Data Warehousing & OLAP

An Introduction

What you will learn …

At the end of this lecture you will be able to:

- Explain what a data warehouse is
- Identify the components of a data warehouse
- Explain what OLAP is
- Explain the use of databases in data warehousing & OLAP

Two Types of Databases

Operational Databases

- Record **day-to-day** operations
- Also called Online Transaction Processing (**OLTP**) databases, transactional databases or production databases
- Transactions should be recorded **accurately** and **immediately**

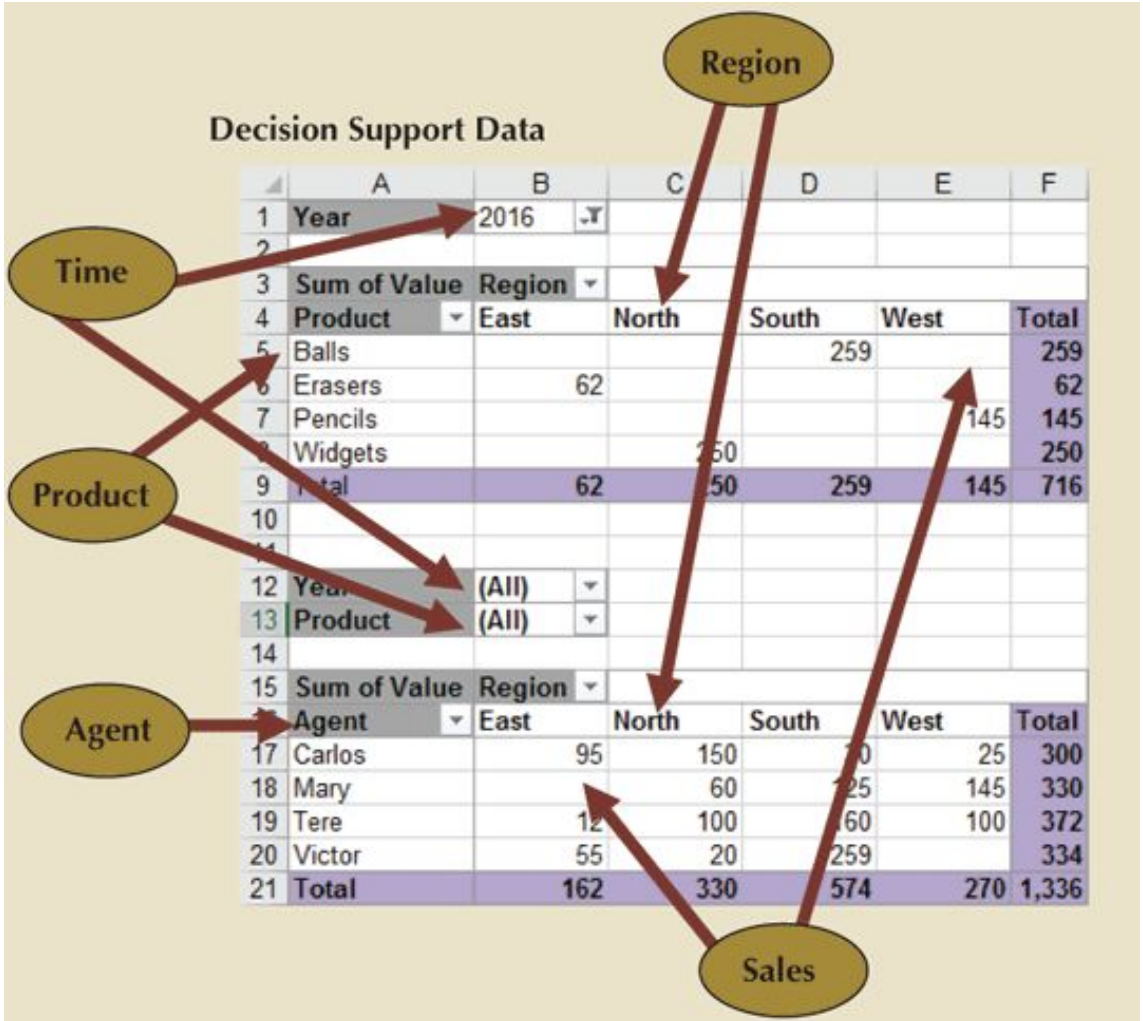
Analytical Databases

- Longer-term, historical data & business metrics to support **decision making**
- Two main parts: data warehouse + Online Analytical Processing (**OLAP**) front end.

