


# CP5806

## WEEK 3

---

# DUE DATES

- Assignment 1 due Sunday 26 July, 11:59 pm (Week 3)
- Assignment 2 due Sunday 9 August, 11:59 pm (Week 5)
- Sisi will hold two extra Saturday sessions for A2
  - Week 3 Saturday 25 July 2:30-3:30 pm
  - Week 4 Saturday 1 August 2:30-3:30 pm



JULY							
	M	T	W	T	F	S	S
O Week			1	2	3	4	5
wk 1	6	7	8	9	10	11	12
wk 2	13	14	15	16	17	18	19
wk 3	20	21	22	23	24	25	26
wk 4	27	28	29	30	31		

AUGUST							
	M	T	W	T	F	S	S
wk 4						1	2
wk 5	3	4	5	6	7	8	9
wk 6	10	11	12	13	14	15	16
wk 7	17	18	19	20	21	22	23
O Week	24	25	26	27	28	29	30
wk 1	31						

# WEEK 3 LEARNING OUTCOMES

---

- Analyse star and snowflake schemas identifying your preferred model
- Evaluate dimensional and entity relationship modelling identifying which is more suitable in your organisation
- Identify risks, issues and concerns related to data warehousing and Extraction, Transformation and Loading (ETL) processes
- Compare database keys and identify them in fact and dimension tables
- Analyse the major trends influencing data warehousing
- Develop an appropriate set of fact and dimension tables for a given business scenario
- Select an appropriate approach for updating dimension tables and aggregating fact tables.

# TOPICS FOR WEEK 3

---

- Topic 1: Dimensional modelling basics
- Topic 2: Star Schema
- Topic 3: Dimensional table updates
- Topic 4: Aggregate fact tables
- Topic 5: ETL

# ER MODELLING VS. DIMENSIONAL MODELLING

---

## Entity-Relationship Modeling

Removes data redundancy  
Ensures data consistency  
Expresses microscopic relationships

- ◆ OLTP systems capture details of events or transactions
- ◆ OLTP systems focus on individual events
- ◆ An OLTP system is a window into micro-level transactions
- ◆ Picture at detail level necessary to run the business
- ◆ Suitable only for questions at transaction level
- ◆ Data consistency, non-redundancy, and efficient data storage critical

## Dimensional Modeling

Captures critical measures  
Views along dimensions  
Intuitive to business users

- ◆ DW meant to answer questions on overall process
- ◆ DW focus is on how managers view the business
- ◆ DW reveals business trends
- ◆ Information is centered around a business process
- ◆ Answers show how the business measures the process
- ◆ The measures to be studied in many ways along several business dimensions

# ER MODELLING VS. DIMENSIONAL MODELLING

---

ER modelling	Dimensional modelling
Support OLTP	Support OLAP
Normalised	Denormalised
Removes redundancy	Allows redundancy
A view of data from data processing	A view of data from business processing
Contains both logical and physical model	Contains only a physical model
Current data	Historical data
Many operational users	Small managerial users
MB to GB	GB to TB even PB
Volatile (read and write)	Non-volatile (read-only)

