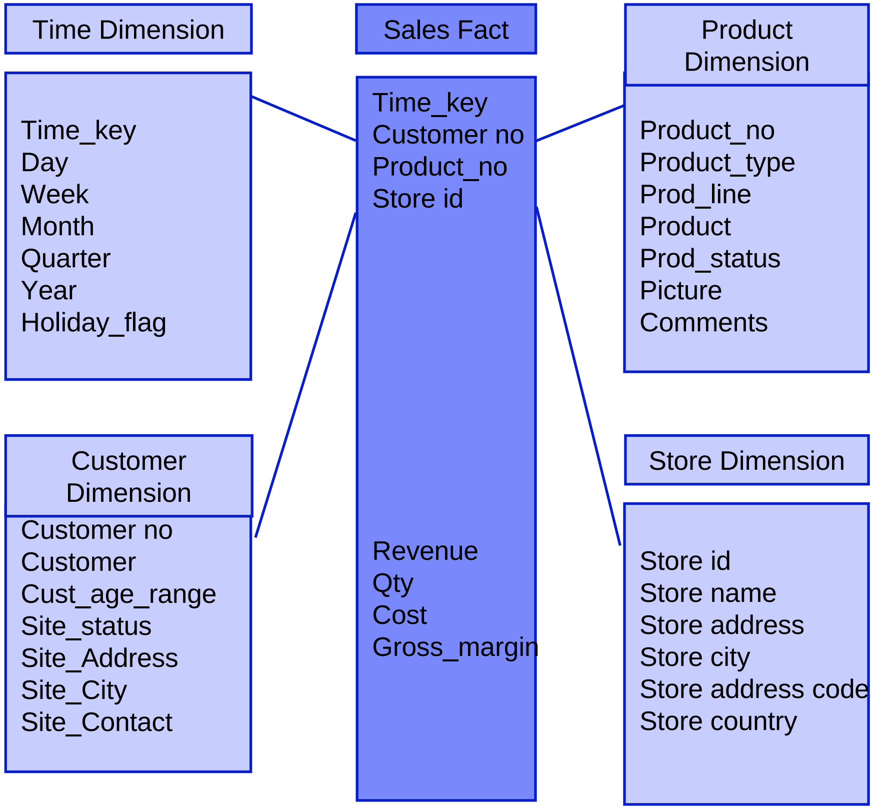
Customer no
Customer
Cust_age_range
Site_status
Site_Address
Site_City
Site_Contact

Store Dimension

Store id
Store name
Store address
Store city
Store address code
Store country

- Definition
  - Fields within the dimension table
  - Describes each item associated with a dimension
- Characteristics
  - Acts as a source of query constraints
  - DW is only as good as the attributes in the dimension table
- Examples
  - month, quarter, year, season, holiday, name, street address, city, brand, category, description
    region, type, manage

DDM: basic terminology



Time Dimension

Time_key
Day
Week
Month
Quarter
Year
Holiday_flag

Sales Fact

Time_key
Customer no
Product_no
Store id

Revenue
Qty
Cost
Gross_margin

Product Dimension

Product_no
Product_type
Prod_line
Product
Prod_status
Picture
Comments

Customer Dimension

Customer no
Customer
Cust_age_range
Site_status
Site_Address
Site_City
Site_Contact

Store Dimension

Store id
Store name
Store address
Store city
Store address code
Store country

- Grain
  - Level of detail contained in fact or dimension table
  - Meaning of a single fact table record
- Hierarchy
  - Represents levels or roll-up of detailed data

Terminology

- **Atomic Layer** - Dimensions and facts at the lowest level of detail (think ODS).
- **Summary Layer** - Dimensions and facts aggregated to intermediate values.
- **Presentation Layer** - Dimensions, facts, and other tables altered specifically for presentation tool limitations.
- **Reporting Layer** - Dimensions, facts, and other tables created or altered to improve reporting capabilities and performance.

