CURRENT TOPICS IN COMPUTER SCIENCE

Business Intelligence Systems and Analytics
DATA WAREHOUSE DESIGN

Trong Nhan Phan, PhD

OUTLINE

- Dimensional data modeling
- Slowly changing dimensions
- Data warehouse and tools
- Data warehouse design lab
- Summary
- References

DATA MODELING

DWH ARCHITECTURE

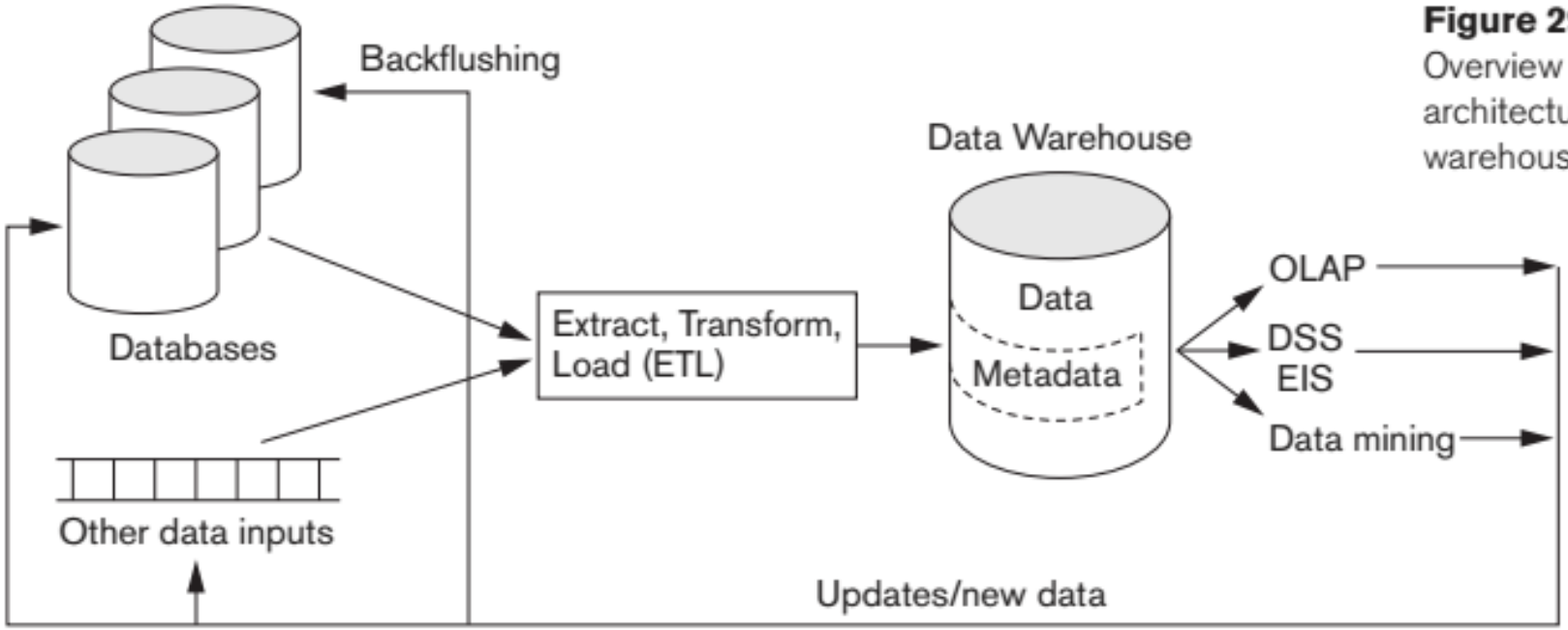


Figure 29.1
Overview of the general architecture of a data warehouse.

[5]

INTEGRATED DATA

- How to place data in data warehouses
- From the operational environment to data warehouse → data integration
  - Same data, different name
  - Different data, same name
  - Incosistent encoding (e.g., 0/1, m/f, male/female, cm, inch, bal, balance, curbal, balcurr)

IMPORTANCE TO MODELING

- A unified view of data
  - Consistency
  - Trust
  - Well-organized abstraction of data
  - Performance (data consolidation was done)
- E.g, date + interval, is_delete='0', delivery experience time by 1st drop-off

TRANSACTIONS VS. ANALYTICS

_id	Mã số sinh viên	Họ tên	Ngày tháng năm sinh	Email	Lớp
6225750748c598abc027bcbb	50501712	Nguyễn Văn A	02-01-86	50501712@hcmut.edu.vn	MT05KH01
6225750748c598abc027bcbc	50503491	Phan Trọng B	05-08-87	50503491@hcmut.edu.vn	MT05KH01
6225750748c598abc027bcbd	50502211	Trần Văn C	04-04-85	50502211@hcmut.edu.vn	MT05KH02

_id	classID	startdate
633fa3ade4411c0cc0774a5e	MT05KH01	01/01/2005
633fa3c8e4411c0cc0774a71	MT05KH02	01/02/2005

```
{
    "_id": ObjectId("62346e38d24cfe35b916477e"),
    "SSN": "123456",
    "Name": "Nguyen Van A",
    "Department": {
        "Dnumber": NumberInt("1"),
        "Dname": "Research",
        "MgrSSN": "456789"
    },
    "hobbies": [
        "football",
        "swimming",
        "chess"
    ]
}
```

DISCUSSION

NORMALIZATION VS. DENORMALIZATION

FOR INSTANCE

AgencyID	AgencyName	ProductID	ProductName	ProductPrice	Quantity	Date
1	A	101	Beauty Soap	7	120	01/01/2022
1	A	102	Tooth Brush	5	100	01/01/2022
2	B	103	Tooth Paste	4	80	01/02/2022
3	C	103	Tooth Paste	4	110	01/02/2022
…	…	…	…	…	…	…

StudentID	StudentName
1001	NVA
1002	NVB
…	…

StudentID	Course
1001	Database Systems
1001	E-commerce
1002	E-commerce
…	…

IMPORTANCE TO MODELING

- Operational vs. analytical modeling
  - ER modeling vs. Dimensional modeling

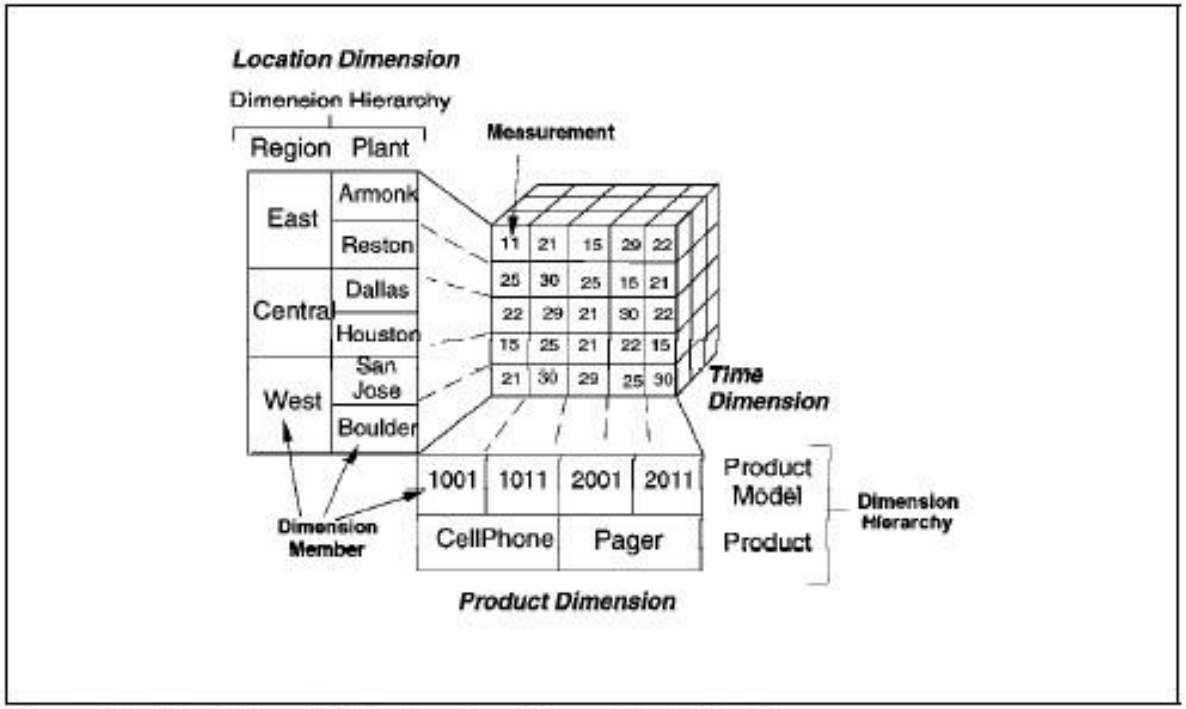
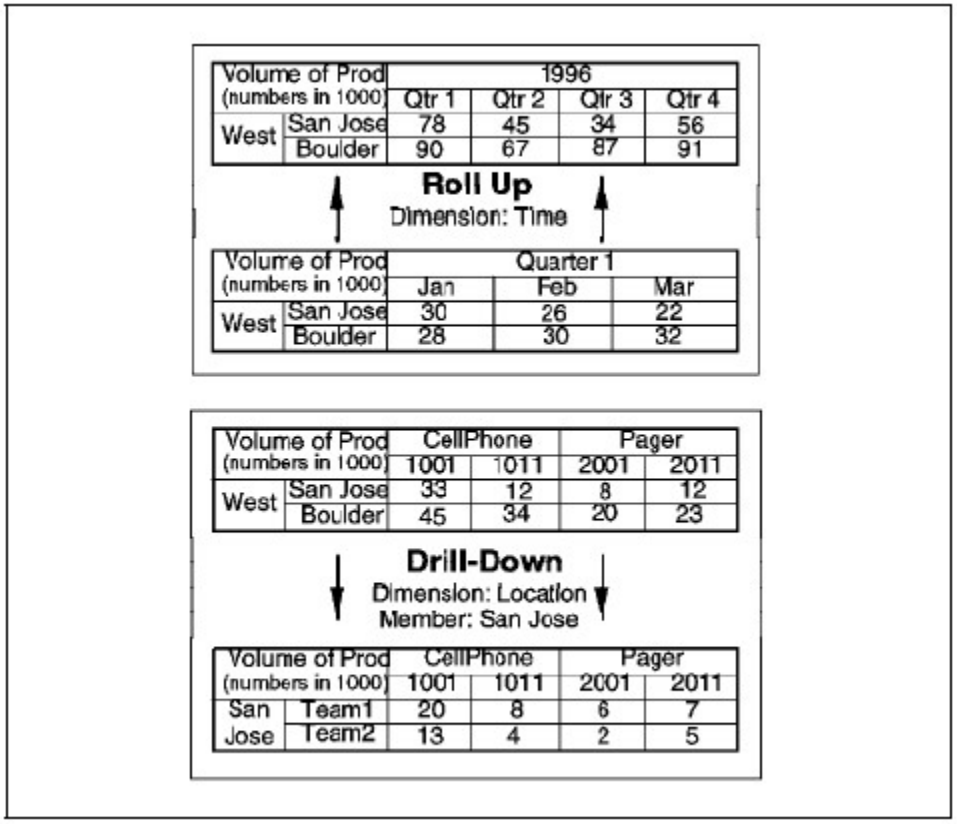


