



COMP8811 Week 2 Day 1

Data Warehousing and Dimensional Modelling

Neda Sakhaee

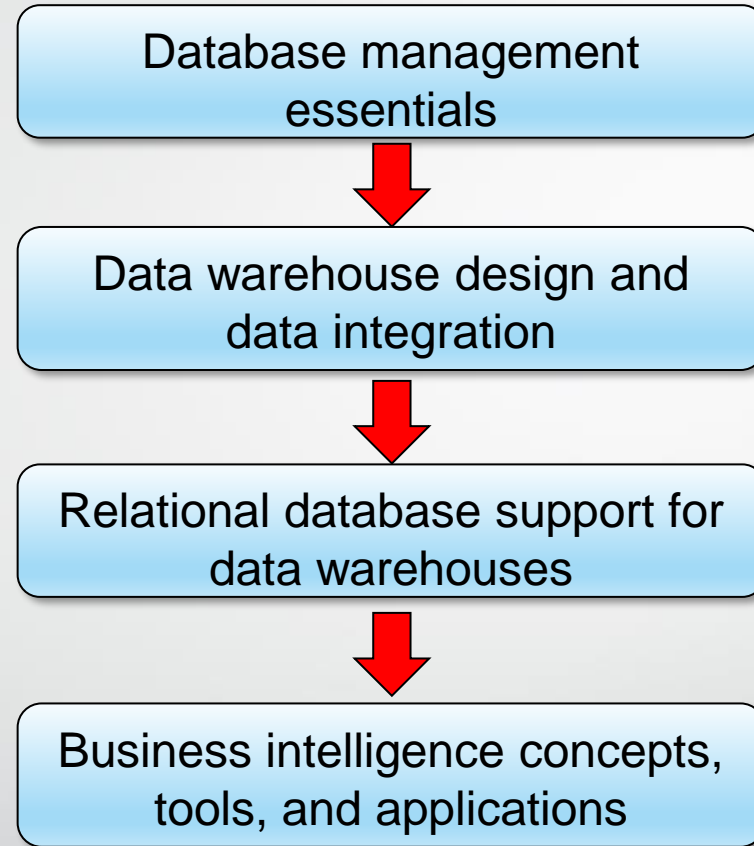
April 2024

Assignment 1

Criteria	Mark
Presentation & Discussion	
- Presentation organization and material	4
- Discussion on findings	3
- Time management	3

Optional
Presentations
15 minutes

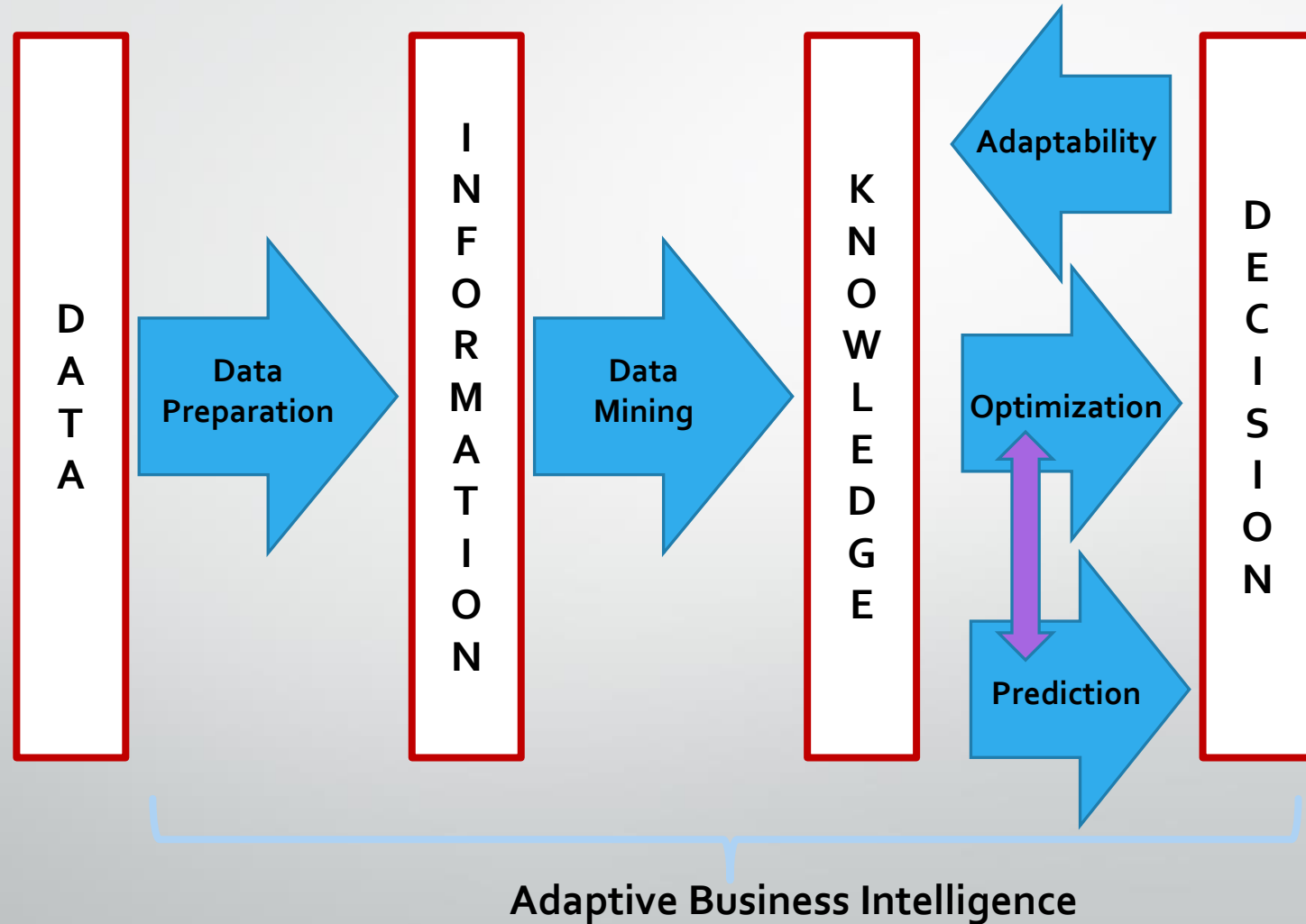
Data Warehousing for Business Intelligence



Motivation

- Databases are crucial for daily operations and decision making in organizations
- Database management technology
 - Major part of software industry
 - Revolutionary evolution over 40 years
 - Foundation for management of long term memory of organizations
- Vibrant field with employment opportunities

Data Analytics and Intelligence



Initial Vocabulary

- Data: raw facts about things and events
- Information: transformed data that has value for decision making
- Essential to organize data for retrieval and maintenance

