Lecture 16: data warehouse

BADM/ACCY 352

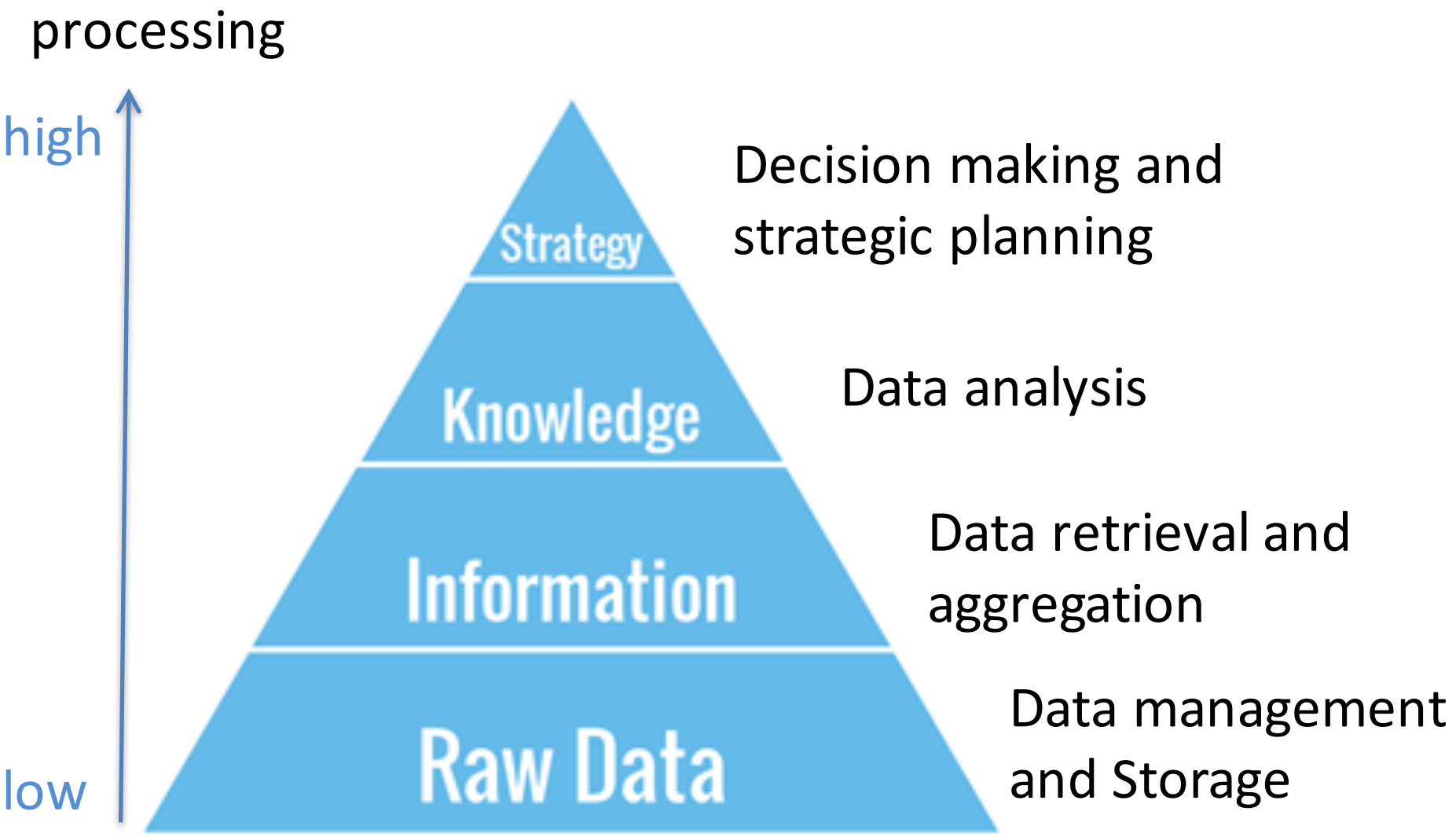
Spring 2017

Instructor: Yi Yang, PhD

Guest speaker takeaways

- In this big data era, companies need to build a scalable database system.
- building a scalable database system is challenging
- one technique is called sharding, which splits and distributes a huge table into smaller tables

Business Intelligence


processing
high
low
Strategy
Knowledge
Information
Raw Data
Decision making and strategic planning
Data analysis
Data retrieval and aggregation
Data management and Storage

identify, develop and create new strategic business opportunities

Are our sales promotions working?
Are we attracting new customers?
Is he/she the potential customer?