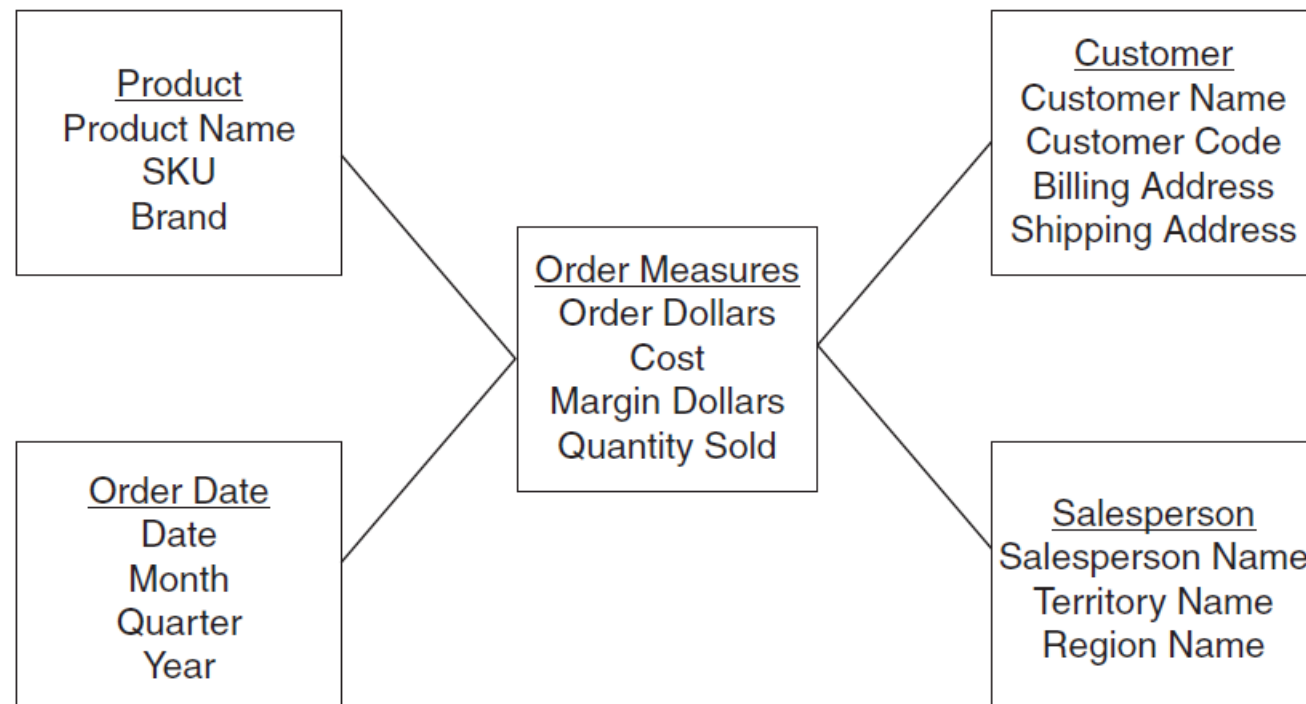
# KEYS USED IN DATABASES

---

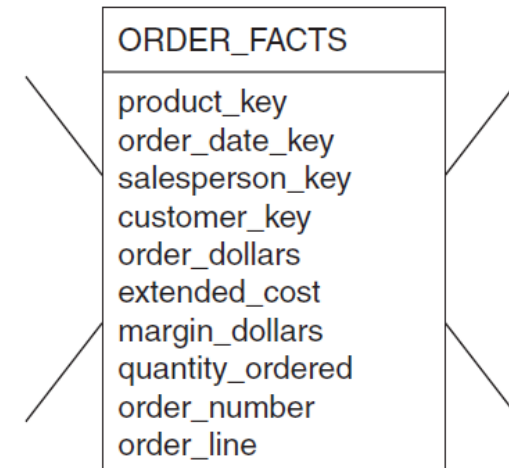
- **Business (natural) keys:** values that can exist outside the database or as part of the business that can uniquely identify each row
- **Candidate keys:** a minimal super key – a set of attributes that can uniquely identify each row
- **Primary keys:** candidate key selected to ensure each row is unique, and used by other tables to reference this one
- **Alternate keys:** all keys that are not primary key
- **Composite (compound) keys:** a key formed from several attributes
- **Surrogate keys:** artificial key, usually auto-incrementing integer
- **Foreign keys:** create a relationship with another table – the primary key from that table

*Reference: <https://www.guru99.com/dbms-keys.html>*

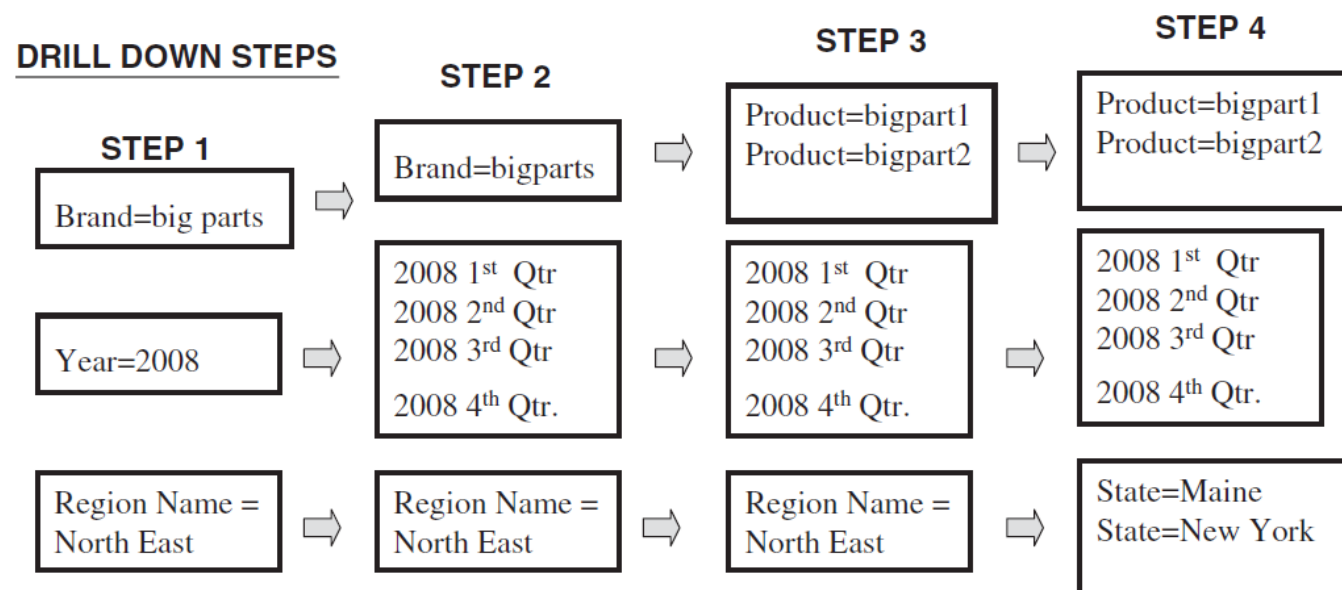
# STAR SCHEMA



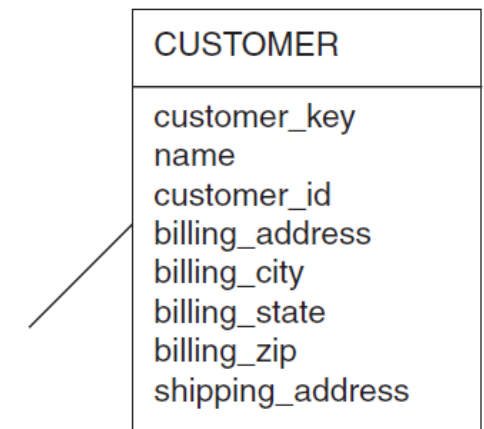
- Concatenated fact table key
- Grain or level of data identified
- Fully additive measures
- Semi-additive measures
- Large number of records
- Only a few attributes
- Sparsity of data
- Degenerate dimensions



**Figure 10-11** Inside a fact table.



- Dimension table key
- Large number of attributes (wide)
- Textual attributes
- Attributes not directly related
- Flattened out, not normalized
- Ability to drill down/roll up
- Multiple hierarchies
- Less number of records