	A	B	C	D	E
1	Year	Region	Agent	Product	Value
2	2016	East	Carlos	Erasers	50
3	2016	East	Tere	Erasers	12
4	2016	North	Carlos	Widgets	120
5	2016	North	Tere	Widgets	100
6	2016	North	Carlos	Widgets	30
7	2016	South	Victor	Balls	145
8	2016	South	Victor	Balls	34
9	2016	South	Victor	Balls	80
10	2016	West	Mary	Pencils	89
11	2016	West	Mary	Pencils	56
12	2017	East	Carlos	Pencils	45
13	2017	East	Victor	Balls	55
14	2017	North	Mary	Pencils	60
15	2017	North	Victor	Erasers	20
16	2017	South	Carlos	Widgets	30
17	2017	South	Mary	Widgets	75
18	2017	South	Mary	Widgets	50
19	2017	South	Tere	Balls	70
20	2017	South	Tere	Erasers	90
21	2017	West	Carlos	Widgets	25
22	2017	West	Tere	Balls	100

Operational Data

- One row for each transaction
- Narrow **time span**
- Low **granularity**
- Single focus
- Difficult to derive useful information



	A	B	C	D	E	F
1	Year	2016				
2						
3	Sum of Value	Region				
4	Product	East	North	South	West	Total
5	Balls			259		259
6	Erasers	62				62
7	Pencils				145	145
8	Widgets		250			250
9	Total	62	250	259	145	716
10						
11						
12	Year	(All)				
13	Product	(All)				
14						
15	Sum of Value	Region				
16	Agent	East	North	South	West	Total
17	Carlos	95	150	30	25	300
18	Mary		60	125	145	330
19	Tere	12	100	160	100	372
20	Victor	55	20	259		334
21	Total	162	330	574	270	1,336

Decision Support Data

- Broader time span
- High level of granularity
- Multiple **dimensions** make it possible to see:
  - Sales by product, region, agent, etc
  - Sales for all years or a few selected years
  - Sales for all products or a few selected products

TABLE 13.6

TEN-YEAR SALES HISTORY FOR A SINGLE DEPARTMENT, IN MILLIONS OF DOLLARS

YEAR	SALES
2008	8,227
2009	9,109
2010	10,104
2011	11,553
2012	10,018
2013	11,875
2014	12,699
2015	14,875
2016	16,301
2017	19,986

TABLE 13.6

TEN-YEAR SALES HISTORY FOR A SINGLE DEPARTMENT, IN MILLIONS OF DOLLARS

YEAR	SALES
2008	8,227
2009	9,109
2010	10,104
2011	11,553
2012	10,018
2013	11,875
2014	12,699
2015	14,875
2016	16,301
2017	19,986

Topic

Aggregated and Summarized data

TABLE 13.7

YEARLY SALES SUMMARIES, TWO STORES AND TWO DEPARTMENTS PER STORE, IN MILLIONS OF DOLLARS

YEAR	STORE	DEPARTMENT	SALES
2008	A	1	1,985
2008	A	2	2,401
2008	B	1	1,879
2008	B	2	1,962
…	…	…	…
2014	A	1	3,912
2014	A	2	4,158
2014	B	1	3,426
2014	B	2	1,203
…	…	…	…
2017	A	1	7,683
2017	A	2	6,912
2017	B	1	3,768
2017	B	2	1,623

Notice the **rapid increase** of number of columns and rows and the resulted **redundancy**

TABLE 13.7

YEARLY SALES SUMMARIES, TWO STORES AND TWO DEPARTMENTS PER STORE, IN MILLIONS OF DOLLARS

YEAR	STORE	DEPARTMENT	SALES
2008	A	1	1,985
2008	A	2	2,401
2008	B	1	1,879
2008	B	2	1,962
…	…	…	…
2014	A	1	3,912
2014	A	2	4,158
2014	B	1	3,426
2014	B	2	1,203
…	…	…	…
2017	A	1	7,683
2017	A	2	6,912
2017	B	1	3,768
2017	B	2	1,623

Subjects of Interest

Topic

Redundant Data

Notice the **rapid increase** of number of columns and rows and the resulted **redundancy**

Decision Support Database Requirements

DB schema

- Should support complex (non-normalized), aggregated, summarized data
- Should support (read-only) queries that extract multi-dimensional time slices