More terminology …

- Facts
- Dimensions
- Attributes
- Grain
- Hierarchies
- Keys
- Referential Integrity
- Sparsity
- Numeric Fields as attributes, not facts
- Slowly Changing Dimensions
- Calculated Facts
- Status Indicators/Flags/Events
- Ranges
- Levels
- Counts/Occurrences
- Conformed Tables
- History Roll-Off
- Causal Dimensions
- Huge Dimensions and Mini-Dimensions
- Star Schema/Snowflake Schema
- Heterogeneous Products
- Factless facts
- Additive, semi-, and non-additive facts
- Degenerate dimensions

Agenda

- Basic Terminology
- Dimensional Model Schemas
- Types of Dimensions
- Types of Facts
- Dimensional Modeling Process

Dimensional Model Schemas

- Dimensional Data Models fall into three types of models:
  - Star Schema
  - Snowflake Schema
  - Multi-dimensional Schema
- Several factors influence schema choice:
  - Presentation restrictions
  - Inconsistency of data
  - Complex queries and analysis

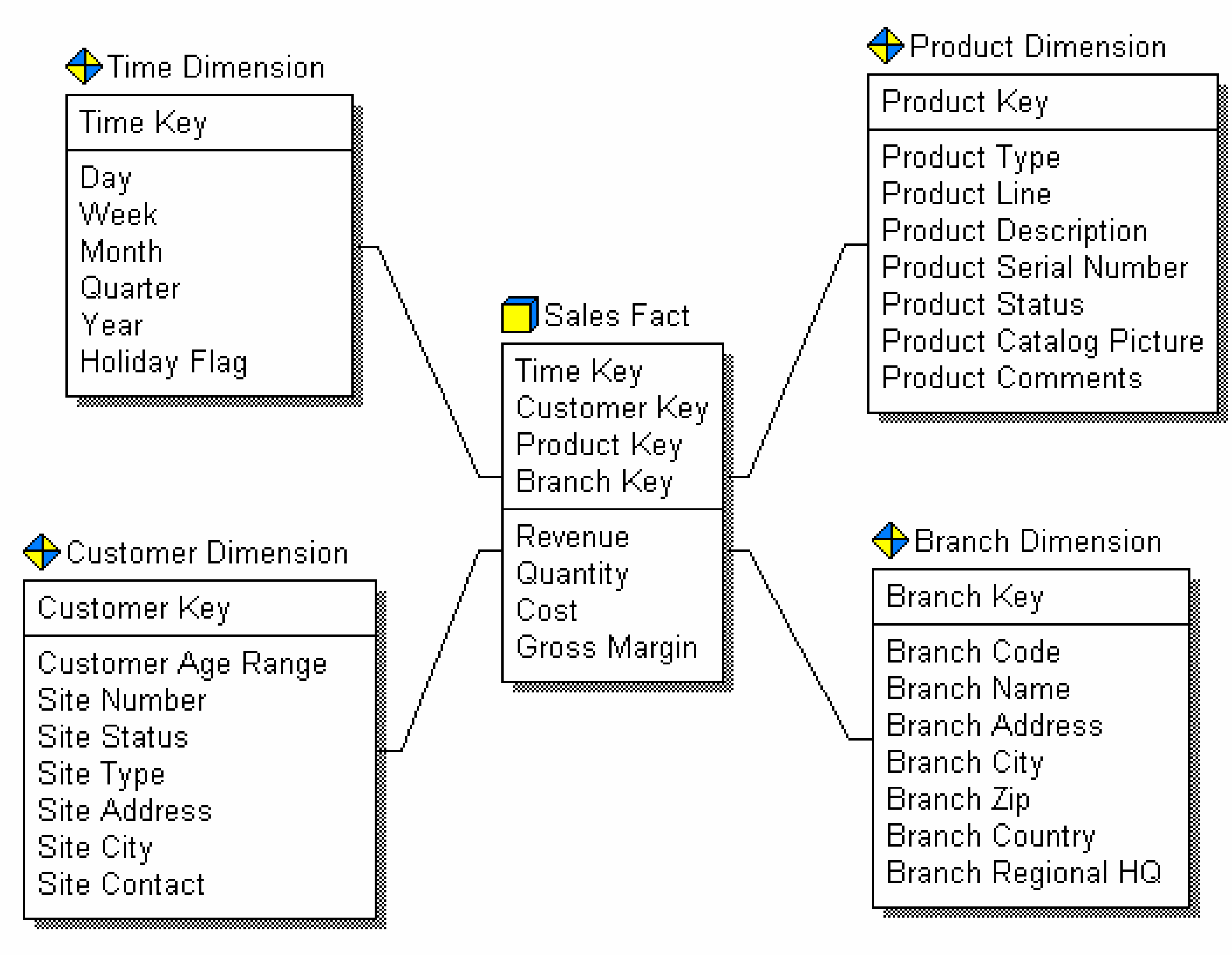
Star Schema

STAR SCHEMA

A database design that stores a central fact table surrounded by multiple dimension tables.

- Star schema represents a compromise between the fully normalized model and the denormalized model.
- Descriptive ‘dimension’ information is maintained in a set of denormalized dimension tables.