Database Characteristics



Persistent

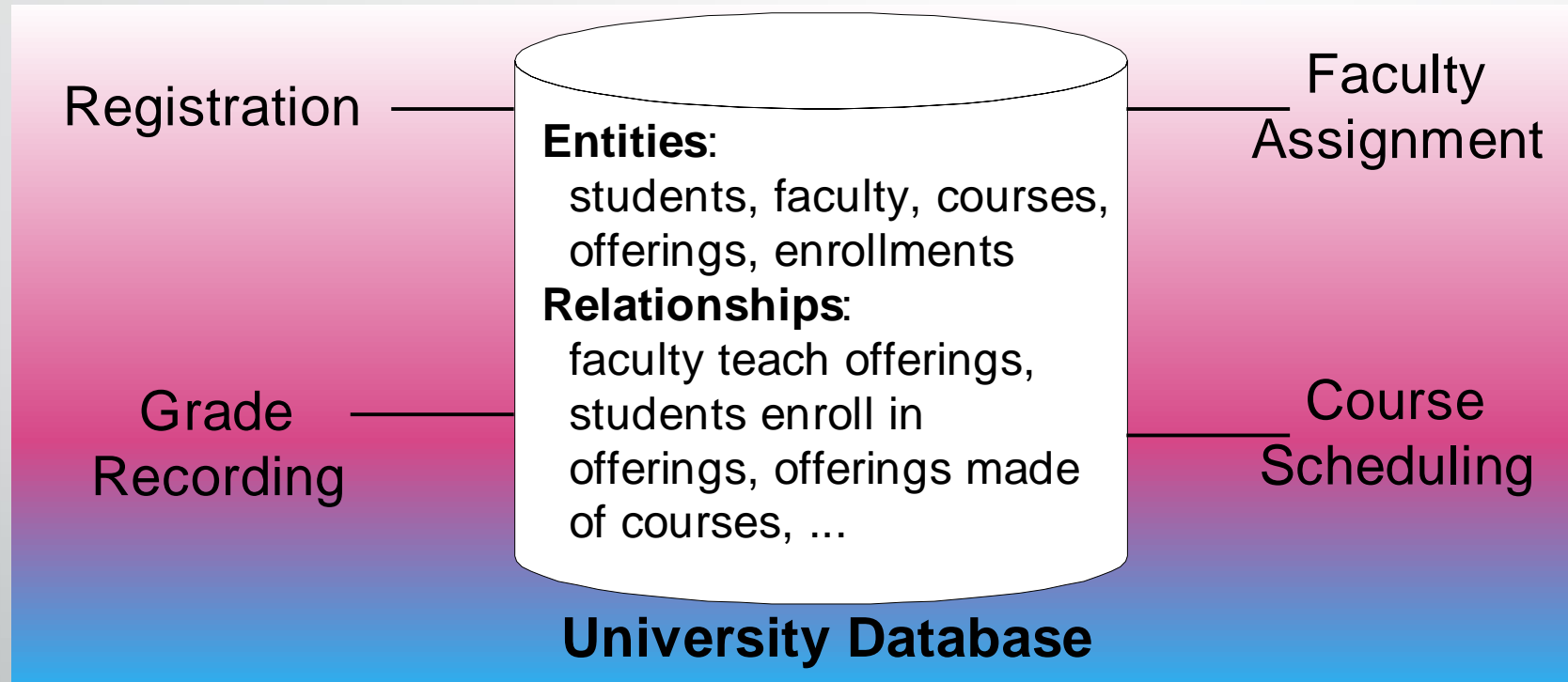


Inter-related

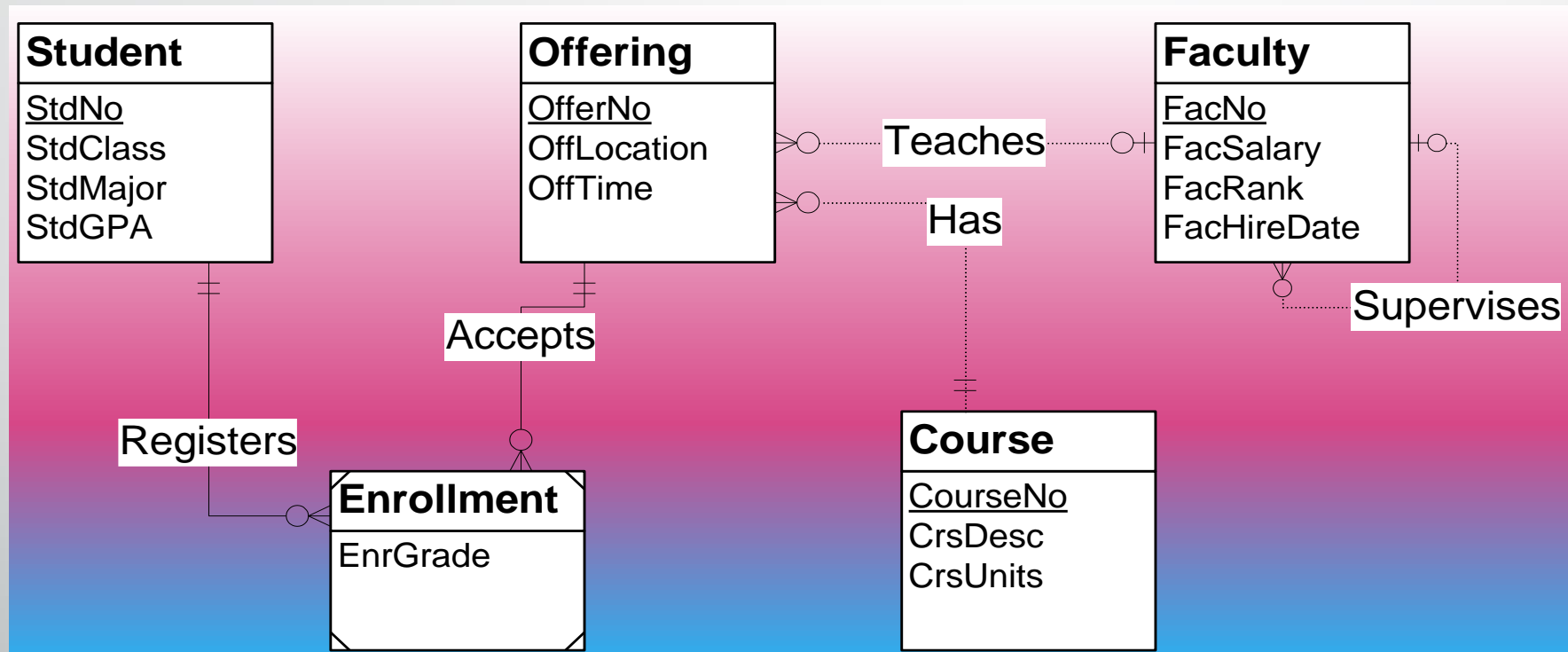


Shared

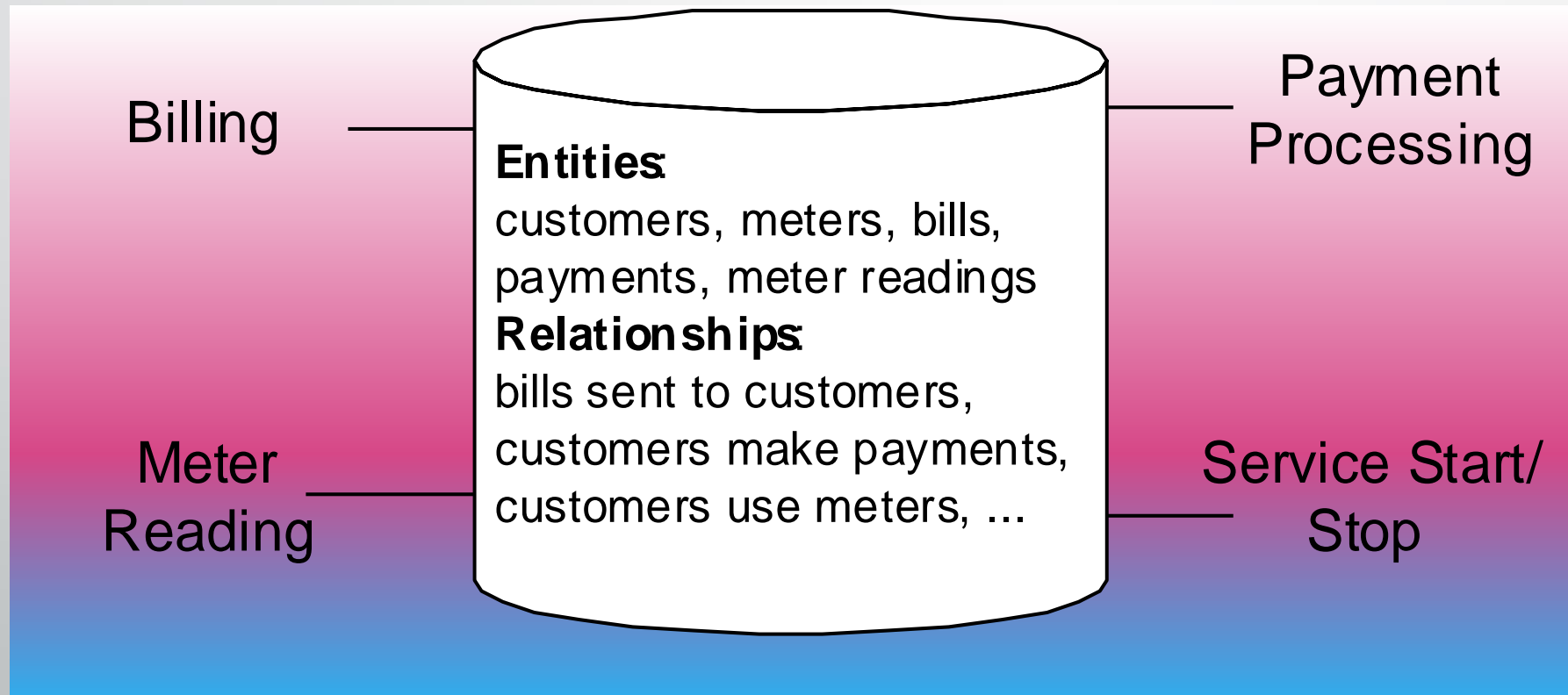
University Database



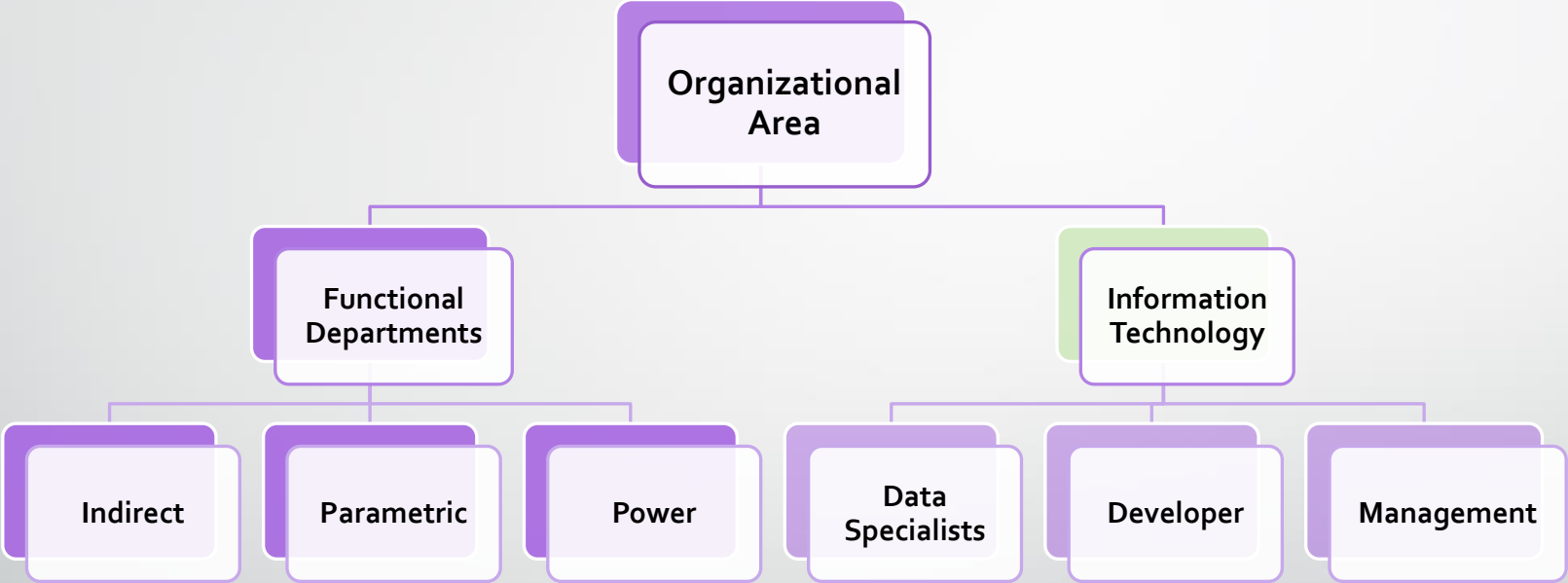
University Database (ERD)



Water Utility Database



Organizational Roles



Database Specialists

- Database administrator (DBA)
 - More technical
 - DBMS specific skills
- Data administrator
 - Less technical
 - Planning role

Summary

- Databases and database technology vital to modern organizations
- Database technology supports daily operations and decision making
- Emphasize structured data
- Essential characteristics of shared, inter-related, and persistent
- Active working with database technology as developer, data specialist, or power user
- Many opportunities to work with databases

List of clues:

1. What does ETL stand for, and what are the three main steps involved?
2. Name two common types of data sources used in data warehousing.
3. Define the term "dimension" in the context of data warehousing.
4. What is a surrogate key, and why is it used in data warehousing?
5. Explain the difference between a star schema and a snowflake schema.
6. What is meant by "data aggregation" in data warehousing?
7. Name two popular tools used for data extraction in ETL processes.
8. Define the term "fact table" in the context of data warehousing.
9. What is the purpose of a slowly changing dimension (SCD) in data warehousing?
10. Explain the concept of "data denormalization" and its benefits.
11. Name two techniques for handling data quality issues in data warehousing.
12. What role does a data warehouse play in business intelligence?
13. Define the term "data mart" and its relationship to a data warehouse.
14. What is the difference between OLAP and OLTP databases?
15. Explain the concept of "surrogate key" and provide an example.
16. How does data warehousing contribute to historical trend analysis?
17. Define the term "star schema" and provide a diagram of its structure.
18. Name two common challenges in designing a data warehouse.
19. Explain the process of data transformation in ETL.
20. What is the purpose of an ETL tool in data warehousing?
21. Define the term "ETL process" and provide an overview of its stages.
22. How can data partitioning improve the performance of a data warehouse?
23. Explain the difference between "full load" and "incremental load" in ETL.
24. Name two benefits of using surrogate keys in data warehousing.
25. What is the role of a "data steward" in maintaining data quality in a data warehouse?

Data Detective Class Activity

4 people in a team

You will be given a CLUE to solve!



Detectives: you will lead the discussion and guide your team to solve the clues. You will work together to figure out the answers based on your collective knowledge.



Scribe: you will write down the answers as your group comes up with them.



Presenter: Once your clue is solved, you will prepare a short presentation summarizing your answers and explaining the reasoning behind each one.



Timekeeper: You will keep time and notify your team at any stage.

Assignment 1

Criteria	Mark
Presentation & Discussion	
- Presentation organization and material	4
- Discussion on findings	3
- Time management	3