Data Extraction & Filtering

- Data extraction from both operational data and external sources
- Should check for inconsistent data, support data validation rules and solve data formatting conflicts

Database Size

- Should support very large database (VLDB)

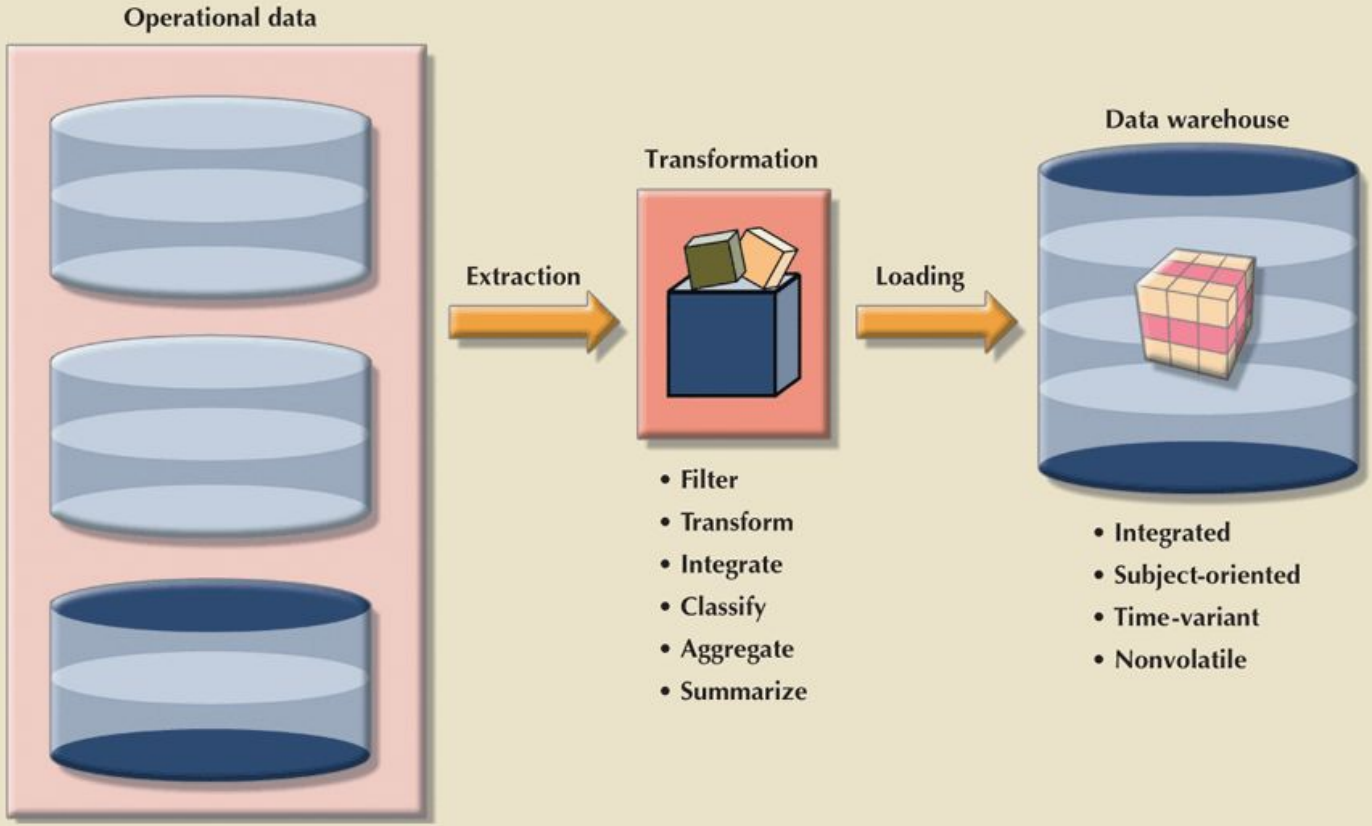
Data Warehouse

The new type of data repository that supports the requirements of Decision Support Systems

Through the ETL process data from operational (OLTP) sources are brought into a single warehouse for OLAP analysis

Data warehouses are **integrated**, **subject-oriented**, **time-variant** and **nonvolatile** collections of data

FIGURE 13.4 THE ETL PROCESS



OLTP vs. OLAP

OLTP -- Online Transaction Processing

- Short transactions
- Simple queries
- Touch small portions of data
- Frequent updates