Star Schema


Time Dimension
Time Key
Day
Week
Month
Quarter
Year
Holiday Flag
Product Dimension
Product Key
Product Type
Product Line
Product Description
Product Serial Number
Product Status
Product Catalog Picture
Product Comments
Sales Fact
Time Key
Customer Key
Product Key
Branch Key
Revenue
Quantity
Cost
Gross Margin
Customer Dimension
Customer Key
Customer Age Range
Site Number
Site Status
Site Type
Site Address
Site City
Site Contact
Branch Dimension
Branch Key
Branch Code
Branch Name
Branch Address
Branch City
Branch Zip
Branch Country
Branch Regional HQ

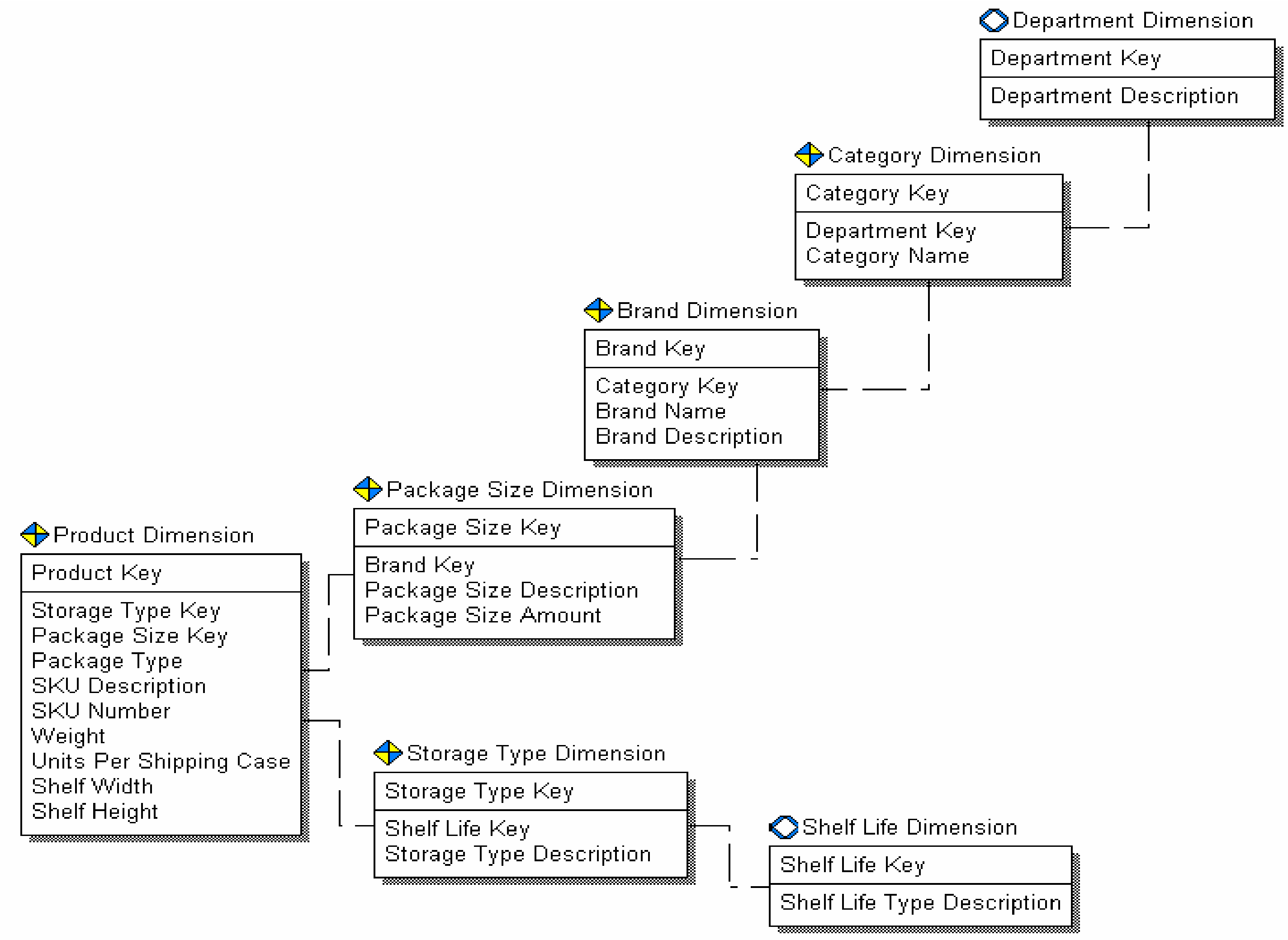
Snowflake Schema

SNOWFLAKE SCHEMA

A database design that stores a central fact table surrounded by multiple dimension tables decomposed or normalized into one or more hierarchies.