Optional
Presentations
15 minutes

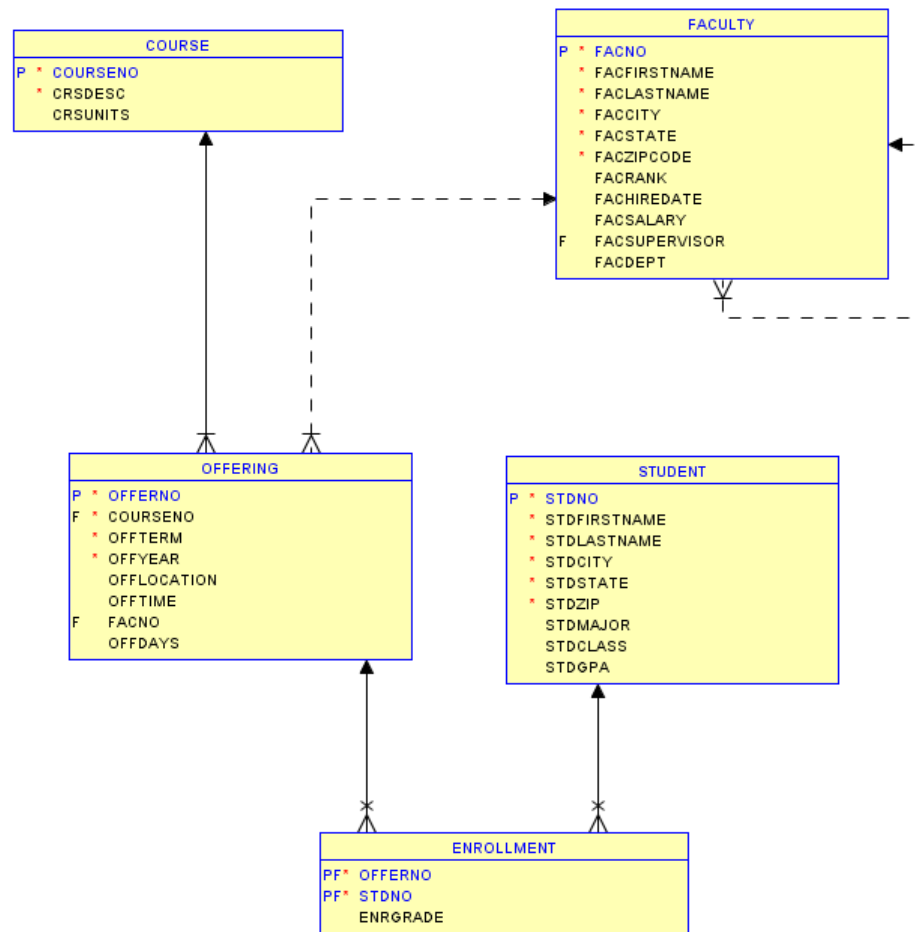
Break 10 min



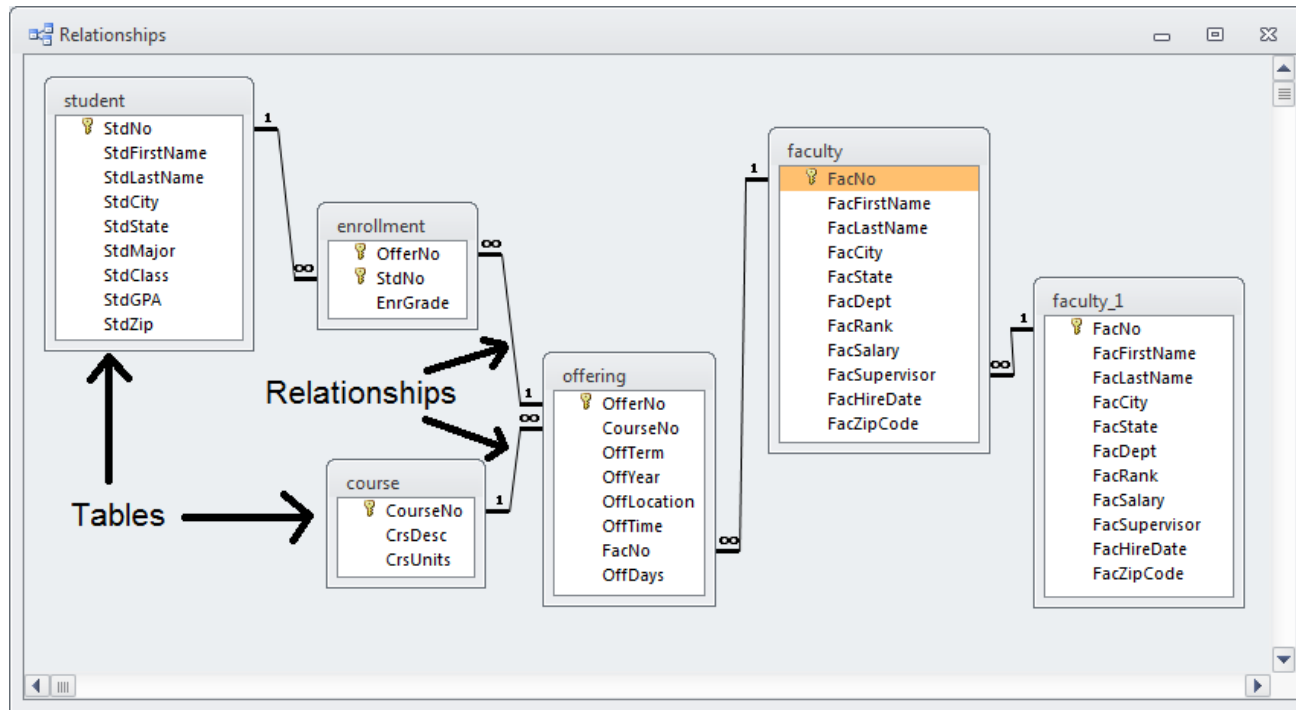
Database Management System (DBMS)

- Collection of components that support data acquisition, dissemination, storage, maintenance, retrieval, and formatting
- Product variations
 - Enterprise DBMSs
 - Desktop DBMSs
 - Embedded DBMSs
- Major part of information technology infrastructure

Oracle Relational Diagram



Microsoft Access Database Diagram



Transaction Definition

- Supports daily operations of an organization
- Collection of database operations
- Reliably and efficiently processed as one unit of work
- No lost data
 - Interference among multiple users
 - Failures

Airline Transaction Example

- START TRANSACTION
 - Display greeting
 - Get reservation preferences from user
 - SELECT departure and return flight records
 - If reservation is acceptable then
 - UPDATE seats remaining of departure flight record
 - UPDATE seats remaining of return flight record
 - INSERT reservation record
 - Print ticket if requested
 - End If
 - On Error: ROLLBACK
- COMMIT

ATM Transaction Example

- START TRANSACTION
 - Display greeting
 - Get account number, pin, type, and amount
 - SELECT account number, type, and balance
 - If balance is sufficient then
 - UPDATE account by posting debit
 - UPDATE account by posting debit
 - INSERT history record
 - Display message and dispense cash
 - Print receipt if requested
 - End If
 - On Error: ROLLBACK
- COMMIT

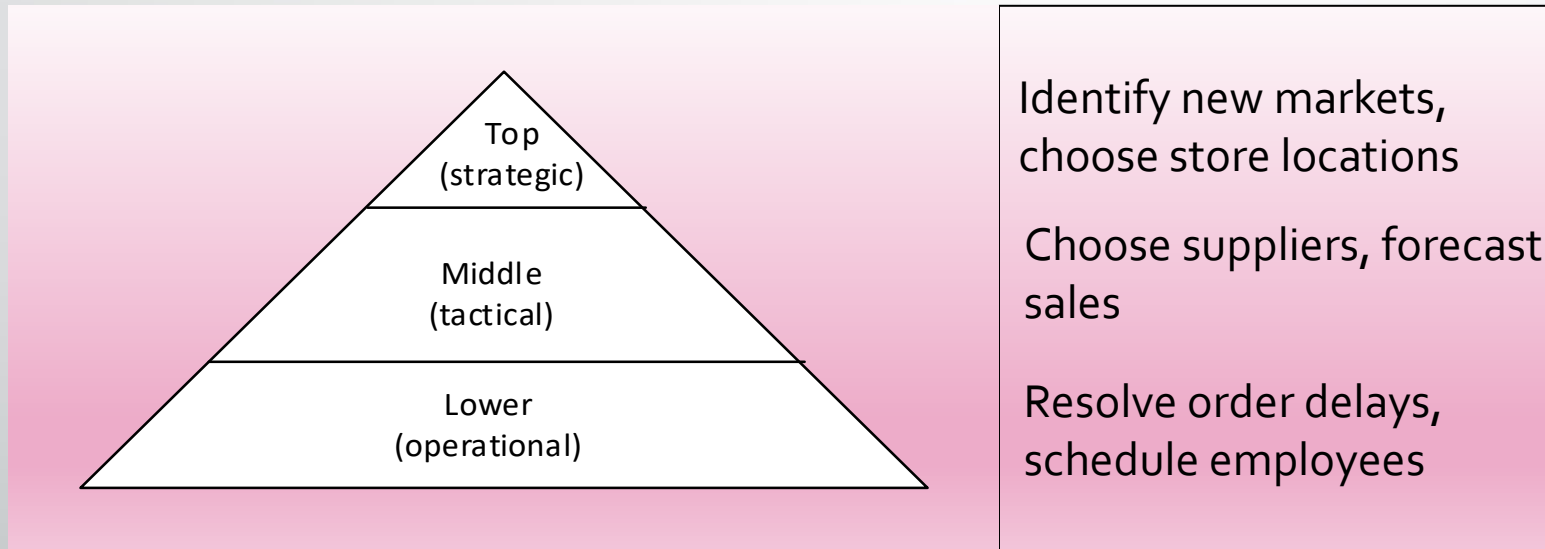
Transaction Processing

- Reliable and efficient processing of transactions
 - Control simultaneous users
 - Recover from failures
- Internal features for enterprise DBMSs
 - Concurrency control manager
 - Recovery manager (data cannot be lost and recovery is possible)
 - Transparent services for application developers

Decision Making Hierarchy

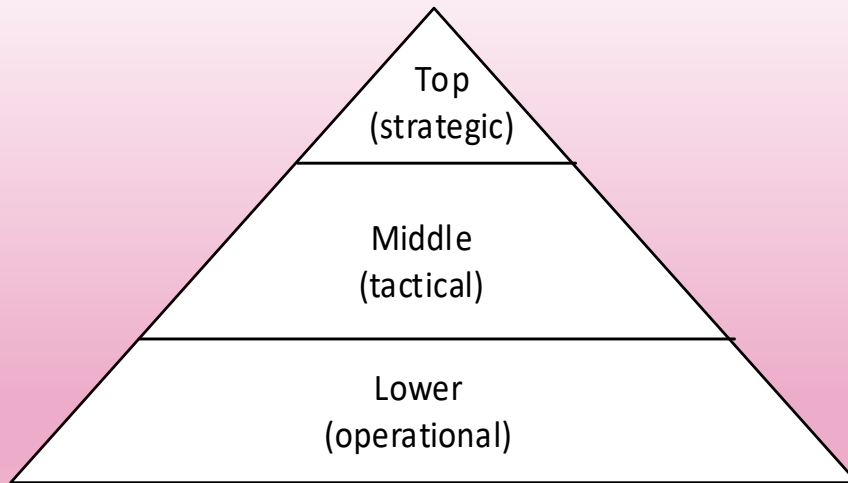
Decision making hierarchy

Typical decisions

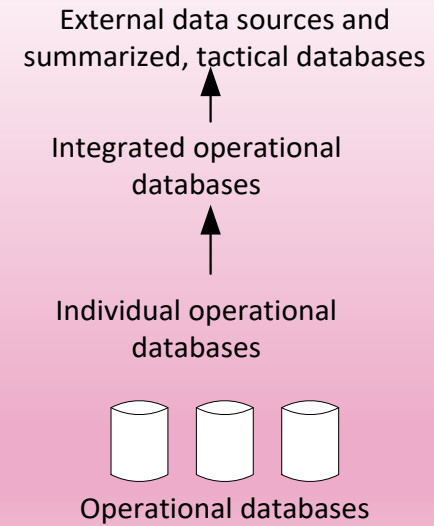


Database Support

Decision making hierarchy



Database support



Data Warehouse Characteristics

- Essential part of infrastructure for business intelligence
- Logically centralized repository for decision making
 - Populated from operational databases and external data sources
 - Integrated and transformed data
 - Optimized for reporting