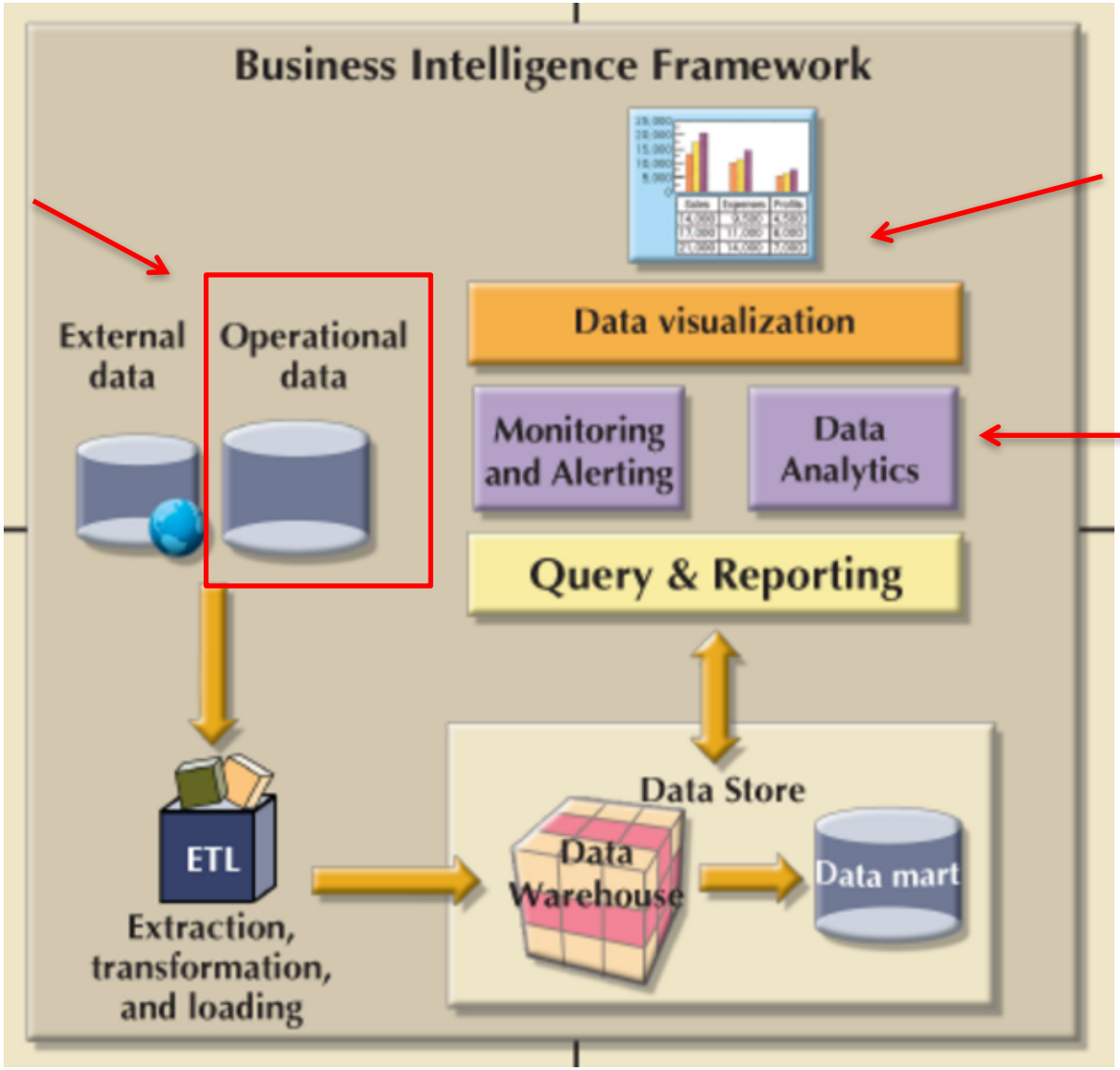
Business Intelligence

- BI refers to “the set of techniques and tools for the transformation of raw data into meaningful and useful information for business analysis purposes”.
  - It handles large amounts of data
  - It helps identify, develop and create new strategic business opportunities
  - It provides businesses with a competitive market advantage and long-term stability.

What does BI do

- Collecting and storing operational data
- Aggregating the operational data into data warehouse data
- Analyzing data warehouse data to generate information
- Presenting information to the end user to support business decisions
- Making business decisions, which in turn generate more data that are collected, stored, and so on (restarting the process)
- Monitoring results to evaluate outcomes of the business decisions, which again provides more data to be collected, stored, and so on
- Predicting future behaviors and outcomes with a high degree of accuracy