- Snowflake schemas are most often used when dealing with large hierarchies that are static.
- Snowflaked tables (look-up tables) may increase the speed of queries depending on the presentation tool (i.e. MicroStrategy)

Snowflake Schema



Department Dimension

Department Key

Department Description

Category Dimension

Category Key

Department Key
Category Name

Brand Dimension

Brand Key

Category Key
Brand Name
Brand Description

Package Size Dimension

Package Size Key

Brand Key
Package Size Description
Package Size Amount

Product Dimension

Product Key

Storage Type Key
Package Size Key
Package Type
SKU Description
SKU Number
Weight
Units Per Shipping Case
Shelf Width
Shelf Height

Storage Type Dimension

Storage Type Key

Shelf Life Key
Storage Type Description

Shelf Life Dimension

Shelf Life Key

Shelf Life Type Description

Multi-Dimensional Schemas

MULTI-DIMENSIONAL SCHEMA

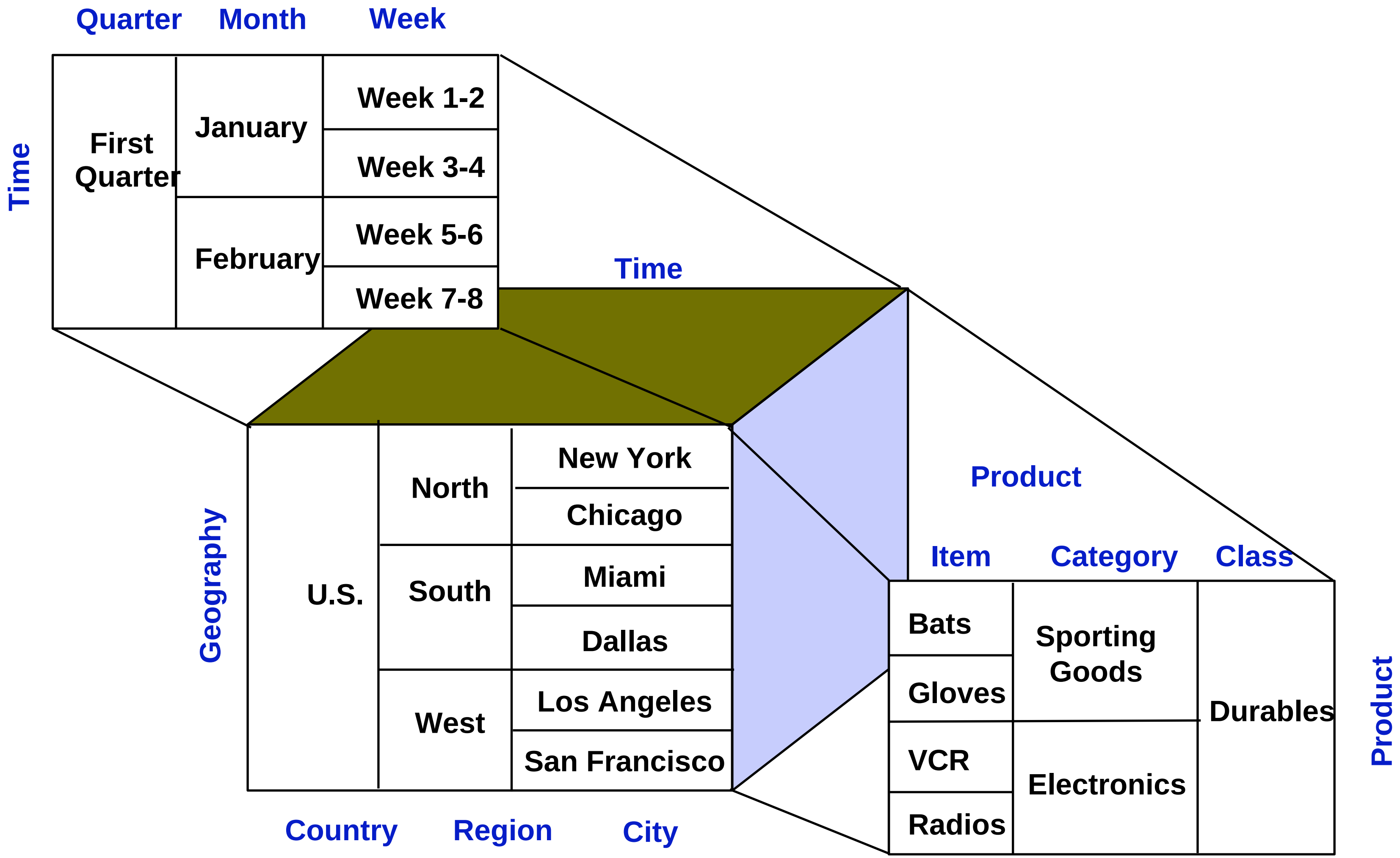
Hierarchical databases that consists of only one structure - a multi-dimensional array - that contains all the summarized data at higher levels in the array.