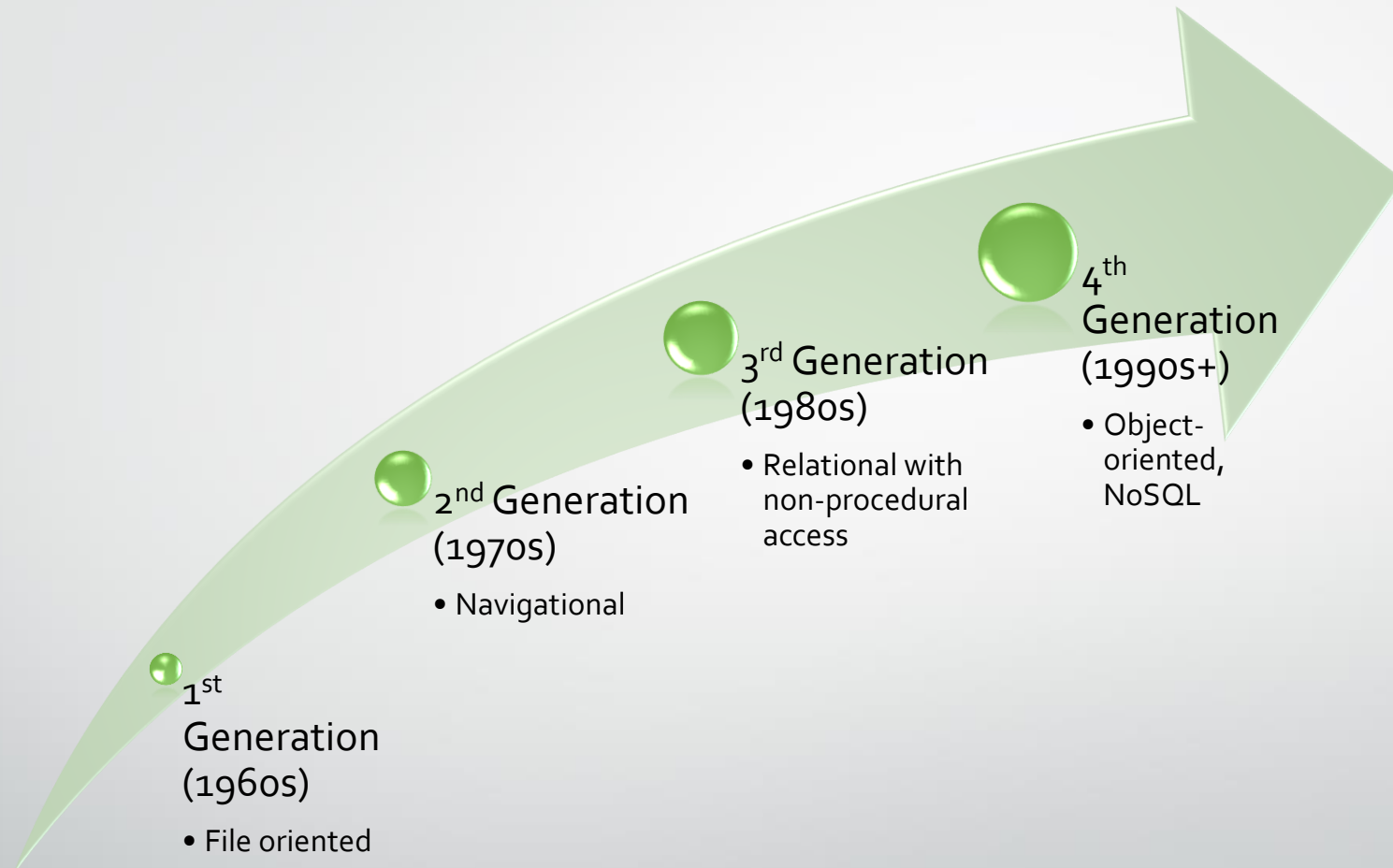
Comparison of Environments

- Transaction processing
 - Primary data in operational databases
 - Large volumes of transactions with relatively small amounts of data per transaction
 - Some reporting requirements for operations
- Business intelligence processing
 - Secondary data from operational databases
 - Substantial processing for transformations and integration
 - Large volumes of data for reporting

Summary

- Data warehouse processing supports tactical and strategic decision making
- Different DBMS features for business intelligence support

DBMS Product Generations



Recent Database Technology Developments

- Business intelligence processing
 - Data integration
 - Storage/retrieval of summary data
- Cloud computing
 - No fixed costs of ownership
 - Data and software
- Optimization for big data demands
 - Demands from smart phones, automotive technology, RFID tags, digitized media
 - NoSQL: simplified models for high performance

DBMS Marketplace

- Enterprise DBMS
 - Oracle: dominates in Unix; strong in Windows
 - SQL Server: strong in Windows
 - DB2: strong in MVS and VM environments
 - Teradata: usage as a data warehouse platform
 - Amazon Web Services
 - SAP Sybase: possible challenge to Oracle
 - Significant open source DBMSs: MySQL, PostgreSQL, MongoDB, MariaDB, SQLite, Cassandra
 - Cloud-based and NoSQL: rapidly evolving
- Desktop DBMS
 - Access: dominates
 - LibreOffice Base, Open Office Base, FileMaker Pro

Summary

- Databases and database technology vital to modern organizations
- Remarkable product evolution
- Competitive industry with lots of continuing innovation

Relational Database Basics

- Collection of tables
- Heading: table name and column names
- Body: rows, occurrences of data

Student

StdNo	StdFirstName	StdLastName	StdCity	StdState	StdZip	StdMajor	StdClass	StdGPA
123-45-6789	HOMER	WELLS	SEATTLE	WA	98121-1111	IS	FR	3.00
124-56-7890	BOB	NORBERT	BOTHELL	WA	98011-2121	FIN	JR	2.70
234-56-7890	CANDY	KENDALL	TACOMA	WA	99042-3321	ACCT	JR	3.50

Sample Tables with Matching Values

Student

StdNo	StdFirstName	StdLastName	StdCity	StdState	StdZip	StdMajor	StdClass	StdGPA
123-45-6789	HOMER	WELLS	SEATTLE	WA	98121-1111	IS	FR	3.00
124-56-7890	BOB	NORBERT	BOTHELL	WA	98011-2121	FIN	JR	2.70
234-56-7890	CANDY	KENDALL	TACOMA	WA	99042-3321	ACCT	JR	3.50

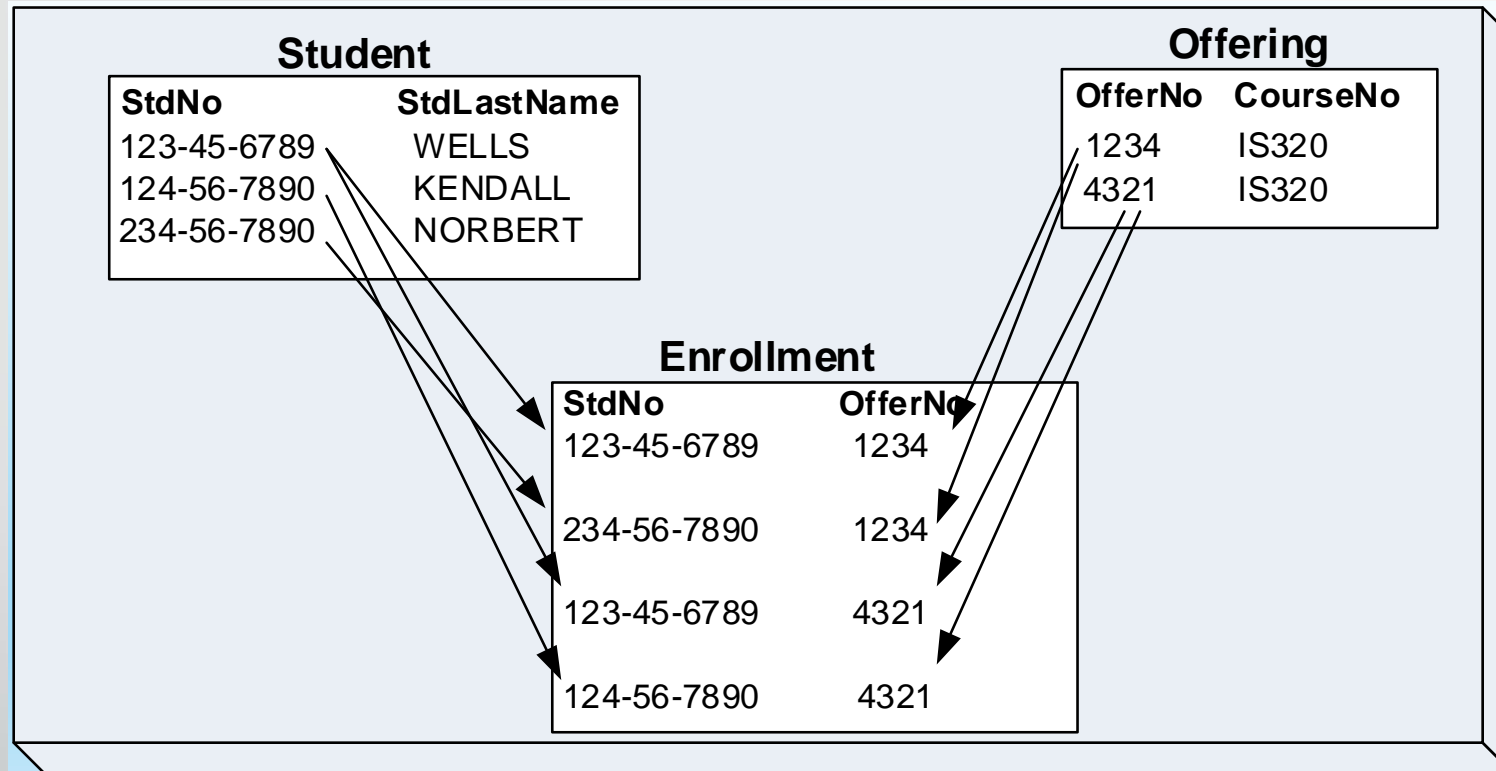
Offering

OfferNo	CourseNo	OffTerm	OffYear	OffLocation	OffTime	FacNo	OffDays
1111	IS320	SUMMER	2013	BLM302	10:30 AM		MW
1234	IS320	FALL	2012	BLM302	10:30 AM	098-76-5432	MW
4321	IS320	FALL	2012	BLM214	3:30 PM	098-76-5432	TTH

Enrollment

OfferNo	StdNo	EnrGrade
1234	123-45-6789	3.3
1234	234-56-7890	3.5
4321	123-45-6789	3.5
4321	124-56-7890	3.2

Graphical Depiction of Matching Values



Alternative Terminology

Table-Oriented	Set-Oriented	Record-Oriented
Table	Relation	Record Type, File
Row	Tuple	Record
Column	Attribute	Field

Definitions

Null value

- Absence of a value (missing value)
- Actual value unknown or not applicable for a row

Primary key (PK)

- Column or combination of columns with unique values in each row
- No extraneous columns (minimal)

Foreign key (FK)

- Column or combination of columns
- Related to a primary key in a related table
- Same data type and often same name as related PK

Integrity Rules

Entity Integrity

- Primary key for each table
- No missing (null) values for primary keys
- Ensures traceable entities

Referential Integrity

- Two kinds of values for a foreign key in a row
- Match a primary key value of a related table (usual)
- Null value (unusual)
- Ensures valid references among tables

Integrity Rule Violations

Student

<u>StdNo</u>	<u>StdLastName</u>
123-45-6789	WELLS
124-56-7890	KENDALL
234-56-7890	NORBERT
--	JONES

Offering

<u>OfferNo</u>	<u>CourseNo</u>
1234	IS320
4321	IS320

Enrollment

<u>StdNo</u>	<u>OfferNo</u>
123-45-6789	1234
234-56-7890	1234
123-45-6789	4321
124-56-7890	4321
234-56-7890	6789
--	4321