OLAP -- Online Analytical Processing

- Long transactions
- Complex queries
- Touch large portions of the data
- Infrequent updates

OLTP vs. OLAP

OLTP -- Online Transaction Processing

- Short transactions
- Simple queries
- Touch small portions of data
- Frequent updates

OLAP -- Online Analytical Processing

- Long transactions
- Complex queries
- Touch large portions of the data
- Infrequent updates

- Two ends of an espectrum
- Relational database systems traditionally are tuned for OLTP
- Special techniques are developed for OLAP

Star Schema

A data **modeling technique** to map multi-dimensional data into a relational DB

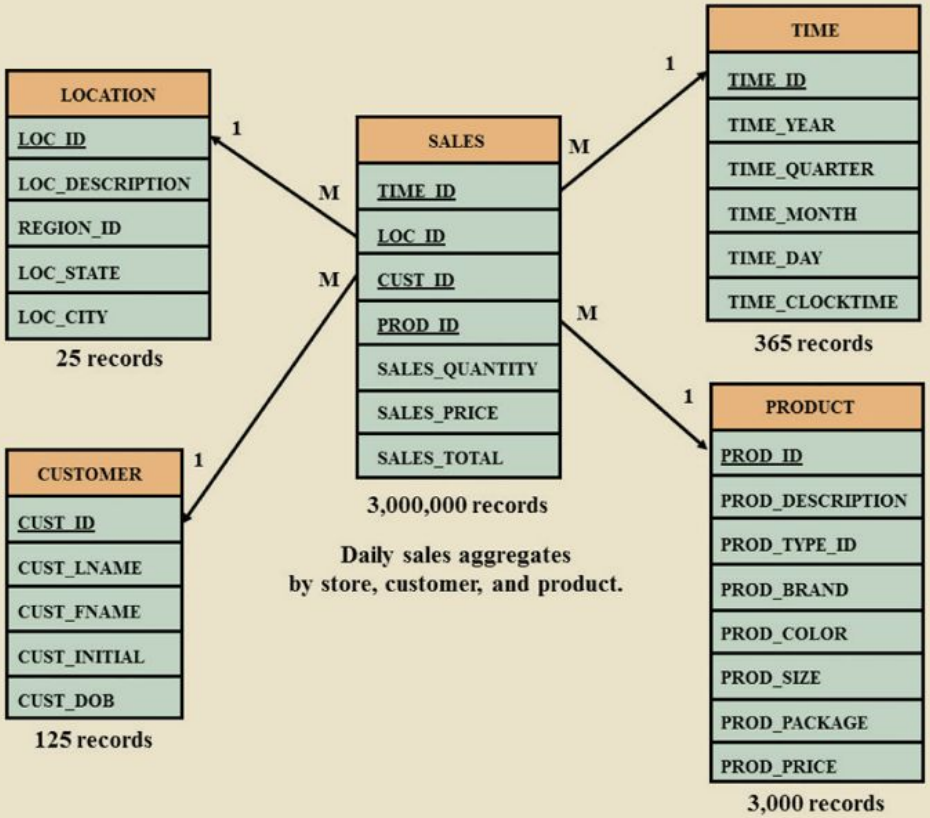
Fact Table

- Updated frequently, often append-only, very large
- e.g. daily sales transactions

