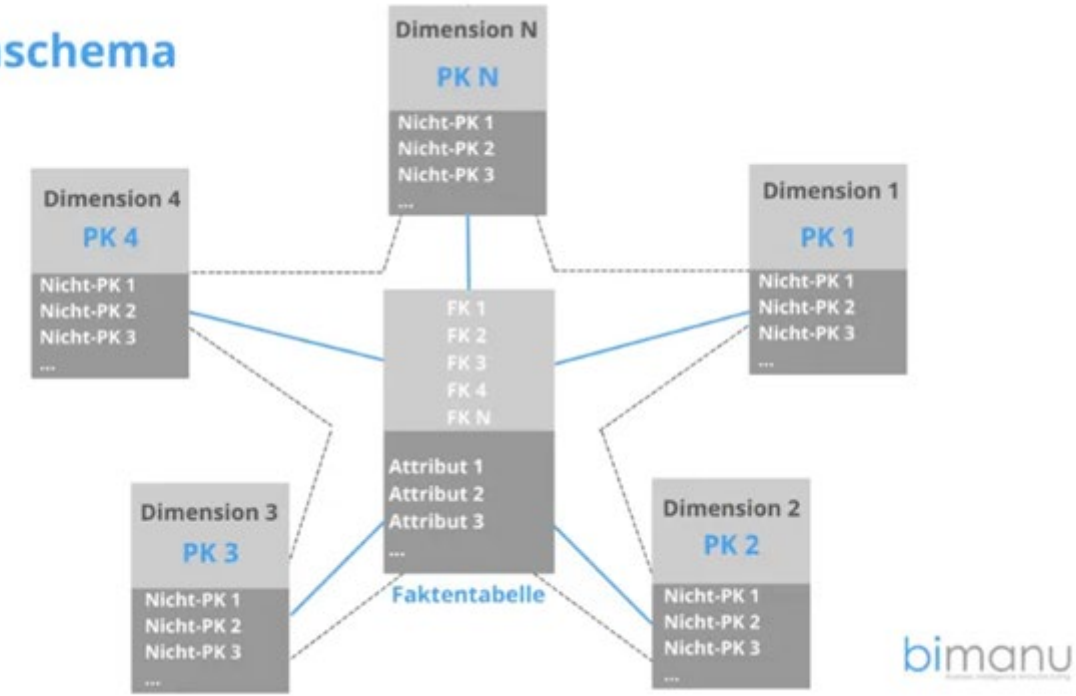
Data Engineer

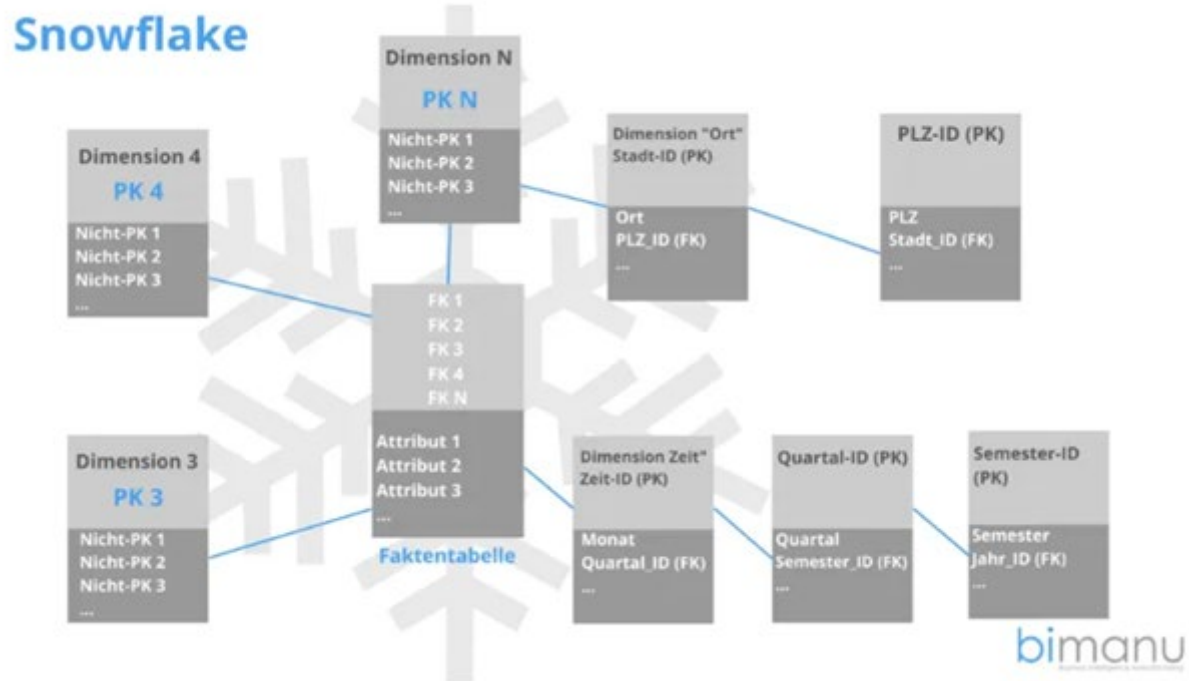
Snowflake – Schema:

Snowflake Schema Grundlagen, Datenmodellierung, Erstellung Snowflake Schema in RDBMS

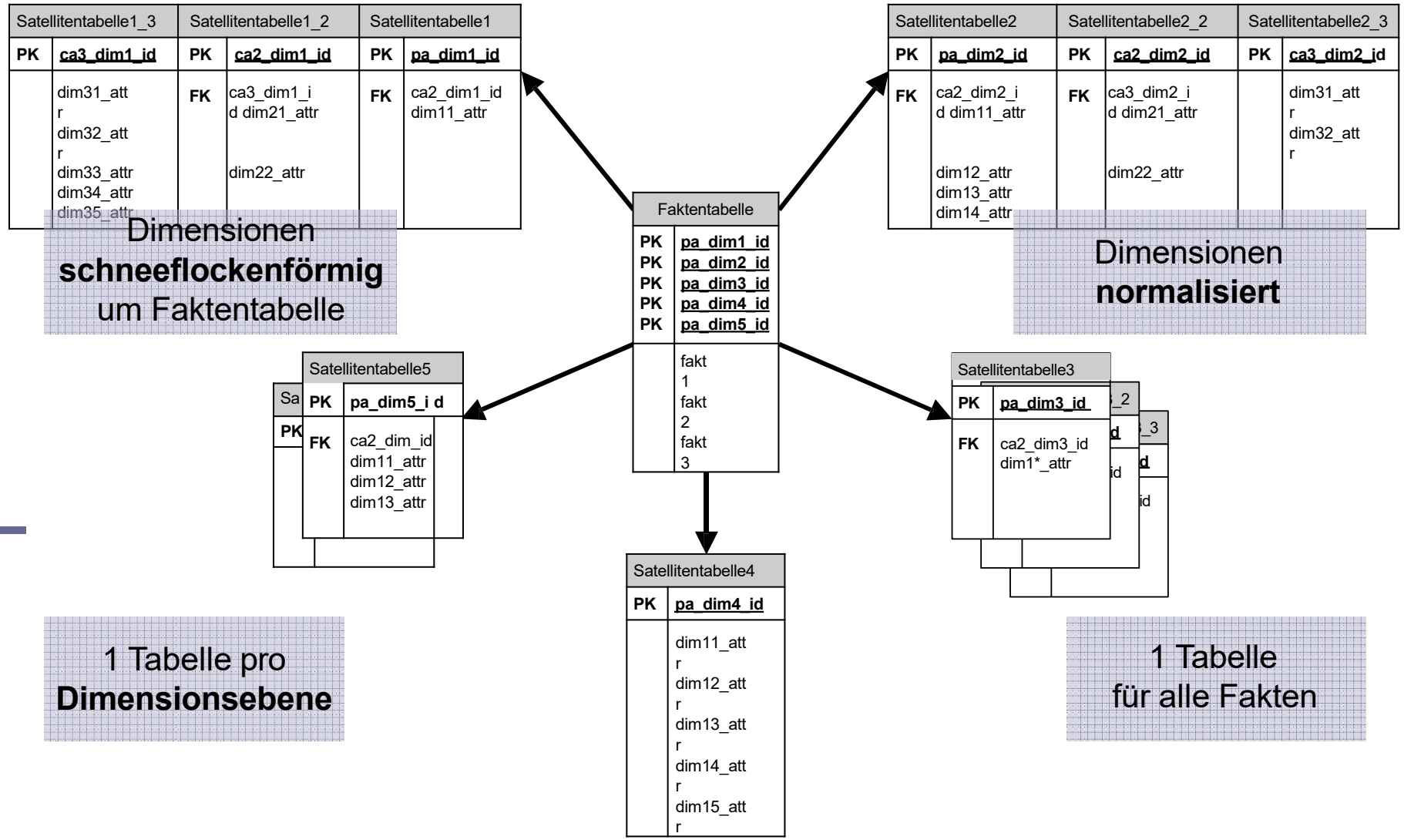
Sternschema



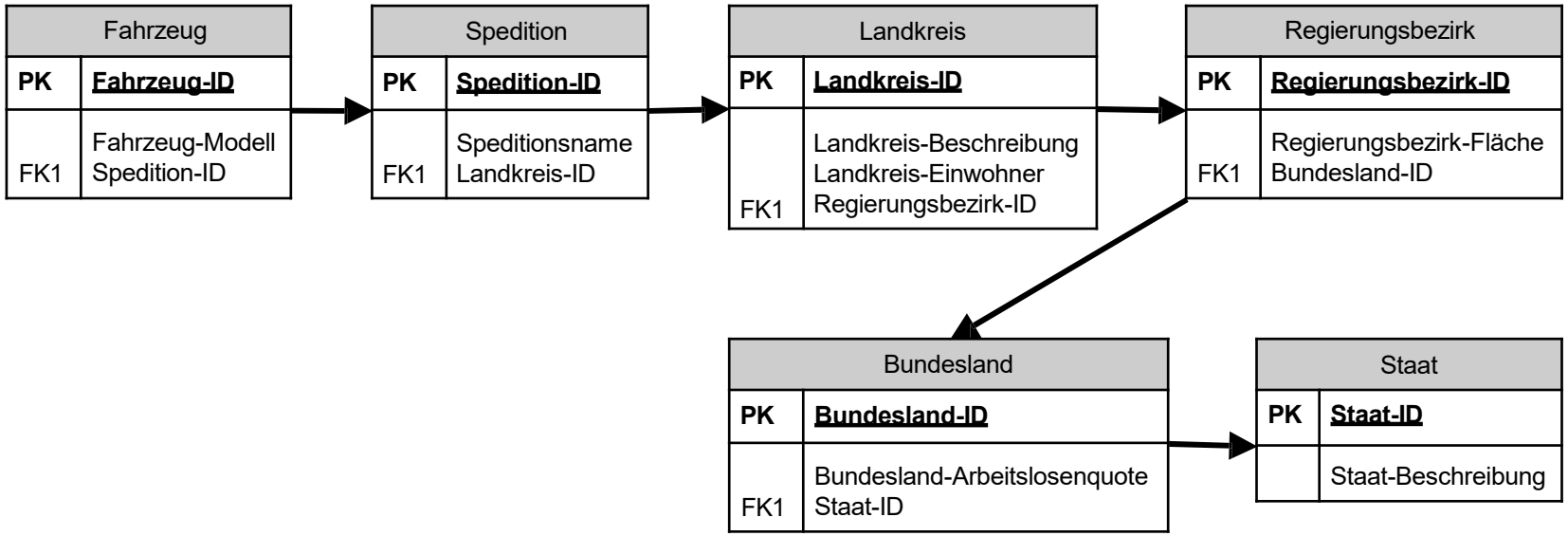
bimanu


Snowflake
Dimension N
PK N
Nicht-PK 1
Nicht-PK 2
Nicht-PK 3
...
Dimension 4
PK 4
Nicht-PK 1
Nicht-PK 2
Nicht-PK 3
...
Dimension "Ort"
Stadt-ID (PK)
Ort
PLZ_ID (FK)
...
PLZ-ID (PK)
PLZ
Stadt_ID (FK)
...
FK 1
FK 2
FK 3
FK 4
FK N
Attribut 1
Attribut 2
Attribut 3
...
Faktentabelle
Dimension 3
PK 3
Nicht-PK 1
Nicht-PK 2
Nicht-PK 3
...
Dimension Zeit"
Zeit-ID (PK)
Monat
Quartal_ID (FK)
...
Quartal-ID (PK)
Quartal
Semester_ID (FK)
...
Semester-ID
(PK)
Semester
Jahr_ID (FK)
...
bimanu

Snowflake-Schema



Snowflake-Dimension: Beispiel



Snowflake-Schema

- Warum muss jetzt alles ID heißen?
  - Eindeutigkeit der Referenzkette in der Klassifkationshierarchie:
    - Eine Monats-ID aus der Faktentabelle gehört zu einem bestimmten Quartal in einem bestimmten Jahr
    - „2. Quartal“ reicht dafür nicht, weil das Jahr nicht mehr herausgefunden werden kann

→ gegebenfalls sind zu den IDs in den Satellitentabellen noch Beschreibungen zu ergänzen (Speditionsname, Landkreis-Beschreibung, …)

Snowflake-Schema

- Probleme
  - weniger übersichtlich
  - weniger performante Anfragebearbeitung
- Eigenschaften
  - **normalisierte Dimensionstabellen**
  - Dimensionsklassifikationen explizit modelliert

Star- vs. Snowflake-Schema

- Welche Variante ist besser?
  - → Welches Schema geeignet ist, hängt vom Anwendungsprofil ab!
  - → Dennoch wird sehr häufig ein Star-Schema oder ein leicht modifiziertes Star-Schema verwendet.

OLAP- Architekturen

- **ROLAP:** Relational On Line Analytical Processing,
  relationale Datenspeicherung – Tabellenform.

- **MOLAP:** Multidimensional On Line Analytical Processing, multidimensional Datenspeicherung, n-dimensionaler Würfel.

- **HOLAP:** Hybrid On Line Analytical Processing (HOLAP)
  Speicherung eines Teils des DWH's in Form von Würfeln.