BADM352



BADM 395 : Section DSA – Data Science Analytics

BADM453: Decision Support Systems

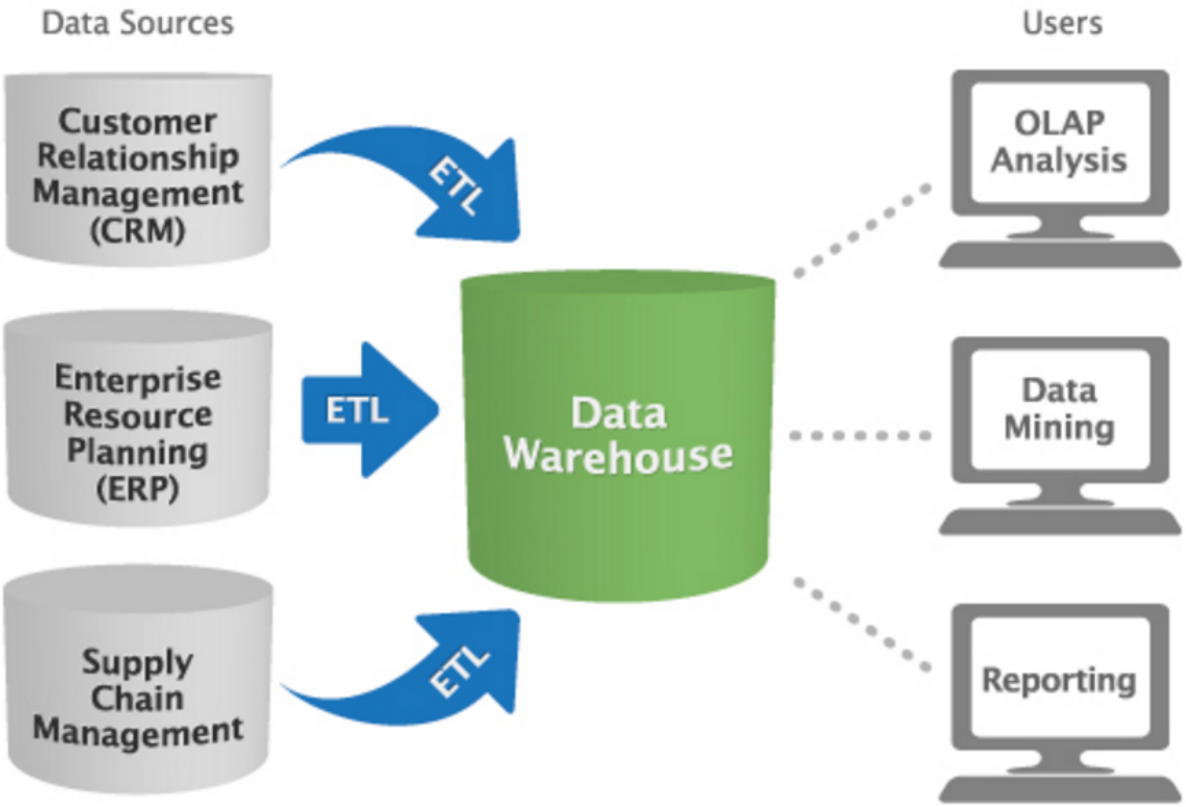
Operational database

- It is (mostly) stored in relational database.
- The database is highly normalized.
- It supports transactions that represent daily business operations.
  - Eg. Each time an item is sold, customer, inventory, invoice tables will be updated.
- It's efficient/secure for data modification, but not efficient for analytics.
  - Eg. To retrieve an invoice, you have to join several tables, but to insert an invoice, only a small amount of data in those tables is affected.

Data warehouse

- A data warehouse ***is*** a (relational) database.
- Data in warehouse is summarized and aggregated. It is designed for query and analysis rather than for transaction processing.
- It contains historical data derived from operational database, but it can include data from other sources.
- The data are snapshot data captured at a given point in time.
- It is an enterprise level data repository. It provides a global picture of the business.

Architecture of a Data Warehouse



Extract, Transform, Load

- Extracts data from different data sources
  - Other databases, files, unstructured data, cloud, sensor, mobile, etc.
- Transforms the data for storing it in proper format or structure for querying and analysis purpose
  - Using standard format. Eg. Student can be “freshman”, “sophomore” in one source, or “FR”, “SO” from other source.
- Loads the data into data warehouse

Extract **T**ransform **L**oad

Data warehouse data vs. operational database data

	Data warehouse data	Operational Data
Time span	Long time frame. E.g. the sales during the last month	Short time frame. One transaction for example.
Granularity	Different levels of aggregation. E.g. by region, by city, by store.	No aggregation
Dimensionality	High dimension data. E.g. how product X fared relative to product Y during the past month by region, state, city, store.	Single, individual transaction

Data warehouse vs. operational database

Database	Data Warehouse
The tables and joins in the DB are complex since they are normalized	Tables and joins are simple since they are de-normalized
ER Modeling techniques are used for database design	Dimension modeling techniques are used for database design
Optimized for write operation	Optimized for read operations
Performance is slow for analytical queries	High performance for analytical queries

Question

There is no concurrency issue in data warehouse.

a) True
b) False

Why do we need data warehouse

- The primary reason is, for a company to get extra edge over its competitors.
- The edge can be gained by taking smarter decisions

Strategic questions

Q: how do we increase the market share by 5%

Q: which product is not doing well in the market

Q: What is the quality of the customer service provided and what improvements are needed?

Benefits

- It standardizes data across an organization
- It improves data quality.
- It makes decision–support queries easier to write, and it delivers excellent query performance.
- It restructures the data so that it makes sense to the business users.
- Reduce costs & increase revenue