Dimension Tables

- Updated infrequently, not as large
- e.g. location, customer, time, product

FIGURE 13.10 STAR SCHEMA FOR SALES



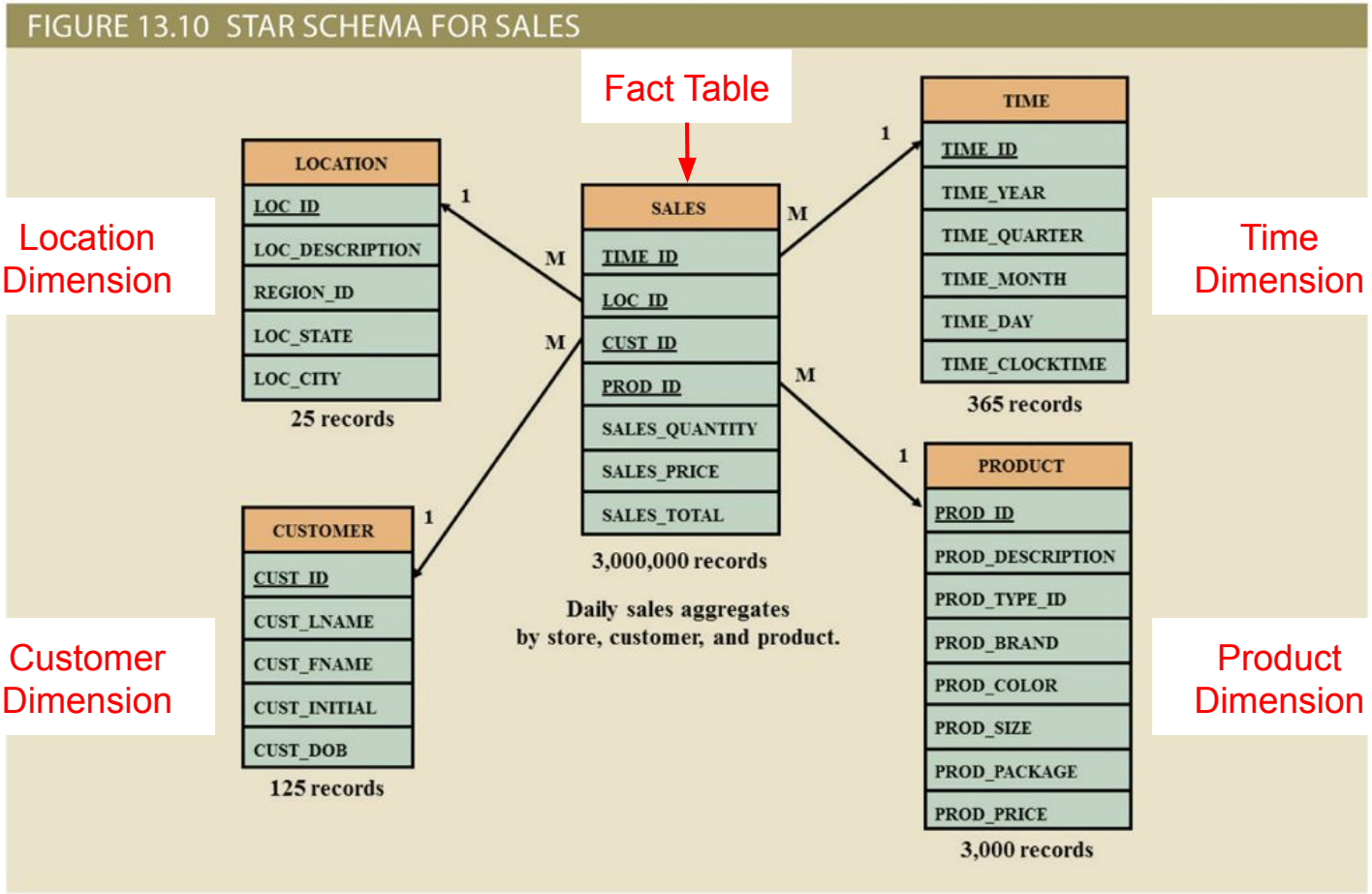
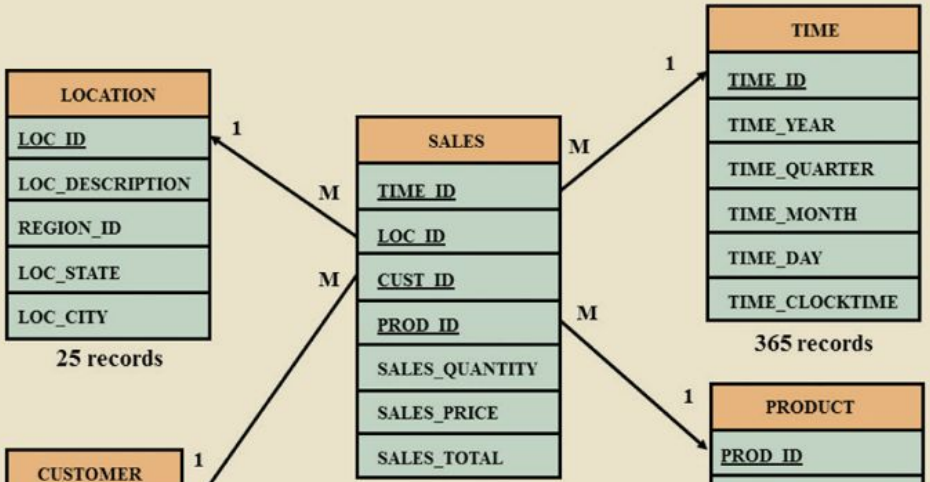