**Figure 10-10** Inside a dimension table.



# STAR SCHEMA

---

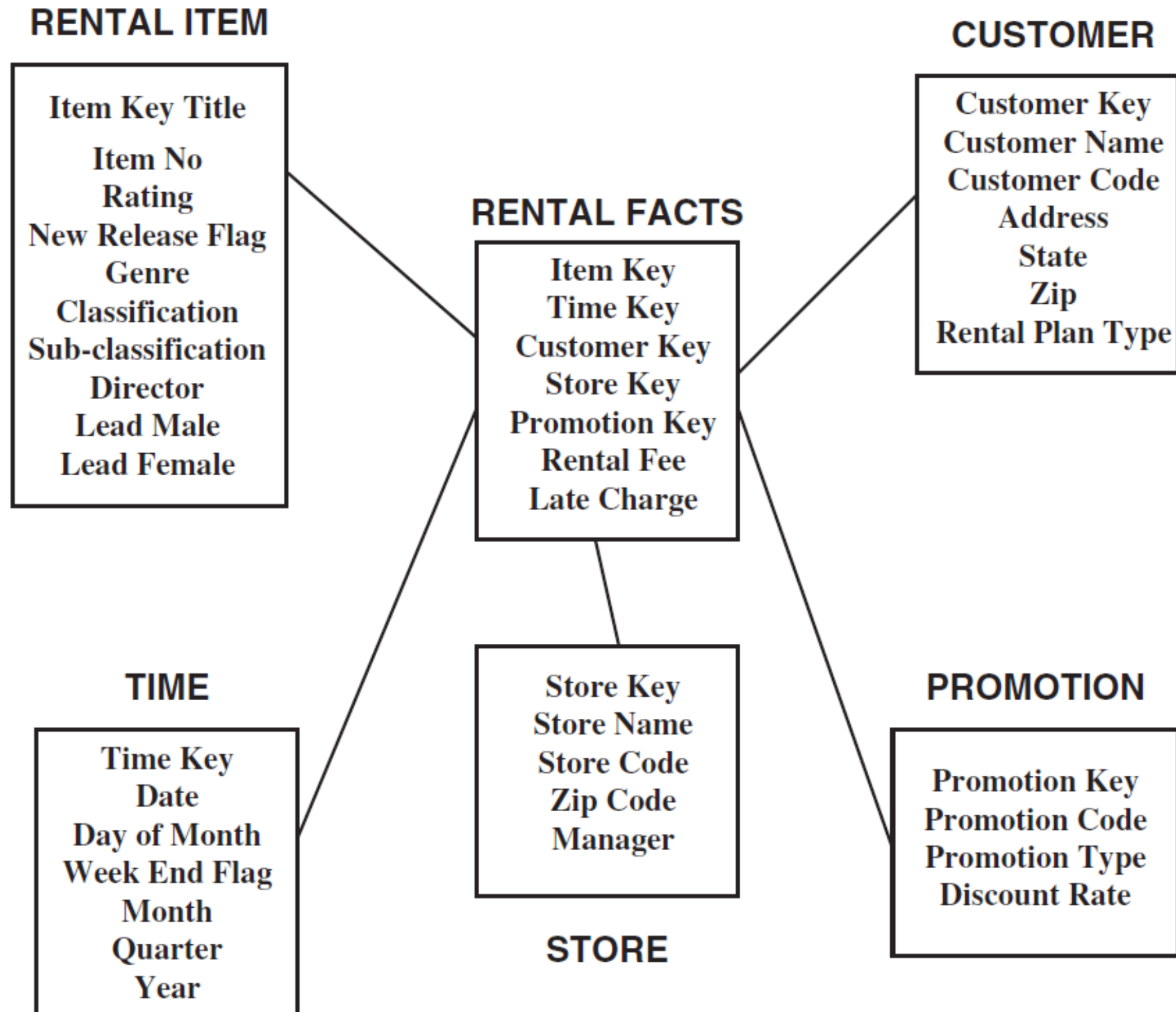


Figure 10-15 STAR schema example: video rental.

# DIMENSION TABLE VS. FACT TABLE

---

Dimension table	Fact table
Descriptive attributes	Quantitative measurement
Surrogate	Composite key built from foreign keys
Less records and more attributes	More records and less attributes
Grows horizontally	Grows vertically
Created first	Created later
More number of tables	Less number of tables
Shorter key	Longer key

# DIMENSIONAL MODELLING

---

- What do you think would be possible fact and dimension tables for a hospital?

# STAR SCHEMA VS. SNOWFLAKE SCHEMA

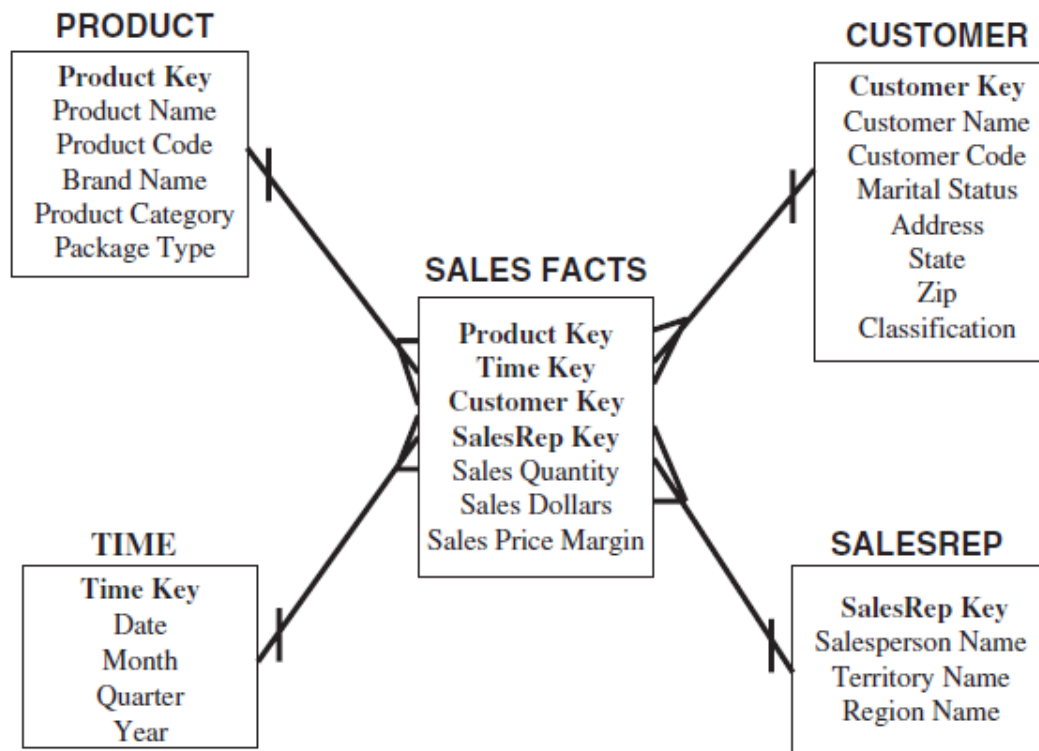


Figure 11-7 Sales: a simple STAR schema.