Figure 15. The Cube: A Metaphor for a Dimensional Model

https://www.empiredatasystems.com/data-warehouse-concept.html

IMPORTANCE TO MODELING

- Operational vs. analytical modeling
  - ER modeling vs. Dimensional modeling
    - Drill down vs. Roll up



Volume of Prod (numbers in 1000)		1996			
		Qtr 1	Qtr 2	Qtr 3	Qtr 4
West	San Jose	78	45	34	56
	Boulder	90	67	87	91

Roll Up
Dimension: Time

Volume of Prod (numbers in 1000)		Quarter 1		
		Jan	Feb	Mar
West	San Jose	30	26	22
	Boulder	28	30	32

Volume of Prod (numbers in 1000)		CellPhone		Pager	
		1001	1011	2001	2011
West	San Jose	33	12	8	12
	Boulder	45	34	20	23

Drill-Down
Dimension: Location
Member: San Jose

Volume of Prod (numbers in 1000)		CellPhone		Pager	
		1001	1011	2001	2011
San Jose	Team1	20	8	6	7
	Team2	13	4	2	5

Figure 16. Example of Drill Down and Roll Up

https://www.empiredatasystems.com/data-warehouse-concept.html

IMPORTANCE TO MODELING

- Operational vs. analytical modeling
  - ER modeling vs. Dimensional modeling
    - Slice & Dice

Volume of Prod (numbers in 1000)		CellPhone		Pager	
		1001	1011	2001	2011
West	San Jose	33	12	8	12
	Boulder	45	34	20	23

Dice

Volume of Prod (numbers in 1000)		1996 (CellPhone & Pager)			
		Qtr 1	Qtr 2	Qtr 3	Qtr 4
West	San Jose	78	45	34	56
	Boulder	90	67	87	91

Slice

Volume of Prod (numbers in 1000)		1996 (CellPhone Only)			
		Qtr 1	Qtr 2	Qtr 3	Qtr 4
West	San Jose	53	35	20	48
	Boulder	76	57	40	80

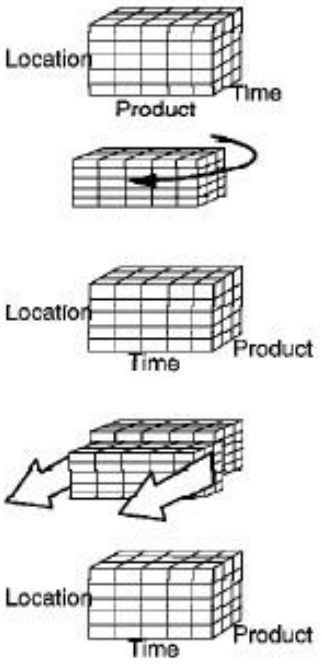


Figure 17. Example of Slice and Dice

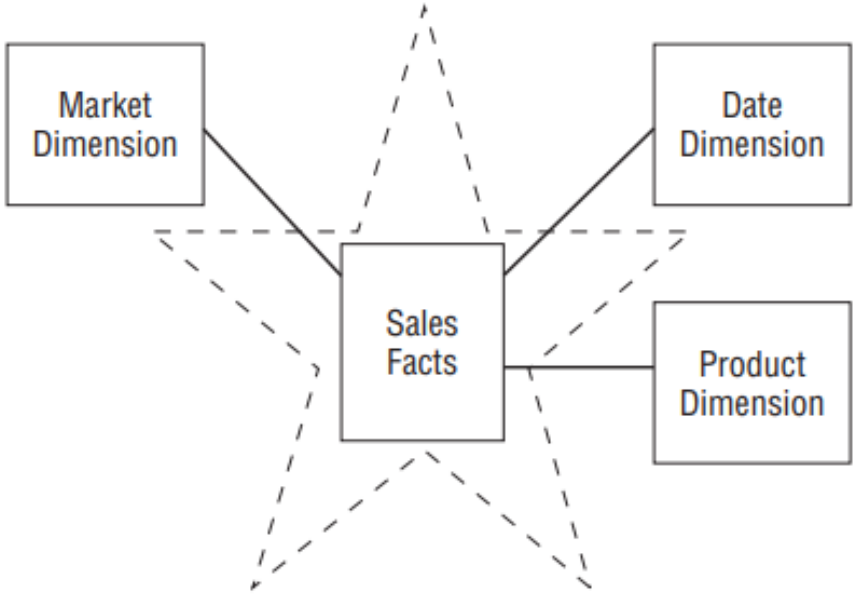
https://www.empiredatasystems.com/data-warehouse-concept.html

IMPORTANCE TO MODELING

- Operational vs. analytical modeling
  - ER modeling vs. Dimensional modeling
- Dimensional modeling is widely accepted
  - Deliver data that's understandable to the business users.
  - Deliver fast query performance.

DIMENSIONAL MODELING

- Dimensional models implemented in relational database management systems are referred to as star schemas.
- Dimensional models implemented in multidimensional database environments are referred to as online analytical processing (OLAP) cubes.



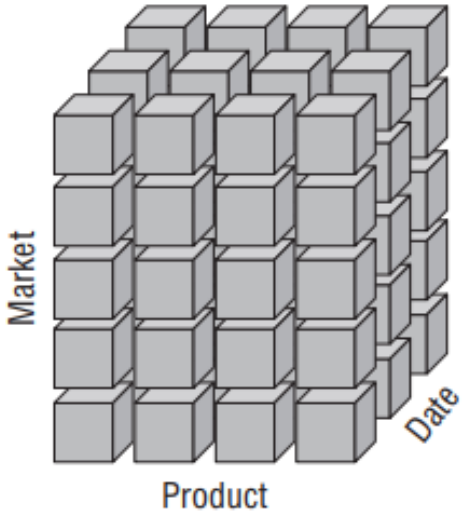


Figure 1-1: Star schema versus OLAP cube.

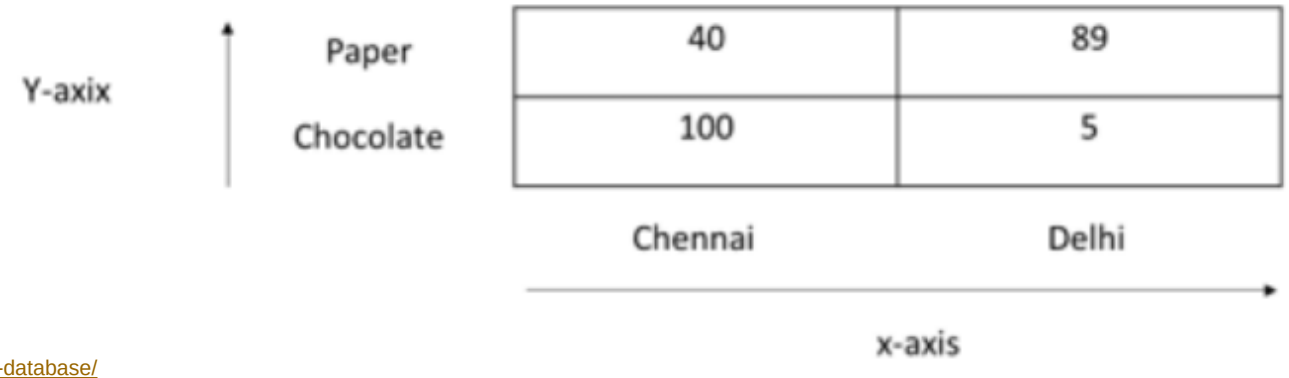
[6]

DISCUSSION

MULTIDIMENSIONAL DATABASES

TWO DIMENSIONS

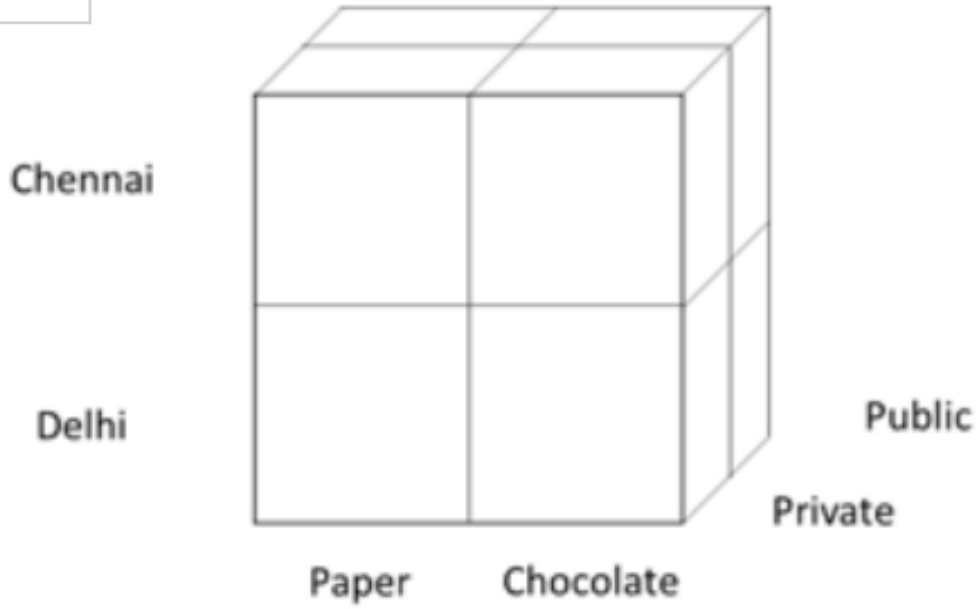
Item	Store Location	Quantity
Paper, A4	Chennai	40
Chocolate, Munch	Delhi	5
Paper, A3	Delhi	89
Chocolate, 5Star	Chennai	100