- Also known as MOLAP databases
- Stores and aggregates data at multiple levels in a hierarchy.
- Utilizes drill-up and drill-down to move around the hierarchy.

Multi-Dimensional Schemas

- Multi-Dimensional Schemas:
  - Provide user with a cross-dimensional perspective allowing analysis across dimensions
  - Specialized programmer must create database
  - Data explosion becomes an issue because each additional dimension results in an exponential increase in the number of dimension intersections (cells).

Multi-dimensional data – a pictorial view



MDM’s and Sparsity

- Sparsity relates to the unpopulated cells in a table.
- It results from every combination of attributes not having a value or an entry associated with it.
- May be reduced if users are satisfied with more summarized than atomic level data.
- Addressing the issue may be tool-dependent.

Agenda

- Basic Terminology
- Dimensional Model Schemas
- Types of Dimensions
- Types of Facts
- Dimensional Modeling Process

Types of Dimensions

- Slowly Changing Dimensions
  - Type 1, 2 or 3
- Rapidly Changing or Volatile Dimensions
- Huge Dimensions and Mini-Dimensions
- Causal Dimensions
- Dirty Dimensions
- Degenerate Dimensions

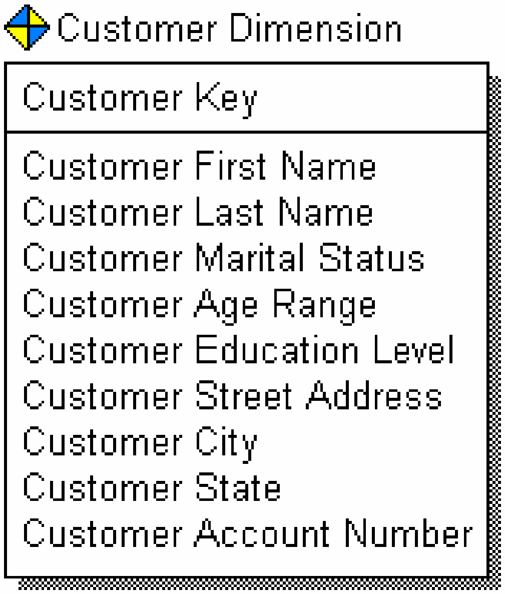
Slowly-Changing Dimensions

- Most dimensions change over time.
  - Products change offered coverage or limits and deductibles.
  - Employees are promoted, fired, or change departments.
  - Customers change names and addresses.
- What are our choices for tracking these changes over time?

Slowly-Changing Dimensions

- There are three types of slowly changing dimensions:
  - **Type 1**: *Overwrites the old data* for a record with new data. This eliminates the ability to track history over time.
  - **Type 2**: *Creates a new record* with the new data at the type of the change. Accurately tracks history, but requires generalized key.
  - **Type 3**: *Tracks new and original values* in separate fields at time of change. Intermediate values are lost.

Type 1 - Overwrite Old Values



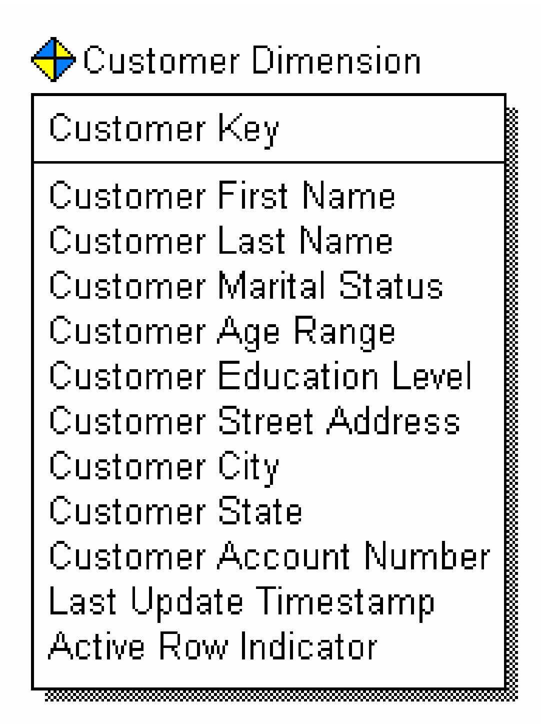
Customer Dimension

Customer Key

Customer First Name
Customer Last Name
Customer Marital Status
Customer Age Range
Customer Education Level
Customer Street Address
Customer City
Customer State
Customer Account Number

- Customer Lynnette Groves is changing her name to ?
- If there is no value in tracking this change, we will overwrite the First Name and Last Name fields with the new values.
- ‘UPDATE’ statement; 1 record is maintained.