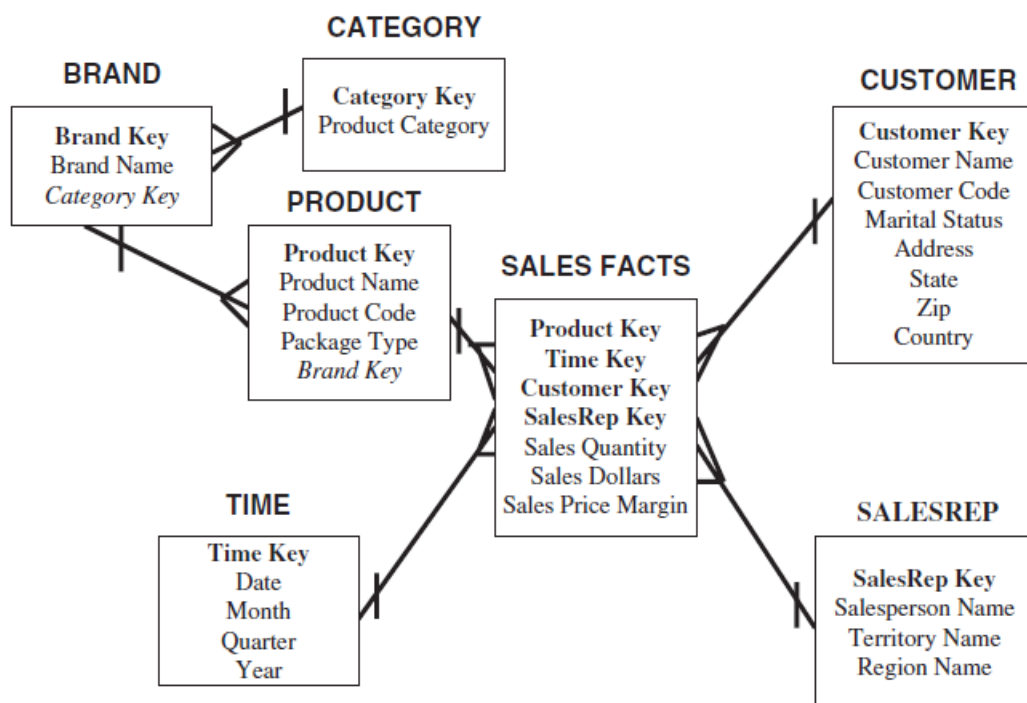


Figure 11-8 Product dimension: partially normalized.

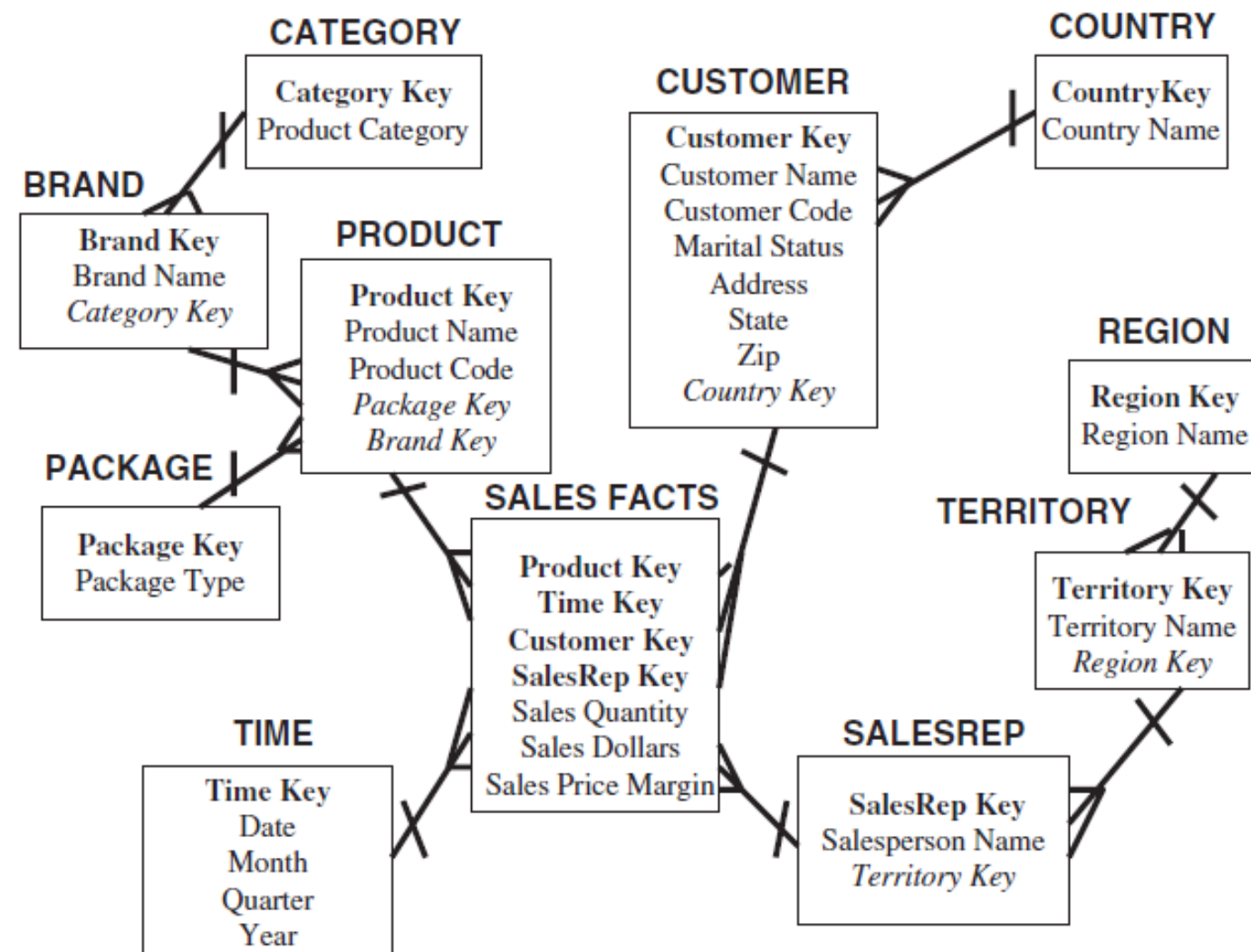


Figure 11-9 Sales: the "snowflake" schema.

# STAR SCHEMA VS. SNOWFLAKE SCHEMA

---

Star schema	Snowflake schema
Simple	Complex
Denormalised	Normalised
One join to fetch the data	Many joins to fetch the data
Faster response time	Slower response time
Redundant data	No redundancy
Top-down approach	Bottom-up approach
One dimension table for each dimension	More than one table for each dimension
Easy to understand	Less easy to understand



# STAR SCHEMA VS. SNOWFLAKE SCHEMA

---



## Modelling choice

**Time:** Set aside 30 minutes to contribute to the discussion.

**Task:** Use the *Modelling choice* discussion forum to discuss the following:

- Which schema do you prefer and why? The star or snowflake schema?
- Which table is more important in modelling? Fact table or dimensional table?
- Which modelling is more important for your company? ER modelling or dimensional modelling?