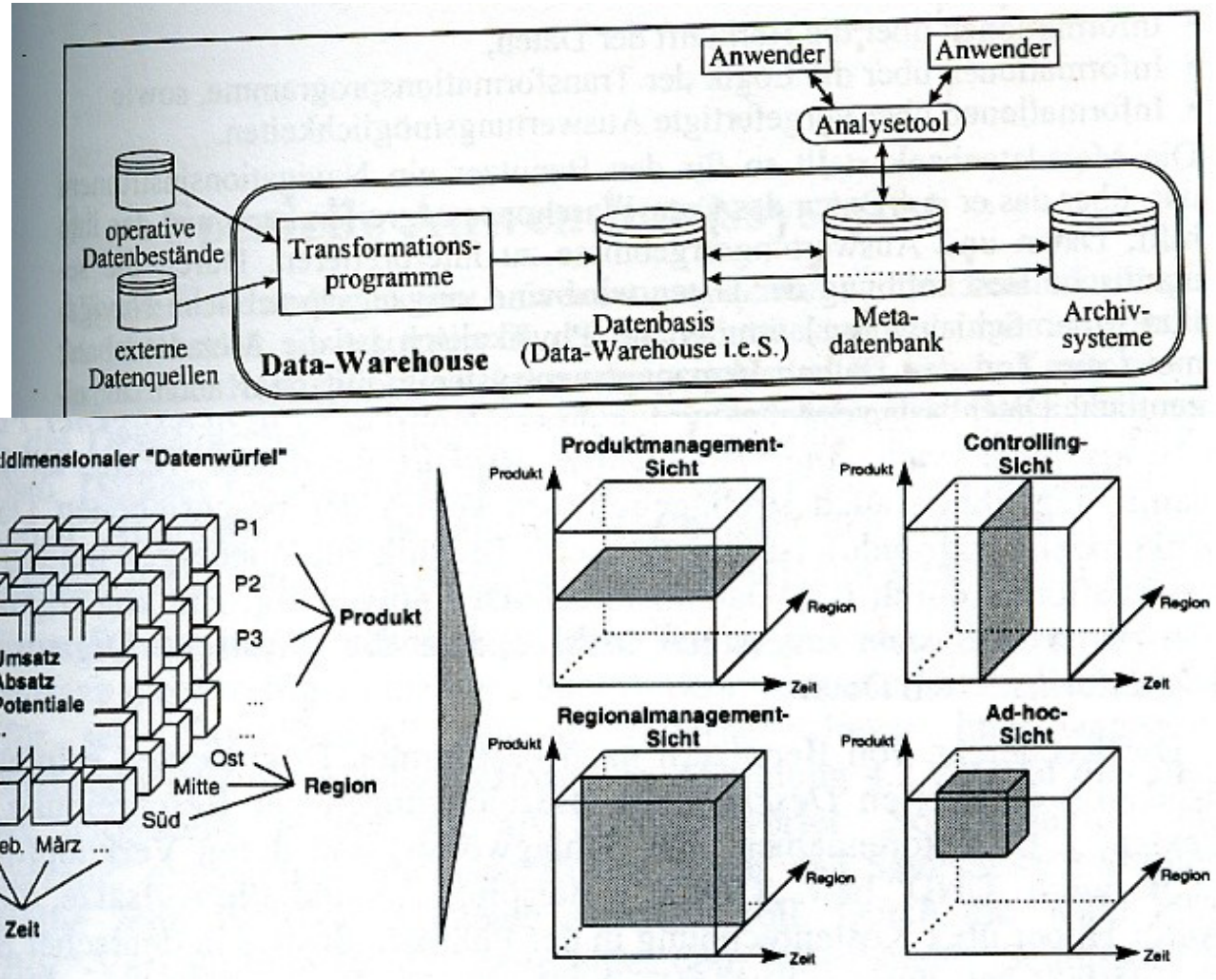
Data Warehouse

Grundkonzept Datenbank/Datenmodellierung

- Ziel: Bereitstellung aller relevanten Daten für betriebswirtschaftliche Analysen des Managements

- Ausprägungen
  - global = Data-Warehouse
  - spezifisch = Data-Mart

- Schritte
  - Modellierung
  - Data-Mining
  - Präsentation



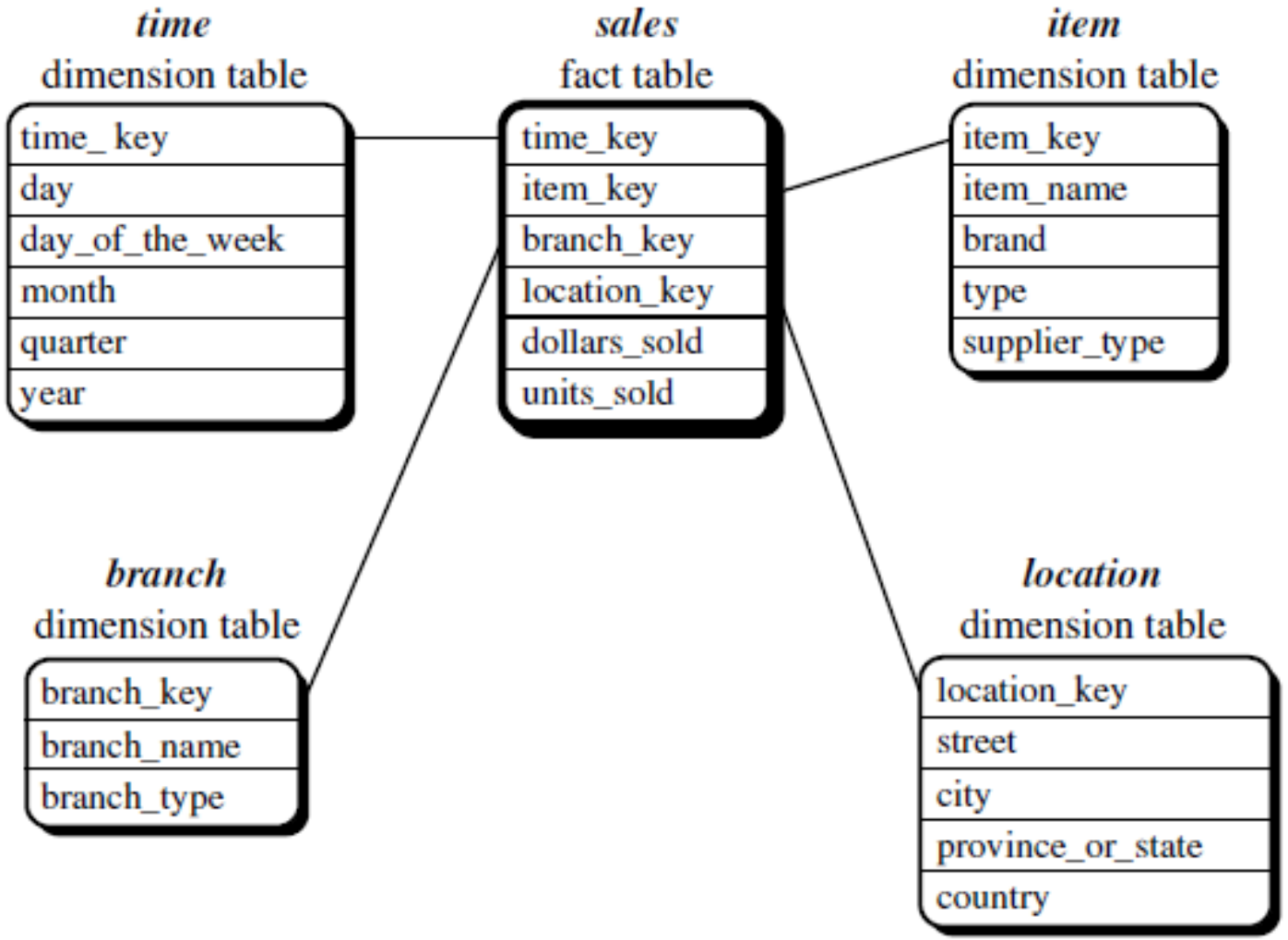
Multi-dimensional Schemas (2)

- Two common multi-dimensional schemas are
  - **Star Schema:**
    - Consists of a fact table with a single table for each dimension
  - **Snowflake Schema:**
    - It is a variation of Star Schema, in which the dimensional tables from a Snowflake Schema are organized into a hierarchy by normalizing them.

Conceptual Modeling of Data Warehouses

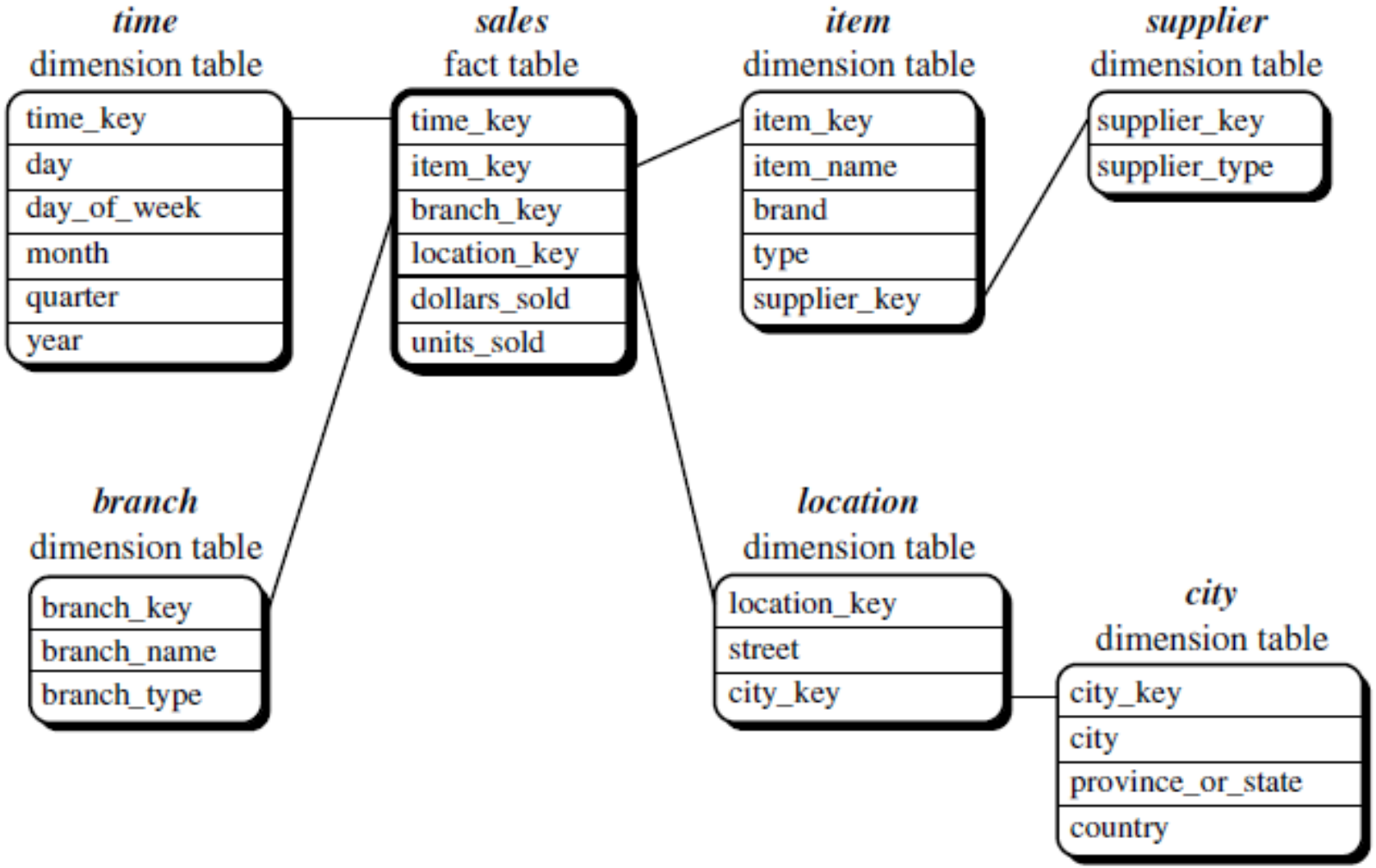
- Modeling data warehouses: multidimensional model
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake

Example of Star Schema



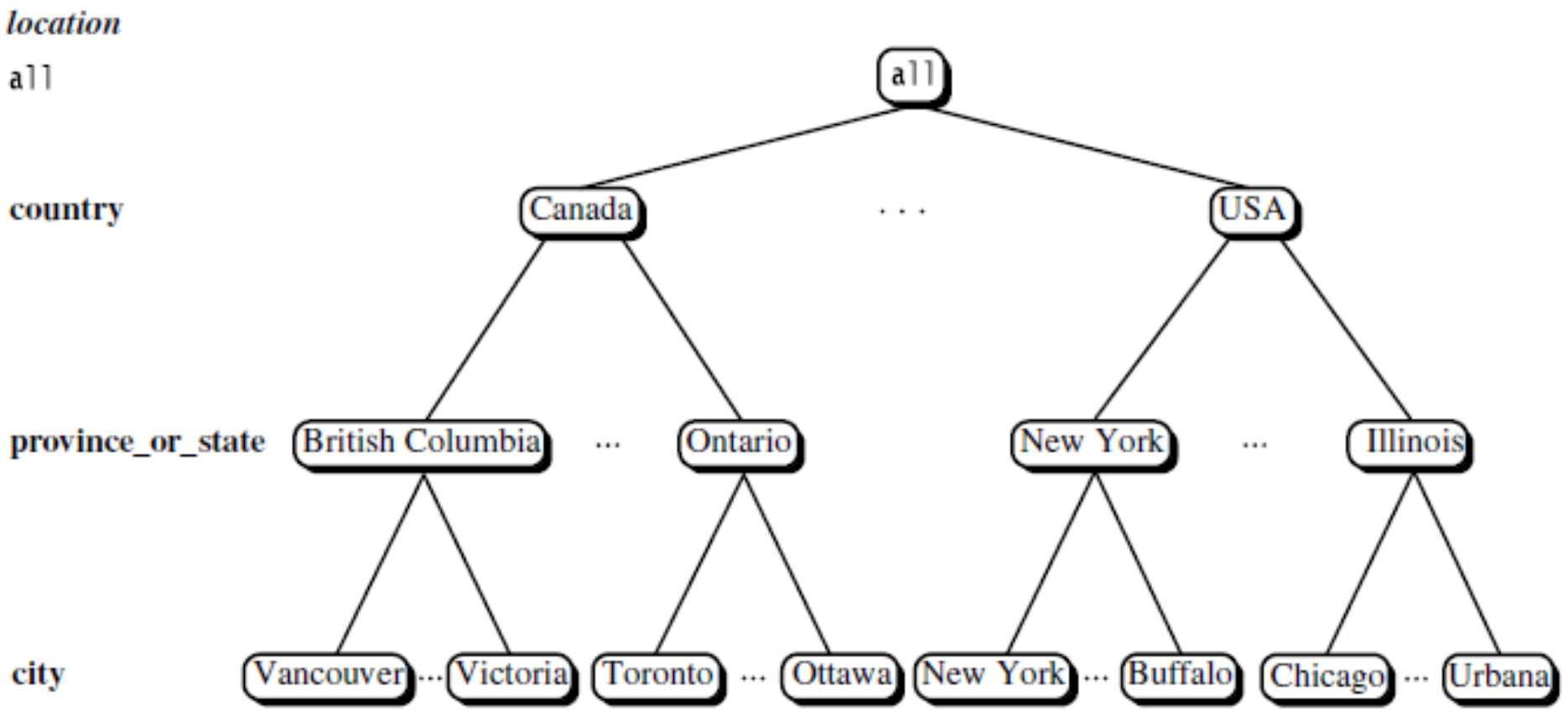
Star schema of a data warehouse for sales

Example of Snowflake Schema



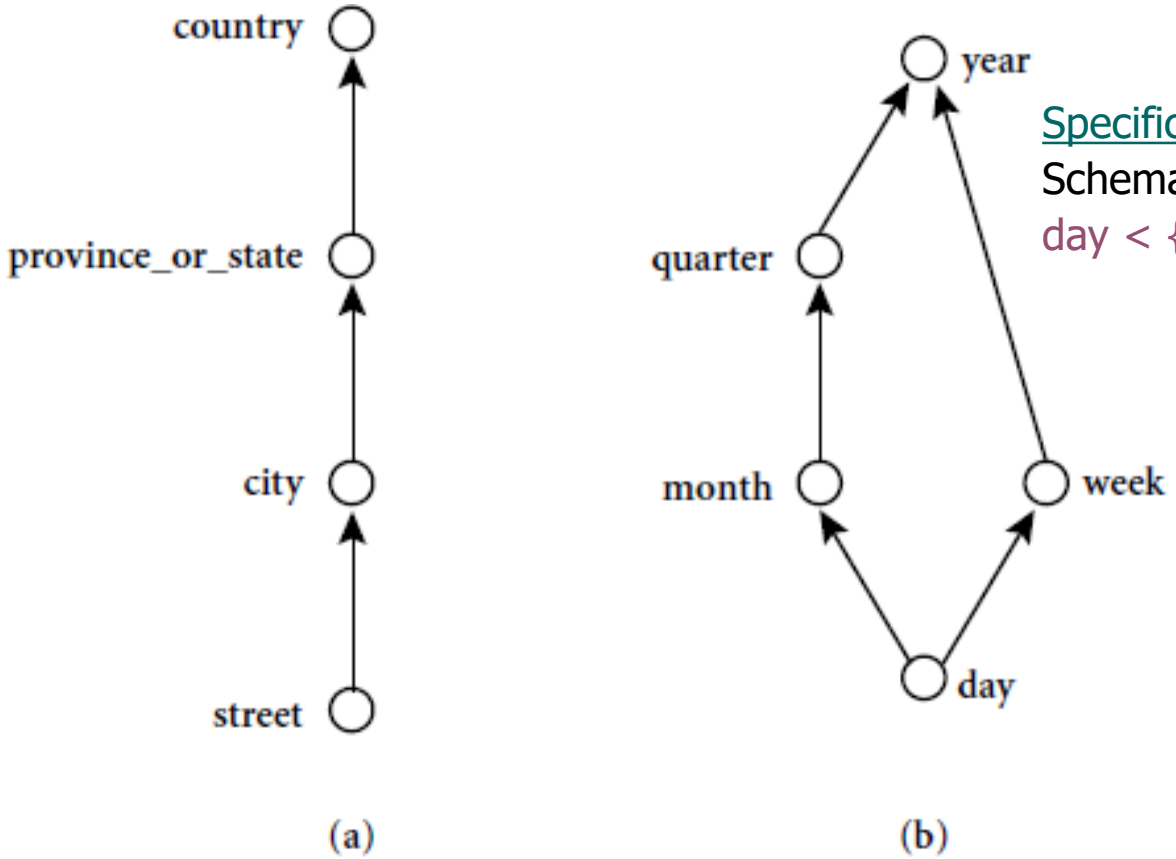
Snowflake schema of a data warehouse for sales