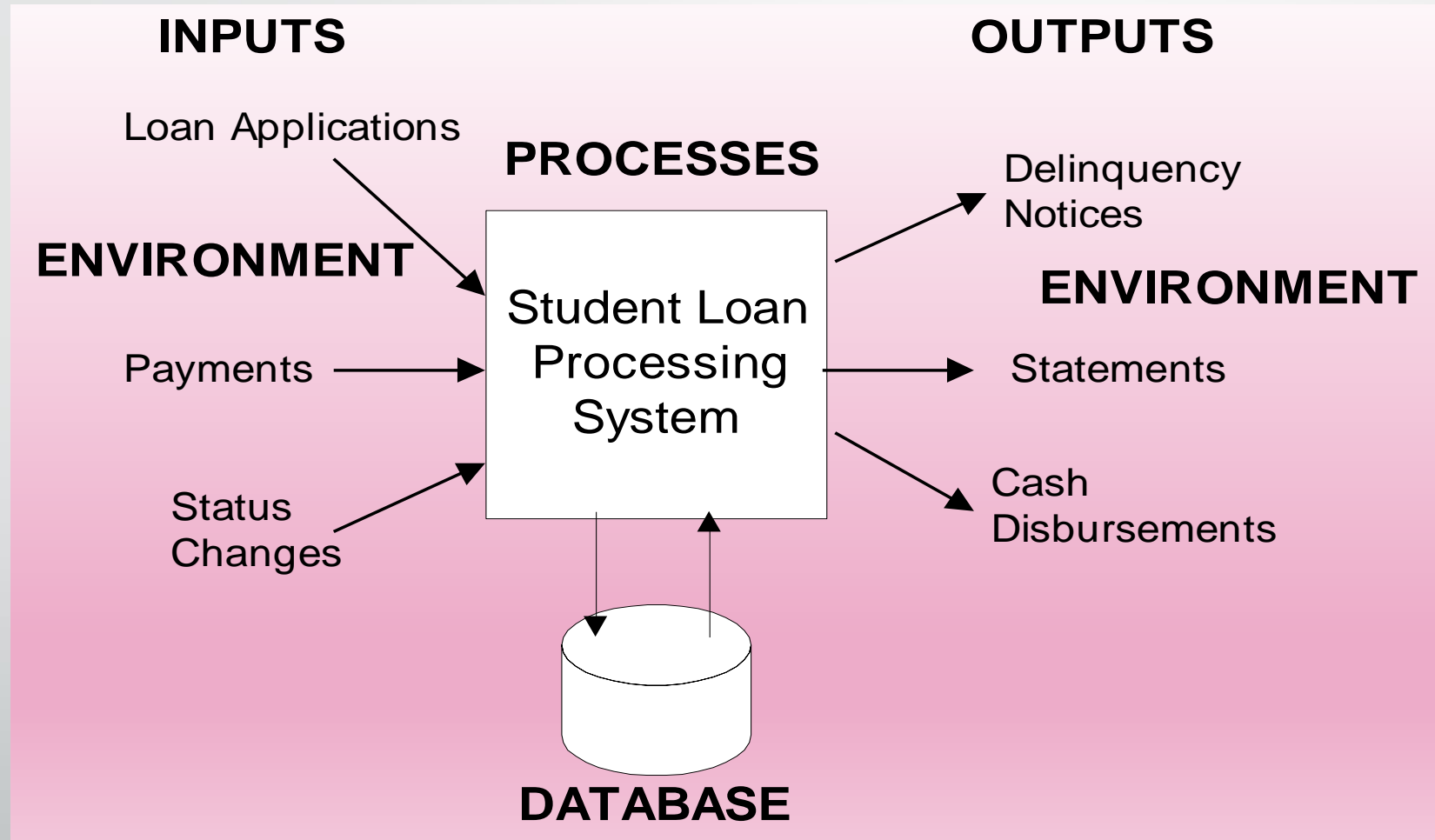
Summary

- Identify primary keys and foreign keys
- Visualize relationships
- Understanding existing databases is crucial to query formulation

Break 10 min



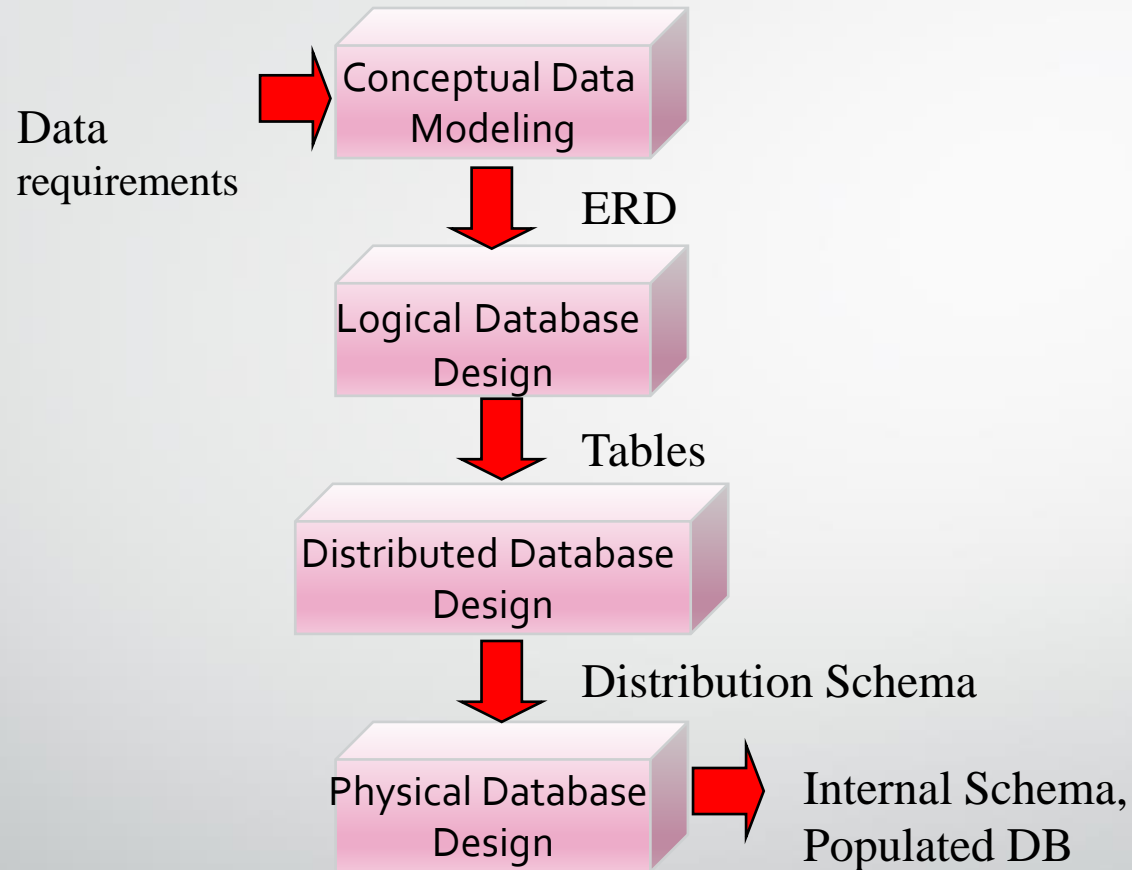
Information System



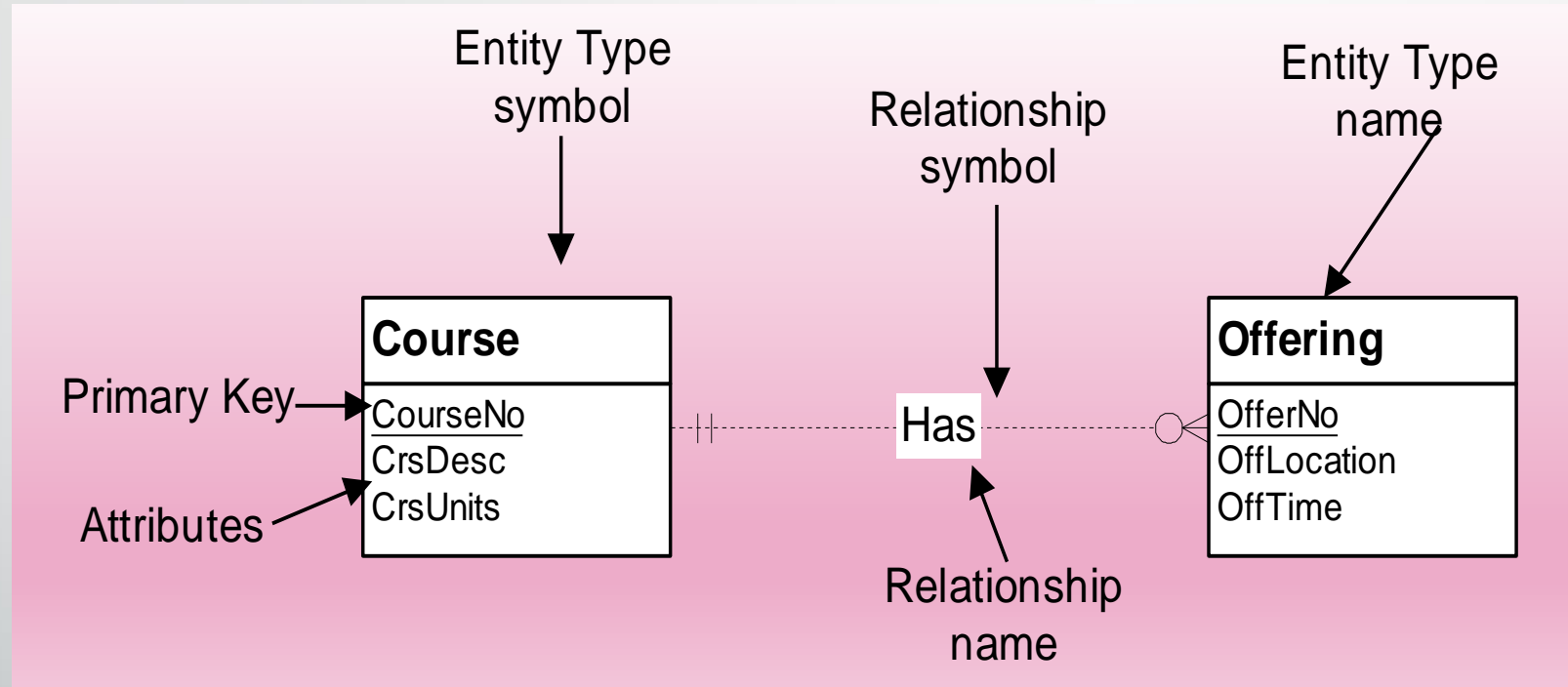
Broad Goals of Database Development

- 
- Develop a common vocabulary
 - Define business rules
 - Ensure data quality
 - Provide efficient implementation