# DIMENSIONAL TABLE UPDATES

---

- **Slowly Changing Dimensions (SCD):** dimension tables are more stable and less volatile. A dimension table does not change just through the increase in the number of rows, but also through changes to the attributes themselves.
- SCD types:
  - Type 1 Changes: Correction of Errors
  - Type 2 Changes: Preservation of History
  - Type 3 Changes: Tentative Soft Revisions

# TYPE 1 CHANGES: CORRECTION OF ERRORS

---

## ➤ Nature of Type 1 Changes:

- Usually, the changes relate to **correction of errors** in source systems.
- Sometimes the **change** in the source system **has no significance**.
- The **old value** in the source system needs **to be discarded**.
- The **change** in the source system need **not be preserved** in the data warehouse.

## ➤ Applying Type 1 Changes to the Data Warehouse:

- Overwrite the attribute value in the dimension table row with the new value.
- The old value of the attribute is not preserved.
- No other changes are made in the dimension table row.
- The key of this dimension table or any other key values are not affected.
- This type is easiest to implement.



# TYPE 2 CHANGES: PRESERVATION OF HISTORY

---

## ➤ Nature of Type 2 Changes:

- They usually **relate to true changes** in source systems.
- There is a need to **preserve history** in the data warehouse.
- This type of **change partitions the history** in the data warehouse.
- Every **change** for the same attribute **must be preserved**.

## ➤ Applying Type 2 Changes to the Data Warehouse:

- Add a new dimension table row with the new value of the changed attribute.
- An effective date field may be included in the dimension table.
- There are no changes to the original row in the dimension table.
- The key of the original row is not affected.
- The new row is inserted with a new surrogate key.

# TYPE 3 CHANGES: TENTATIVE SOFT REVISIONS

---

## ➤ Nature of Type 3 Changes:

- They usually relate to **“soft” or tentative changes** in the source systems.
- There is a need to **keep track of history with old and new values** of the changed attribute.
- They are used to **compare performances** across the transition.
- They provide the ability to **track forward and backward**.

## ➤ Applying Type 3 Changes to the Data Warehouse:

- Add an “old” field in the dimension table for the affected attribute.
- Push down the existing value of the attribute from the “current” field to the “old” field.
- Keep the new value of the attribute in the “current” field.
- Also, you may add a “current” effective date field for the attribute.
- The key of the row is not affected.
- No new dimension row is needed.
- The existing queries will seamlessly switch to the “current” value.
- Any queries that need to use the “old” value must be revised accordingly.
- The technique works best for one “soft” change at a time.
- If there is a succession of changes, more sophisticated techniques must be devised.

# DIMENSIONAL MODELLING TECHNIQUES

---

Dimension	Major function
Mini dimension	Isolate rapidly changing attributes
Role playing dimension	Multiple fact table foreign keys pointing to the same dimension (e.g. multiple dates)
Outrigger dimension	Dimension references another dimension (e.g. bank account references a date)
Snowflaked dimension	Normalised to 3rd normal form
Conformed dimension	The same meaning to every fact table with which it relates
Degenerate dimension	No associated dimension table
Junk dimension	Catch all for one-off attributes

# MINI-DIMENSION

---

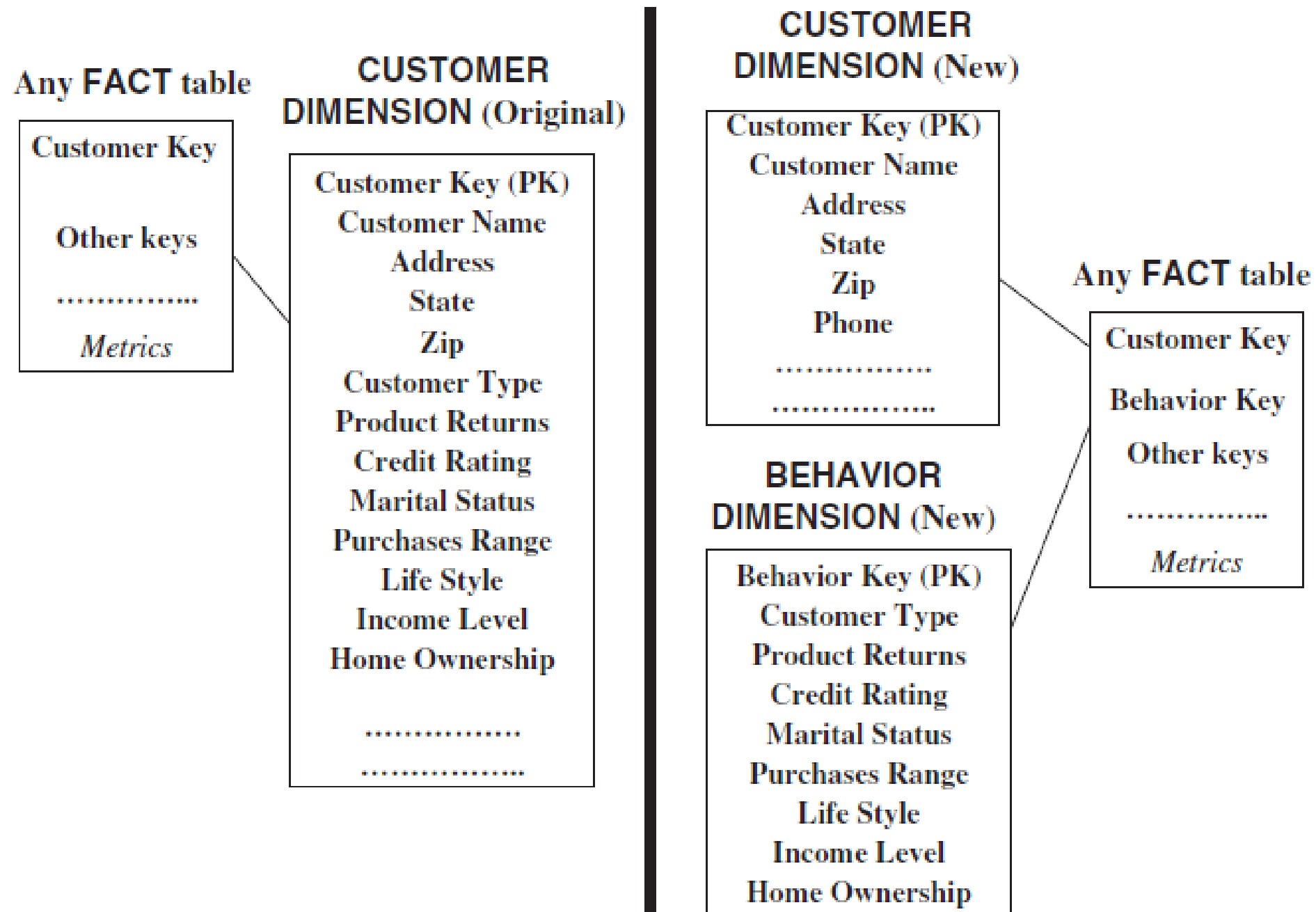


Figure 11-6 Dividing a large, rapidly changing dimension table.