A Concept Hierarchy: Dimension (location)



A concept hierarchy for the dimension *location*

Concept Hierarchies



Hierarchical and lattice structures of attributes in warehouse dimensions: (a) a hierarchy for *location*; (b) a lattice for *time*

Specification of hierarchies

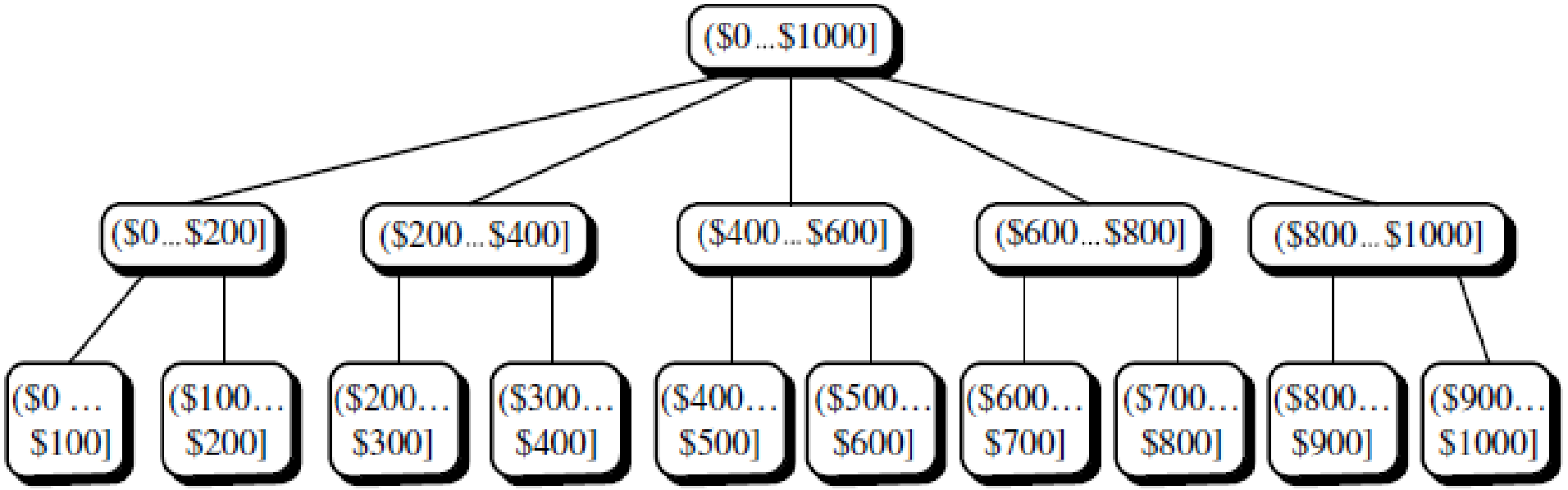
Schema hierarchy

day < {month < quarter; week} < year

Concept Hierarchies

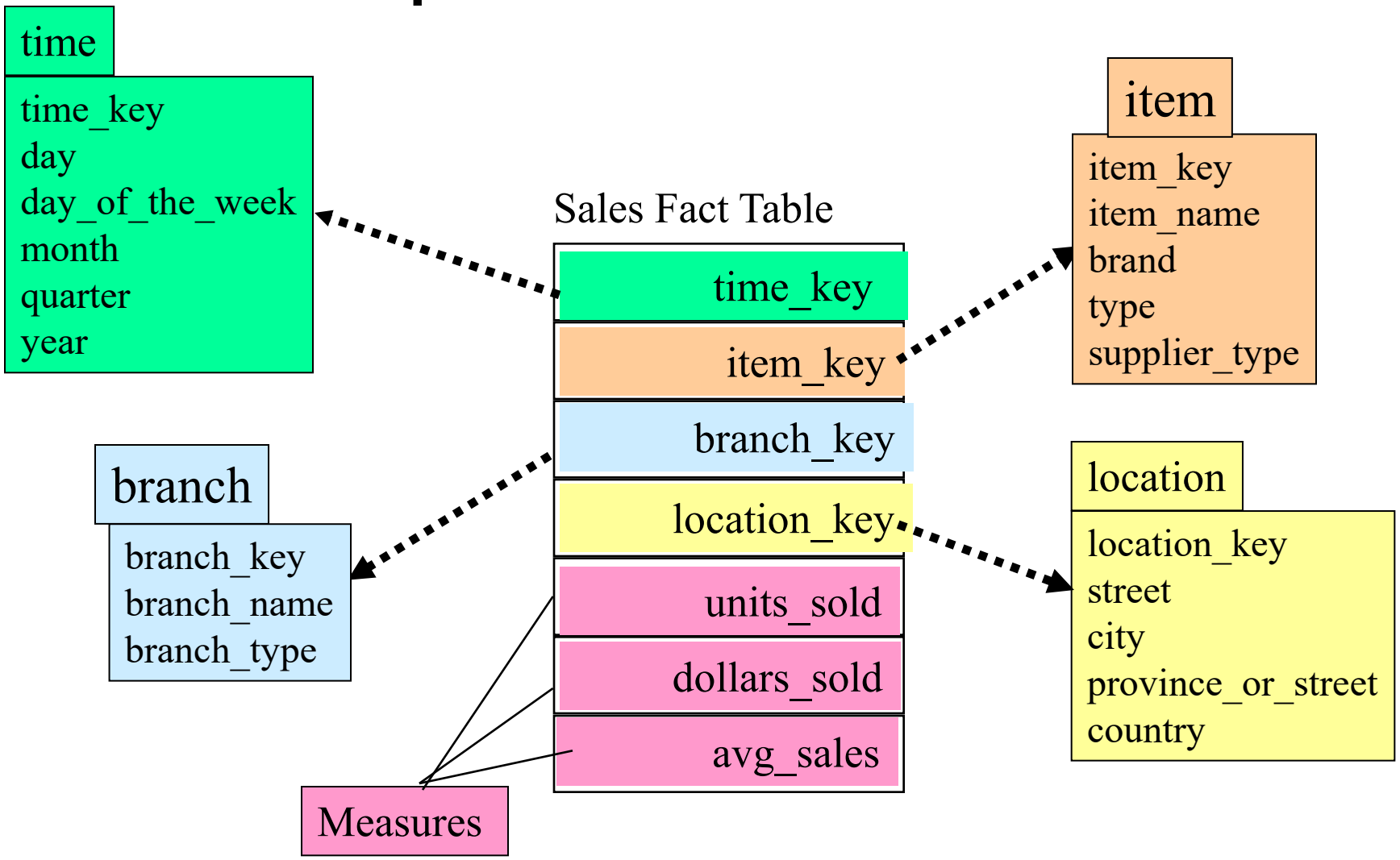
Specification of hierarchies

Set_grouping hierarchy

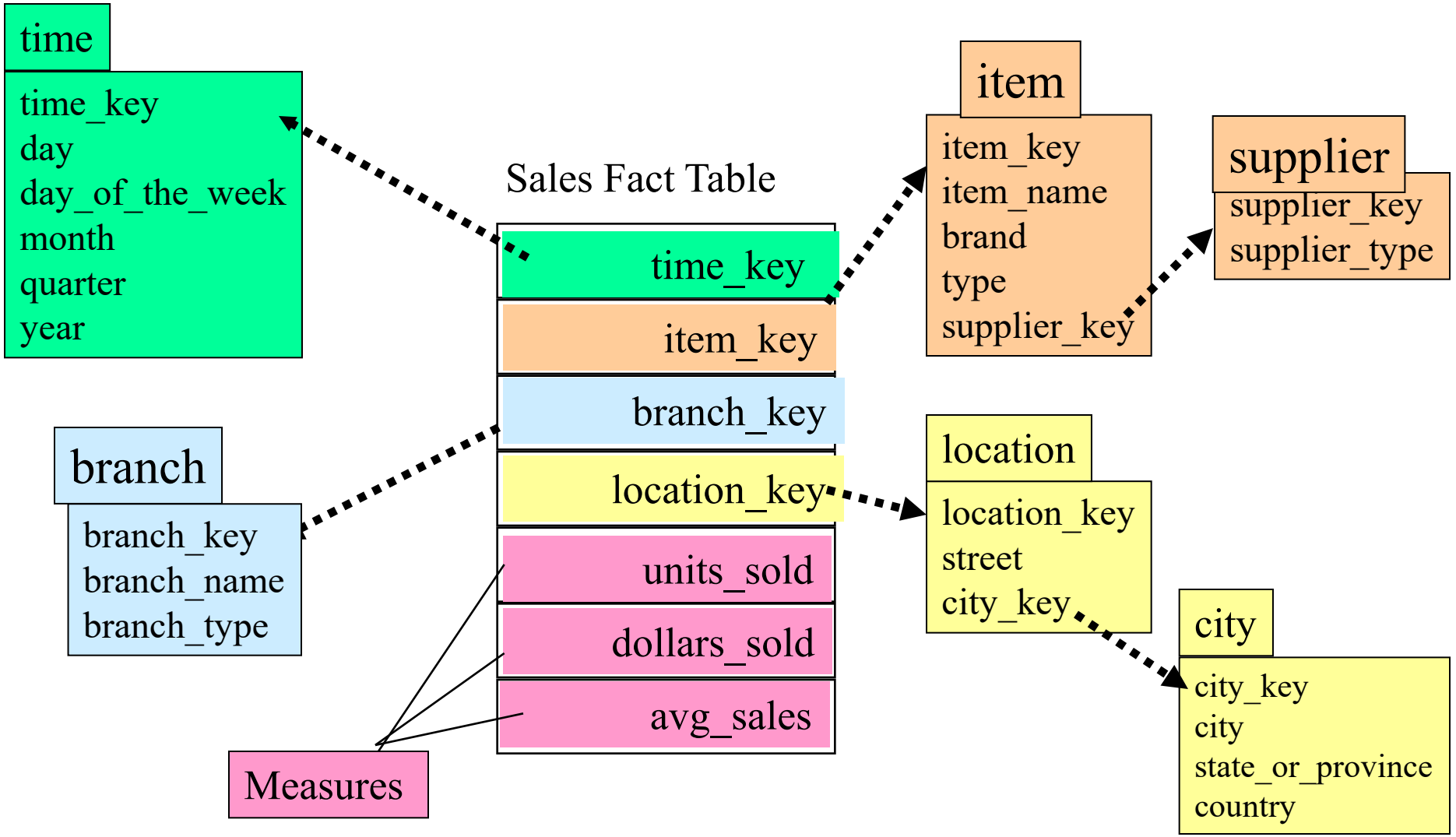


A concept hierarchy for the attribute *price*

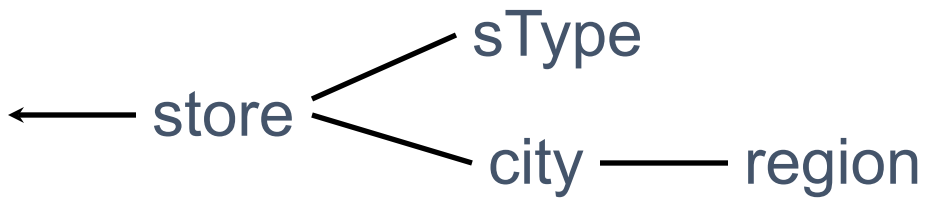
Example of Star Schema



Example of **Snowflake Schema**



Dimension Hierarchies



store	storeId	cityId	tId	mgr
	s5	sfo	t1	joe
	s7	sfo	t2	fred
	s9	la	t1	nancy

sType	tId	size	location
	t1	small	downtown
	t2	large	suburbs

city	cityId	pop	regId
	sfo	1M	north
	la	5M	south

region	regId	name
	north	cold region
	south	warm region

➔ snowflake schema
➔ constellations

Cube

Fact table view:

sale	prodId	storeId	amt
	p1	c1	12
	p2	c1	11
	p1	c3	50
	p2	c2	8

Multi-dimensional cube:

	c1	c2	c3
p1	12		50
p2	11	8	

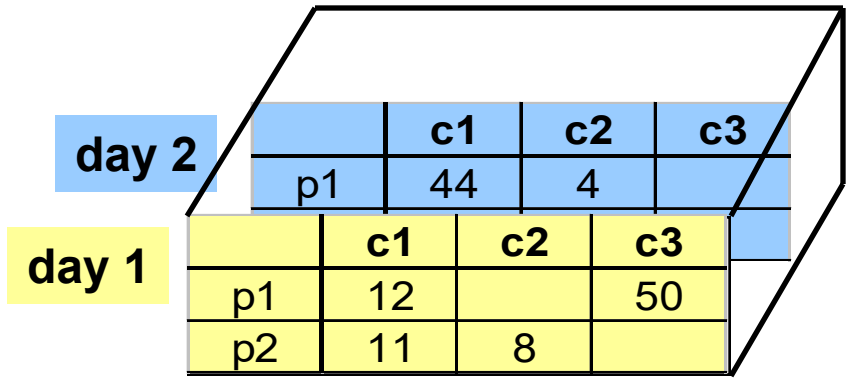
dimensions = 2

3-D Cube

Fact table view:

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

Multi-dimensional cube:

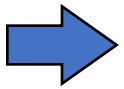


dimensions = 3

Aggregates

- Add up amounts for day 1
- In SQL: SELECT sum(amt) FROM SALE
  WHERE date = 1

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

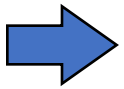


81

Aggregates

- Add up amounts by day
- In SQL: SELECT date, sum(amt) FROM SALE
  GROUP BY date

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

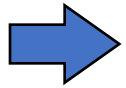


ans	date	sum
	1	81
	2	48

Another Example

- Add up amounts by day, product
- In SQL: SELECT date, sum(amt) FROM SALE
  GROUP BY date, prodId