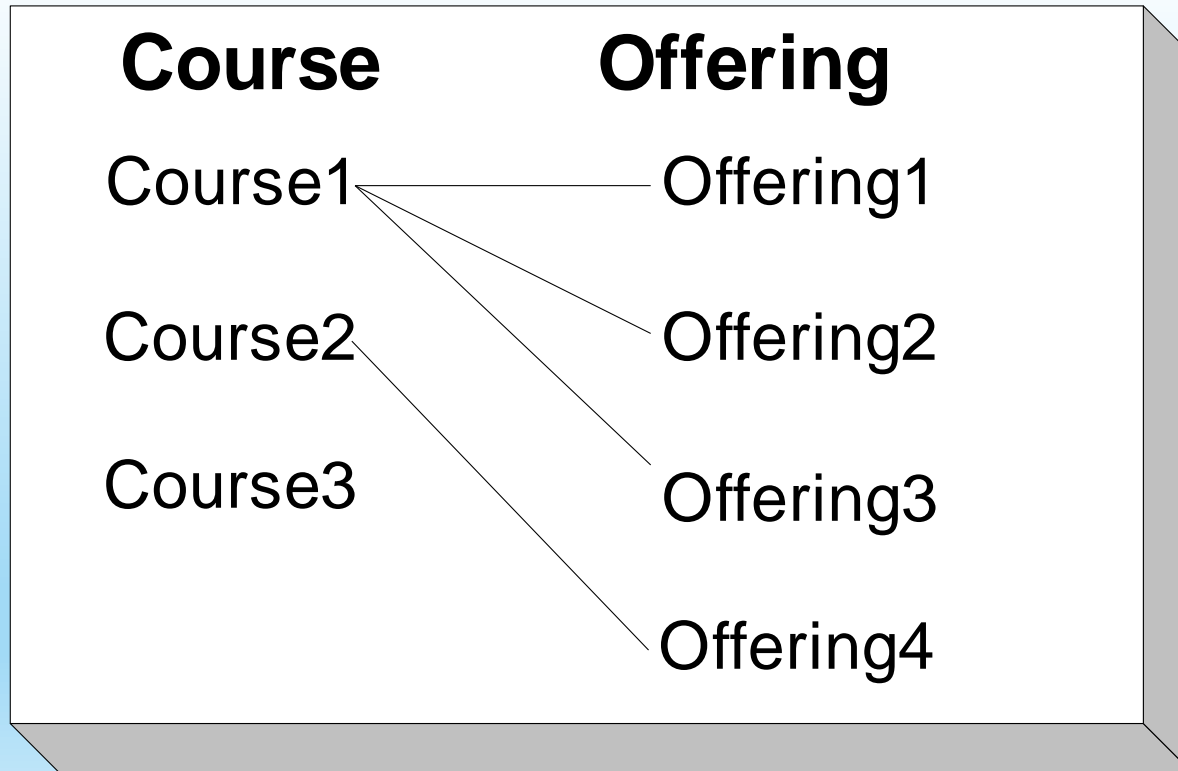
Database Development Phases



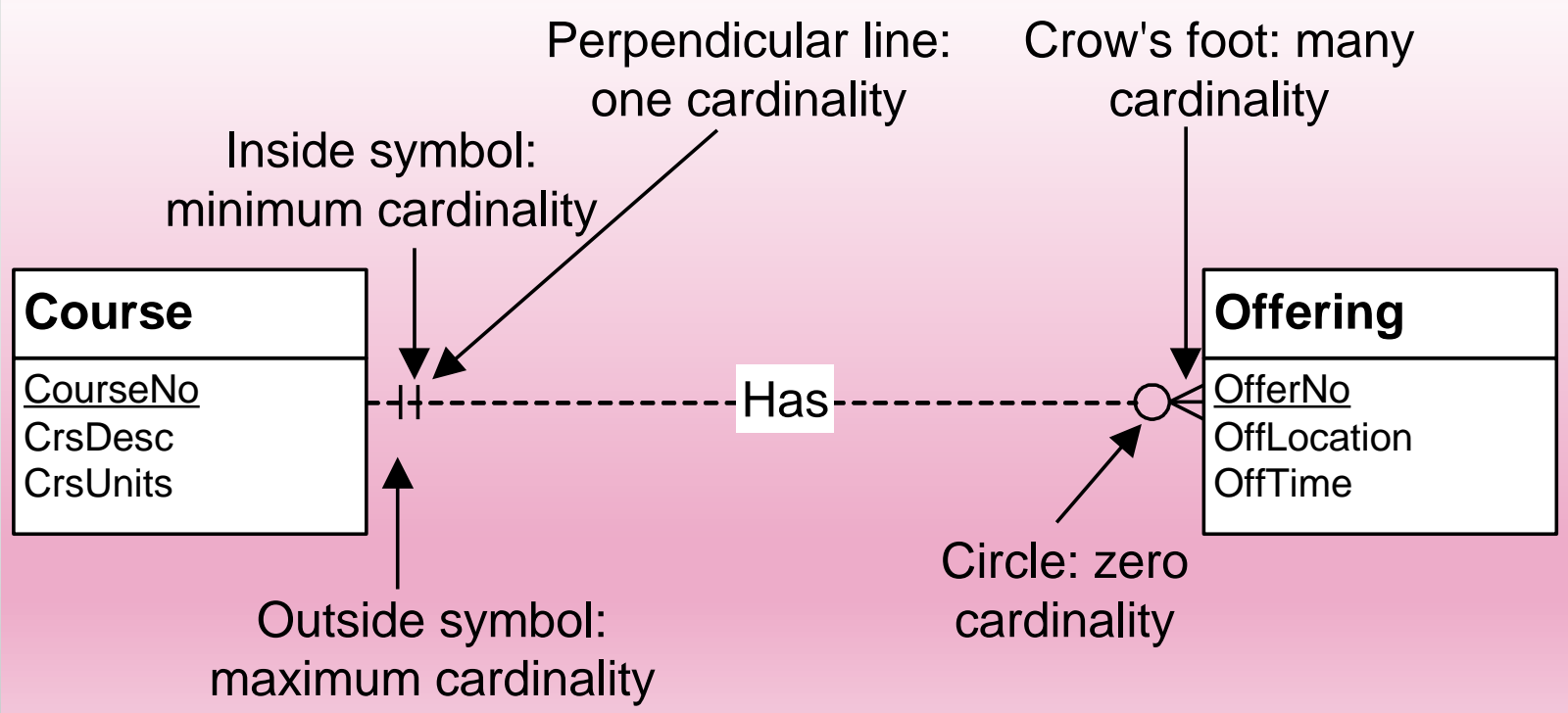
Basic Symbols



Cardinalities



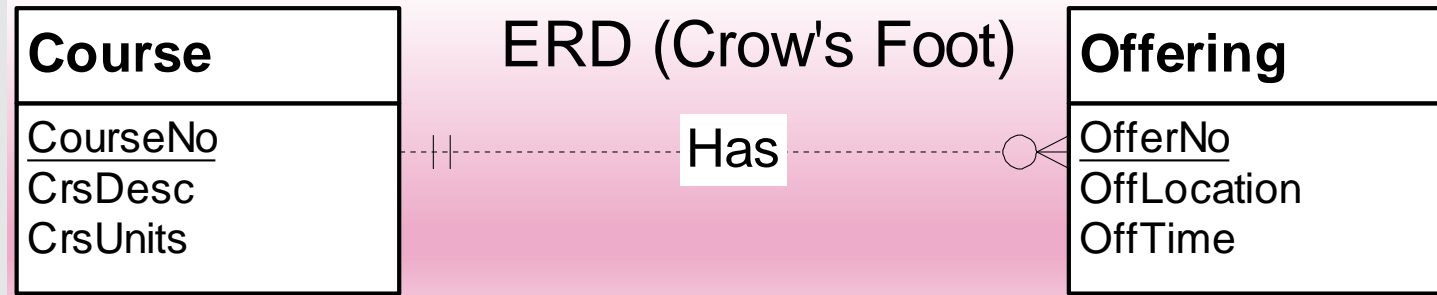
Cardinality Notation



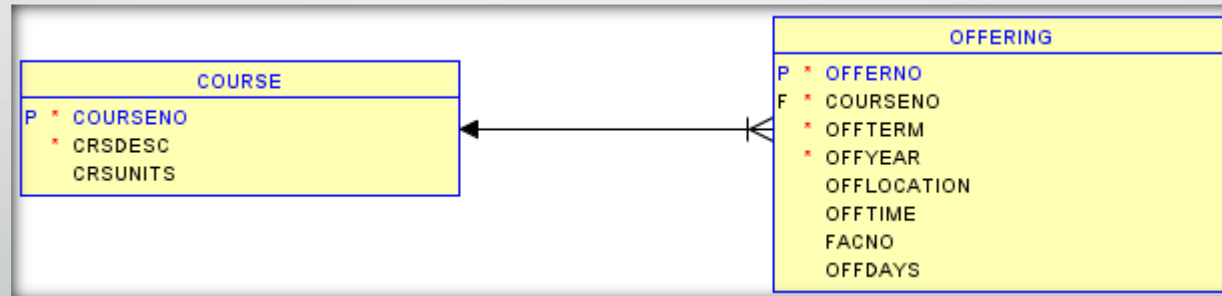
Important Cardinalities

Classification	Cardinality Restrictions
Mandatory	Minimum cardinality ≥ 1
Optional	Minimum cardinality = 0
Functional or single-valued	Minimum cardinality = 1
1-M	Maximum cardinality = 1 in one direction; maximum cardinality > 1 in the other direction
M-N	Maximum cardinality > 1 in both directions
1-1	Maximum cardinality = 1 in both directions

Comparison to Oracle Notation



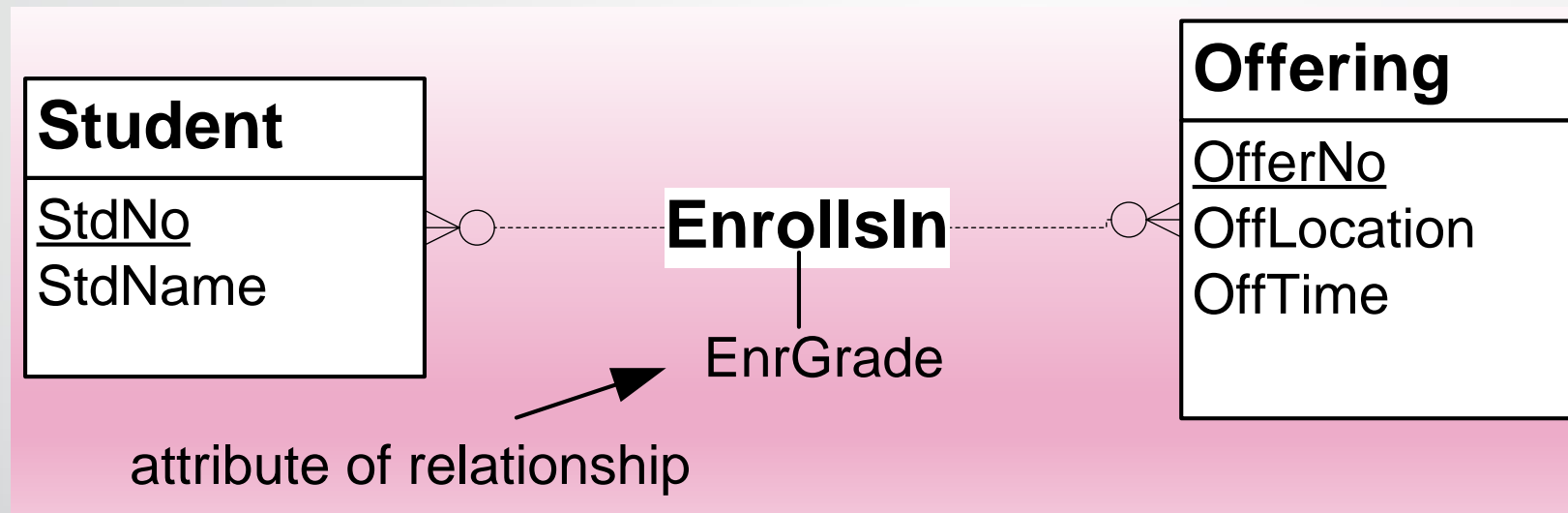
Oracle Relational Model Diagram



Summary

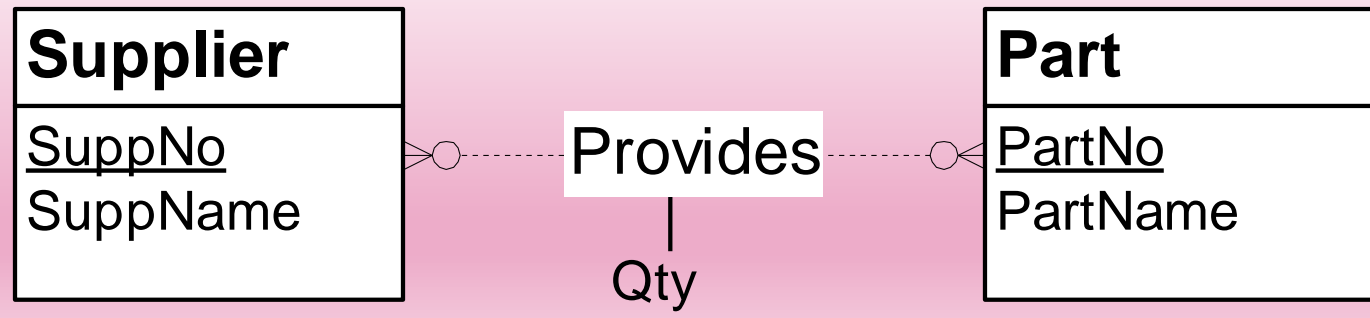
- Crow's Foot ERD notation is widely used
- Use notation precisely
- Differentiate ERD notation from Relational Data Model
- Understanding the ERD notation is a prerequisite to applying the notation on business problems