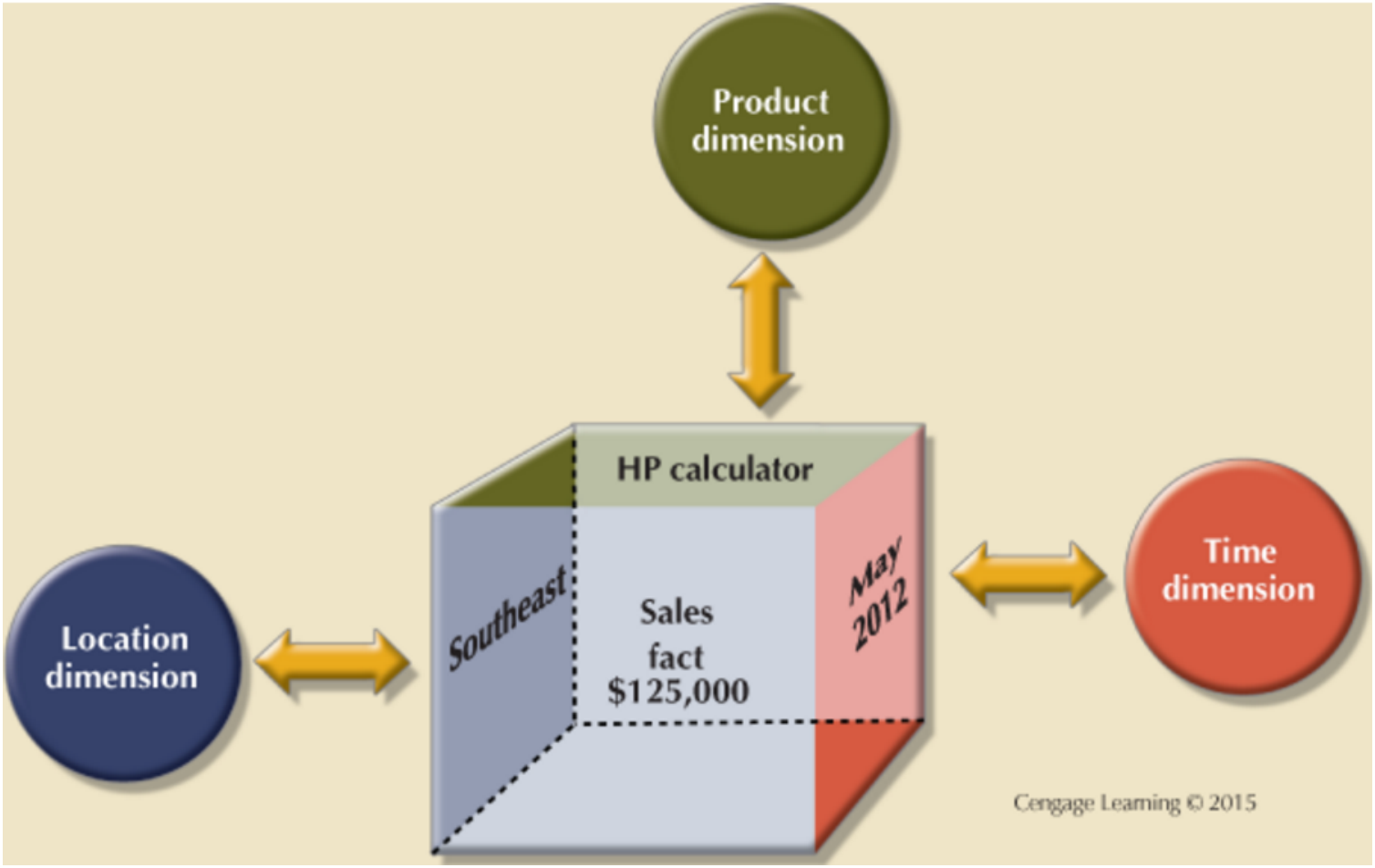
However

- Creating a data warehouse requires time, money, and considerable managerial effort.
- Business managers may be reluctant to embrace this strategy.

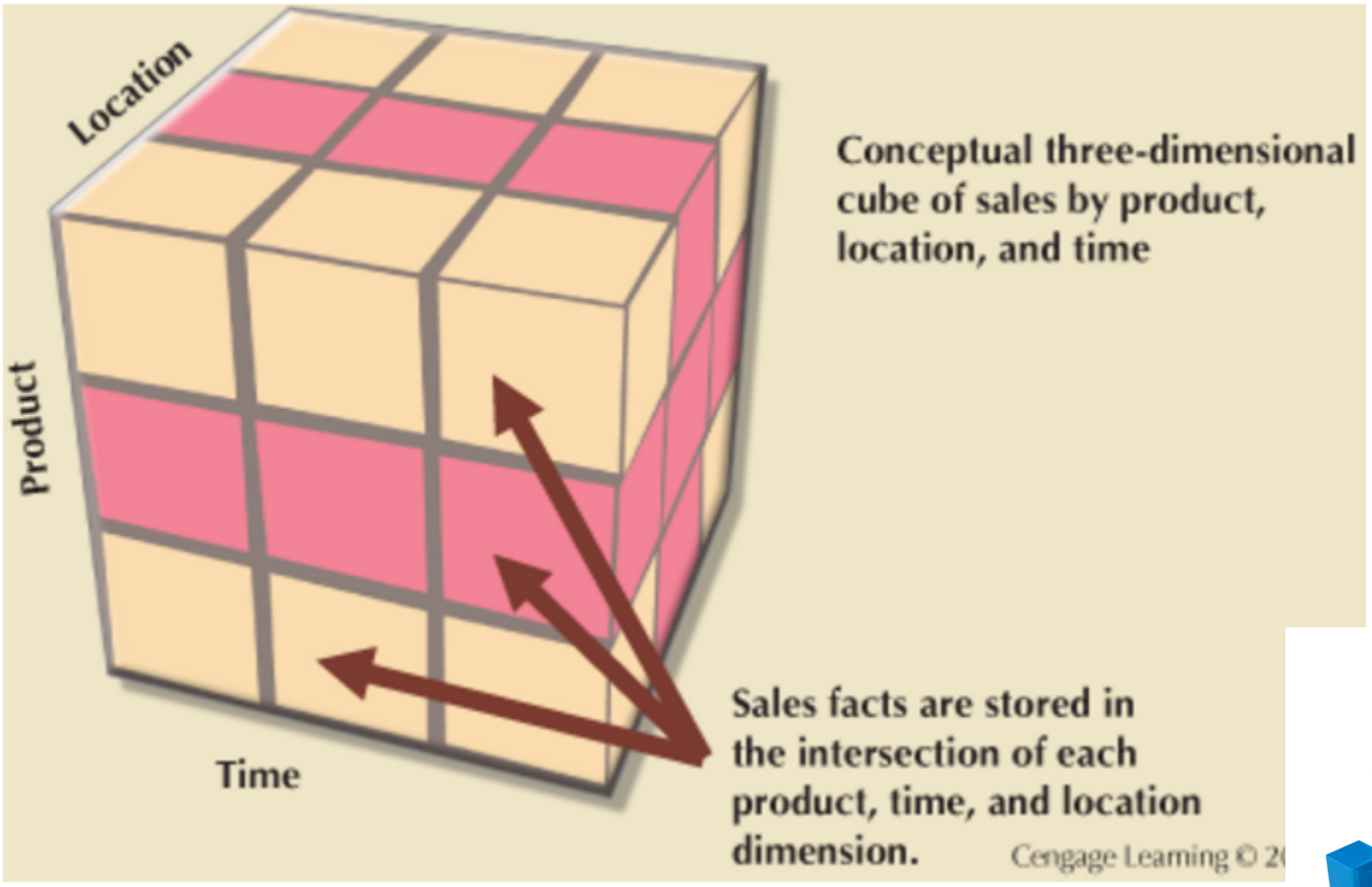
Star schema

- It is a data modeling technique used to map multidimensional data into a (relational) database. Simply put, it's data modeling method for data warehouse.
- It provides an easier way to query multidimensional data.
- It contains a set of fact tables and dimension tables.