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4



sale	prodId	date	amt
	p1	1	62
	p2	1	19
	p1	2	48

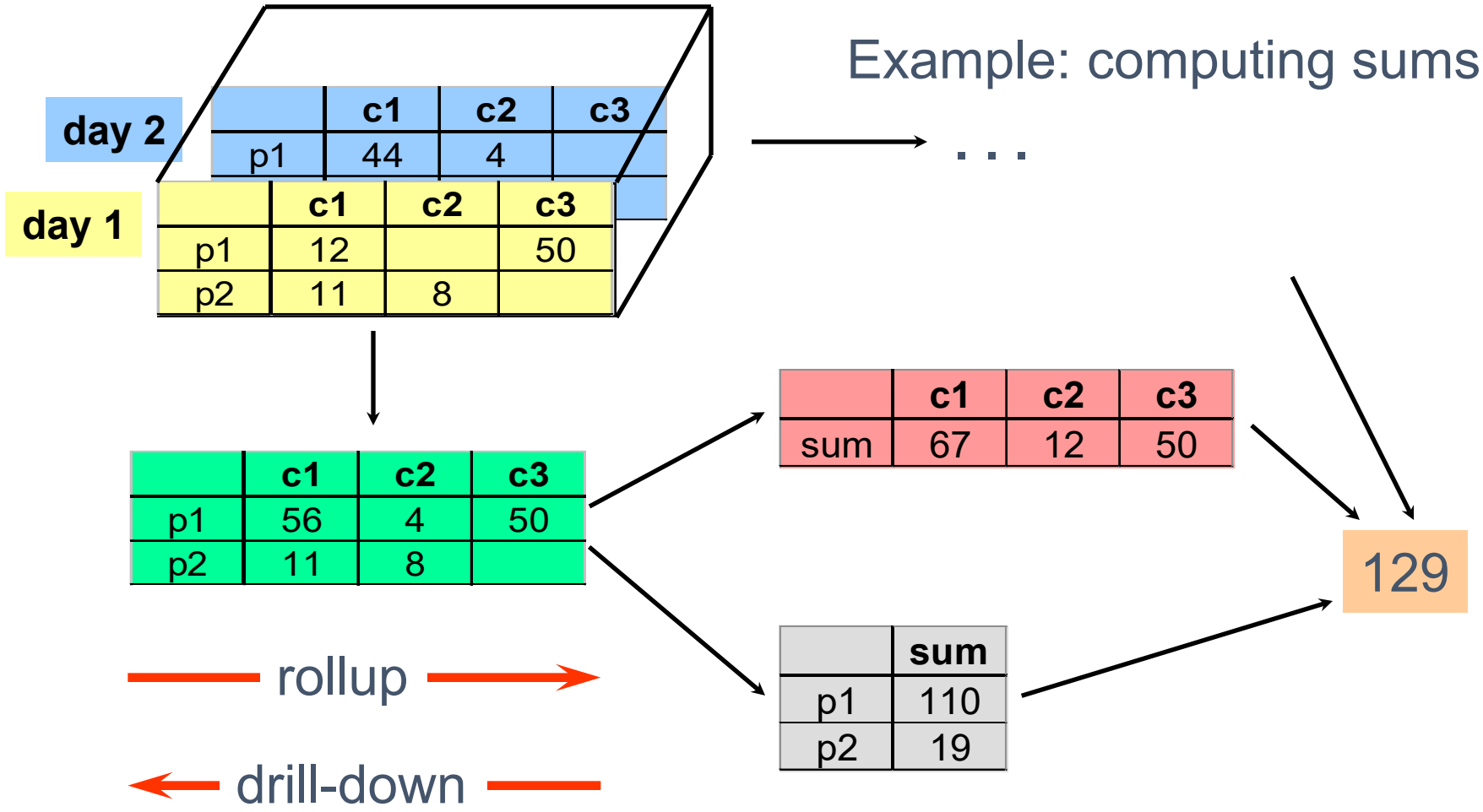
rollup →

← drill-down

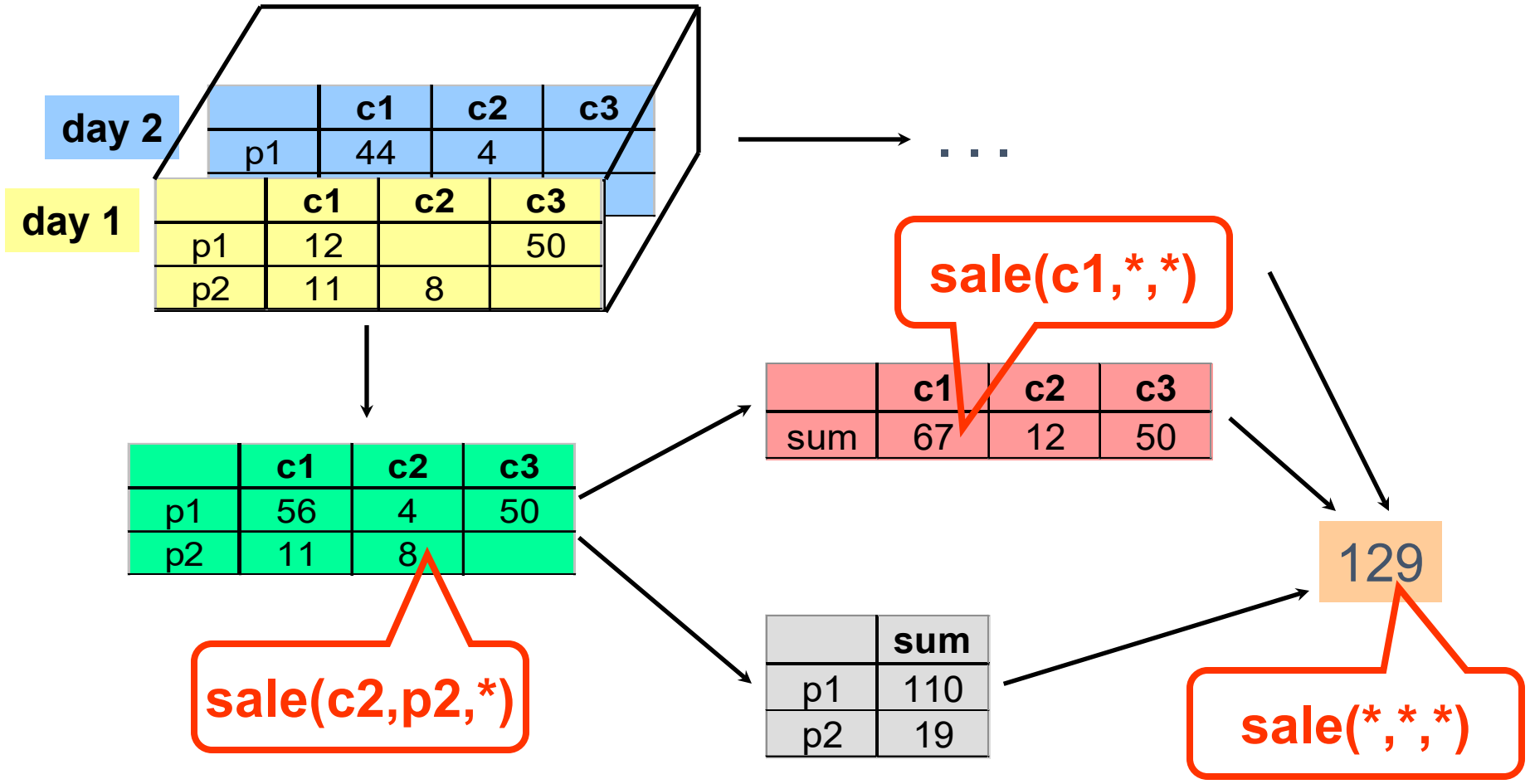
Aggregates

- Operators: sum, count, max, min, median, ave
- “Having” clause
- Using dimension hierarchy
  - average by region (within store)
  - maximum by month (within date)

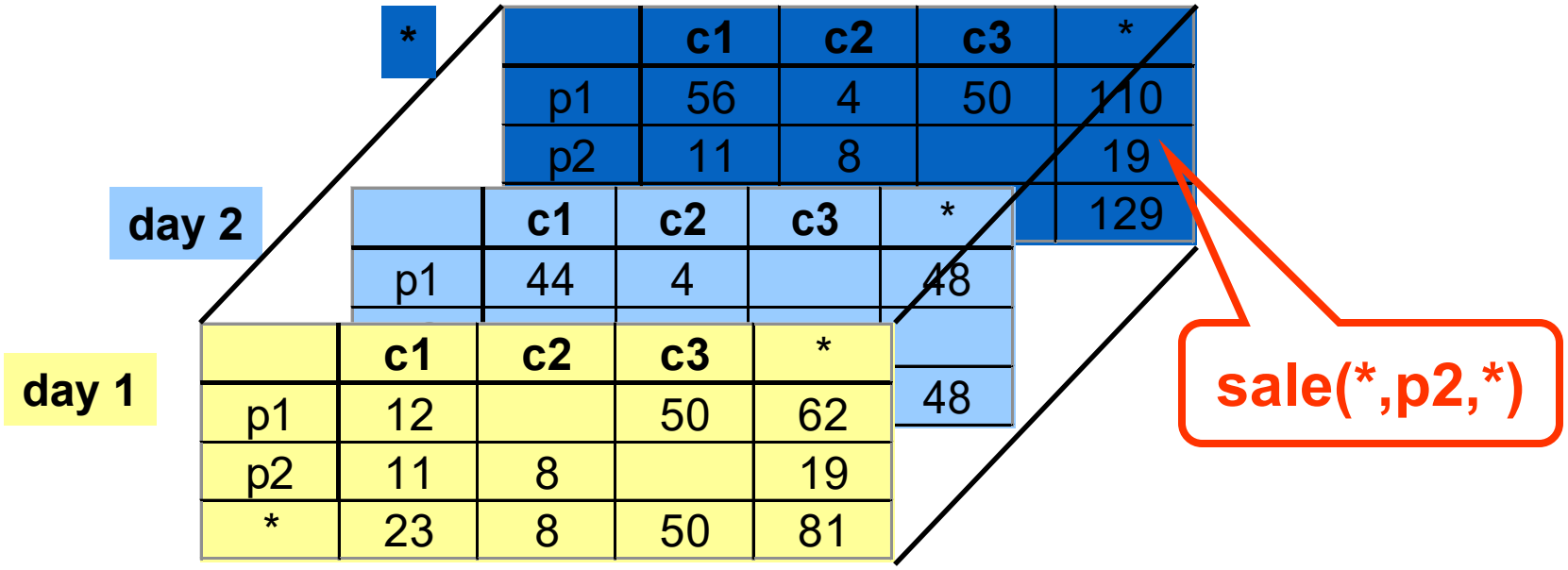
Cube Aggregation



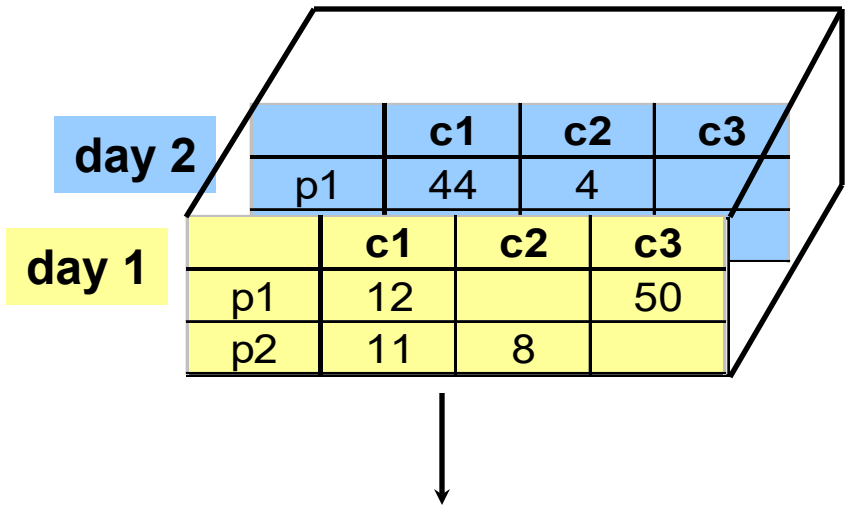
Cube Operators



Extended Cube



Aggregation Using Hierarchies



	region A	region B
p1	56	54
p2	11	8

customer

region

country

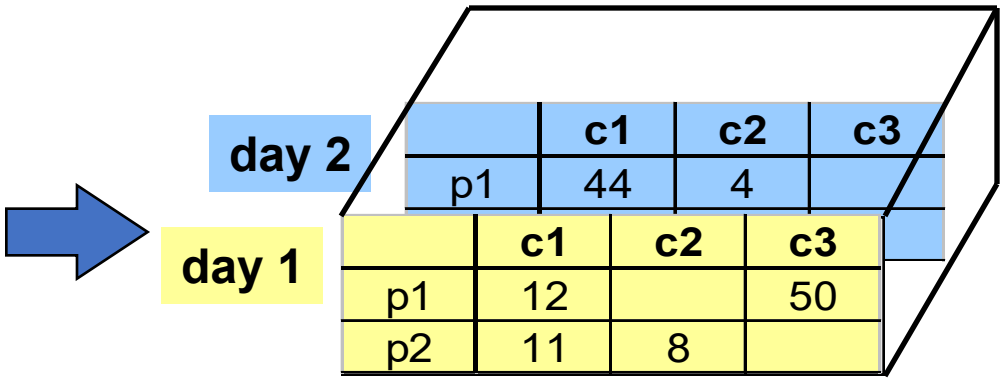
(customer c1 in Region A;
customers c2, c3 in Region B)

Pivoting

Fact table view:

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

Multi-dimensional cube:

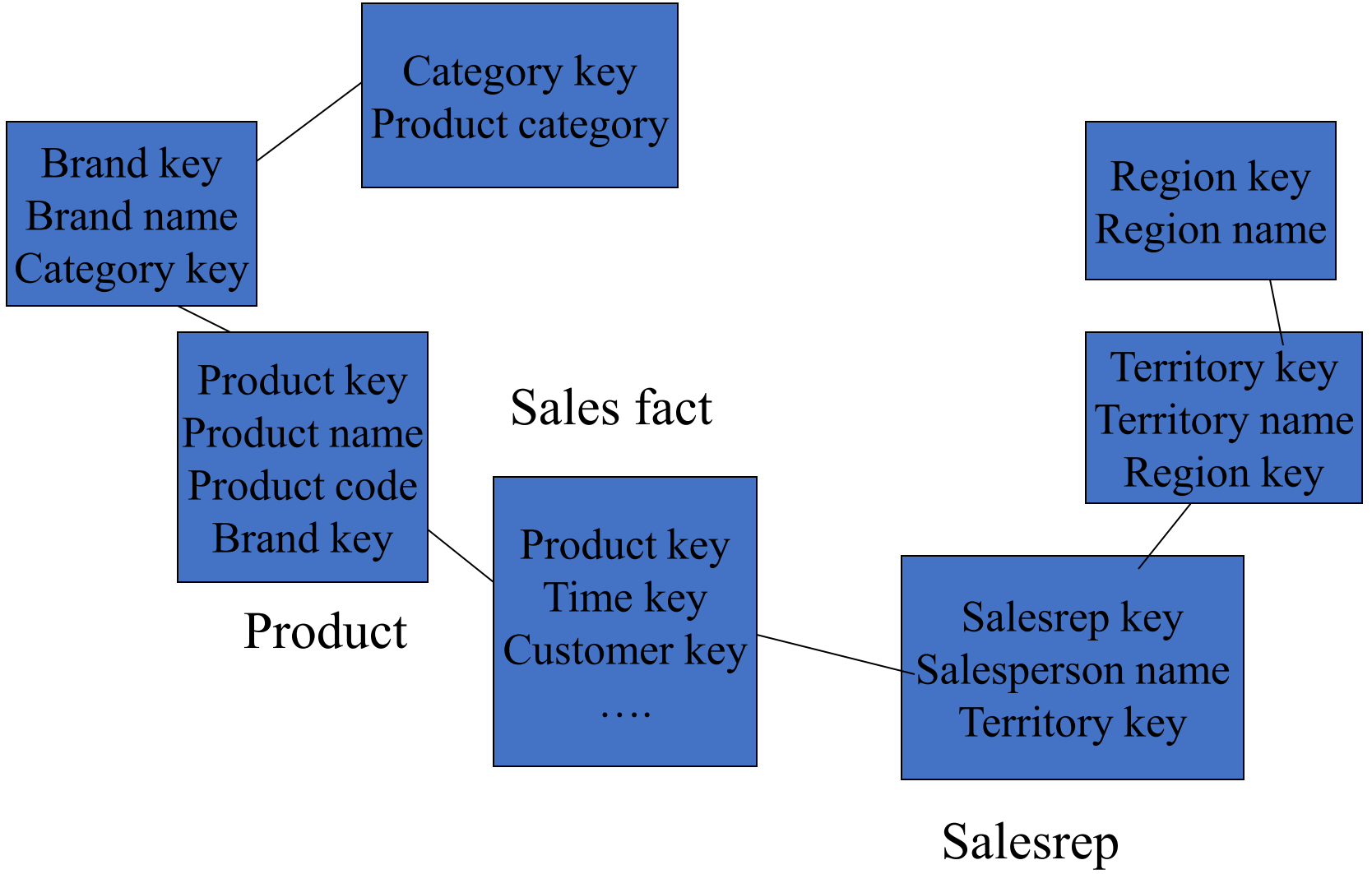


	c1	c2	c3
p1	56	4	50
p2	11	8	

Snowflake Schema

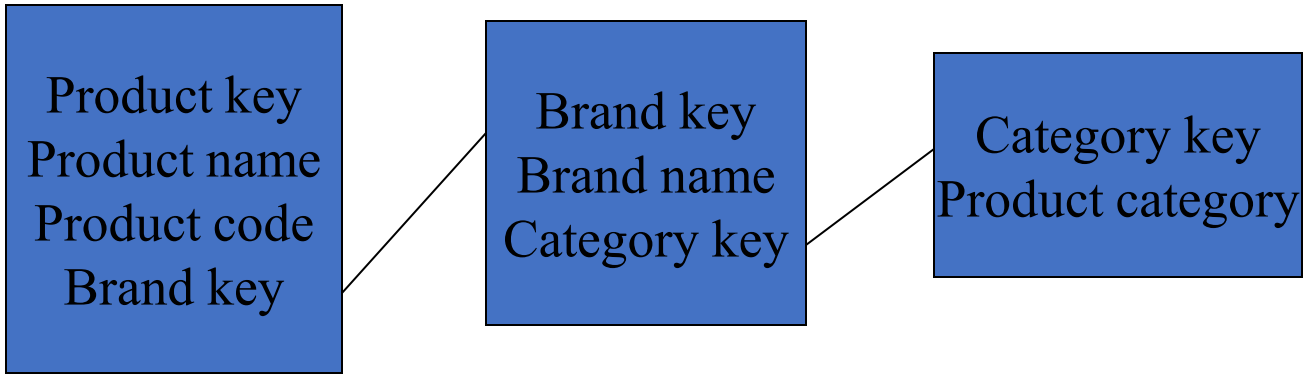
- Snowflake schema is a type of star schema but a more complex model.
- “Snowflaking” is a method of normalizing the dimension tables in a star schema.
- The normalization eliminates redundancy.
- The result is more complex queries and reduced query performance.

Sales: Snowflake Schema



Snowflaking

- The attributes with low cardinality in each original dimension table are removed to form separate tables. These new tables are linked back to the original dimension table through artificial keys.

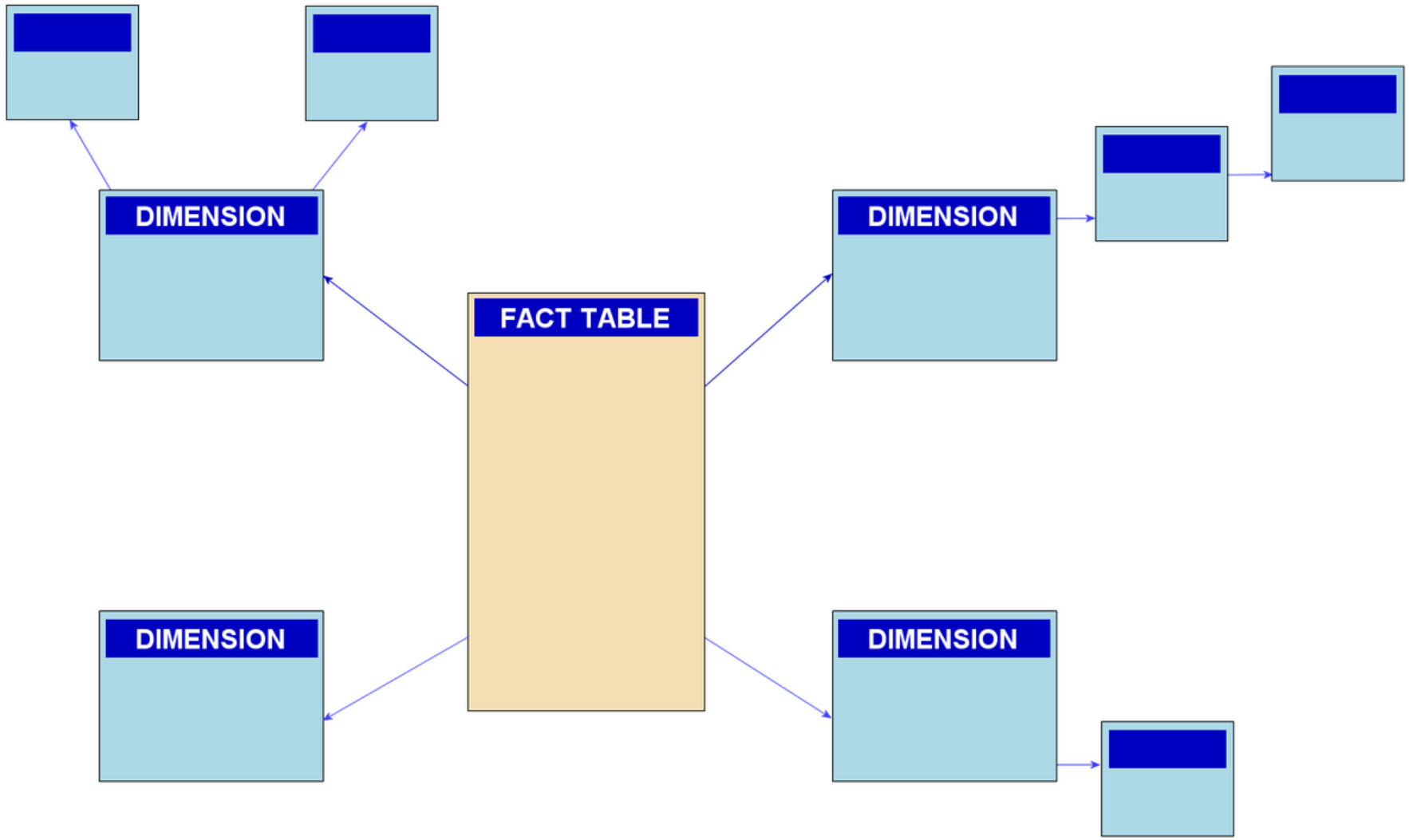


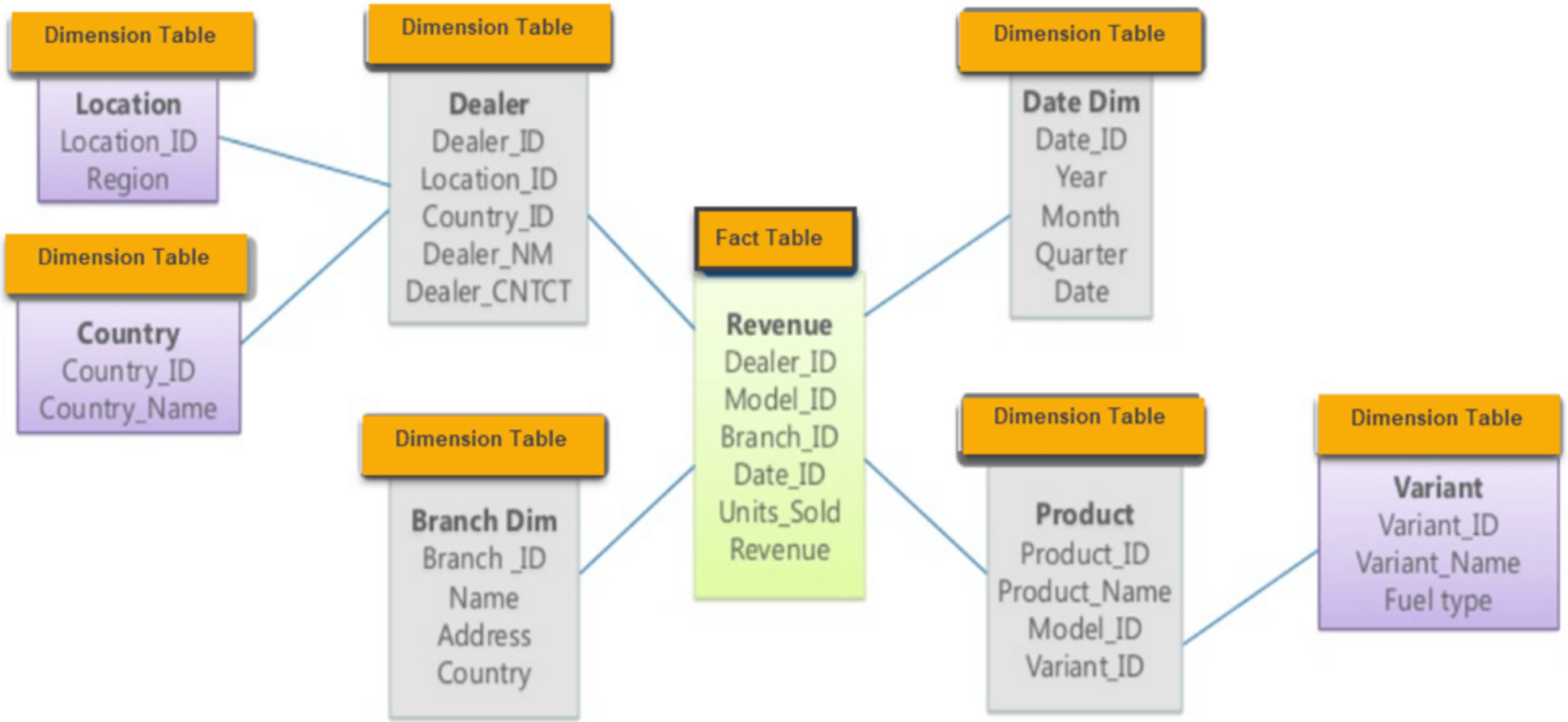
Snowflake Schema

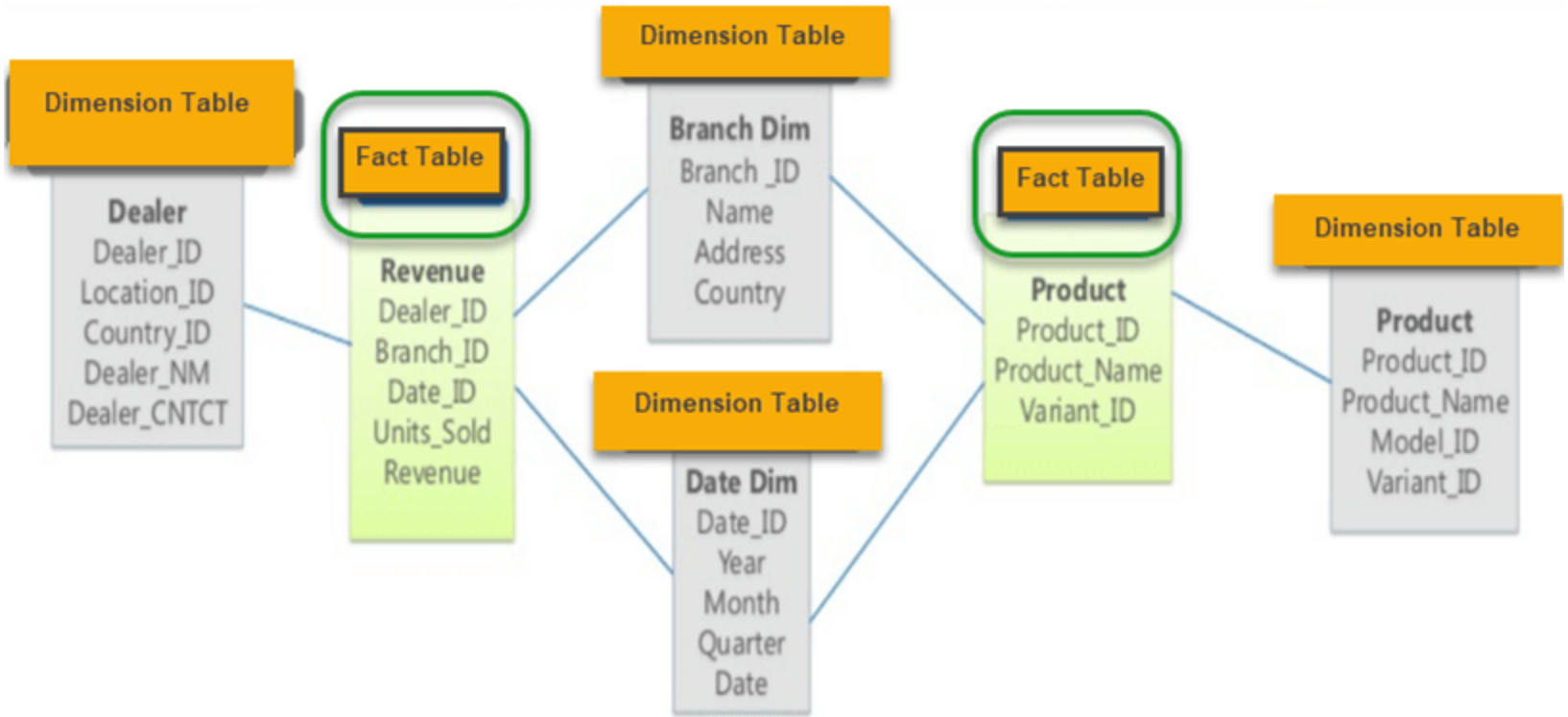
- Advantages:
  - Small saving in storage space
  - Normalized structures are easier to update and maintain
- Disadvantages:
  - Schema less intuitive and end-users are put off by the complexity
  - Ability to browse through the contents difficult
  - Degrade query performance because of additional joins

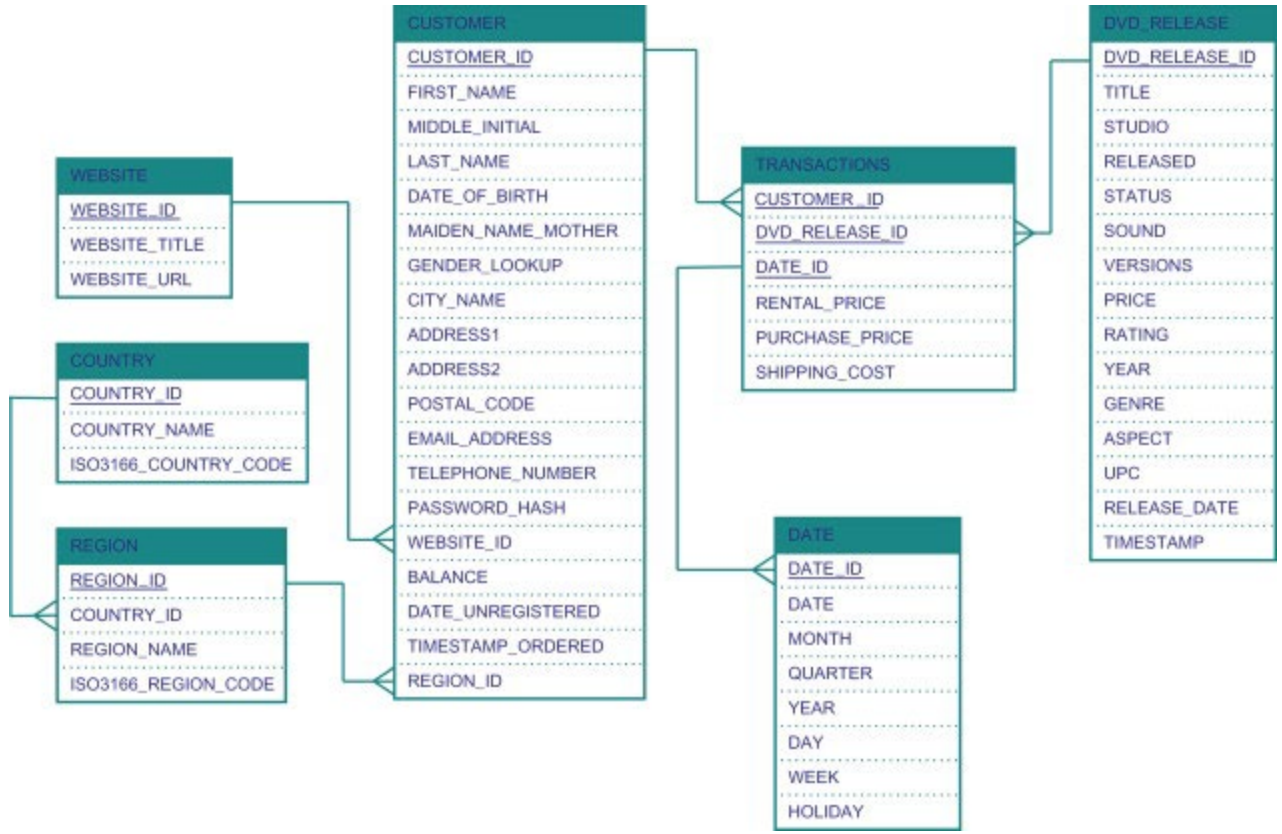
What is the Best Design?