https://www.educba.com/multidimensional-database/

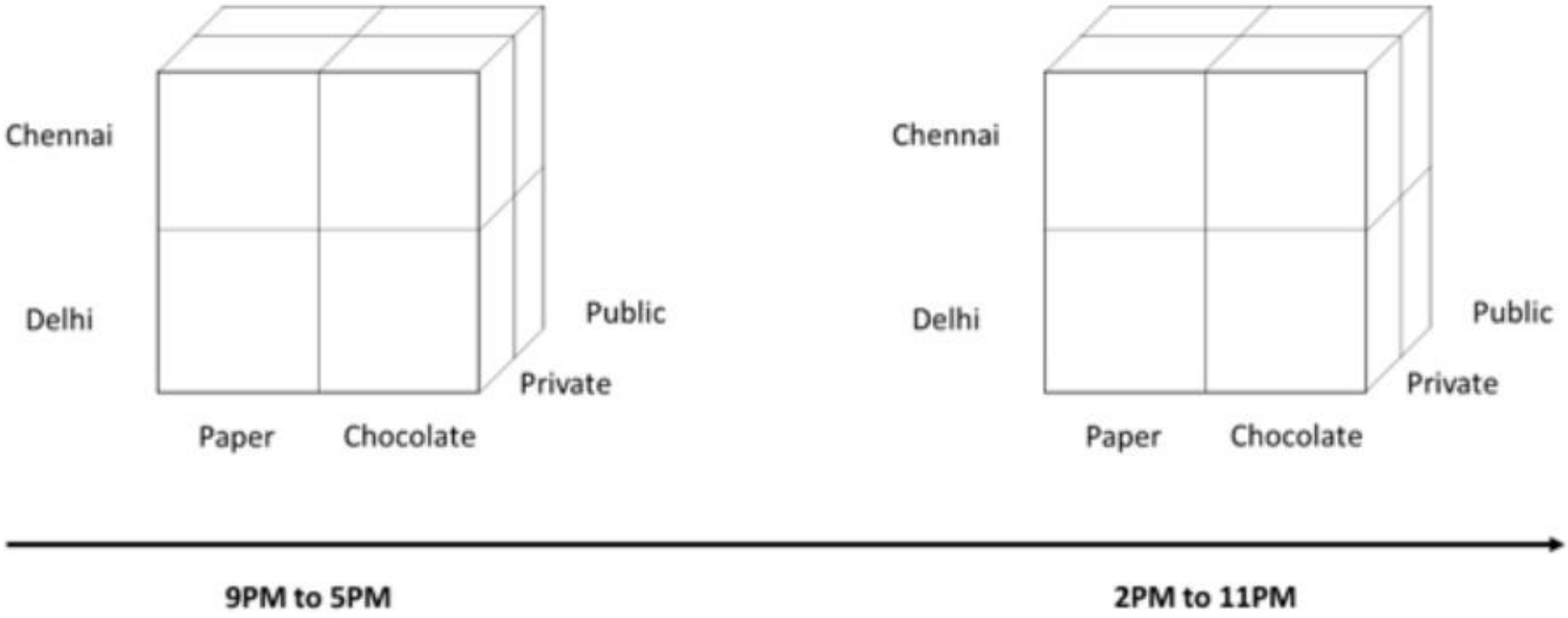
THREE DIMENSIONS

Item	Store Location	Customer	Quantity
Paper, A4	Chennai	Public	40
Chocolate, Munch	Delhi	Private	5
Paper, A3	Delhi	Public	89
Chocolate, 5Star	Chennai	Private	100



https://www.educba.com/multidimensional-database/

FOUR DIMENSIONS



https://www.educba.com/multidimensional-database/

DISCUSSION

DATA WAREHOUSE DESIGN APPROACH

Top-down (Bill Inmon)

Bottom-up (Ralph Kimball)

DIMENSIONAL MODELING

- Star model
- Snowflake model
- Variants such as constellation model, multistar model

DIMENSIONAL MODELING

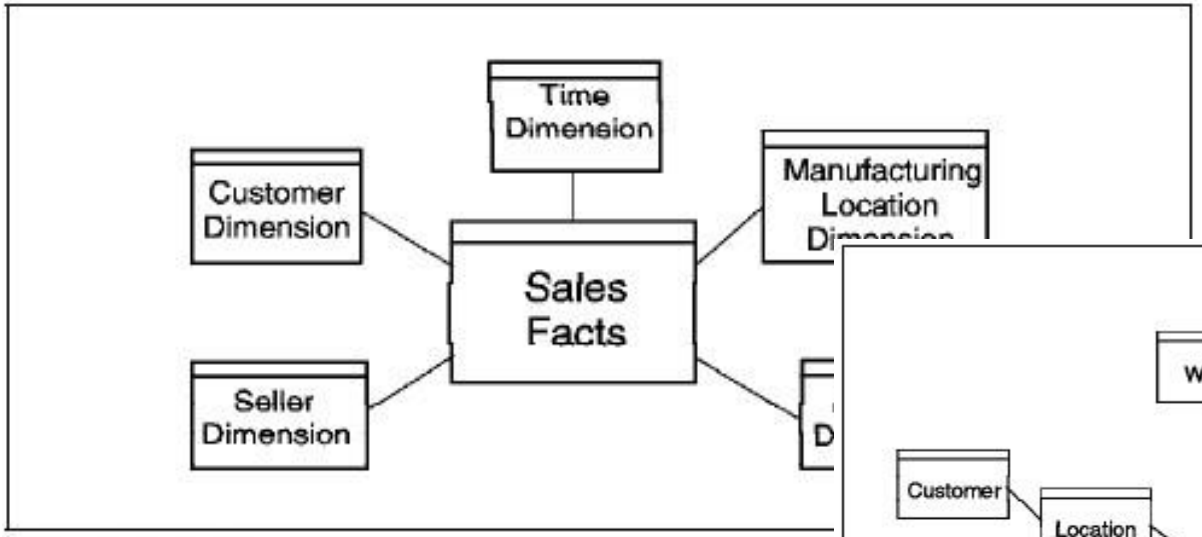


Figure 18. Star Model.

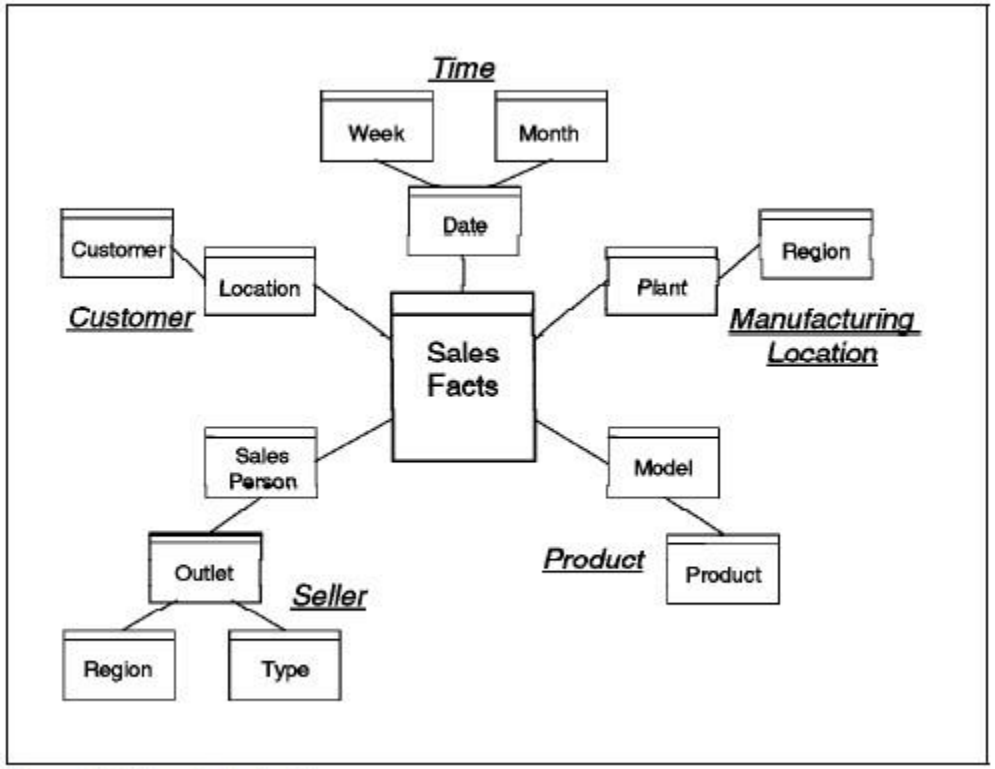


Figure 19. Snowflake Model

https://www.empiredatasystems.com/data-warehouse-concept.html

FACT TABLE

- The fact table in a dimensional model stores the performance measurements.
- Composite key and referential integrity
- The most useful facts are numeric and additive.
  - Additive vs. semi-additive vs. non-additive

Retail Sales Facts
Date Key (FK)
Product Key (FK)
Store Key (FK)
Promotion Key (FK)
Customer Key (FK)
Clerk Key (FK)
Transaction #
Sales Dollars
Sales Units

[6]

DIMENSION TABLE

- The dimension tables contain the textual context associated with a business process measurement event.
- Dimension tables often represent hierarchical relationships.

Product Key	Product Description	Brand Name	Category Name
1	PowerAll 20 oz	PowerClean	All Purpose Cleaner
2	PowerAll 32 oz	PowerClean	All Purpose Cleaner
3	PowerAll 48 oz	PowerClean	All Purpose Cleaner
4	PowerAll 64 oz	PowerClean	All Purpose Cleaner
5	ZipAll 20 oz	Zippy	All Purpose Cleaner
6	ZipAll 32 oz	Zippy	All Purpose Cleaner
7	ZipAll 48 oz	Zippy	All Purpose Cleaner
8	Shiny 20 oz	Clean Fast	Glass Cleaner
9	Shiny 32 oz	Clean Fast	Glass Cleaner
10	ZipGlass 20 oz	Zippy	Glass Cleaner
11	ZipGlass 32 oz	Zippy	Glass Cleaner

Product Dimension
Product Key (PK)
SKU Number (Natural Key)
Product Description
Brand Name
Category Name
Department Name
Package Type
Package Size
Abrasive Indicator
Weight
Weight Unit of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
...