A simple star schema



Three dimensional view of sales



Facts table

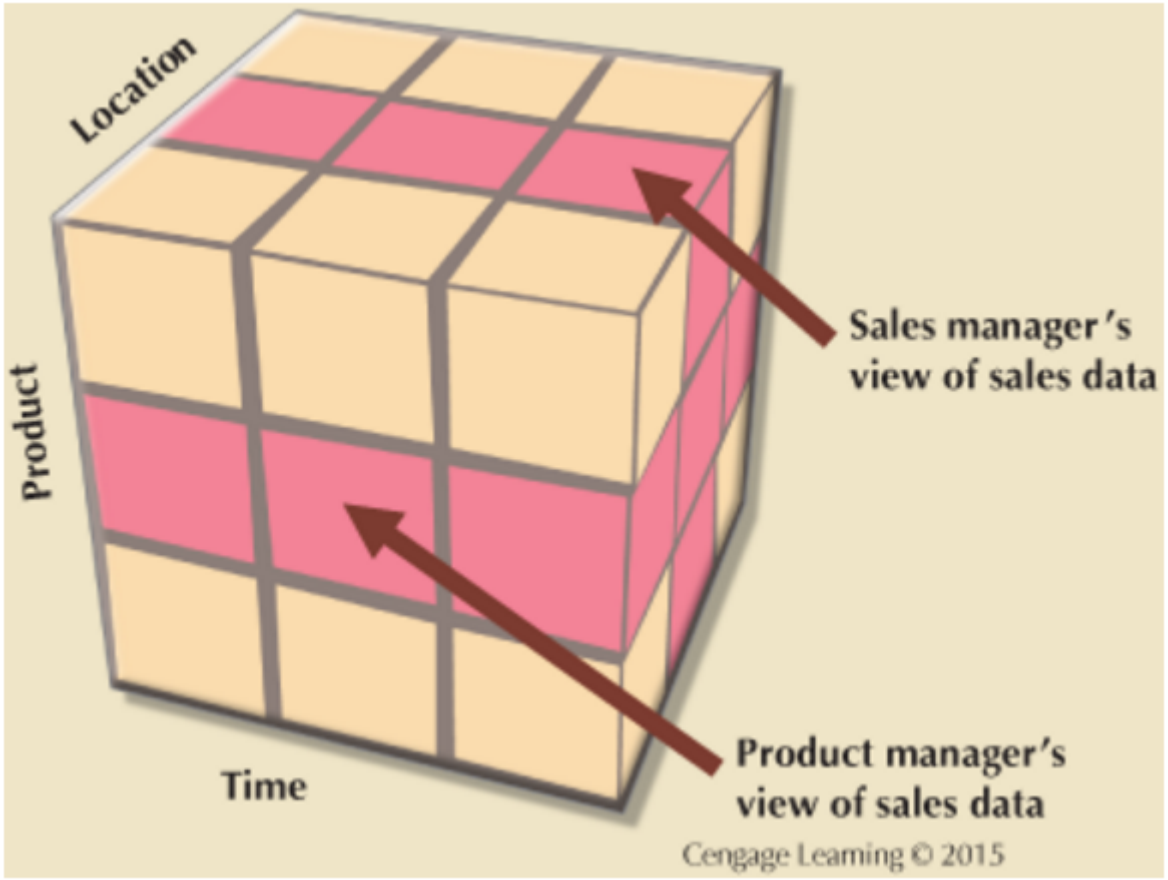
- Fact is a numeric measurement/value that represent a specific business activity. Facts are usually units, costs, prices, and revenues.
- For example, sales figures are numeric measurements that represent product sales.
- A fact table contains fact measurement and linked dimensions (via PK/FK)

Dimensions

- Dimensions are characteristics that provide additional perspectives to a given fact.
- A dimension table contains dimension attributes. For example, a location dimension table contains region, state, city attributes.
- Multiple dimensions: time, product, location, sales representatives, etc