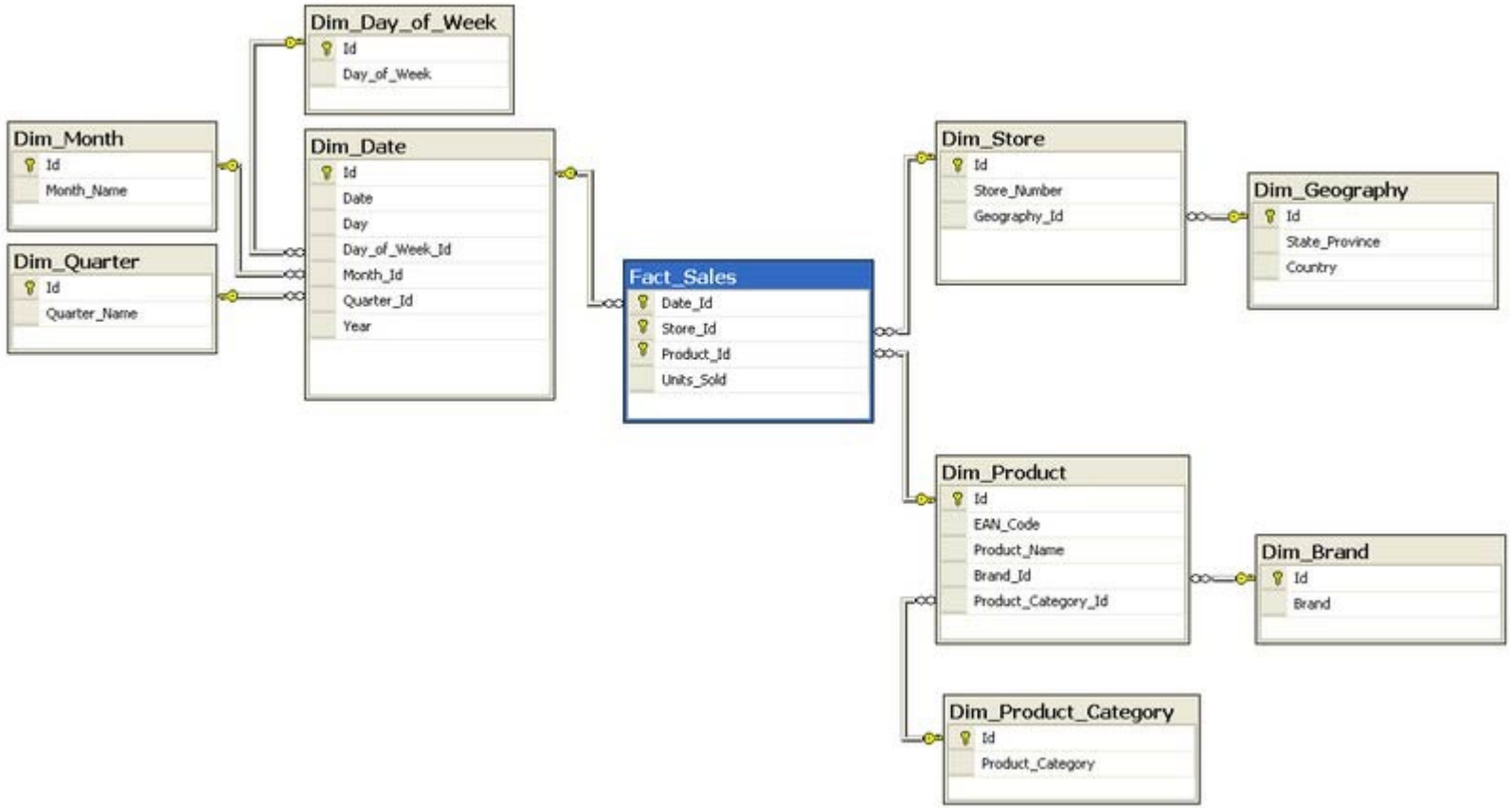
- Performance benchmarking can be used to determine what is the best design.
- Snowflake schema: easier to maintain dimension tables when dimension tables are very large (reduce overall space). It is not generally recommended in a data warehouse environment.
- Star schema: more effective for data cube browsing (less joins): can affect performance.

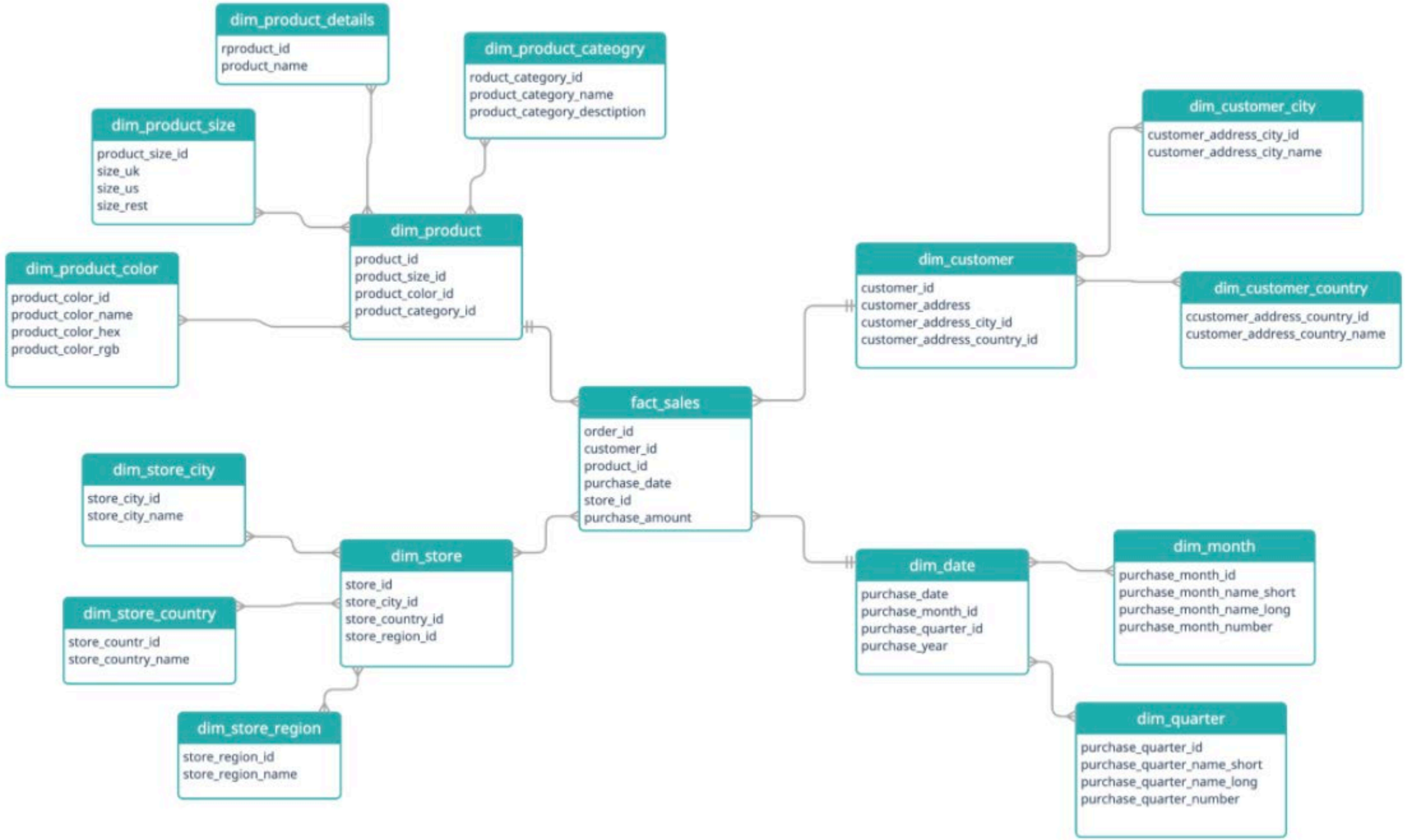

DIMENSION
DIMENSION
FACT TABLE
DIMENSION
DIMENSION

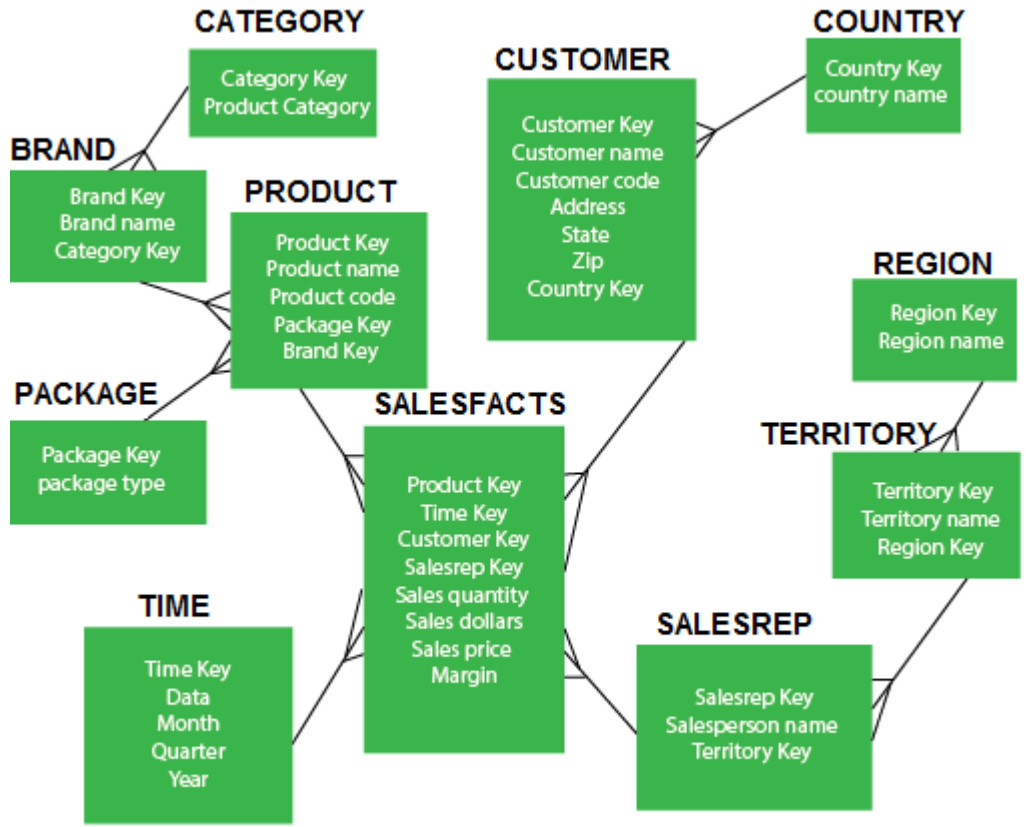

Dimension Table
Location
Location_ID
Region
Dimension Table
Country
Country_ID
Country_Name
Dimension Table
Dealer
Dealer_ID
Location_ID
Country_ID
Dealer_NM
Dealer_CNTCT
Dimension Table
Branch Dim
Branch _ID
Name
Address
Country
Fact Table
Revenue
Dealer_ID
Model_ID
Branch_ID
Date_ID
Units_Sold
Revenue
Dimension Table
Date Dim
Date_ID
Year
Month
Quarter
Date
Dimension Table
Product
Product_ID
Product_Name
Model_ID
Variant_ID
Dimension Table
Variant
Variant_ID
Variant_Name
Fuel type


Dimension Table
Dealer
Dealer_ID
Location_ID
Country_ID
Dealer_NM
Dealer_CNTCT
Fact Table
Revenue
Dealer_ID
Branch_ID
Date_ID
Units_Sold
Revenue
Dimension Table
Branch Dim
Branch _ID
Name
Address
Country
Dimension Table
Date Dim
Date_ID
Year
Month
Quarter
Date
Fact Table
Product
Product_ID
Product_Name
Variant_ID
Dimension Table
Product
Product_ID
Product_Name
Model_ID
Variant_ID

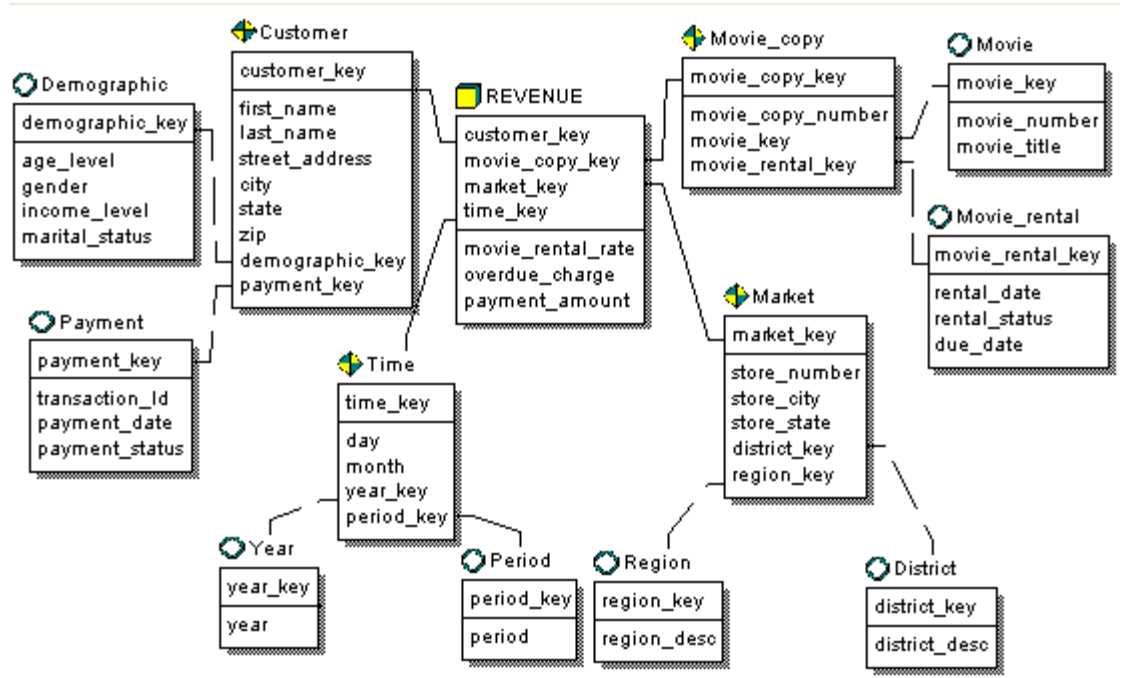

CUSTOMER
CUSTOMER_ID
FIRST_NAME
MIDDLE_INITIAL
LAST_NAME
DATE_OF_BIRTH
MAIDEN_NAME_MOTHER
GENDER_LOOKUP
CITY_NAME
ADDRESS1
ADDRESS2
POSTAL_CODE
EMAIL_ADDRESS
TELEPHONE_NUMBER
PASSWORD_HASH
WEBSITE_ID
BALANCE
DATE_UNREGISTERED
TIMESTAMP_ORDERED
REGION_ID
WEBSITE
WEBSITE_ID
WEBSITE_TITLE
WEBSITE_URL
COUNTRY
COUNTRY_ID
COUNTRY_NAME
ISO3166_COUNTRY_CODE
REGION
REGION_ID
COUNTRY_ID
REGION_NAME
ISO3166_REGION_CODE
TRANSACTIONS
CUSTOMER_ID
DVD_RELEASE_ID
DATE_ID
RENTAL_PRICE
PURCHASE_PRICE
SHIPPING_COST
DVD_RELEASE
DVD_RELEASE_ID
TITLE
STUDIO
RELEASED
STATUS
SOUND
VERSIONS
PRICE
RATING
YEAR
GENRE
ASPECT
UPC
RELEASE_DATE
TIMESTAMP
DATE
DATE_ID
DATE
MONTH
QUARTER
YEAR
DAY
WEEK
HOLIDAY