# AGGREGATE FACT TABLES

---

- Multi-way aggregate fact tables
  - One-way aggregate
  - Two-way aggregate
  - Three-way aggregate
- Effect of sparsity on aggregation: it is suggested that each aggregate table row summarizes at least 10 rows in the lower level table.
- Aggregate or summary tables improve performance. Formulate a strategy for building aggregate tables.
- A set of related STAR schemas make up a family of STARS. Examples are snapshot and transaction tables, core and custom tables, and tables supporting a value chain or a value circle. A family of STARS relies on conformed dimension tables and standardized fact tables.

# AGGREGATE FACT TABLES

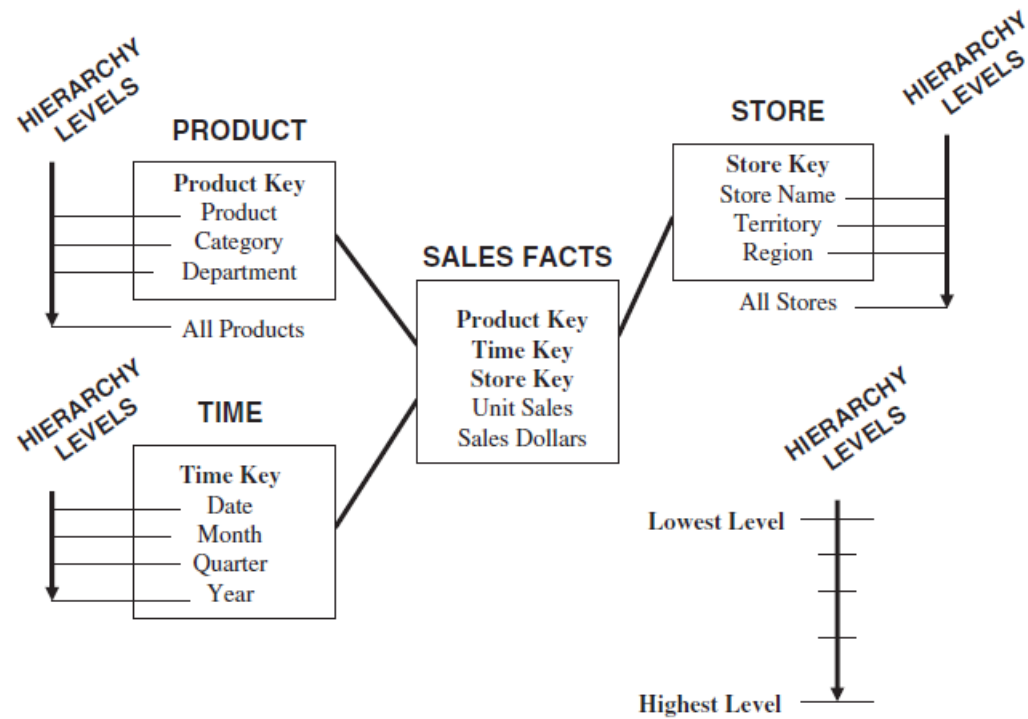


Figure 11-13 Dimension hierarchies.