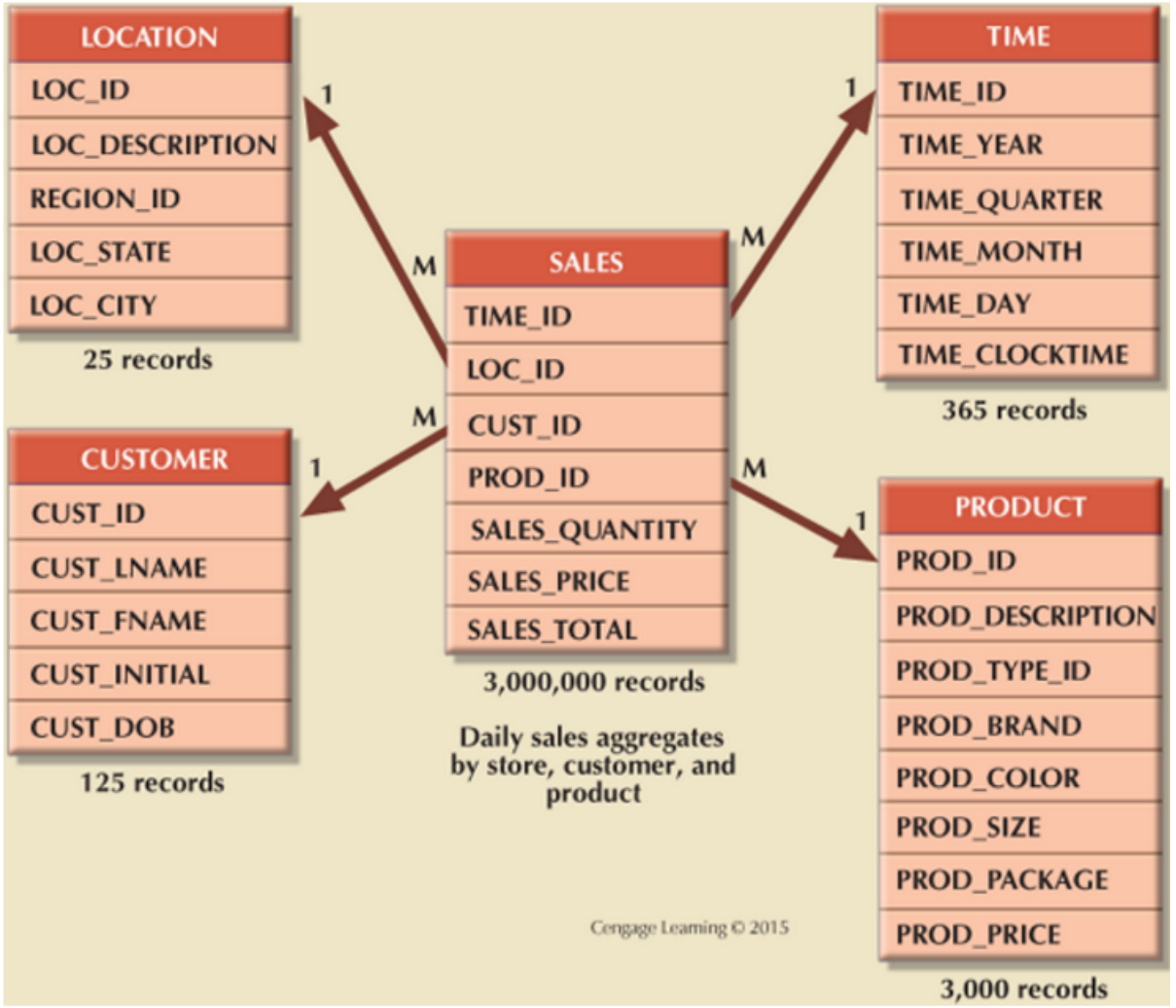
- What are the fact and dimensions of student?

Why is data retrieval and analysis easier?