M-N Relationships with Attributes

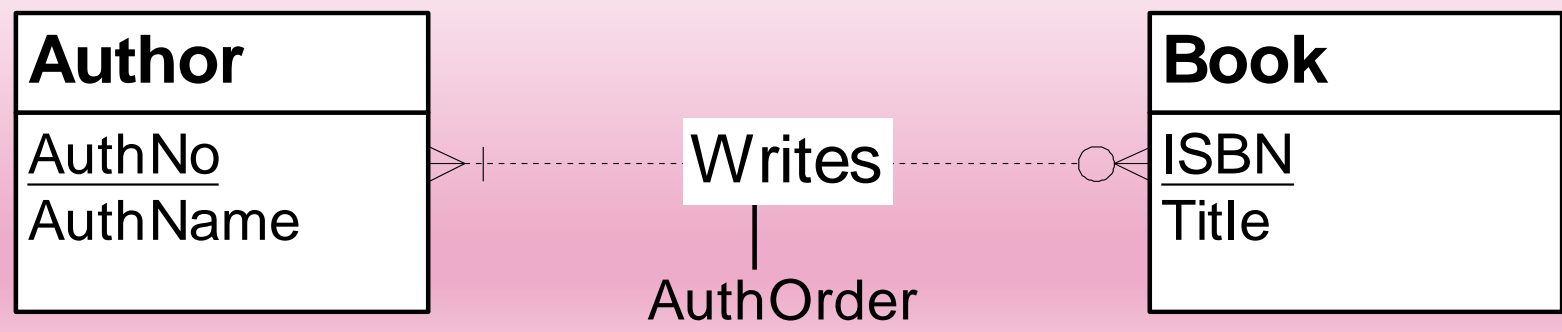


M-N Relationships with Attributes (II)

a) *Provides* relationship



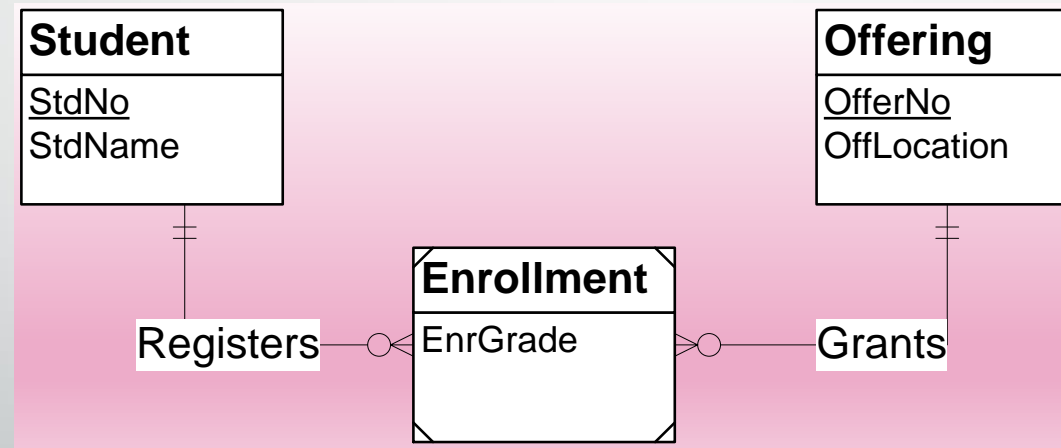
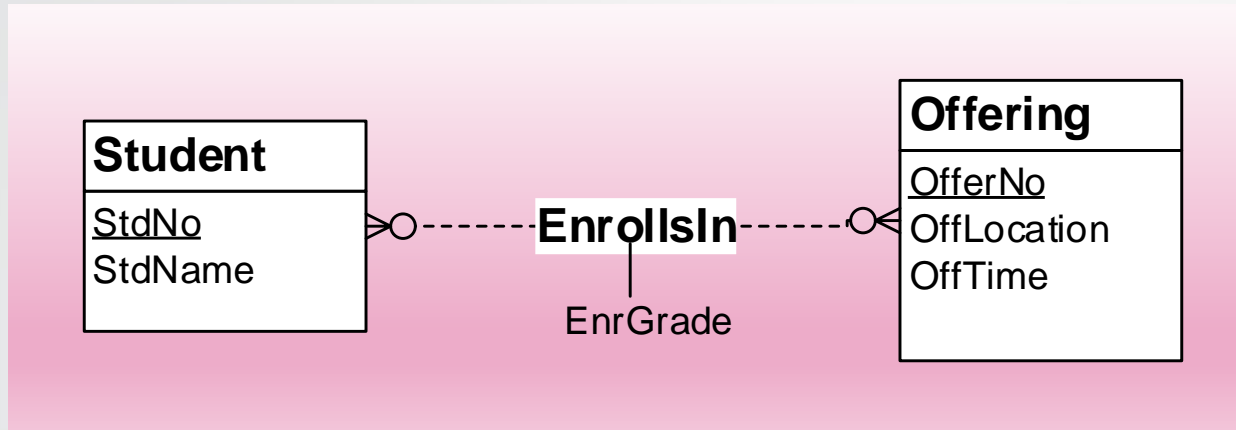
b) *Writes* relationship



M-N Relationship Equivalency Rule

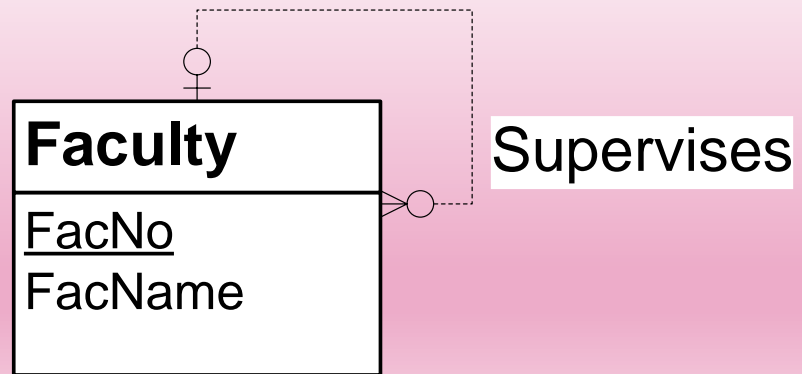
- Replace M-N relationship
 - Associative entity type
 - Two identifying 1-M relationships
- M-N relationship versus associative entity type
 - Largely preference
 - Associative entity type is more flexible in some situations

Relationship Equivalency Example

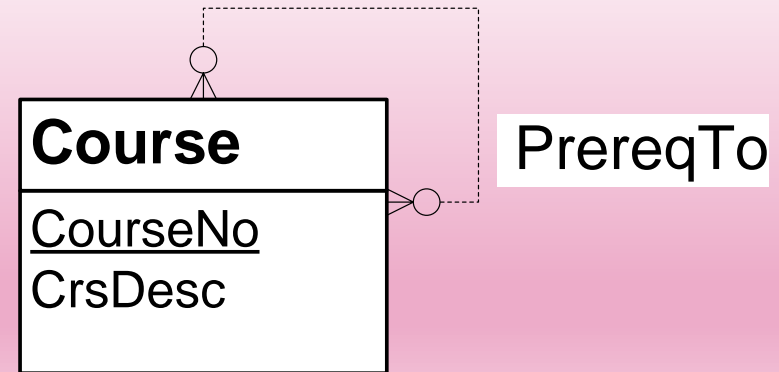


ERD Notation for Self-Referencing Relationships

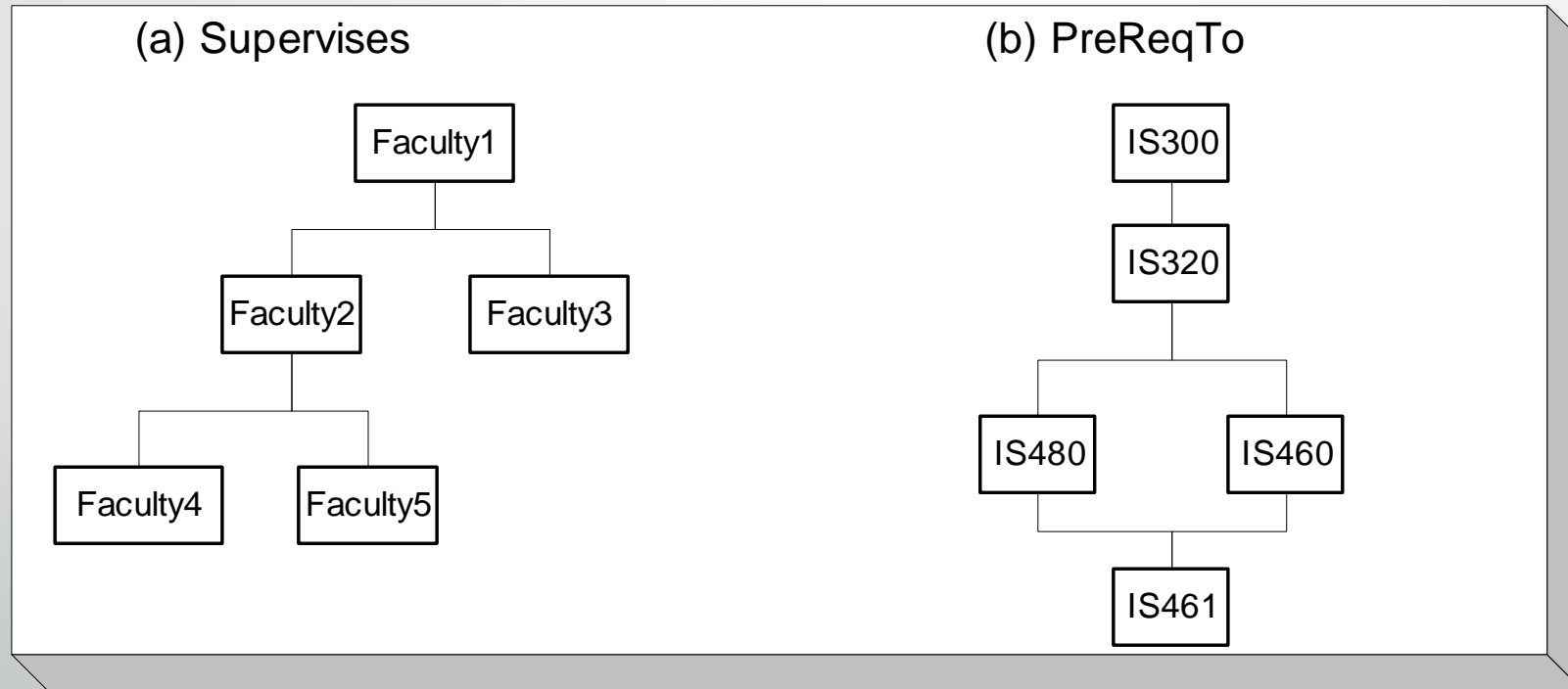
a) manager-subordinate



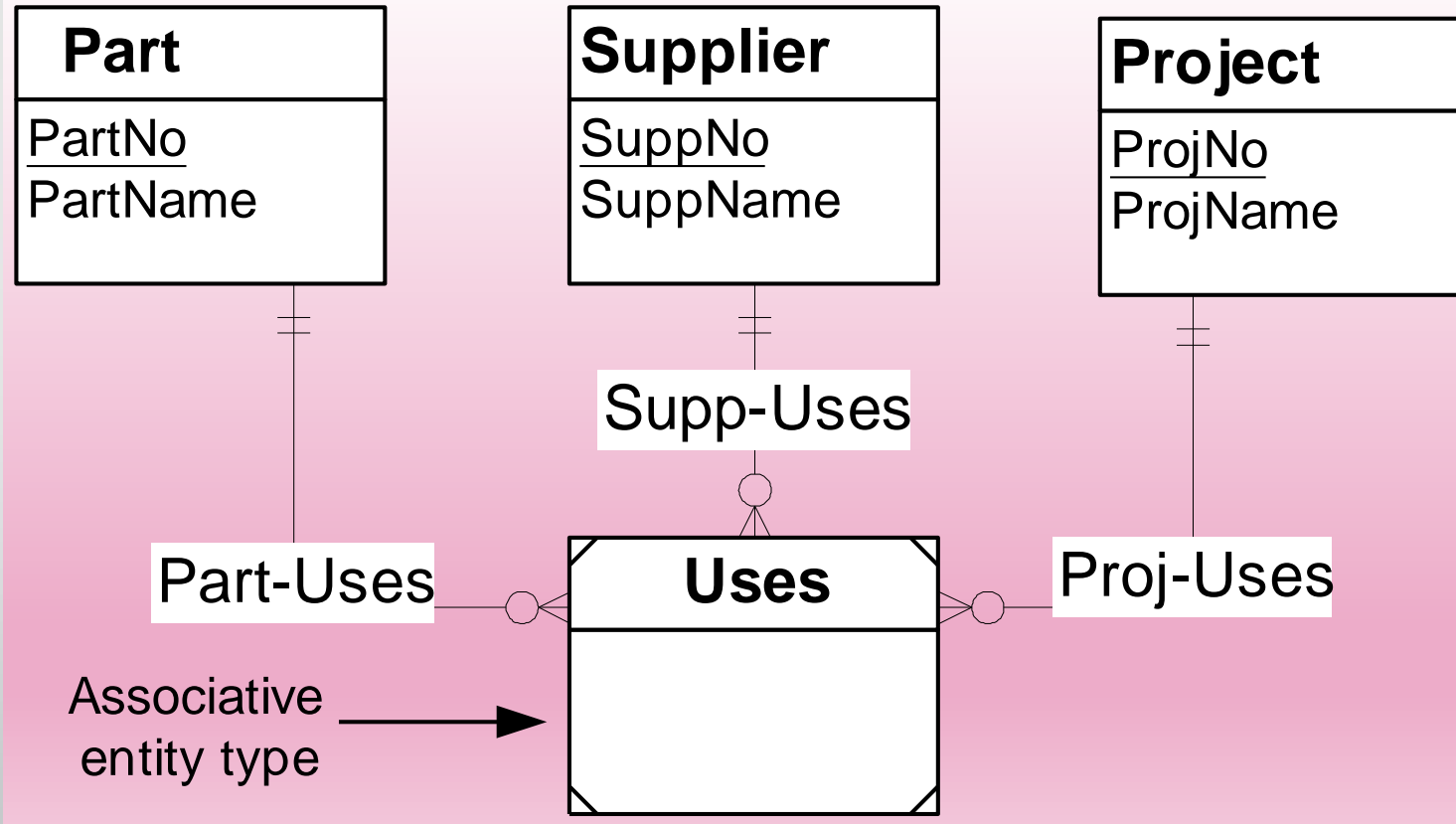
b) course prerequisites



Instance Diagrams for Self-Referencing Relationships



Associative Entity Types for M-way Relationships



Summary

- Specialized relationships are not common but important when in some situations
- Do not overuse specialized relationships
- Avoid notation errors with specialized relationships