[6]

A SIMPLE REPORT

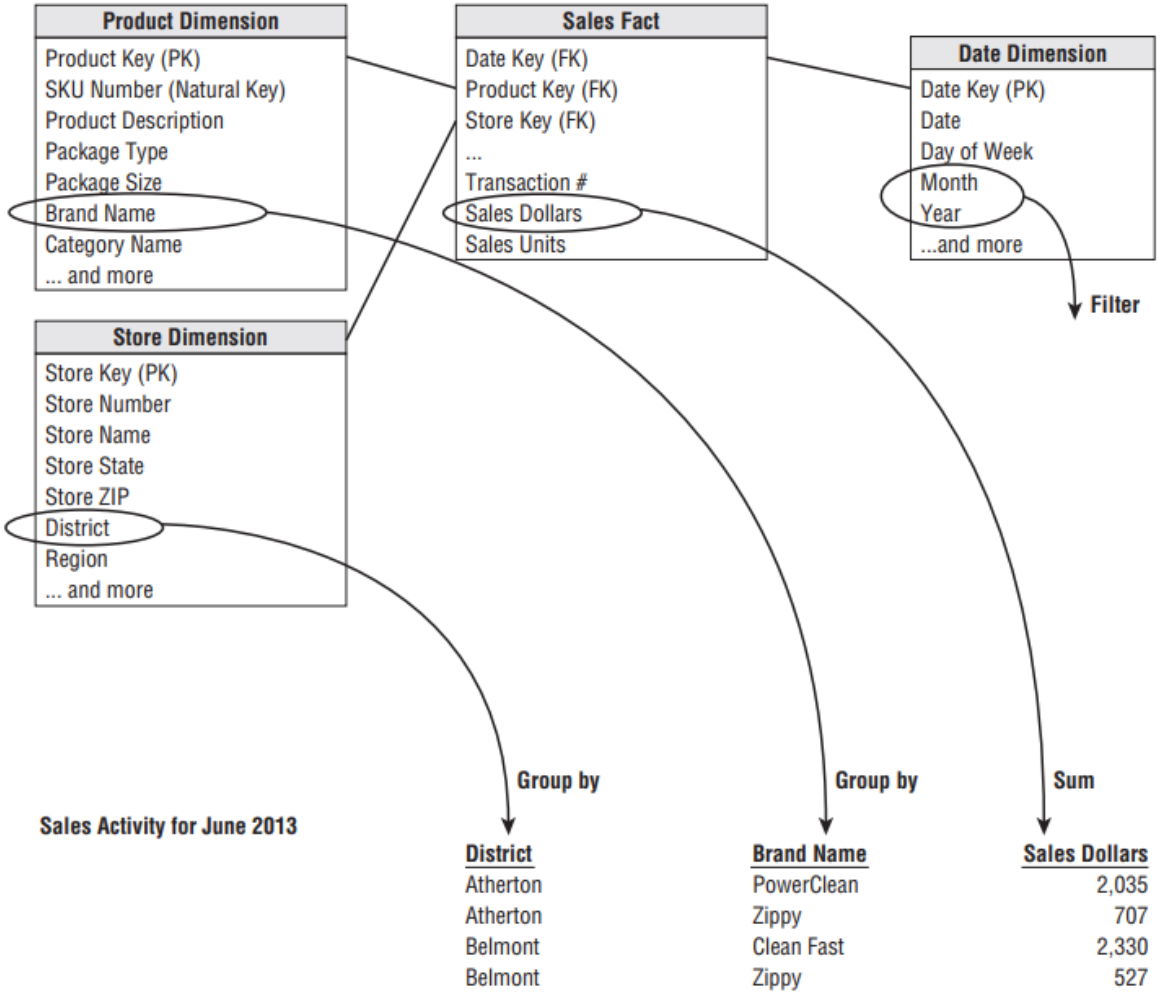


Figure 1-6: Dimensional attributes and facts form a simple report.

[6]

THE QUERY BEHIND

```
SELECT
     store.district_name,
     product.brand,
     sum(sales_facts.sales_dollars) AS "Sales Dollars"
FROM
     store,
     product,
     date,
     sales_facts
WHERE
     date.month_name="June" AND
     date.year=2013 AND
     store.store_key = sales_facts.store_key AND
     product.product_key = sales_facts.product_key AND
     date.date_key = sales_facts.date_key
GROUP BY
     store.district_name,
     product.brand
```

[6]

DIMENSIONAL DESIGN PROCESS

for bottom up approach in building a database

1. Select the business process.
2. Declare the grain.
3. Identify the dimensions.
4. Identify the facts.

Collaborative Dimensional Modeling

THE MATRIX

- The matrix plan
  - Vertical: first-level data marts
  - Horizontal: dimensions

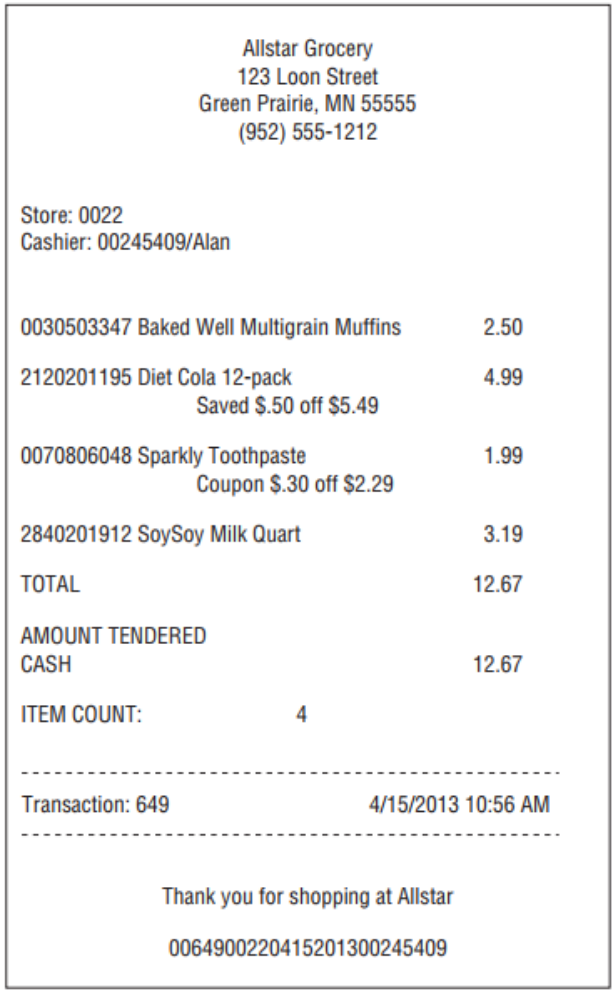
Business Process / Event	Time	Customer	Service	Rate Category	Local Svc Provider	Calling Party	Called Party	Long Dist Provider	Internal Organization	Employee	Location	Equipment Type	Supplier	Item Shipped	Account Status
Customer Billing	X	X	X	X	X			X			X				X
Service Orders	X	X	X		X			X	X	X	X	X			X
Trouble Reports	X	X	X		X	X		X	X	X	X	X	X	X	X
Yellow Page Ads	X	X		X		X			X	X	X				X
Customer Inquiries	X	X	X	X	X	X		X	X	X	X				X
Promotions & Communication	X	X	X	X	X	X		X	X	X	X	X	X	X	X
Billing Call Detail	X	X	X	X	X	X	X	X	X		X	X	X	X	X
Network Call Detail	X	X	X	X	X	X	X	X	X		X	X	X	X	X
Customer Inventory	X	X	X	X	X			X	X		X	X	X	X	X
Network Inventory	X		X						X	X	X	X	X	X	
Real Estate	X								X	X	X	X			
Labor & Payroll	X								X	X	X				
Computer Charges	X	X	X		X			X	X	X	X	X	X	X	
Purchase Orders	X								X	X	X	X	X	X	
Supplier Deliveries	X								X	X	X	X	X	X	

https://www.kimballgroup.com/1999/12/the-matrix/

CASE STUDY – RETAIL SALES

- Select a business process
  - Customer purchase in POS

which products are selling
in which stores
on which days
under what promotional conditions
in which transactions



Allstar Grocery
123 Loon Street
Green Prairie, MN 55555
(952) 555-1212

Store: 0022
Cashier: 00245409/Alan

0030503347 Baked Well Multigrain Muffins	2.50
2120201195 Diet Cola 12-pack Saved $.50 off $5.49	4.99
0070806048 Sparkly Toothpaste Coupon $.30 off $2.29	1.99
2840201912 SoySoy Milk Quart	3.19
TOTAL	12.67
AMOUNT TENDERED CASH	12.67
ITEM COUNT: 4	

Transaction: 649 4/15/2013 10:56 AM

Thank you for shopping at Allstar

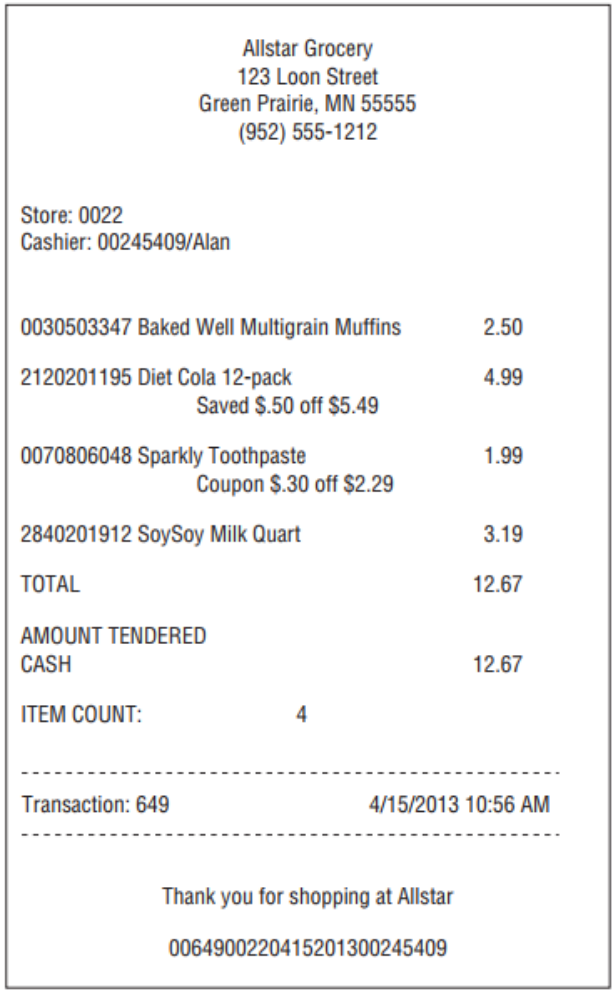
00649002204152013002454O9

Figure 3-2: Sample cash register receipt.

[6]

CASE STUDY – RETAIL SALES

- Declare the grain
  - Sales for
    - A given product
    - within a shopping cart
    - into a single line item



Allstar Grocery
123 Loon Street
Green Prairie, MN 55555
(952) 555-1212

Store: 0022
Cashier: 00245409/Alan

0030503347 Baked Well Multigrain Muffins	2.50
2120201195 Diet Cola 12-pack Saved $.50 off $5.49	4.99
0070806048 Sparkly Toothpaste Coupon $.30 off $2.29	1.99
2840201912 SoySoy Milk Quart	3.19
TOTAL	12.67
AMOUNT TENDERED CASH	12.67
ITEM COUNT: 4	

Transaction: 649 4/15/2013 10:56 AM

Thank you for shopping at Allstar

00649002204152013002454O9

Figure 3-2: Sample cash register receipt.

[6]

CASE STUDY – RETAIL SALES

- Identify the dimensions
  - By date
  - By store
  - By promotion
  - By cashier
  - By payment method
  - Etc.