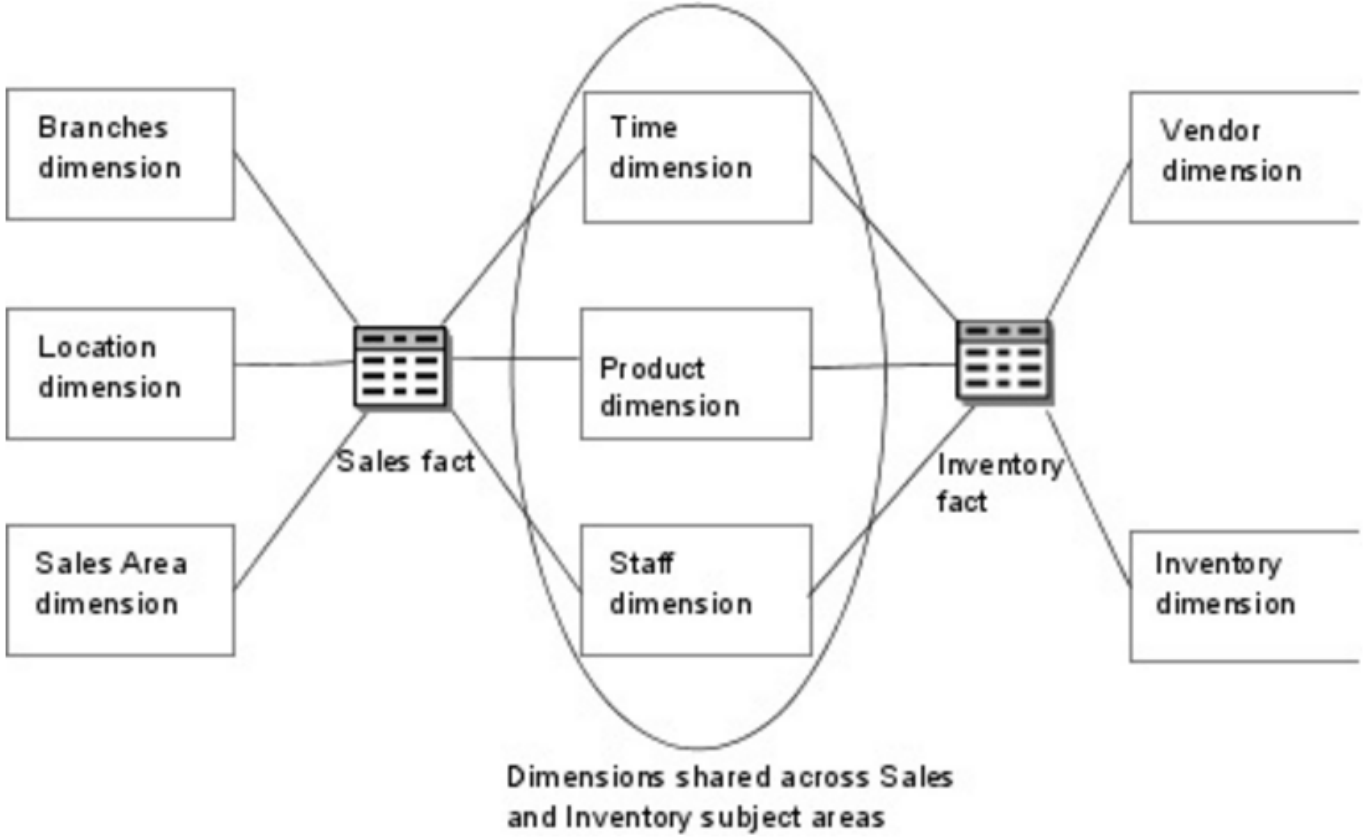


Slice-and-dice operation

More complete star schema for sales



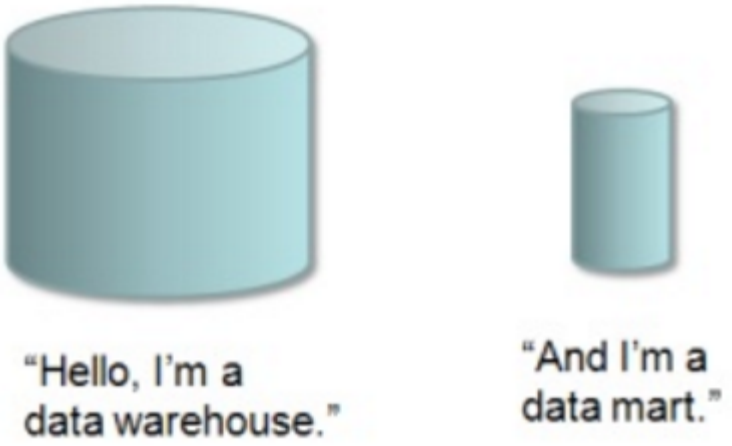
Multiple fact tables with conforming dimensions (as marked)



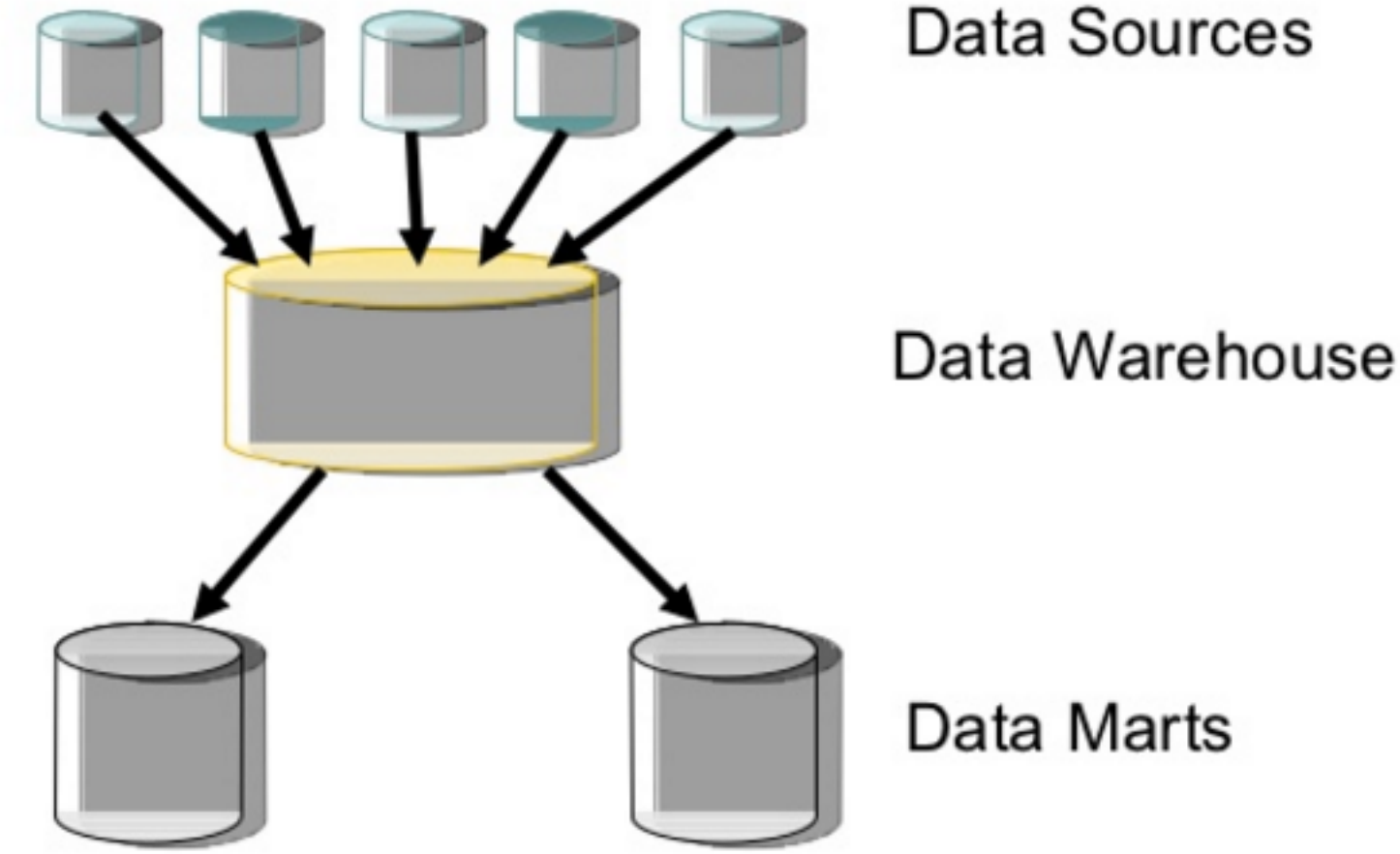
Data warehouse data vs. operational database data