Type 2 - Create New Record



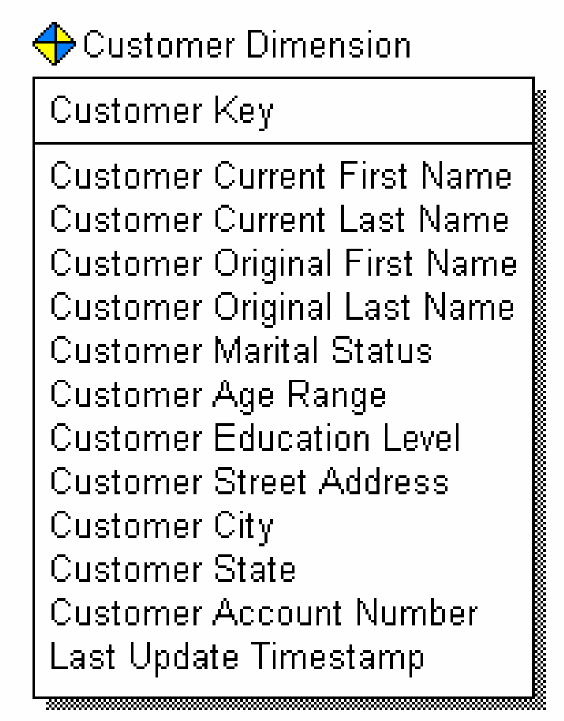
Customer Dimension

Customer Key

Customer First Name
Customer Last Name
Customer Marital Status
Customer Age Range
Customer Education Level
Customer Street Address
Customer City
Customer State
Customer Account Number
Last Update Timestamp
Active Row Indicator

- Lynnette Groves is changing her name and we want to track both values
- Add a second record with a new Customer Key and make it the active row
- ‘INSERT’ statement for new, ‘UPDATE’ for active; 2 records are maintained
- New record for each change up to n records

Type 3 - Original and Current



Customer Dimension

Customer Key

Customer Current First Name
Customer Current Last Name
Customer Original First Name
Customer Original Last Name
Customer Marital Status
Customer Age Range
Customer Education Level
Customer Street Address
Customer City
Customer State
Customer Account Number
Last Update Timestamp

- We decide that no matter how many times she changes her name, we only want to track the original and the current.
- Before any changes, original and current are the same. Any name change updates ‘current’ fields.
- UPDATE’ statement; 1 record is maintained

Volatile Dimensions

- What if a dimension’s values change frequently?
- Price would naturally be an attribute of product and would change semi-frequently.
- Few products have prices that remain constant over many months or years.
- To capture these changes over time, we can capture these values in the fact table rather than treating it as a slowly changing dimension.

A General Rule...

- Fact tables contain counts, amounts, and other numerical information.
- Dimensions describe the business with textual fields and dates in time.
- As a general rule, one should question numerical information that occurs in the dimension tables as well as textual and data fields that occur in the fact table.

Huge Dimensions and Mini-Dimensions

- Product and Customer dimensions with millions and tens of millions of entries are not unusual for retailers, telecommunications companies, insurance companies, or financial service institutions.
- These dimensions can have hundreds of attributes and complex,multiple hierarchies that can exist simultaneously.

Huge Dimensions

HUGE DIMENSIONS

Dimensions with millions or tens of millions of entries, such as customer, that take too long to browse among relationships due to volume.

- The customer dimension in financial institutions, telecommunications companies, and catalog retailers hold data for customers on an individual basis.
- Over time, these can grow to tens of millions of rows.

Huge Dimensions and Mini-Dimensions

- The heavily-used fields in the Customer dimension consist of demographic information: age, sex, number of children, income level, education level, and other purchasing behavior information.
- These fields are also compared together to select an interesting subset of the market base for analysis.

Huge Dimensions and Mini-Dimensions

- The most effective technique for handling this situation is to separate one or more sets of these attributes into demographic mini-dimensions.
- If five or six of the demographic variables are isolated into a separate table, we need only to store the distinct combinations of information that actually occur.