Allstar Grocery
123 Loon Street
Green Prairie, MN 55555
(952) 555-1212

Store: 0022
Cashier: 00245409/Alan

0030503347 Baked Well Multigrain Muffins	2.50
2120201195 Diet Cola 12-pack Saved $.50 off $5.49	4.99
0070806048 Sparkly Toothpaste Coupon $.30 off $2.29	1.99
2840201912 SoySoy Milk Quart	3.19
TOTAL	12.67
AMOUNT TENDERED CASH	12.67
ITEM COUNT: 4	

Transaction: 649 4/15/2013 10:56 AM

Thank you for shopping at Allstar

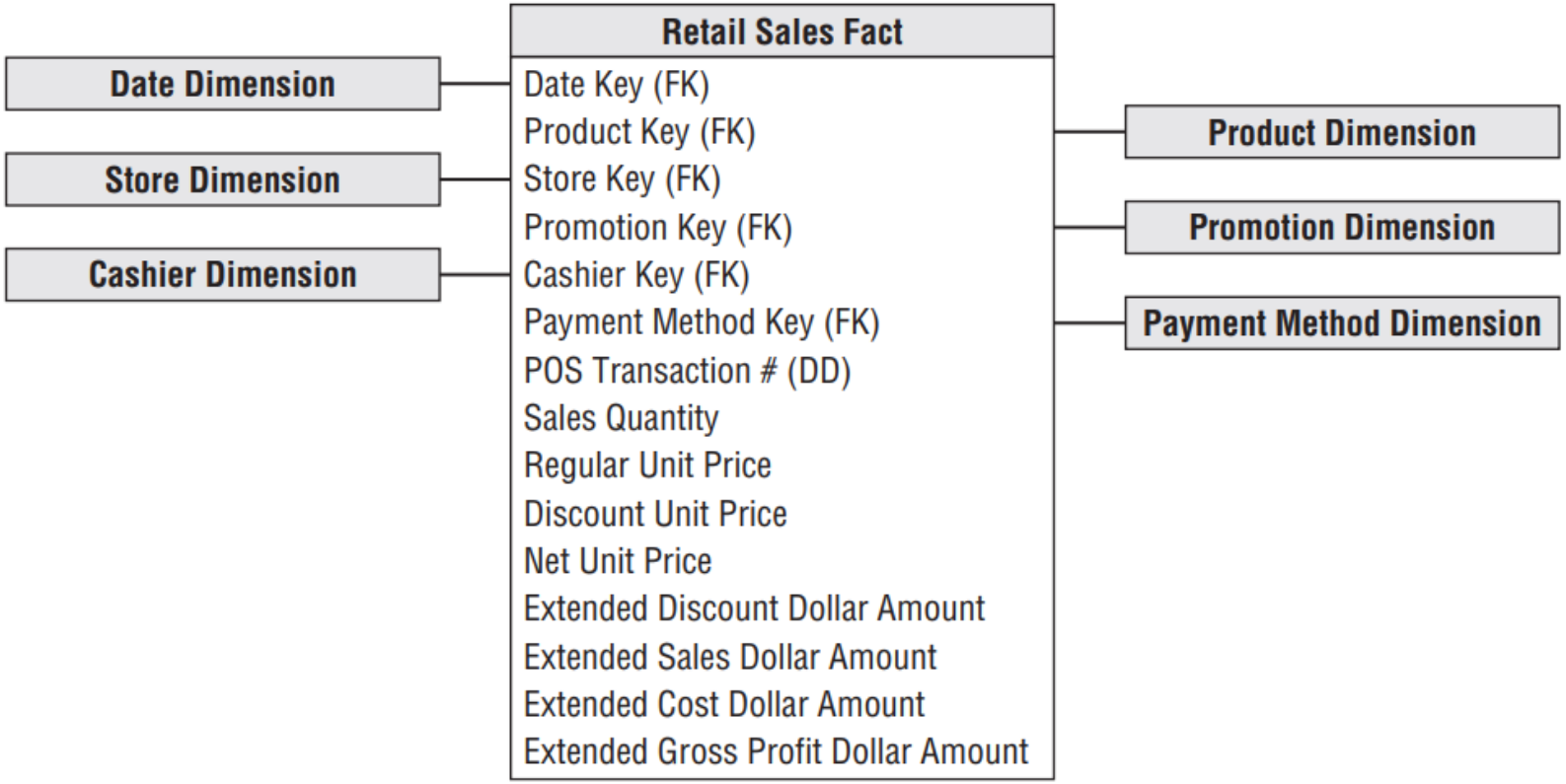
00649002204152013002454O9

Figure 3-2: Sample cash register receipt.

[6]

CASE STUDY – RETAIL SALES

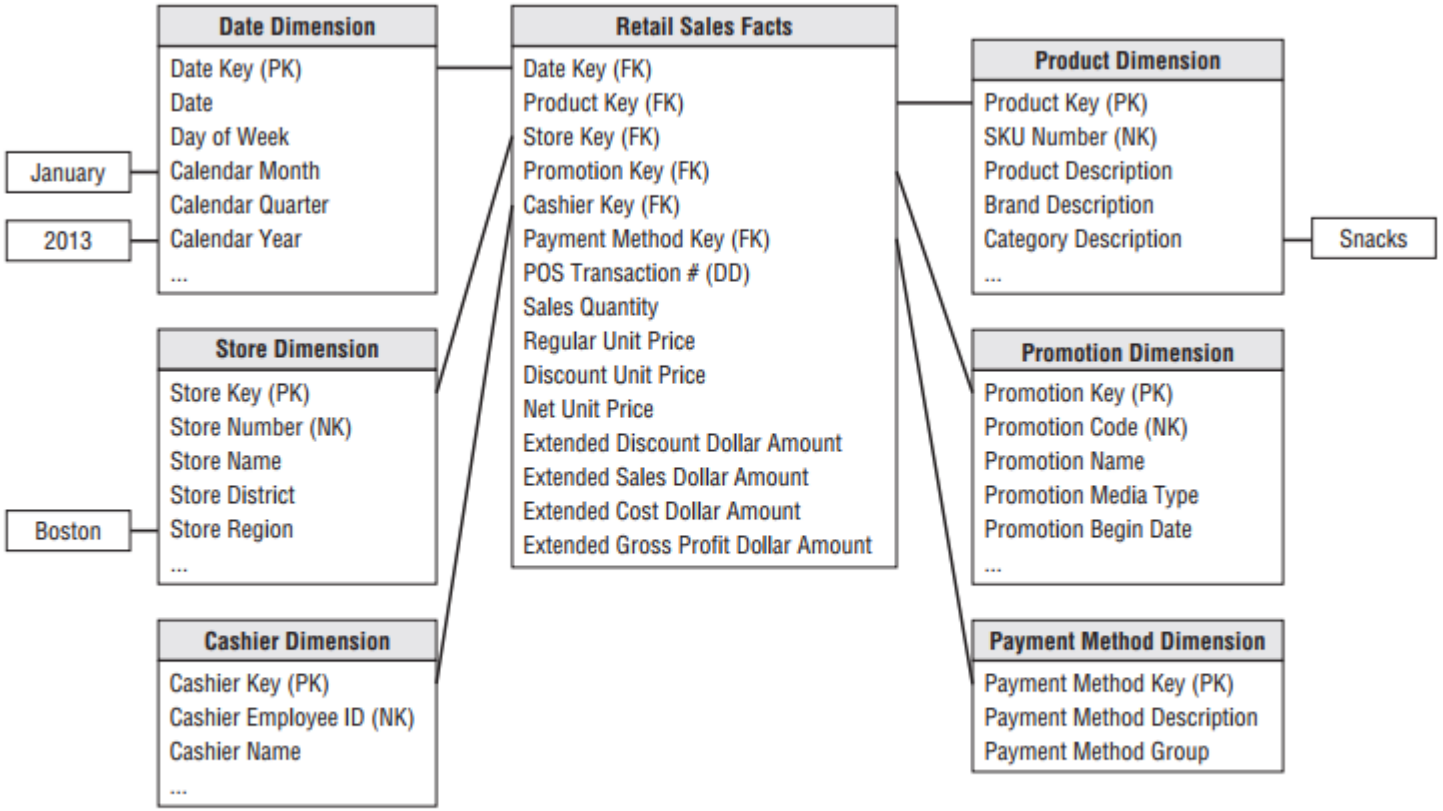
- Identify the facts



[6]

CASE STUDY – RETAIL SALES

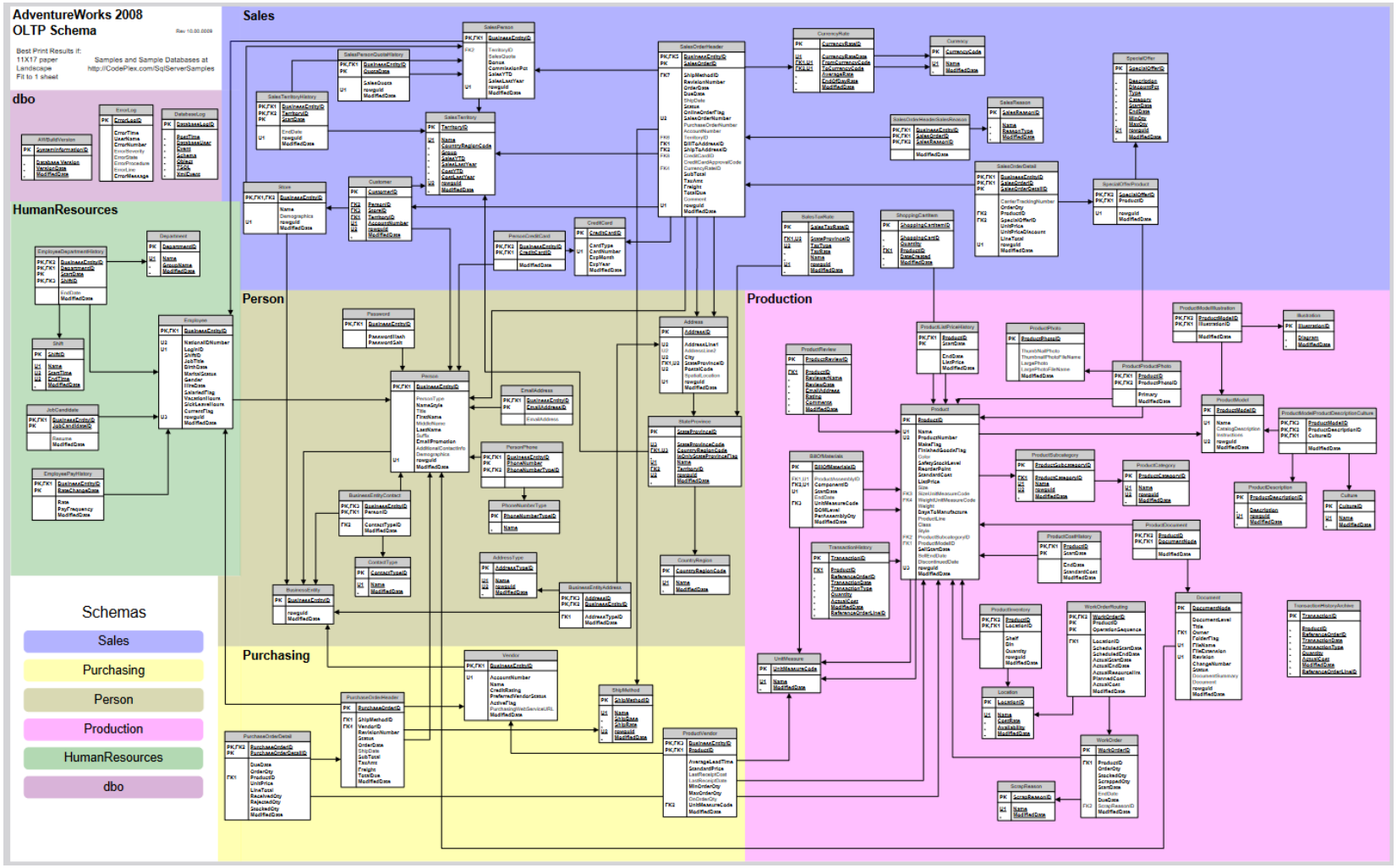
- Detail the dimensions



[6]