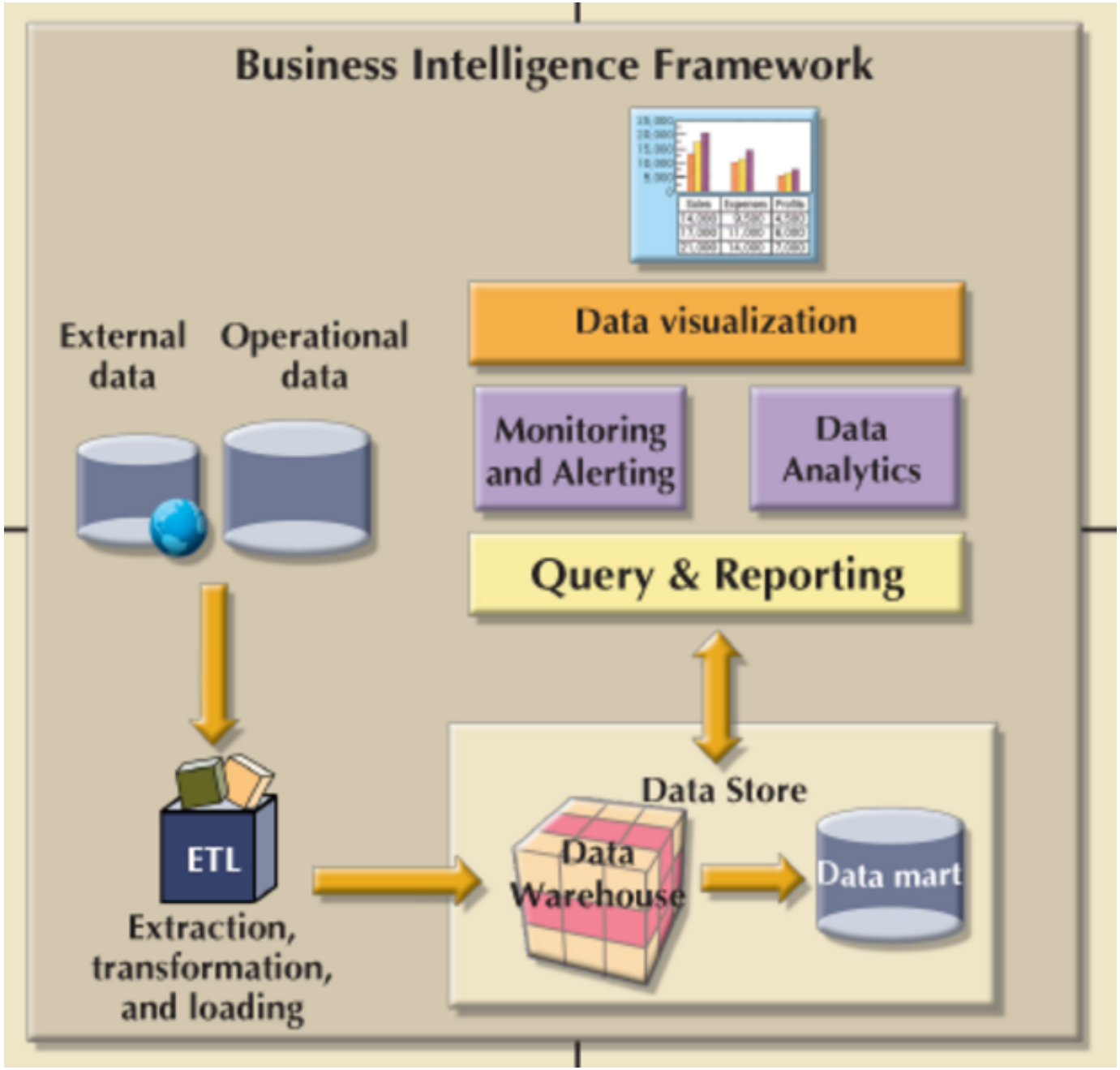
	Data warehouse data	Operational Data
Time span	Long time frame. E.g. the sales during the last month	Short time frame. One transaction for example.
Granularity	Different levels of aggregation. E.g. by region, by city, by store.	No aggregation
Dimensionality	High dimension data. E.g. how product X fared relative to product Y during the past month by region, state, city, store.	Single, individual transaction

Data mart vs data warehouse



The data mart is a subset of the data warehouse which is usually oriented to a specific business area or team (Finance has their data mart, marketing has theirs, sales has theirs and so on).


Data Sources
Data Warehouse
Data Marts


Business Intelligence Framework
Data visualization
External data
Operational data
Monitoring and Alerting
Data Analytics
Query & Reporting
Data Store
ETL
Data Warehouse
Data mart
Extraction, transformation, and loading

Summary

- Business Intelligence
- Data warehouse
- Learning objective
  - Difference between operational database and data warehouse
  - Star schema