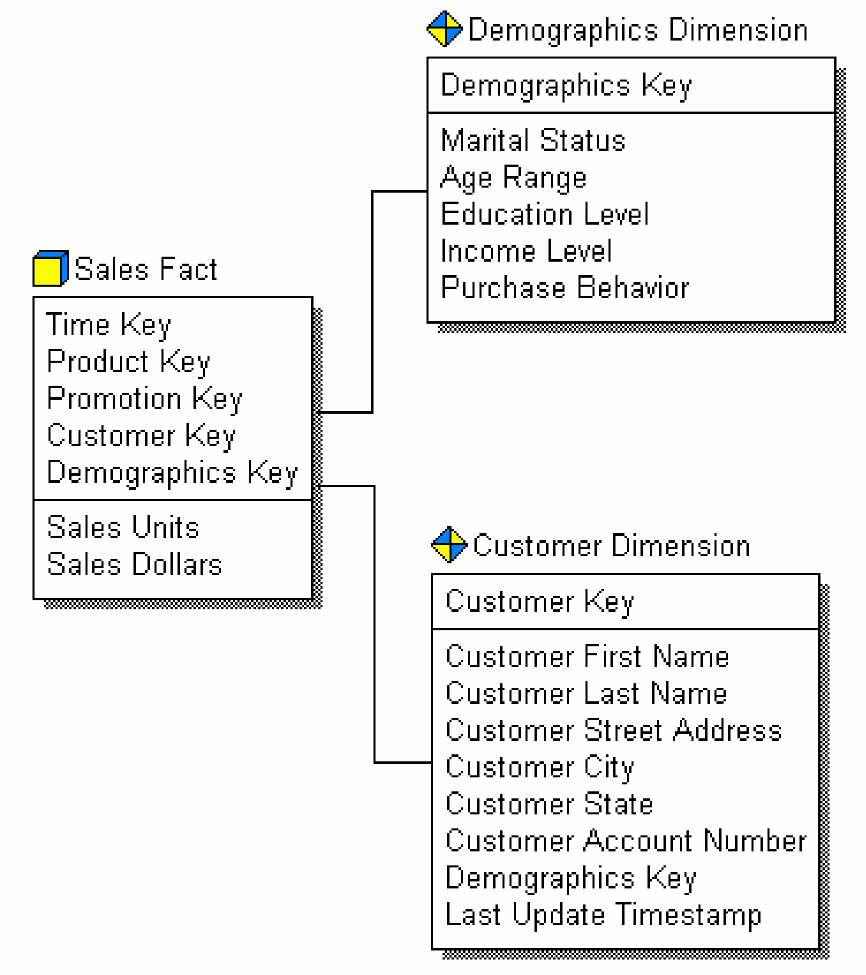
Mini-Dimensions

MINI-DIMENSIONS

Groups of related attributes separated into separate dimensions that create significant gains in performance and decreased volatility in the parent dimension.

- Typically, demographic information changes at a different rate than other customer information.
- Marketing can analyze different segments of the customer base for purchasing habits and other information.

Huge Dimensions and Mini-Dimensions



Demographics Dimension

Demographics Key

Marital Status
Age Range
Education Level
Income Level
Purchase Behavior

Sales Fact

Time Key
Product Key
Promotion Key
Customer Key
Demographics Key

Sales Units
Sales Dollars

Customer Dimension

Customer Key

Customer First Name
Customer Last Name
Customer Street Address
Customer City
Customer State
Customer Account Number
Demographics Key
Last Update Timestamp

- Demographic dimension can be joined directly to the fact itself or ‘snowflaked’ to the customer dimension.
- Demographics Key is included in the Customer Dimension to browse data interactively.

Causal Dimensions

CAUSAL DIMENSIONS

Causal dimensions describe factors that are thought to cause a change in the performance of a measure or fact, such as advertising or promotion.

- Causal Dimensions track conditions that may influence sales, counts, or revenue.
- Promotions, holidays, and weather conditions may influence the behavior of fact data.

Causal Dimensions

- Causal dimensional attributes can be placed in a single dimension table or separated into different tables by subject.
- A ‘Promotion’ dimension could include price reduction type, ad type, display type, and promo start and end dates.
- The trade-offs include efficient browsing vs. more understandable tables for the user community.

Causal Dimensions

- Single table design:
  - What type of conditions are being tracked
  - Generalizes all conditions into one table
  - Multiple causal conditions may need to be stored on the same record.
- Multiple table design:
  - Different dimensions for holidays, marketing campaigns, and weather conditions.
  - Increases sparsity of fact when all conditions do not apply.