FIGURE 13.10 STAR SCHEMA FOR SALES
Fact Table
Location Dimension
Time Dimension
Customer Dimension
Product Dimension
LOCATION
LOC_ID
LOC_DESCRIPTION
REGION_ID
LOC_STATE
LOC_CITY
25 records
SALES
TIME_ID
LOC_ID
CUST_ID
PROD_ID
SALES_QUANTITY
SALES_PRICE
SALES_TOTAL
3,000,000 records
Daily sales aggregates
by store, customer, and product.
TIME
TIME_ID
TIME_YEAR
TIME_QUARTER
TIME_MONTH
TIME_DAY
TIME_CLOCKTIME
365 records
CUSTOMER
CUST_ID
CUST_LNAME
CUST_FNAME
CUST_INITIAL
CUST_DOB
125 records
PRODUCT
PROD_ID
PROD_DESCRIPTION
PROD_TYPE_ID
PROD_BRAND
PROD_COLOR
PROD_SIZE
PROD_PACKAGE
PROD_PRICE
3,000 records
1
M
1
M
1
M
1
M

FIGURE 13.10 STAR SCHEMA FOR SALES



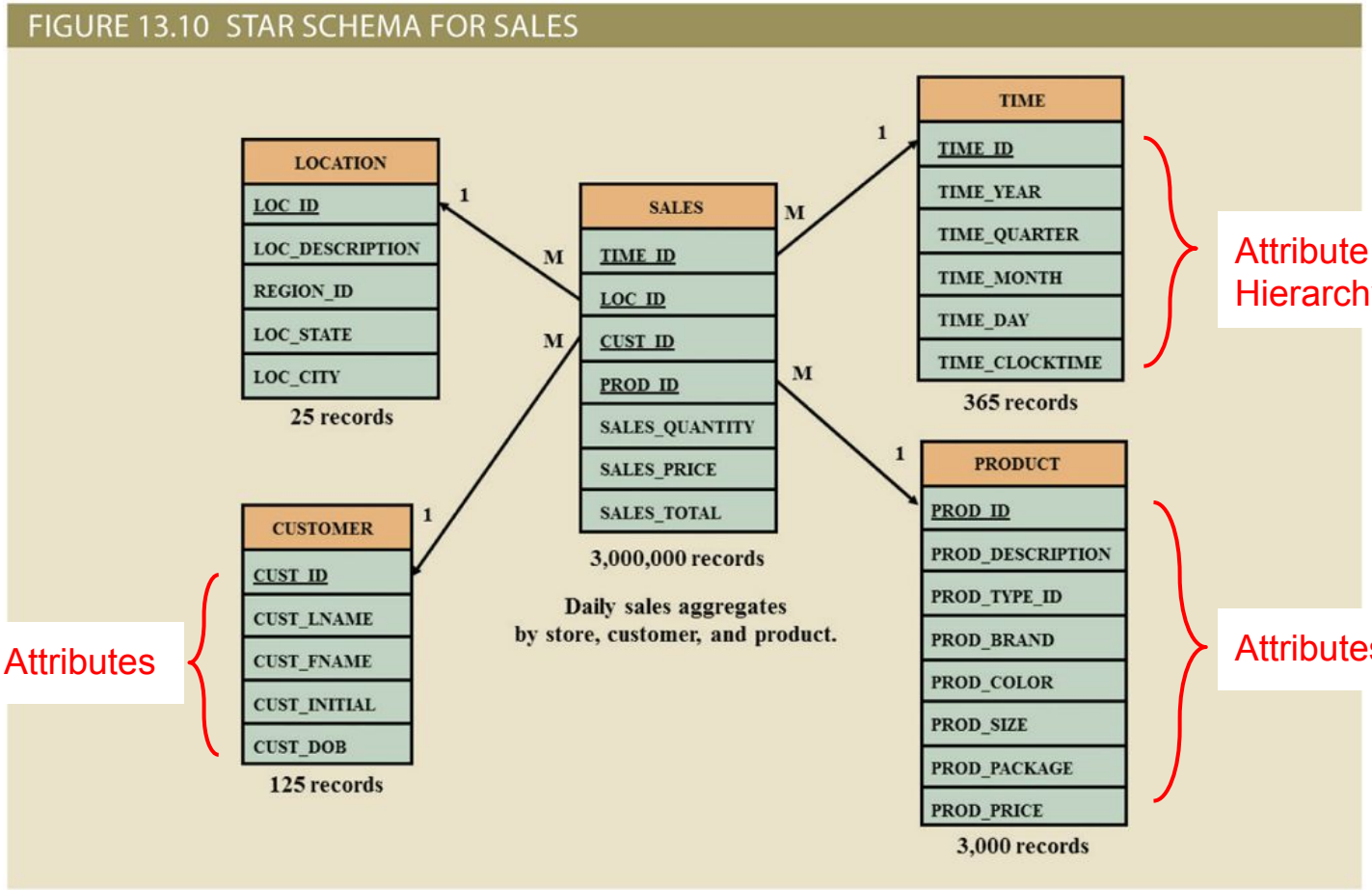


FIGURE 13.10 STAR SCHEMA FOR SALES
LOCATION
LOC_ID
LOC_DESCRIPTION
REGION_ID
LOC_STATE
LOC_CITY
25 records
CUSTOMER
CUST_ID
CUST_LNAME
CUST_FNAME
CUST_INITIAL
CUST_DOB
125 records
SALES
TIME_ID
LOC_ID
CUST_ID
PROD_ID
SALES_QUANTITY