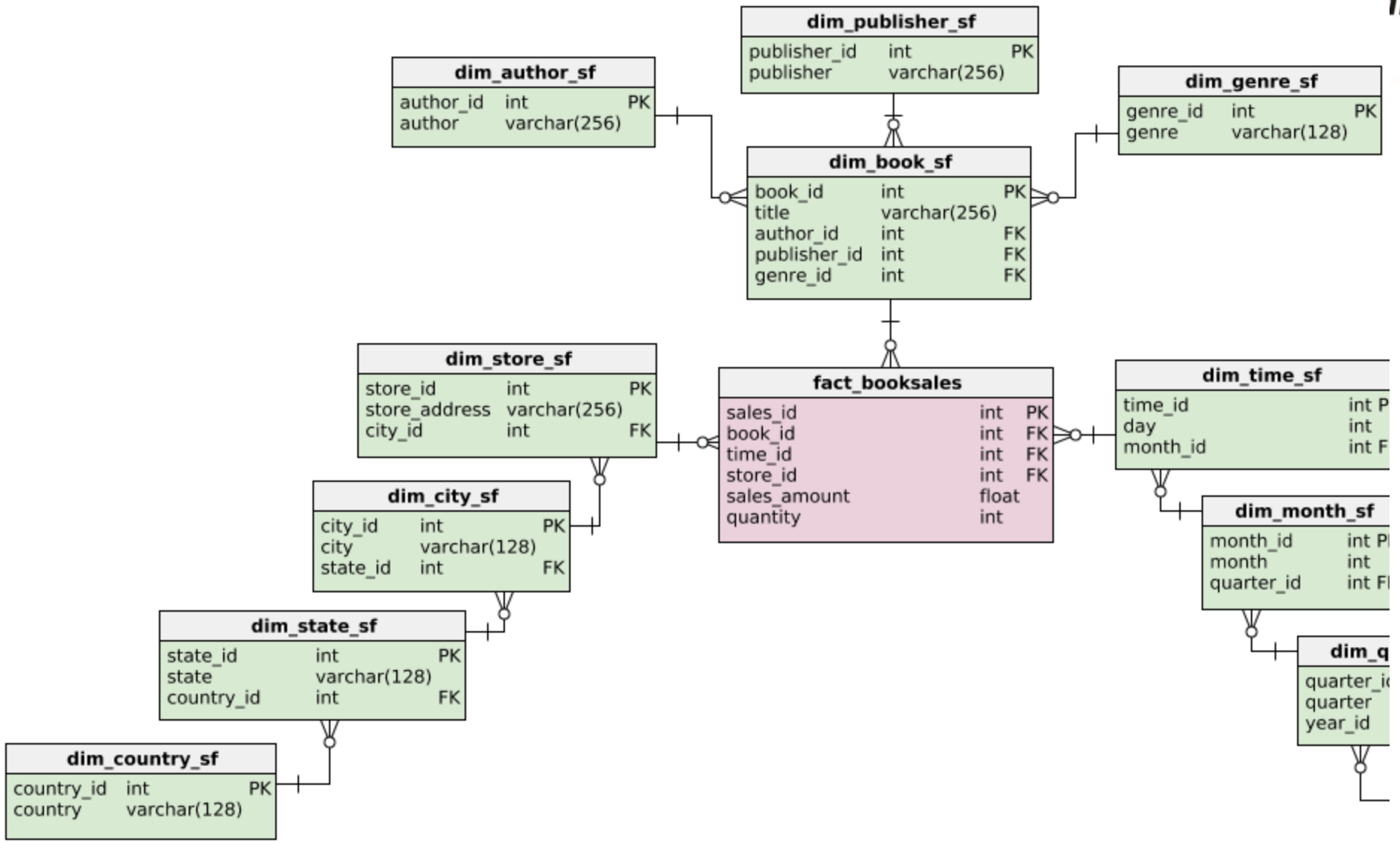

Dim_Day_of_Week
Id
Day_of_Week
Dim_Month
Id
Month_Name
Dim_Date
Id
Date
Day
Day_of_Week_Id
Month_Id
Quarter_Id
Year
Dim_Quarter
Id
Quarter_Name
Fact_Sales
Date_Id
Store_Id
Product_Id
Units_Sold
Dim_Store
Id
Store_Number
Geography_Id
Dim_Geography
Id
State_Province
Country
Dim_Product
Id
EAN_Code
Product_Name
Brand_Id
Product_Category_Id
Dim_Brand
Id
Brand
Dim_Product_Category
Id
Product_Category

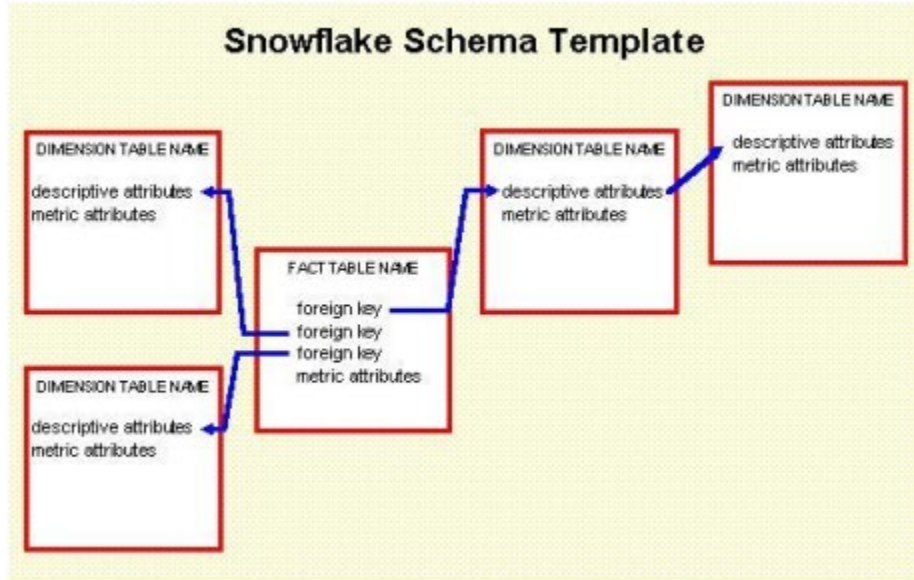

dim_product_details
rproduct_id
product_name
dim_product_cateogry
roduct_category_id
product_category_name
product_category_desctiption
dim_product_size
product_size_id
size_uk
size_us
size_rest
dim_product
product_id
product_size_id
product_color_id
product_category_id
dim_product_color
product_color_id
product_color_name
product_color_hex
product_color_rgb
dim_customer_city
customer_address_city_id
customer_address_city_name
dim_customer
customer_id
customer_address
customer_address_city_id
customer_address_country_id
dim_customer_country
ccustomer_address_country_id
customer_address_country_name
fact_sales
order_id
customer_id
product_id
purchase_date
store_id
purchase_amount
dim_store_city
store_city_id
store_city_name
dim_store
store_id
store_city_id
store_country_id
store_region_id
dim_store_country
store_countr_id
store_country_name
dim_store_region
store_region_id
store_region_name
dim_date
purchase_date
purchase_month_id
purchase_quarter_id
purchase_year
dim_month
purchase_month_id
purchase_month_name_short
purchase_month_name_long
purchase_month_number
dim_quarter
purchase_quarter_id
purchase_quarter_name_short
purchase_quarter_name_long
purchase_quarter_number



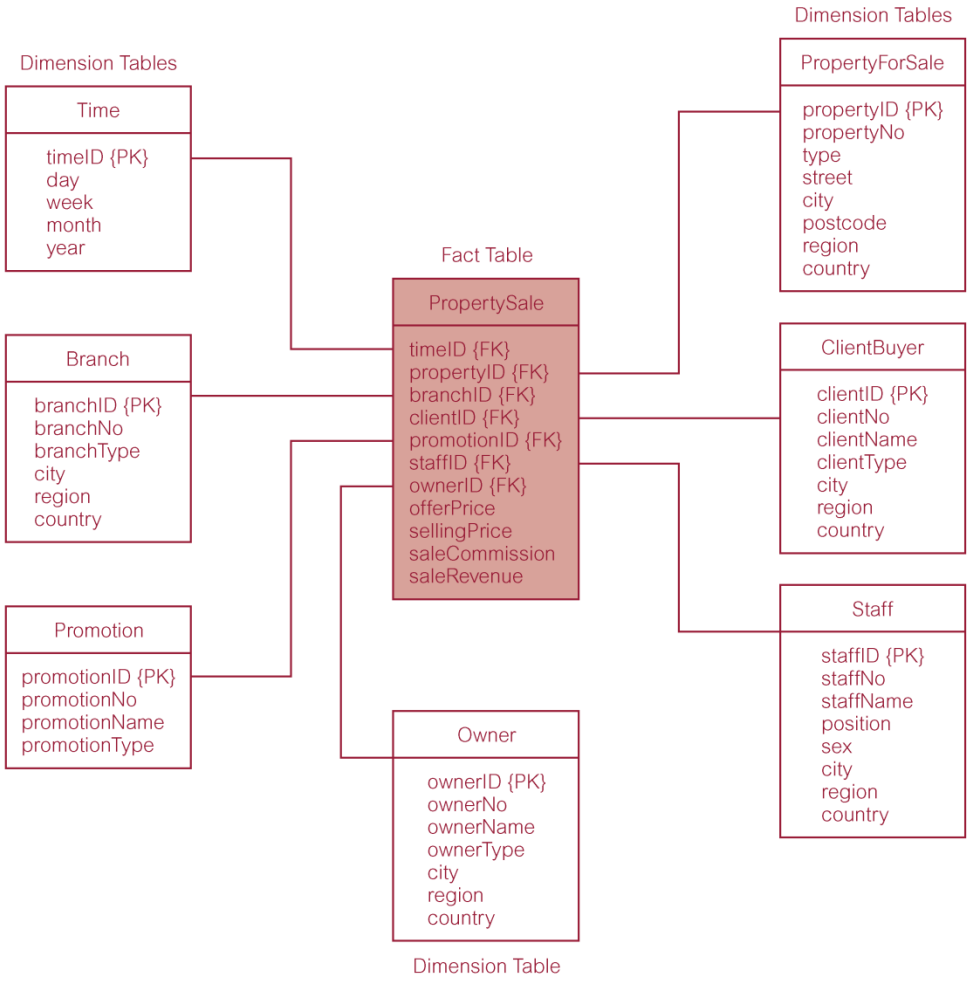
Snowflake Schema


Customer
customer_key
first_name
last_name
street_address
city
state
zip
demographic_key
payment_key
Demographic
demographic_key
age_level
gender
income_level
marital_status
REVENUE
customer_key
movie_copy_key
market_key
time_key
movie_rental_rate
overdue_charge
payment_amount
Movie_copy
movie_copy_key
movie_copy_number
movie_key
movie_rental_key
Movie
movie_key
movie_number
movie_title
Movie_rental
movie_rental_key
rental_date
rental_status
due_date
Payment
payment_key
transaction_Id
payment_date
payment_status
Time
time_key
day
month
year_key
period_key
Market
market_key
store_number
store_city
store_state
district_key
region_key
Year
year_key
year
Period
period_key
period
Region
region_key
region_desc
District
district_key
district_desc

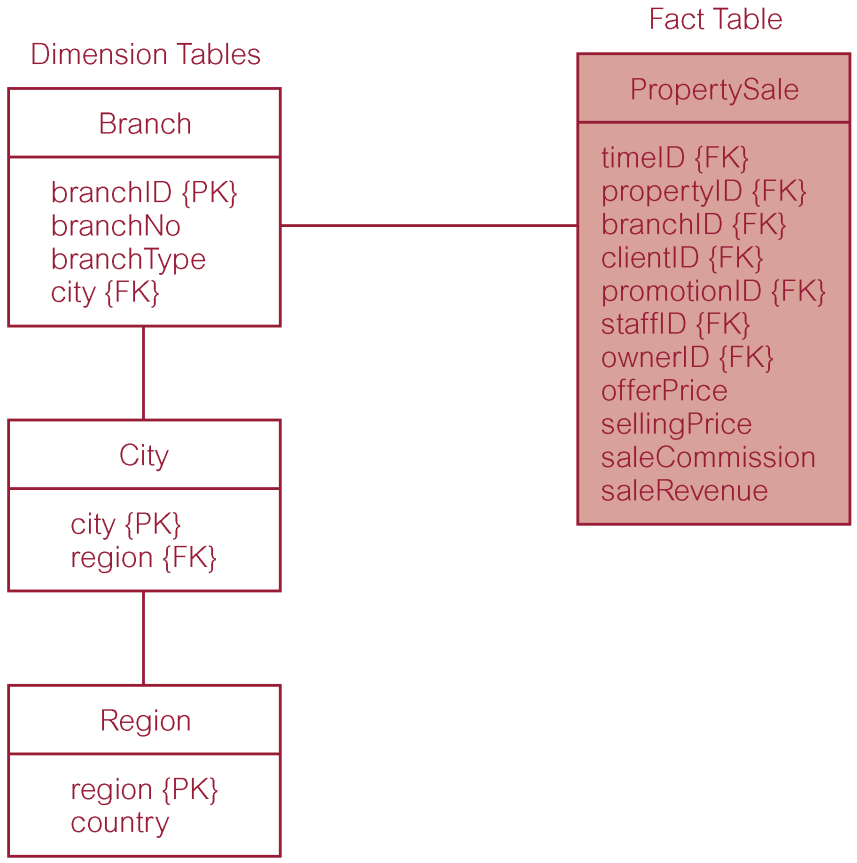

dim_publisher_sf
publisher_id int PK
publisher varchar(256)
dim_author_sf
author_id int PK
author varchar(256)
dim_genre_sf
genre_id int PK
genre varchar(128)
dim_book_sf
book_id int PK
title varchar(256)
author_id int FK
publisher_id int FK
genre_id int FK
dim_store_sf
store_id int PK
store_address varchar(256)
city_id int FK
fact_booksales
sales_id int PK
book_id int FK
time_id int FK
store_id int FK
sales_amount float
quantity int
dim_time_sf
time_id int P
day int
month_id int F
dim_city_sf
city_id int PK
city varchar(128)
state_id int FK
dim_month_sf
month_id int P
month int
quarter_id int F
dim_state_sf
state_id int PK
state varchar(128)
country_id int FK
dim_q
quarter_i
quarter
year_id
dim_country_sf
country_id int PK
country varchar(128)


Snowflake Schema Template
DIMENSION TABLE NAME
descriptive attributes
metric attributes
DIMENSION TABLE NAME
descriptive attributes
metric attributes
FACT TABLE NAME
foreign key
foreign key
foreign key
metric attributes
DIMENSION TABLE NAME
descriptive attributes
metric attributes
DIMENSION TABLE NAME
descriptive attributes
metric attributes