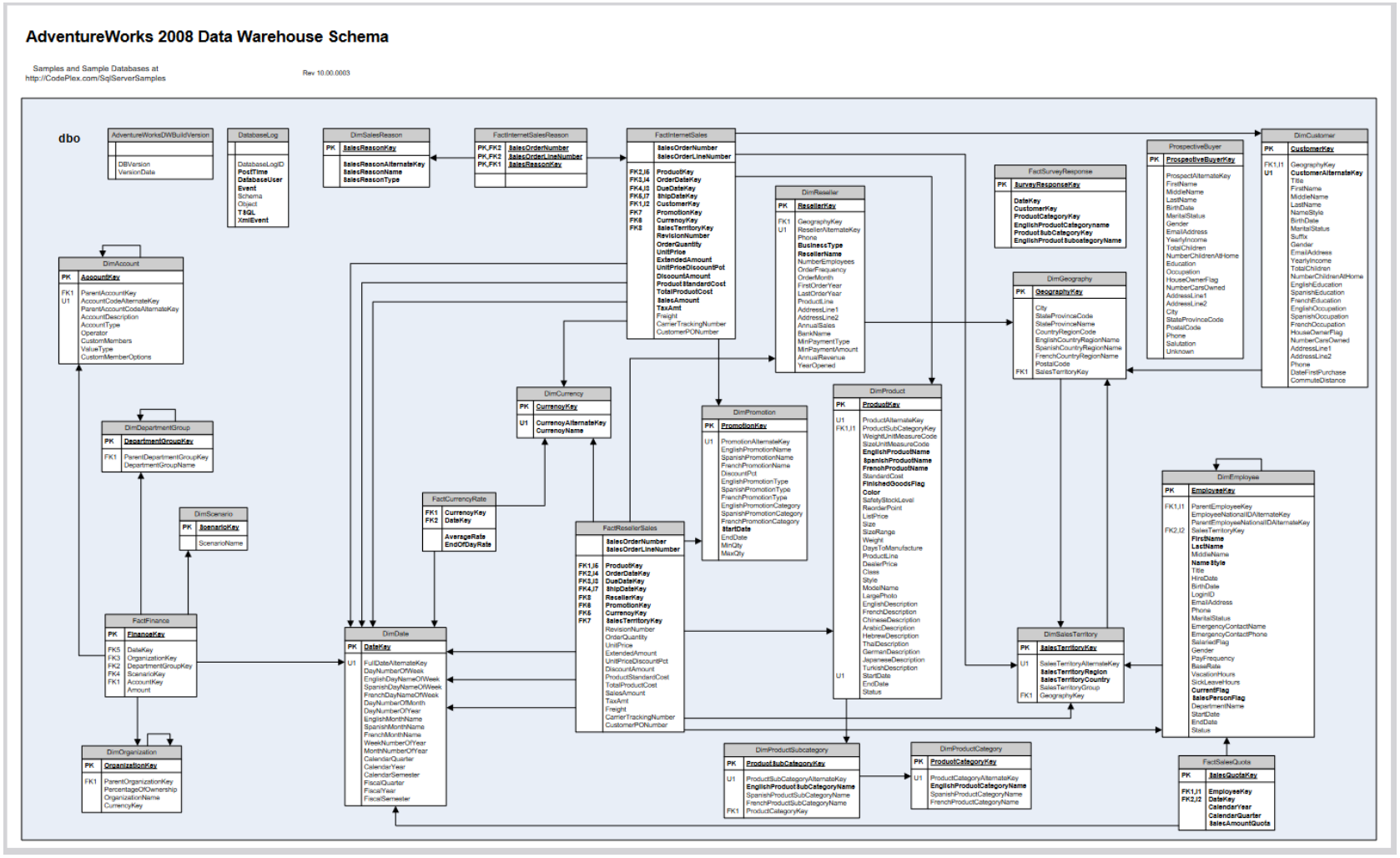
Figure 3-12: Querying the retail sales schema.

SAMPLE OLTP DATABASE



https://akela.mendelu.cz/~jprich/vyuka/db2/AdventureWorks2008_db_diagram.pdf;

SAMPLE DATA WAREHOUSE



https://moodle.usth.edu.vn/pluginfile.php/5907/mod_resource/content/1/AdventureWorksDW2008.pdf

DATA WAREHOUSE AND SLOWLY CHANGING DIMENSIONS

CONCEPT

- A slowly changing dimension (SCD) is a dimension that is able to handle data attributes which change over time (i.e., history preservation).
- For example, a customer dimension may hold attributes such as name, address, and phone number.
  - Over time, a customer's details may change (e.g., addresses, phone number).

SCD TYPES

- Type 0: Retain original
- Type 1: Overwrite
- Type 2: Add new row
- Type 3: Add new attribute
- Type 4: Add mini-dimension
- Type 5: Add mini-dimension and Type 1 outrigger
- Type 6: Add Type 1 attributes to Type 2 dimension
- Type 7: Dual Type 1 and Type 2 dimensions

[6]

SCENARIO

Incoming data

customer_key	phone_number
01	123
02	345
03	999
04	567
05	789

Existing data

customer_key	phone_number
01	123
02	345
03	456

SCD TYPE 0

- No update and history
- Applying to attributes never change (e.g., Full Name, DOB, the date)

Date	Year	Month	Day
2023-09-15	2023	September	Friday
2023-09-16	2023	September	Saturday
2023-09-17	2023	September	Sunday

SCD TYPE 1

- Old values are overwritten with new values
- Update is allowed but not history

customer_key	phone_number
01	123
02	345
03	456

customer_key	phone_number
01	123
02	345
03	999
04	567
05	789

SCD TYPE 2

- Adding a new record
- Creating multiple records with duration (e.g., ValidFrom, ValidTo)
- Update is allowed and history is preserved
- Commong columns needed
  - A surrogate key
  - ValidFrom
  - ValidTo
  - isActive

FOR EXAMPLE

surrogate_key	customer_key	phone_number	is_active	valid_from	valid_to
1	01	123	1	2020-01-01	9999-12-31
2	02	345	1	2020-03-01	9999-12-31
3	03	456	1	2020-08-01	9999-12-31

surrogate_key	customer_key	phone_number	is_active	valid_from	valid_to
1	01	123	1	2020-01-01	9999-12-31
2	02	345	1	2020-03-01	9999-12-31
3	03	456	0	2020-08-01	2021-09-01
4	03	999	1	2021-09-01	9999-12-31
5	04	567	1	2021-09-01	9999-12-31
6	05	789	1	2021-09-01	9999-12-31

SCD TYPE 3

- Adding a new field
- Track changes using separate columns (e.g., CurrentValue, PreviousValue)
- Update is allowed and history is preserved with limitation

FOR EXAMPLE

customer_key	current_phone_number	previous_phone_number
01	123	NULL
02	345	NULL
03	456	NULL

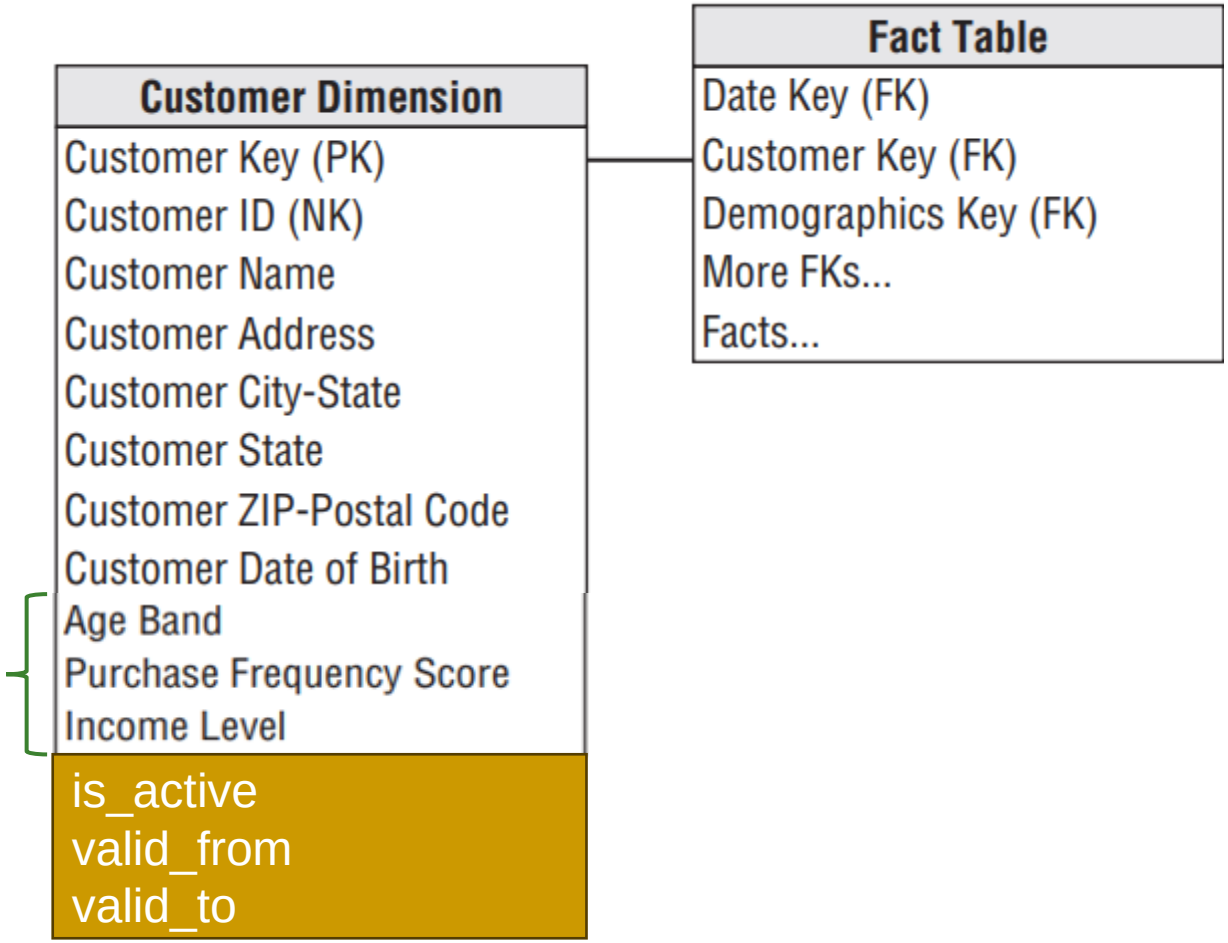
customer_key	current_phone_number	previous_phone_number
01	123	NULL
02	345	NULL
03	999	456
04	567	NULL
05	789	NULL