SALES_PRICE
SALES_TOTAL
3,000,000 records
Daily sales aggregates
by store, customer, and product.
TIME
TIME_ID
TIME_YEAR
TIME_QUARTER
TIME_MONTH
TIME_DAY
TIME_CLOCKTIME
365 records
PRODUCT
PROD_ID
PROD_DESCRIPTION
PROD_TYPE_ID
PROD_BRAND
PROD_COLOR
PROD_SIZE
PROD_PACKAGE
PROD_PRICE
3,000 records
1
M
Attributes
Attributes
Attribute
Hierarchy

OLAP Queries

SALES(TIME_ID, LOC_ID, CUST_ID, PROD_ID, SALES_QUANTITY, SALES_PRICE, SALES_TOTAL**)**

TIME(TIME_ID, TIME_YEAR, TIME_QUARTER, TIME_MONTH, TIME_DAY, TIME_CLOCKTIME**)**

LOCATION(LOC_ID, LOC_DESCRIPTION, REGION_ID, LOC_STATE, LOC_CITY**)**

CUSTOMER(CUST_ID, CUST_LNAME, CUST_FNAME, CUST_INITIAL, CUST_DOB**)**

PRODUCT(PROD_ID, PROD_DESCRIPTION, PROD_TYPE_ID, PROD_BRAND, PROD_COLOR, PROD_SIZE, ...**)**

Join → Filter → Group → Aggregate

Complex queries that touch large portions of a large database

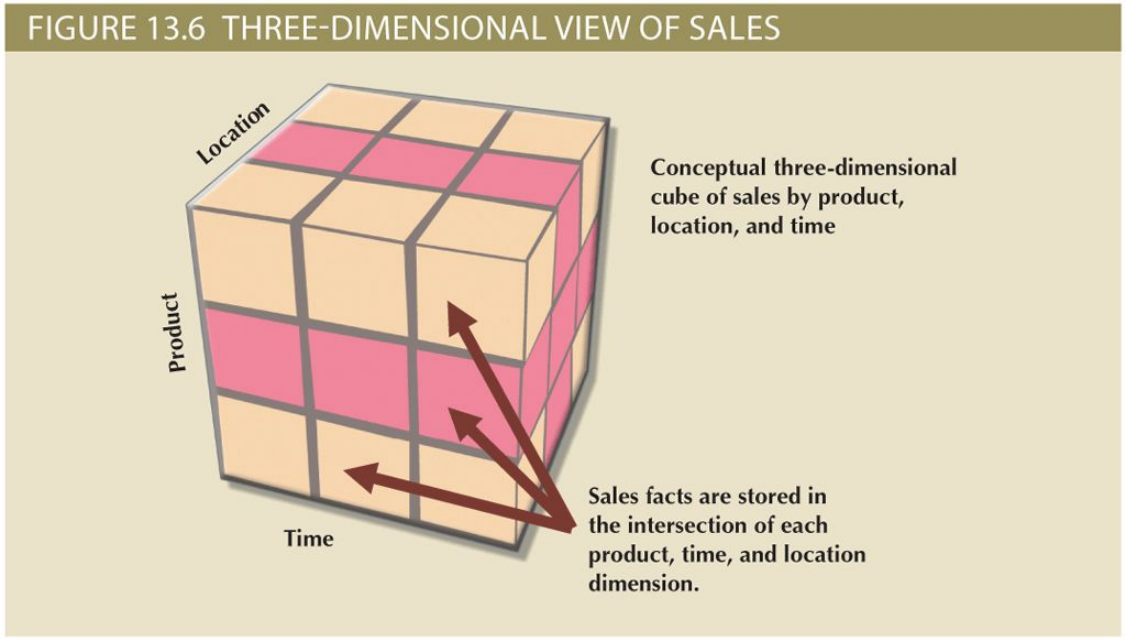
Performance becomes a worry!