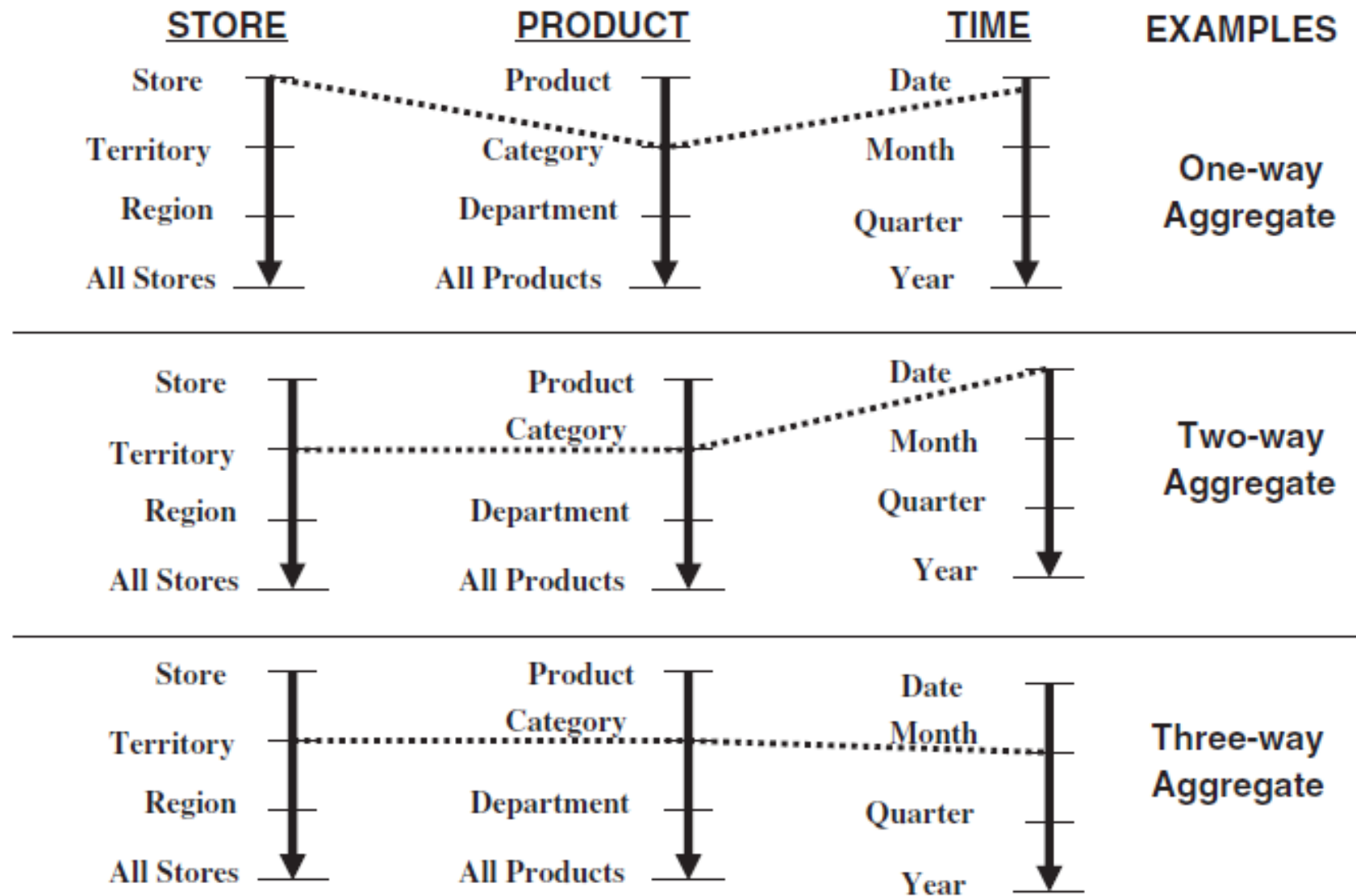
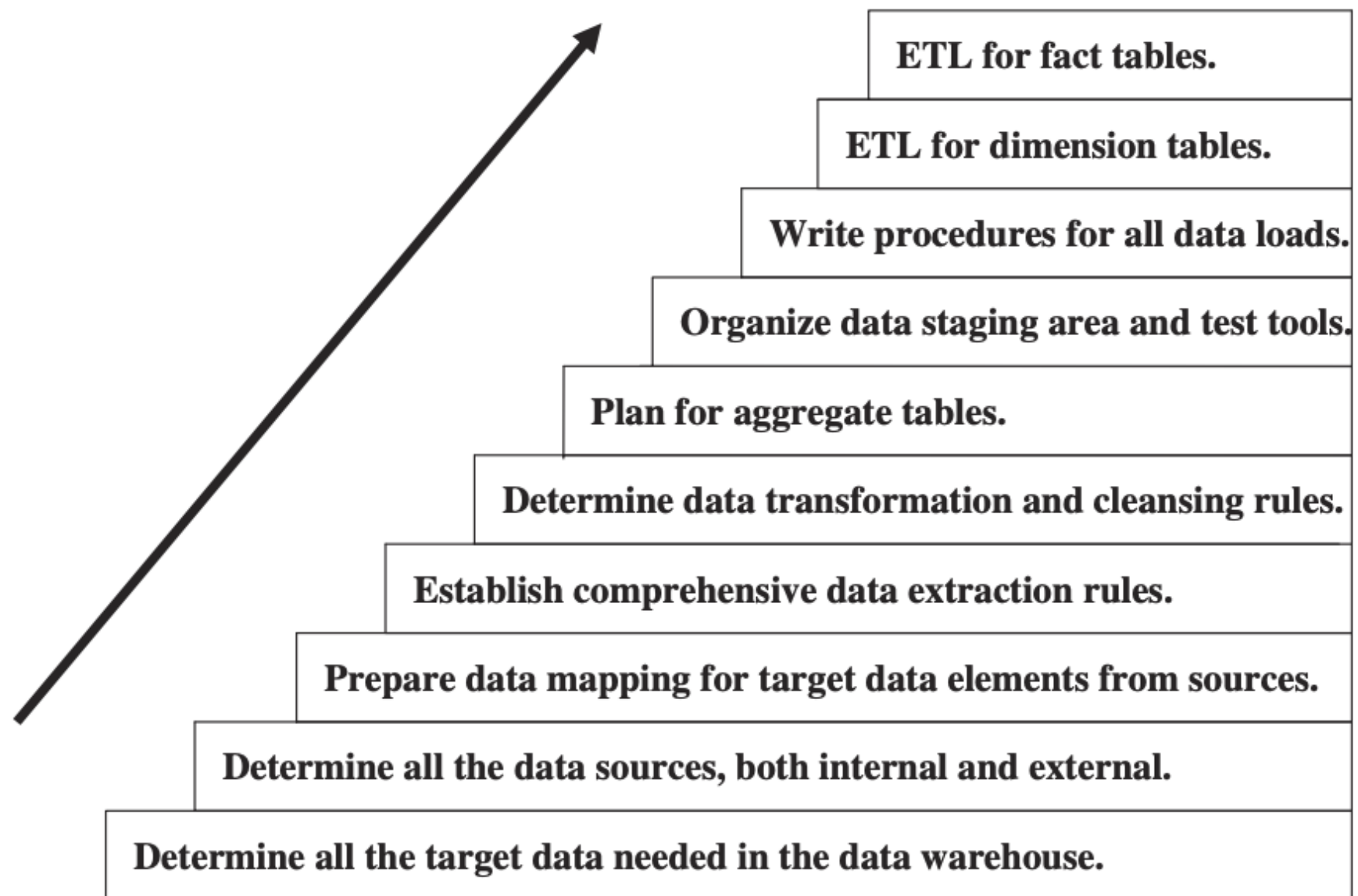


Figure 11-14 Forming aggregate fact tables.

# TOPIC 5: ETL

---



**Figure 12-1** Major steps in the ETL process.

# DATA EXTRACTION

---

- Source identification—identify source applications and source structures.
- Method of extraction—for each data source, define whether the extraction process is manual or tool-based.
- Extraction frequency—for each data source, establish how frequently the data extraction must be done: daily, weekly, quarterly, and so on.
- Time window—for each data source, denote the time window for the extraction process.
- Job sequencing—determine whether the beginning of one job in an extraction job stream has to wait until the previous job has finished successfully.
- Exception handling—determine how to handle input records that cannot be extracted.