SCD TYPE 4

- Adding a mini-dimension
- Dealing with browsing performance and change tracking challenges when dimension grows rapidly due to the frequently changing dimension attributes
- The idea is to break off frequently changing attributes into a separate dimension, referred to as a mini-dimension
- Mini-dimension does not store the historical attributes, but the fact table preserved the history of dimension attribute assignment
- Multiple mini-dimensions may be applied

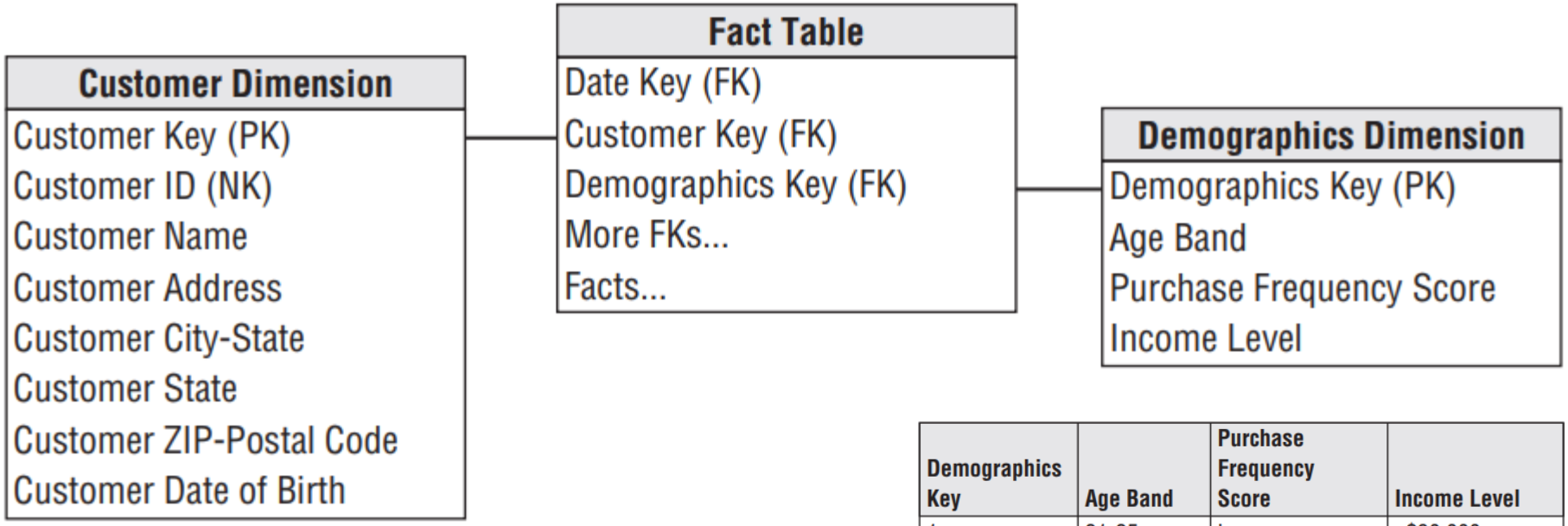
[6]

FOR EXAMPLE



[6]

FOR EXAMPLE



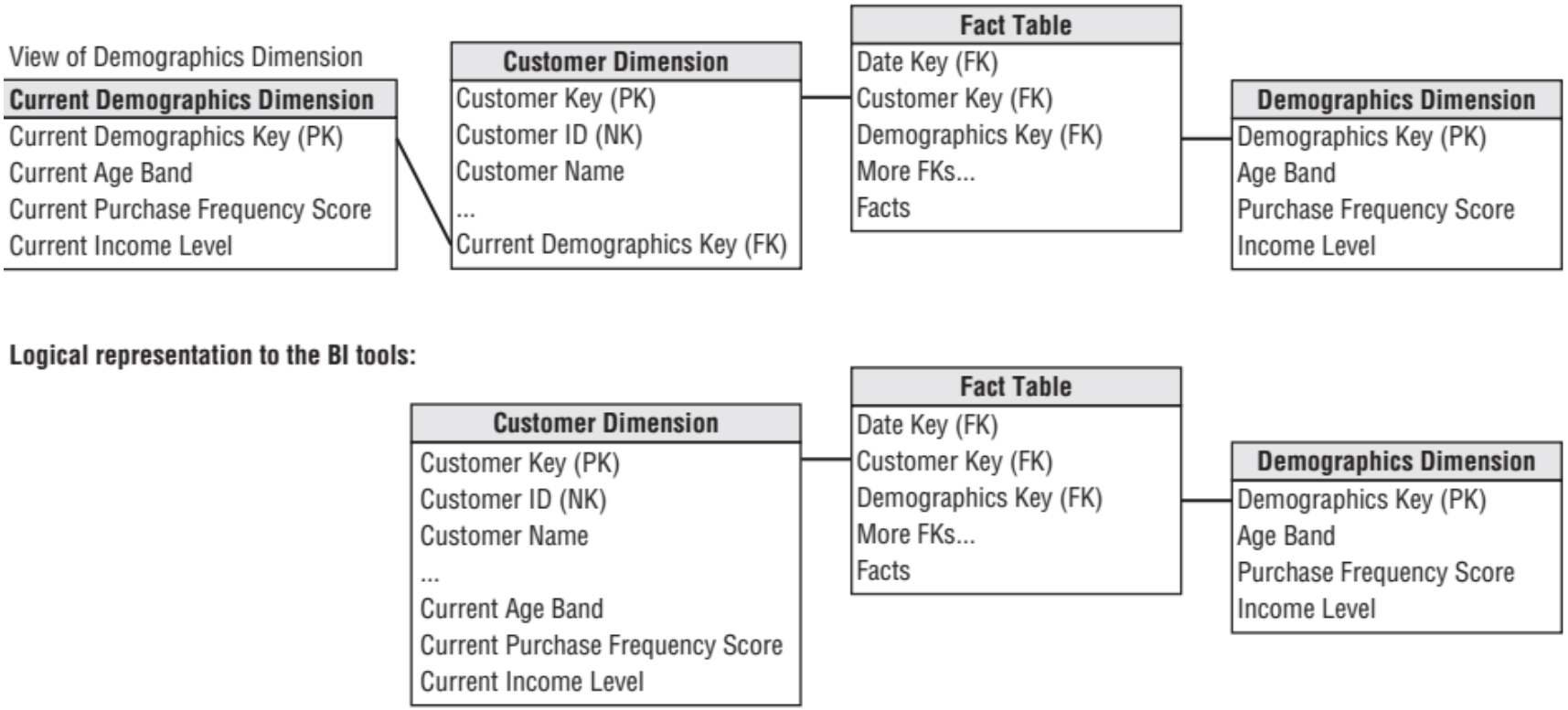
Demographics Key	Age Band	Purchase Frequency Score	Income Level
1	21-25	Low	<$30,000
2	21-25	Medium	<$30,000
3	21-25	High	<$30,000
4	21-25	Low	$30,000-39,999
5	21-25	Medium	$30,000-39,999
6	21-25	High	$30,000-39,999
...	...	...	...
142	26-30	Low	<$30,000
143	26-30	Medium	<$30,000
144	26-30	High	<$30,000
...	...	...	...

[6]

SCD TYPE 5

- The combination of SCD Type 4 and Type 1
- Querying the dimension itself at any one time to find out the value of those attributes, without having to go via the Fact.

FOR EXAMPLE



[6]

SCD TYPE 6

- The combination of SCD Type 2, Type 3 and Type 1
  - It's a type 2 row with a type 3 column that's overwritten as a type 1
- The current attributes are updated on all prior type 2 rows

FOR EXAMPLE

Original row in Product dimension:

Product Key	SKU (NK)	Product Description	Historic Department Name	Current Department Name	Row Effective Date	Row Expiration Date	Current Row Indicator
12345	ABC922-Z	IntelliKidz	Education	Education	2012-01-01	9999-12-31	Current

Rows in Product dimension following first department reassignment:

Product Key	SKU (NK)	Product Description	Historic Department Name	Current Department Name	Row Effective Date	Row Expiration Date	Current Row Indicator
12345	ABC922-Z	IntelliKidz	Education	Strategy	2012-01-01	2012-12-31	Expired
25984	ABC922-Z	IntelliKidz	Strategy	Strategy	2013-01-01	9999-12-31	Current

Rows in Product dimension following second department reassignment:

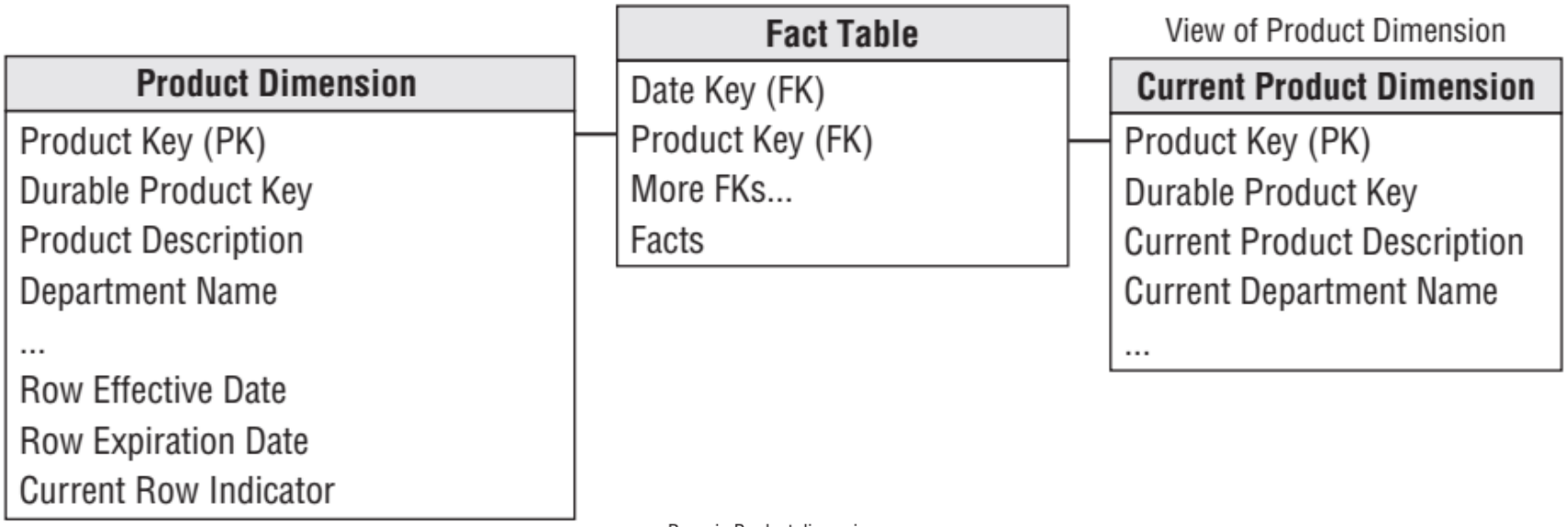
Product Key	SKU (NK)	Product Description	Historic Department Name	Current Department Name	Row Effective Date	Row Expiration Date	Current Row Indicator
12345	ABC922-Z	IntelliKidz	Education	Critical Thinking	2012-01-01	2012-12-31	Expired
25984	ABC922-Z	IntelliKidz	Strategy	Critical Thinking	2013-01-01	2013-02-03	Expired
31726	ABC922-Z	IntelliKidz	Critical Thinking	Critical Thinking	2013-02-04	9999-12-31	Current

[6]

SCD TYPE 7

- Dual Type 1 and Type 2 Dimensions
- This approach delivers the same functionality as type 6, but requires less ETL effort and queries based on current attribute values would be filtering on a smaller dimension table than previously described with type 6

FOR EXAMPLE



Rows in Product dimension:

Product Key	SKU (NK)	Durable Product Key	Product Description	Department Name	...	Row Effective Date	Row Expiration Date	Current Row Indicator
12345	ABC922-Z	12345	IntelliKidz	Education	...	2012-01-01	2013-01-31	Expired
25984	ABC922-Z	12345	IntelliKidz	Strategy	...	2013-02-01	2013-06-30	Expired
31726	ABC922-Z	12345	IntelliKidz	Critical Thinking	...	2013-07-01	9999-12-31	Current

Rows in Product dimension's current view:

Product Key	SKU (NK)	Durable Product Key	Current Product Description	Current Department Name	...
12345	ABC922-Z	12345	IntelliKidz	Critical Thinking	...
25984	ABC922-Z	12345	IntelliKidz	Critical Thinking	...
31726	ABC922-Z	12345	IntelliKidz	Critical Thinking	...

[6]

SCD SUMMARIZATION

SCD Type	Dimension Table Action	Impact on Fact Analysis
Type 0	No change to attribute value.	Facts associated with attribute's original value.
Type 1	Overwrite attribute value.	Facts associated with attribute's current value.
Type 2	Add new dimension row for profile with new attribute value.	Facts associated with attribute value in effect when fact occured.
Type 3	Add new column to preserve attribute's current and prior values.	Facts associated with both current and prior attribute alternative values.
Type 4	Add mini-dimension table containing rapidly changing attributes.	Facts associated with rapidly changing attributes in effect when fact occured.
Type 5	Add type 4 mini-dimension, along with overwritten type 1 mini-dimension key in base dimension.	Facts associated with rapidly changing attributes in effect when fact occurred, plus current rapidly changing attribute values.
Type 6	Add type 1 overwritten attributes to type 2 dimension row, and overwrite all prior dimension rows.	Facts associated with attribute value in effect when fact occurred, plus current values.
Type 7	Add type 2 dimension row with new attribute value, plus view limited to current rows and/or attribute values.	Facts associated with attribute value in effect when fact occurred, plus current values.

[6]

DATA WAREHOUSE AND TOOLS

POPULAR ETL TOOLS

- SSIS
- OpenTalend
- Xplenty
- Informatica
- IBM Infosphere

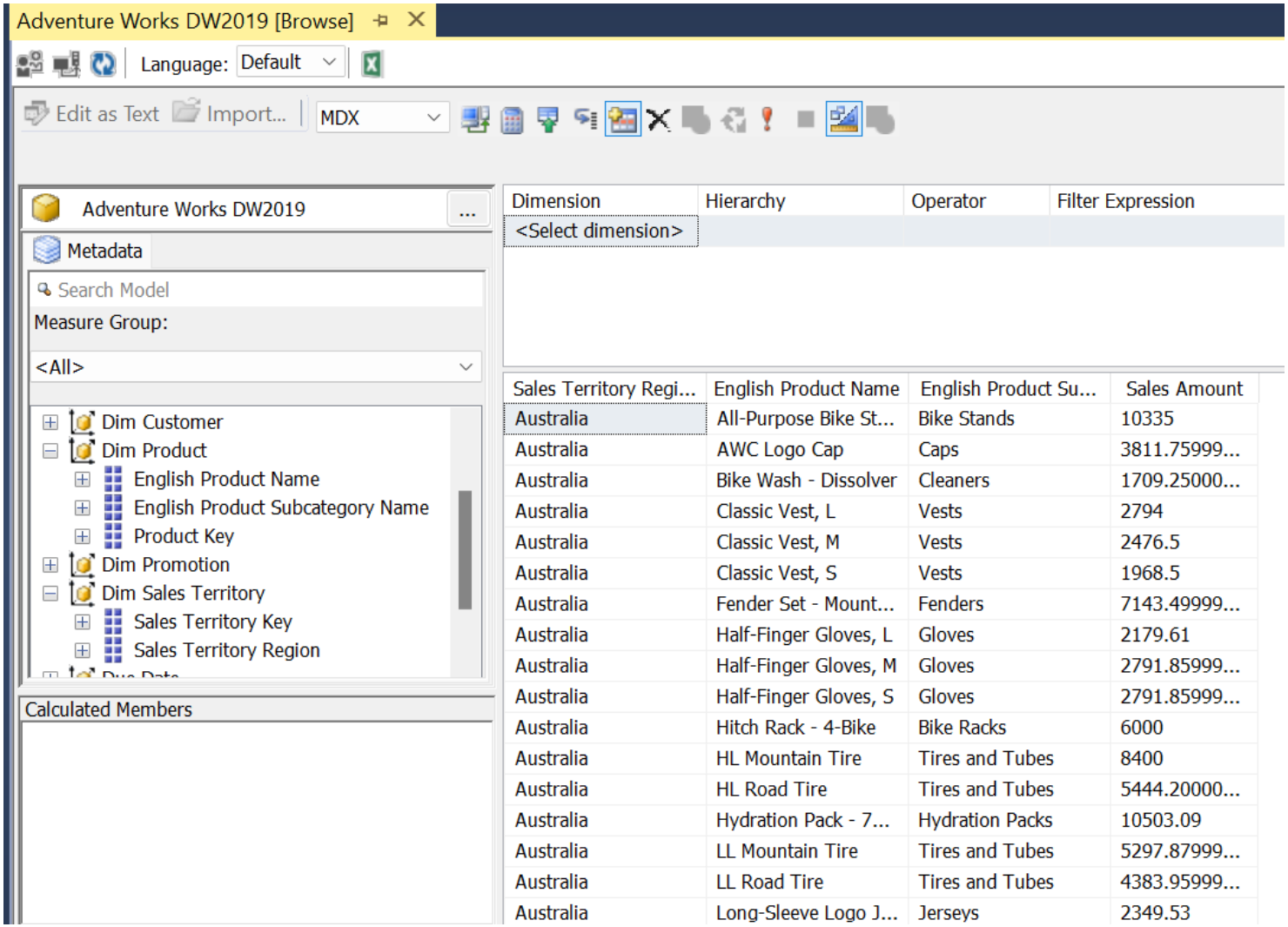
OLAP TYPES

- Relational OLAP
- Multidimensional OLAP
- Hybrid OLAP

DISCUSSION

DATA WAREHOUSE DESIGN LAB

INTERNET SALES (1/2)



INTERNET SALES (2/2)

```
SELECT st.Name ,p.Name ,ps.Name ,sum(od.LineTotal)
  FROM Sales.SalesOrderDetail od
  JOIN Production.Product p ON od.ProductID = p.ProductID
  JOIN sales.SalesOrderHeader oh on od.SalesOrderID =
oh.SalesOrderID
  JOIN sales.SalesTerritory st on oh.TerritoryID =
st.TerritoryID
  JOIN production.ProductSubcategory ps on
p.ProductSubcategoryID = ps.ProductSubcategoryID
  --where st.Name = 'Canada'
  group by st.Name ,p.Name ,ps.Name
  order by st.Name ,p.Name ,ps.Name
;
```

SUMMARY

- Dimensional data modeling
- Slowly changing dimensions
- Data warehouse and tools
- Data warehouse design lab

QUESTIONS AND ANSWERS

Picture from: http://philadelphiasculpturegym.blogspot.com/2013/09/save-date-free-talk-and-q-on-affordable.html

REFERENCES

1. Tobias Zwingmann, “AI-Powered Business Intelligence,” Kindle Edition, O'reilly Press, 2022
2. Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques,” Third Edition, Morgan Kaufmann Publishers, 2012.
3. David L. Olson, Dursun Delen, “Advanced Data Mining Techniques,” Springer-Verlag, 2008.
4. Jeen Su Lim, John Heinrichs, “Digital Business Intelligence Management with Big Data Analytics," Kindle Edition, O'reilly Press, 2021.
5. William H. Inmon, “Building the Data Warehouse,” Fourth Edition, Wiley Publishing, Inc., 2005.
6. R. Kimball, M. Ross, "The Data Warehouse ToolKit," 3rd Edition, Wiley Publishing, Inc., 2013.
7. Turban, E., Aronson,J.E., “Decision Support Systems and Intelligent Systems” - 7th Edition, Prentice-Hall, 2005.
8. Ramesh Sharda, Dursun Delen, Efraim Turban, “Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support,” 7th Edition, Pearson Education, Inc., 2020.