Relational OLAP

ROLAP adds the following functionality to traditional RDBMS technology:

- Multidimensional data schema support within the RDBMS
  - Star schema
- data access language and query performance optimized for multidimensional data.
  - Extends SQL to **differentiate between queries** to normalized tables vs queries to data warehouses
  - Use of **advanced indexing** such as bitmap indexes
  - Extensive use of **materialized views**
- support for very large databases.
  - Decision support data is loaded in bulk

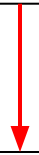
Data Cube / Multidimensional OLAP



- Dimension data forms axes of “cube”
- Fact data in cells
- Aggregated data on sides, edges, corner

Drill-Down & Roll-Up

```
SELECT LOC_STATE, PROD_BRAND, SUM(SALES_QUANTITY * SALES_PRICE)
FROM SALES S, LOCATION L, PRODUCT P
WHERE S.LOC_ID = L.LOC_ID AND S.PROD_ID = P.PROD_ID
GROUP BY LOC_STATE, PROD_BRAND
```



Drill Down

```
SELECT LOC_STATE, PROD_TYPE_ID, PROD_BRAND, SUM(SALES_QUANTITY *
SALES_PRICE)
FROM SALES S, LOCATION L, PRODUCT P
WHERE S.LOC_ID = L.LOC_ID AND S.PROD_ID = P.PROD_ID
GROUP BY LOC_STATE, PROD_TYPE_ID, PROD_BRAND
```

Drill-Down & Roll-Up

```
SELECT LOC_STATE, PROD_BRAND, SUM(SALES_QUANTITY * SALES_PRICE)
FROM SALES S, LOCATION L, PRODUCT P
WHERE S.LOC_ID = L.LOC_ID AND S.PROD_ID = P.PROD_ID
GROUP BY LOC_STATE, PROD_BRAND
```

Roll Up

```
SELECT LOC_STATE, PROD_BRAND, SUM(SALES_QUANTITY * SALES_PRICE)
FROM SALES S, LOCATION L, PRODUCT P
WHERE S.LOC_ID = L.LOC_ID AND S.PROD_ID = P.PROD_ID
GROUP BY LOC_STATE, PROD_BRAND
```

SQL Analytics Functions - The CUBE Extension

```
SELECT dimension-attrs , aggregates
FROM tables
WHERE conditions
GROUP BY CUBE dimension-attrs
```

- Adds faces, edges and corners of cube to the result
- Uses NULL values for attributes not considered

SQL Analytics Functions - The ROLLUP Extension

```
SELECT dimension-attrs , aggregates
FROM tables
WHERE conditions
GROUP BY ROLLUP dimension-attrs
```

- Returns a portion of the GROUP BY CUBE results
- Specifically useful with hierarchical dimensions

Summary

OLTP -- Online Transaction Processing

- Short transactions
- Simple queries
- Touch small portions of data
- Frequent updates

OLAP -- Online Analytical Processing

- Long transactions
- Complex queries
- Touch large portions of the data
- Infrequent updates

Summary

OLTP -- Online Transaction Processing

- Short transactions
- Simple queries
- Touch small portions of data
- Frequent updates

OLAP -- Online Analytical Processing

- Long transactions
- Complex queries
- Touch large portions of the data
- Infrequent updates

- Star Schemas
- Data Cubes
- CUBE & ROLLUP extensions
- Special indexes & query processing techniques