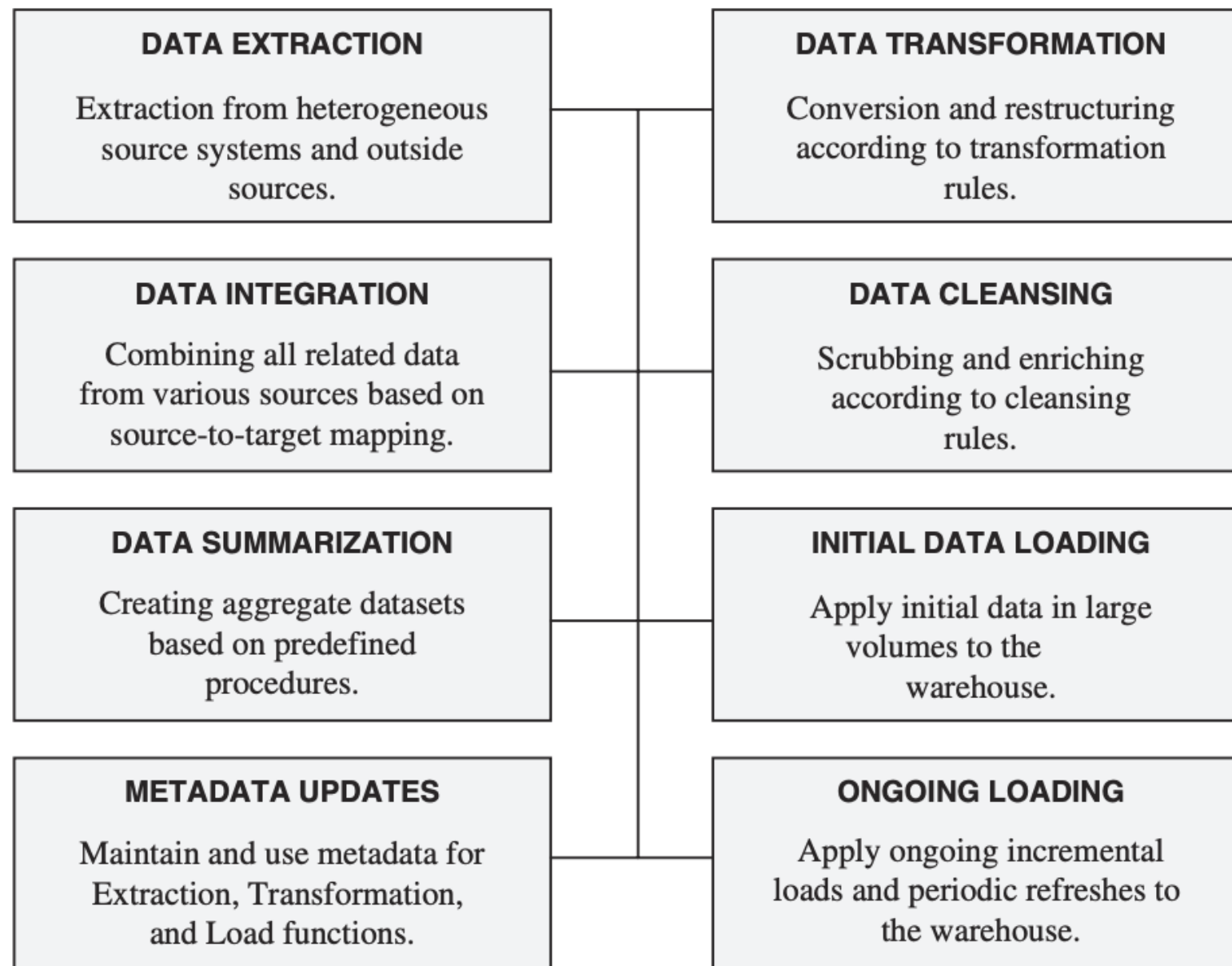
# DATA TRANSFORMATION & LOADING

---

- Data transformation encompasses data conversion, cleansing, consolidation, and integration.
- Implement the transformation function using a combination of specialized tools and in-house developed software.
- The data loading function relates to the initial load, regular periodic incremental loads, and full refreshes from time to time.
- Four methods to apply data are: load, append, destructive merge, and constructive merge.

# ETL SUMMARY

.....



**Figure 12-14** ETL summary.

# NEXT WEEK

---

- Data quality
- OLAP basics
- OLAP operations
- Data warehouse implementation
- Data cube technologies

# ASSESSMENT 1 THE DIKW HIERARCHY

---

► Specifications:

- weight: 25%
- Due: Sunday 26 July 11:59PM (Week 3)
- 1500 words limitation, less than 5 pages and in 12pt Arial
  - Each task should be one page, around 300 words
- APA 6th referencing style

Criteria	HD 85-100%
<b>Readability 1%</b>	<ul style="list-style-type: none"><li>- Excellent progression of topics with a convincing logical flow</li><li>- A highly conventional academic writing style, including the use of appropriate terminology and unbiased language</li><li>- Excellent writing without typographical or grammatical errors</li></ul>
<b>Citation/ Referencing 1%</b>	<ul style="list-style-type: none"><li>- All citations correctly adhere to APA 6<sup>th</sup> ed. referencing conventions</li><li>- Excellent use of in-text citations for all claims and statements</li><li>- Excellent list of references with no errors</li></ul>
<b>Justification 3%</b>	<ul style="list-style-type: none"><li>- Excellent development of decision, supported by logical, valid and comprehensive justifications</li></ul>

# ASSESSMENT 1 THE DIKW HIERARCHY

---

► Specifications:

- weight: 25%
- Due: Sunday 26 July 11:59PM (Week 3)
- 1500 words limitation, less than 5 pages and in 12pt Arial
  - Each task should be one page, around 300 words
- APA 6th referencing style

Criteria	HD 85-100%
<b>Readability 1%</b>	<ul style="list-style-type: none"><li>- Excellent progression of topics with a convincing logical flow</li><li>- A highly conventional academic writing style, including the use of appropriate terminology and unbiased language</li><li>- Excellent writing without typographical or grammatical errors</li></ul>
<b>Citation/ Referencing 1%</b>	<ul style="list-style-type: none"><li>- All citations correctly adhere to APA 6<sup>th</sup> ed. referencing conventions</li><li>- Excellent use of in-text citations for all claims and statements</li><li>- Excellent list of references with no errors</li></ul>
<b>Justification 3%</b>	<ul style="list-style-type: none"><li>- Excellent development of decision, supported by logical, valid and comprehensive justifications</li></ul>

# ASSESSMENT 2 THE DATA WAREHOUSE

---

## ► Specifications:

- weight: 40%
- Due: Sunday 9 August 11:59PM (Week 5)
- 2400 word limit, less than 16 pages and in 12pt Arial
  - Each task should be one page, around 300 words

## • Tasks

1. Business scenario
2. Information package
3. Data design
4. Dimensional modelling
5. Size of fact table
6. Aggregating fact table
7. Lattice of cuboids
8. Data cube computation