Dirty Dimensions

DIRTY DIMENSIONS

Dimensional information that may contain duplicate or extraneous entries due to inconsistent legacy data.

- Financial institutions might have a poor account-to-account correlation of individual’s names.
- Insurance companies may not make a serious attempt to identify previous instances of an insured party or other policies.

Dirty Dimensions

- Some cleaning can be done in ETL process.
- Will influence fact data accuracy.
- All tools that access the data will need to take the possible inaccuracy of data into account. Some tools are designed to alleviate some of the problem to 80% accuracy.
- Level of inaccuracy may influence design of dimensions and facts so that it may be minimized.

Degenerate Dimensions

DEGENERATE DIMENSIONS

Dimensions that are so small and have no attributes of their own that they have been added to the fact table.

- Certain attributes are tracked that don’t necessarily belong in their own dimension - orphan attributes.
- This may occur when fact tables are designed to reflect the actual working document.

Degenerate Dimensions

- Examples include ‘order_number’, ‘bill_of_lading_num’, and ‘invoice_number’.
- While these fields seem very transaction oriented, they are helpful in grouping things such as all line items on an invoice.
- Including these fields on the fact table amounts to denormalizing the attribute due to the granularity of the fact table being the document itself or a line item of the document.

Agenda

- Basic Terminology
- Dimensional Model Schemas
- Types of Dimensions
- Types of Facts
- Dimensional Modeling Process

Types of Facts

FACT

A measurement, generally additive in nature, of the organization.

- We use facts to measure performance based on business questions.
- This data is numeric in nature and is contained in our fact tables by subject and granularity.

Types of Facts

- Understanding which facts can be added across which dimensions is an important data design issue.
- Three Types of Facts:
  - Additive
  - Non-Additive
  - Semi-Additive

Additive Facts

ADDITIVE FACTS

Measurements in a fact table that can be added across all dimensions.

- Since aggregation is a key element in the usefulness of the dimensional model, its best utilized for facts that are additive, numeric values.
- We can add revenue, cost, and quantity sold for all products, all stores, and any time period.

Semi-Additive Facts

SEMI-ADDITIVE FACTS

Measurements in a fact table that can be added across some dimensions but not others.

- We cannot add risk exposure at the coverage level to get the number of policy level exposures.
- We can add coverage level exposures across the customer dimension to determine exposure by gender or age range.

Non-Additive Facts

NON-ADDITIVE FACTS

Measurements in a fact table that cannot be added across any dimensions, like ratios.

- A new value will need to be calculated at each level for each level or for each set of data.
- It should be determined, at what levels, if any, the fact should be stored. Some values may need to be pre-calculated.

Factless Fact Tables

FACTLESS FACT TABLES

Tables that seem like fact tables but are used to represent data or events for which there are no measured facts.

- These tables are used to track events as the simultaneous coming together of a number of dimensions.
- Two major variations: Event Tracking and Coverage tables.

Agenda

- Basic Terminology
- Dimensional Model Schemas
- Types of Dimensions
- Types of Facts
- Dimensional Modeling Process

Dimensional Modeling Process

Step 1: Choose the grain of each fact table.

- Granularity defines the level of detailed data.
- It must be determined prior to going forward in the modeling process.
- Typical grains are individual transactions, time-based aggregation, and/or aggregations along a commonly used dimension.

Dimensional Modeling Process

Step 2: Choose the dimension attributes.

- For example, what should our time dimension look like? Should it have just ‘January for month’, or also ‘Jan’ and ‘1’?
- Should we store the code and the description, just the code, or just the description?
- What values will our users need to filter or report on?

Dimensional Modeling Process

Step 3: Identify dimensional hierarchies.

- A dimension such as time may have days rolling into months and then quarters, as well as days rolling into weeks which may cross months and quarters.
- Sales geography may differ from physical geography.
- Zip codes can cross city boundaries and cities are made up of multiple zip codes.

Dimensional Modeling Process

Step 4: Choose the dimensions that apply to each fact table.

- Typical dimensions include time, product, policyholder, agent, and geography.
- Remember to evaluate granularity when applying dimensions to facts.

Dimensional Modeling Process

Step 5: Choose the measured facts, including precalculated facts.

- Each aggregated and derived fact will need to be evaluated for inclusion in the model or calculation in the application.
- Trade-offs include storage and indexing and must be weighed against the access requirements.

Dimensional Modeling Process

Step 6: Determine slowly changing dimensions

- These are the dimensions that change over time.
- If tracking these changes is important, the method must be decided.
- Options: overwrite the existing record, store all records with effective dates, or a historical and current value tables.