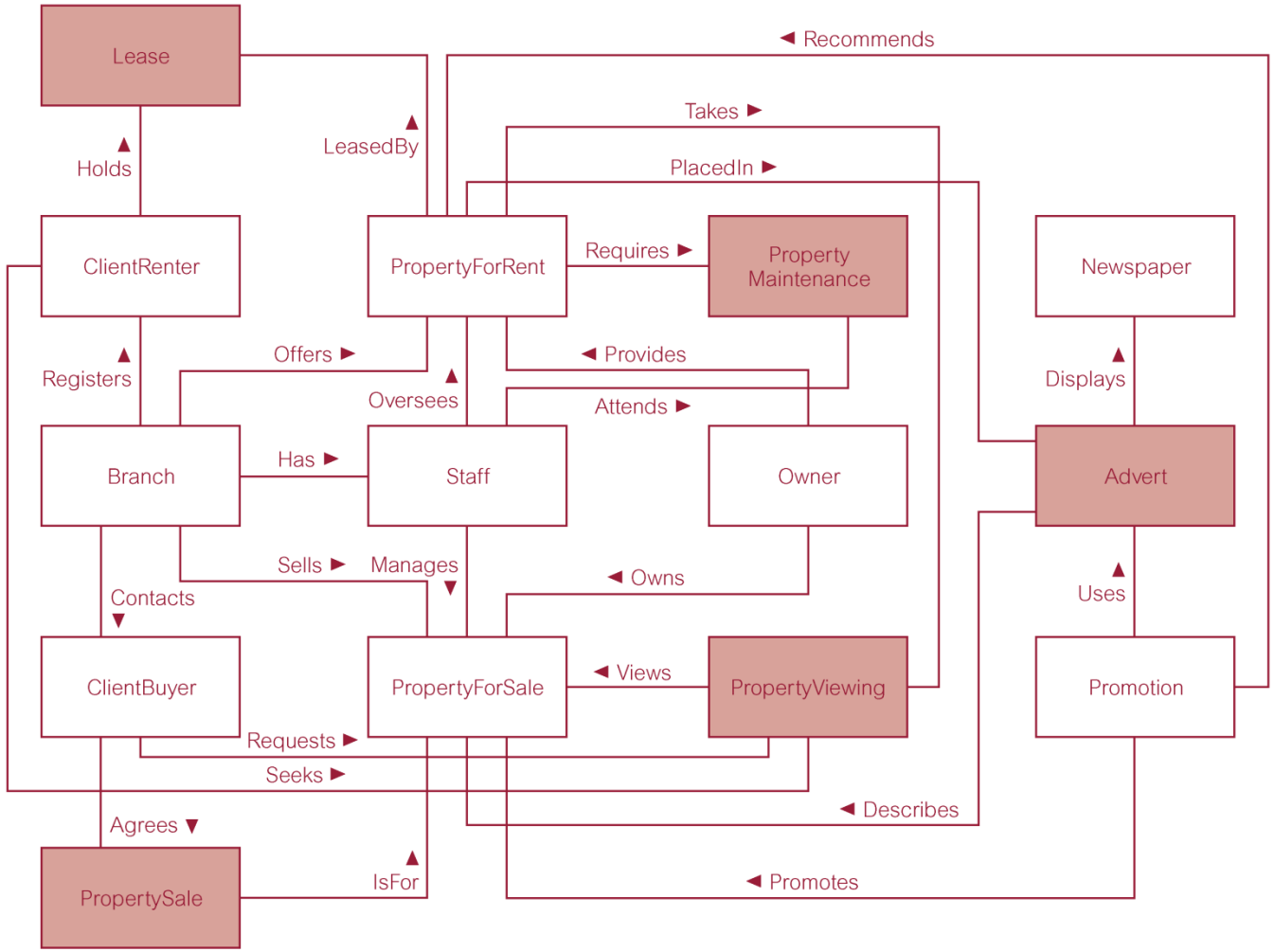
Star schema (dimensional model) for property sales of *DreamHome*



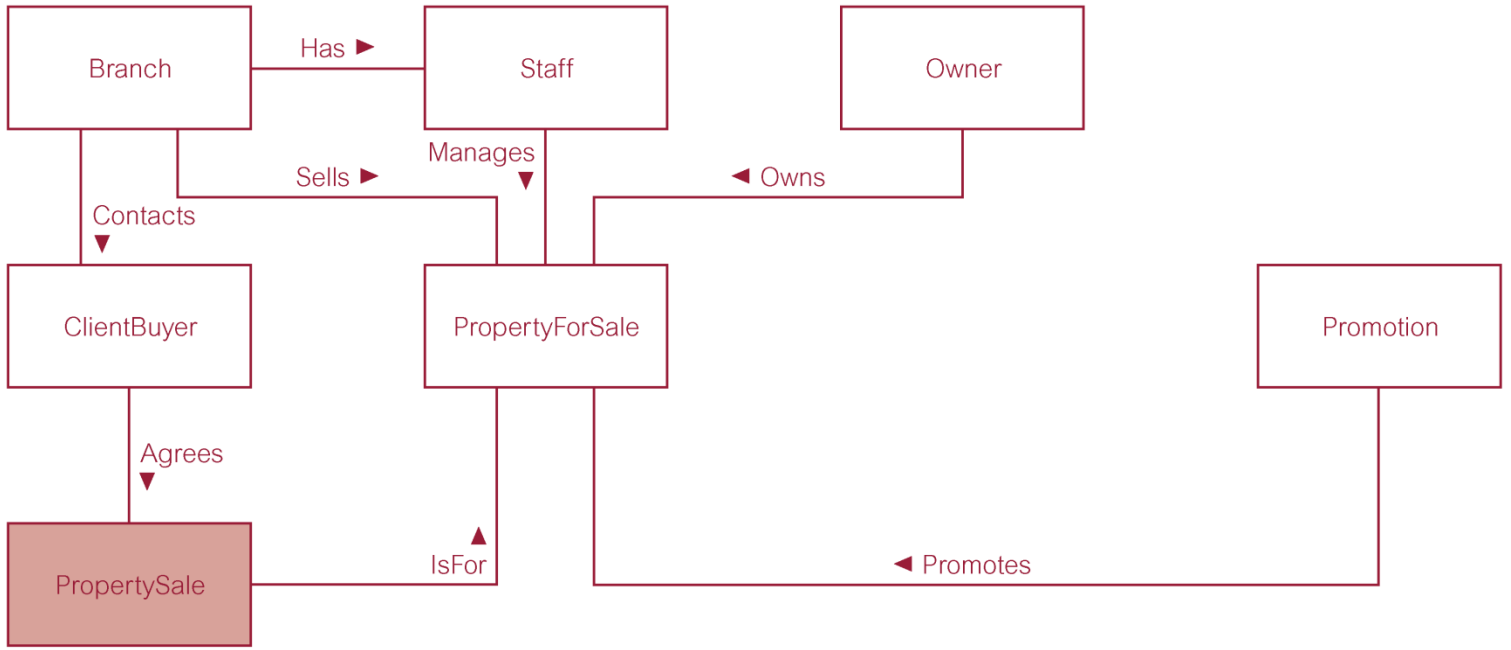
Property sales with normalized version of Branch dimension table



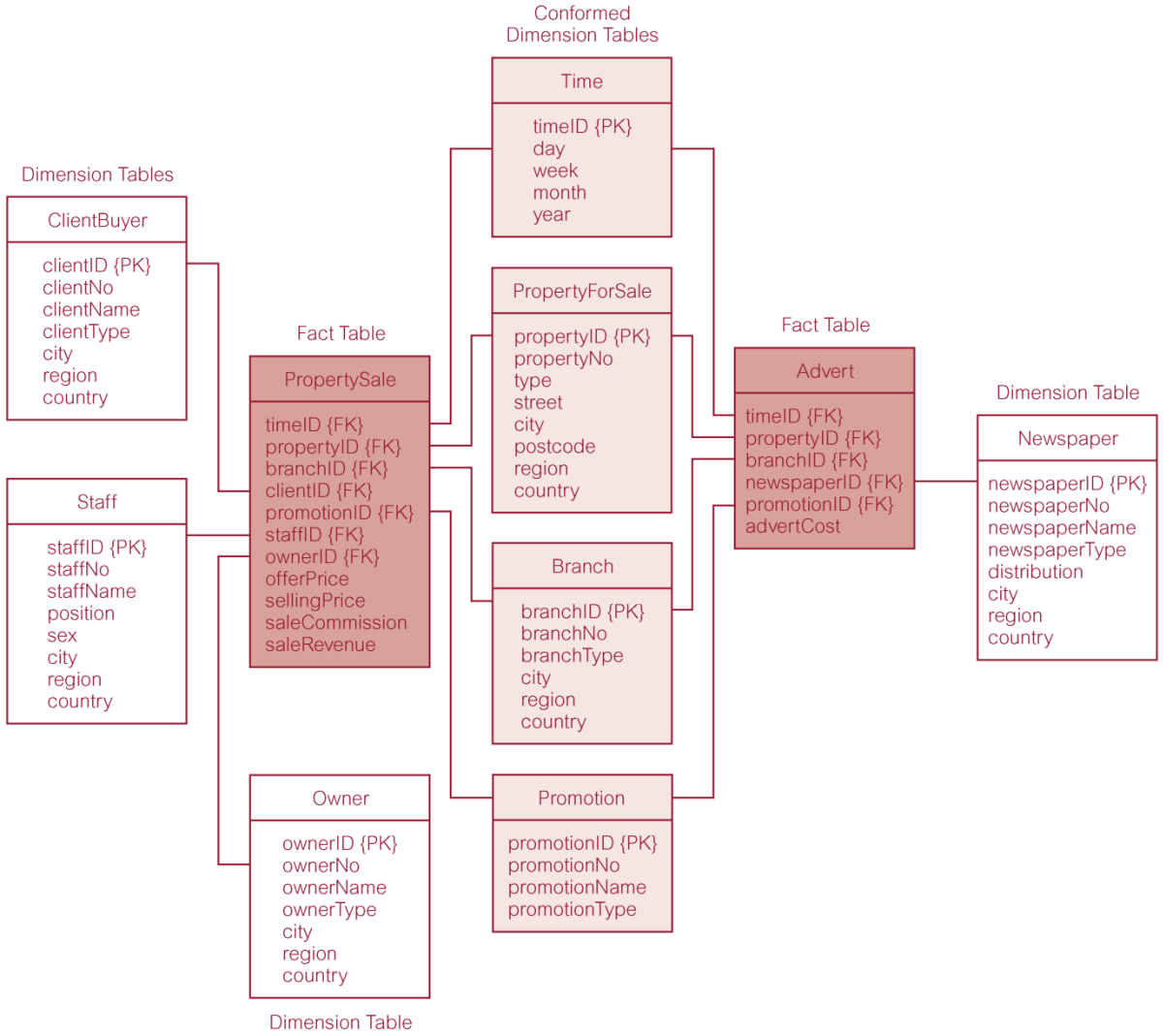
ER model of an extended version of *DreamHome*



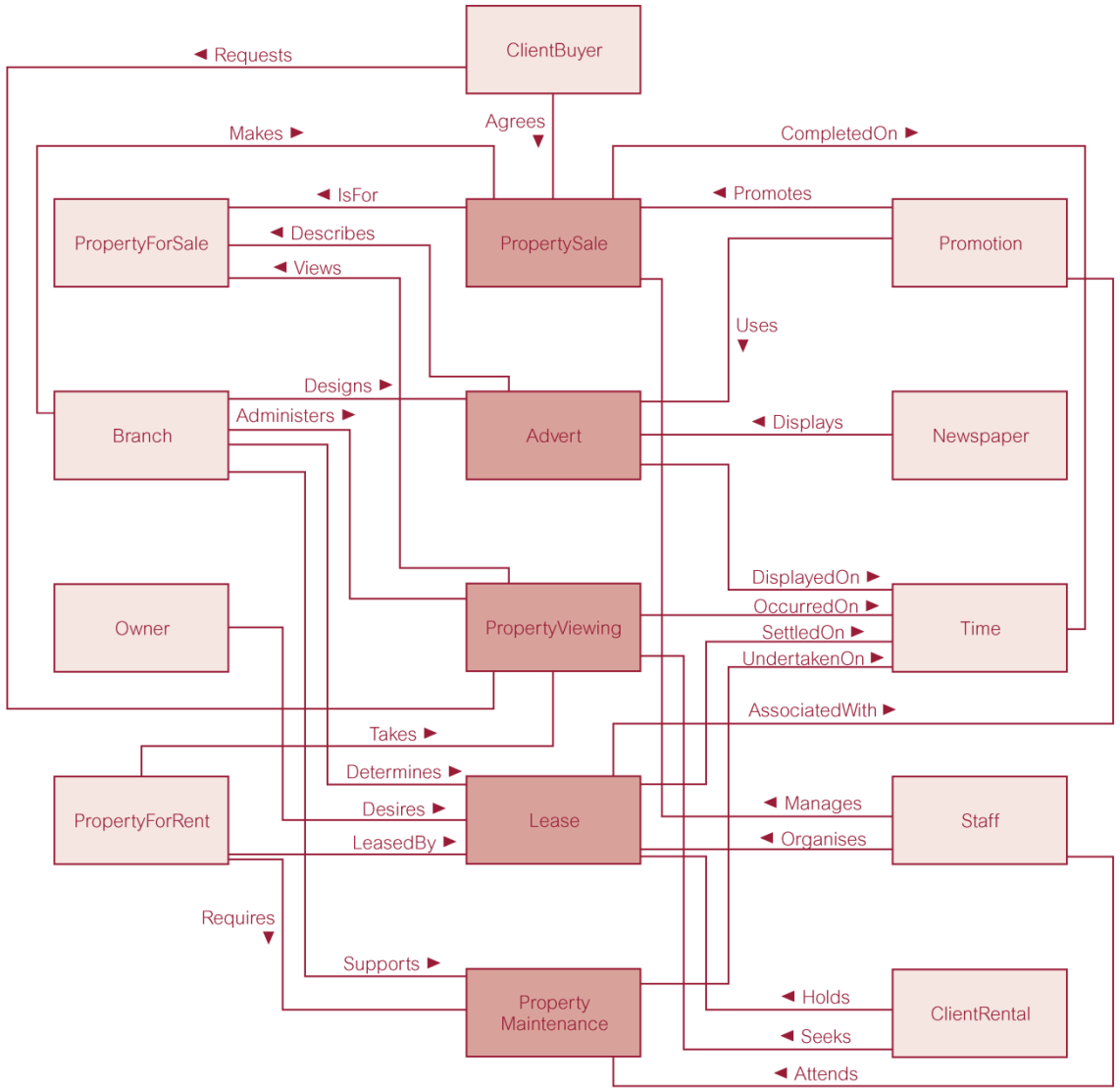
ER model of property sales business process of *DreamHome*



Star schemas for property sales and property advertising



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



Vergleich

Kriterium	Star-Schema	Snowflake-Schema
Struktur	Einfacher und denormalisierter. Ein Faktentabelle in der Mitte, umgeben von Dimensionstabellen.	Komplexer und normalisierter. Ein Faktentabelle in der Mitte, umgeben von Dimensionstabellen, die in kleinere Untertabellen aufgeteilt sind.
Performanz	In der Regel schneller, weil weniger Joins benötigt werden.	Kann langsamer sein, da mehr Joins benötigt werden, um auf die notwendigen Daten zuzugreifen.
Speicherplatz	Benötigt mehr Speicherplatz, da Daten denormalisiert sind und Duplikate existieren können.	Benötigt weniger Speicherplatz, da Daten normalisiert sind und Duplikate reduziert werden.
Einfachheit der Abfragen	Einfacher zu verstehen und zu schreiben, da weniger Joins benötigt werden.	Komplexer, da mehr Joins und eine tiefere Kenntnis der Struktur erforderlich sind.
Flexibilität	Geringer, da Änderungen an den Dimensionstabellen die Faktentabelle beeinflussen können.	Höher, da Änderungen an den Untertabellen die Haupttabellen weniger wahrscheinlich beeinflussen.
Datenaktualisierung	Einfacher, da die Daten weniger normalisiert sind.	Komplexer, da die Daten stärker normalisiert sind.

Vergleich

https://youtu.be/Jir31Ti8jps

Fragen?