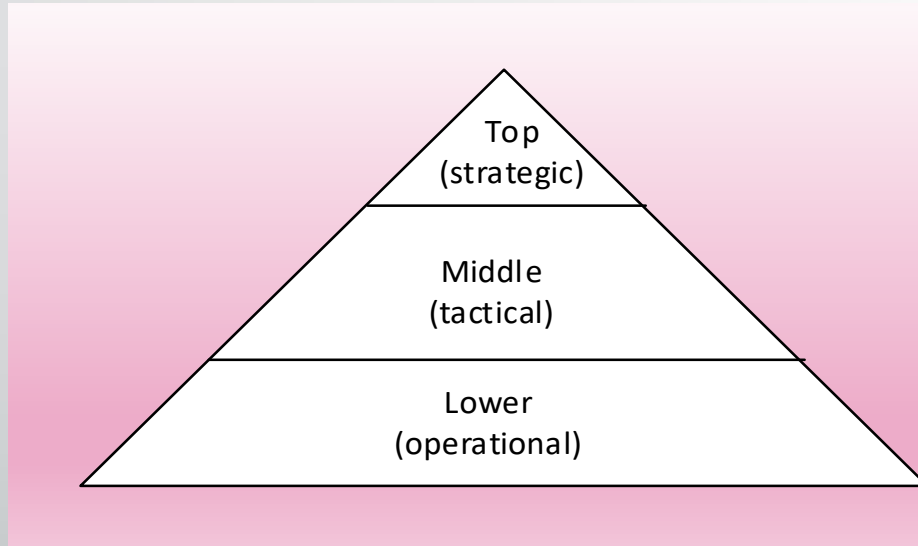
Exercise 1

Break 10 min



Decision Making Hierarchy

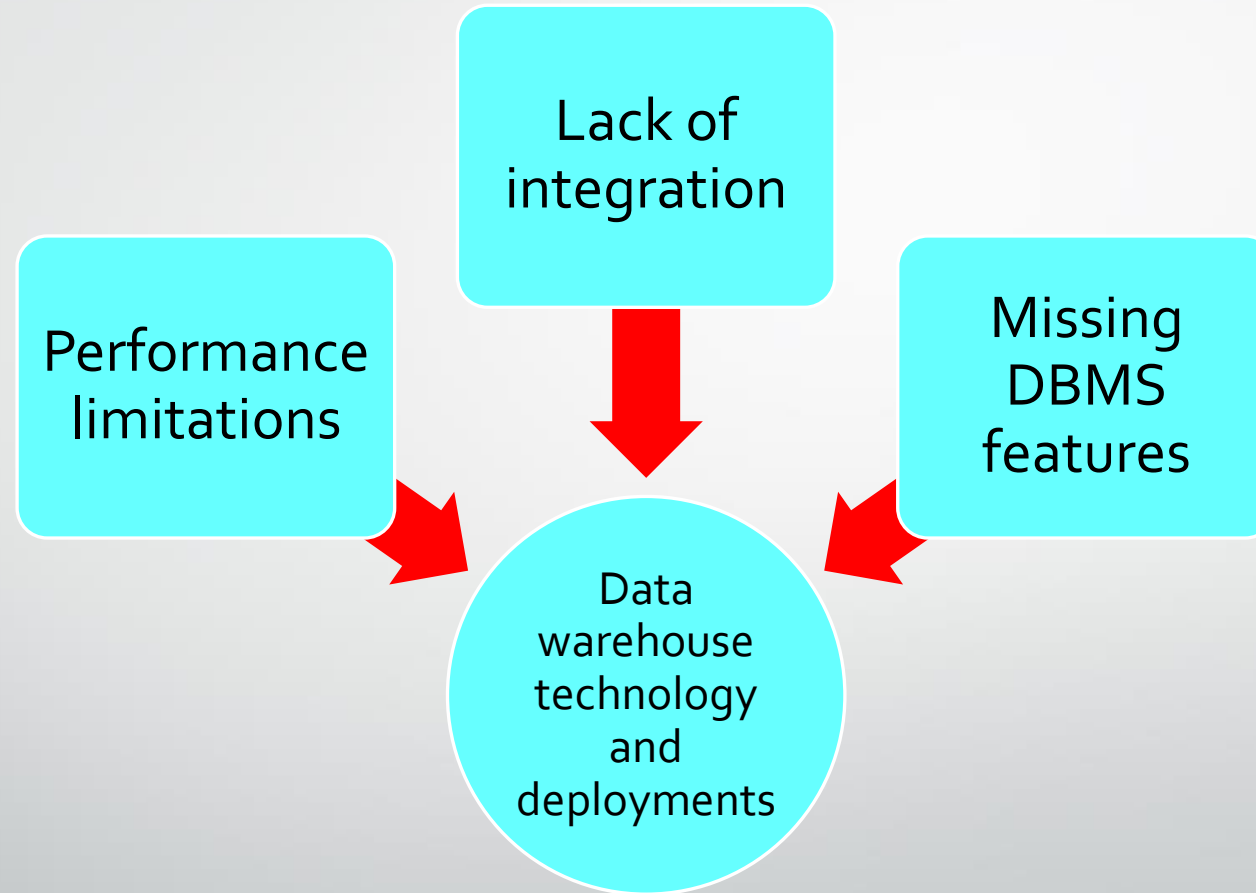
Decision making hierarchy



Typical decisions



Technology and Deployment Limitations



Data Warehouse Characteristics

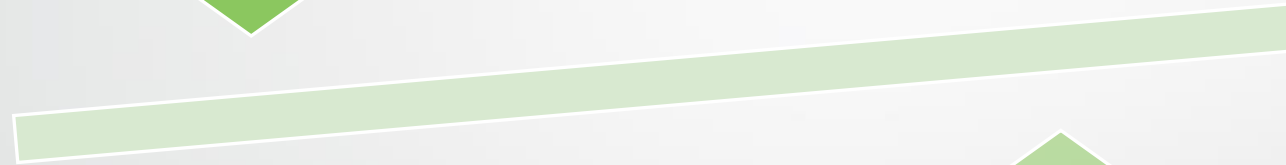
- Essential part of infrastructure for business intelligence
- Logically centralized repository for decision making
 - Populated from operational databases and external data sources
 - Integrated and transformed data
 - Optimized for reporting and periodic integration

Comparison of Processing Environments



Transaction processing

- Primary data from transactions
- Daily operations and short term decisions



Business intelligence processing

- Transformed secondary data
- Medium and long-term decisions



Summary

- Historical reasons for data warehouse development
- Key characteristics of data warehouses
- Differences between operational databases and data warehouses



Challenges in Data Warehouse Projects

- Substantial coordination across organizational units
- Uncertain data quality in data sources
- Difficult to scale data warehouse

Architecture Choices



Top Down

- Enterprise data warehouse
- Higher integration levels
- Logically centralized
- Larger project scope

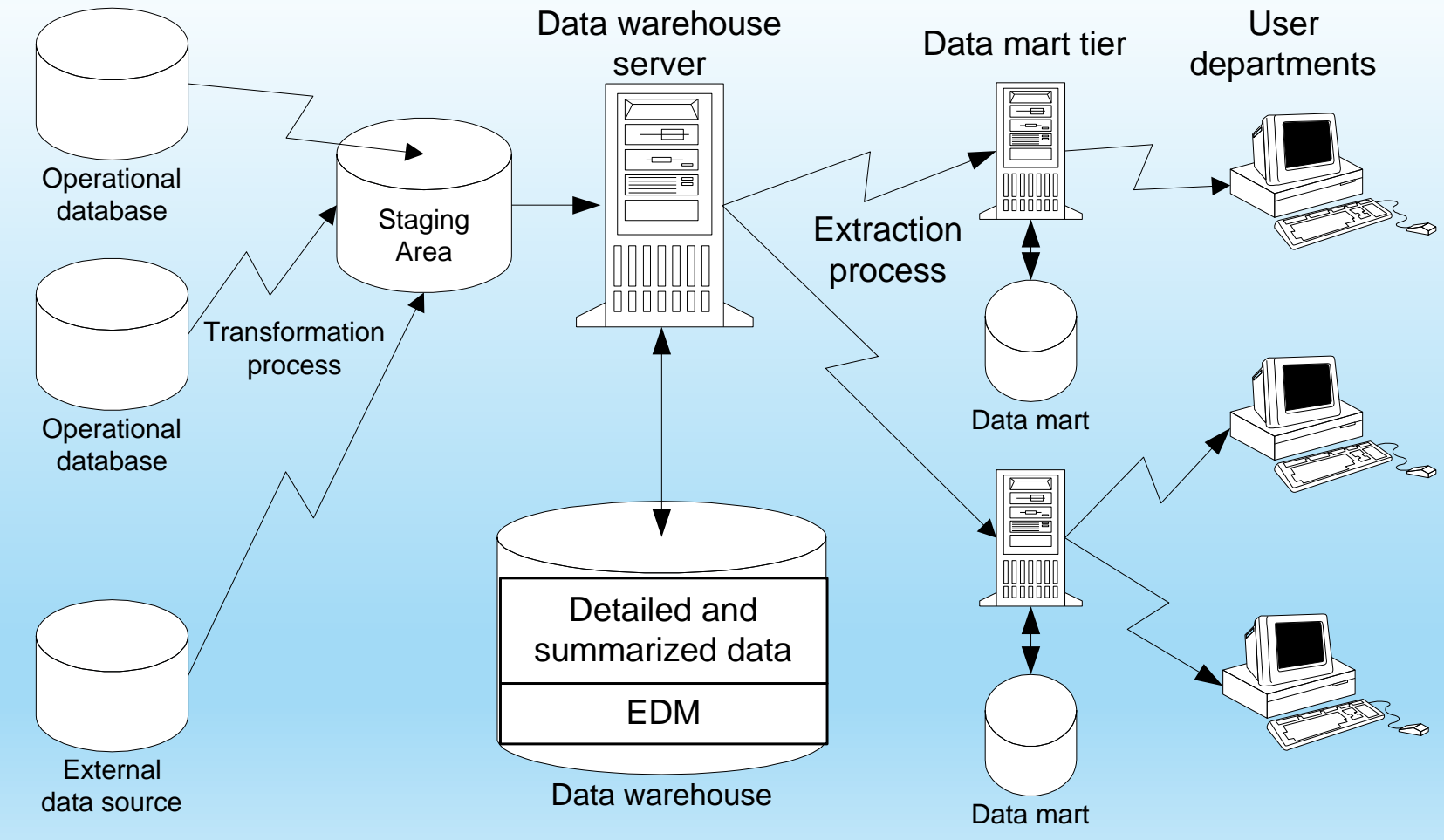


Bottom Up

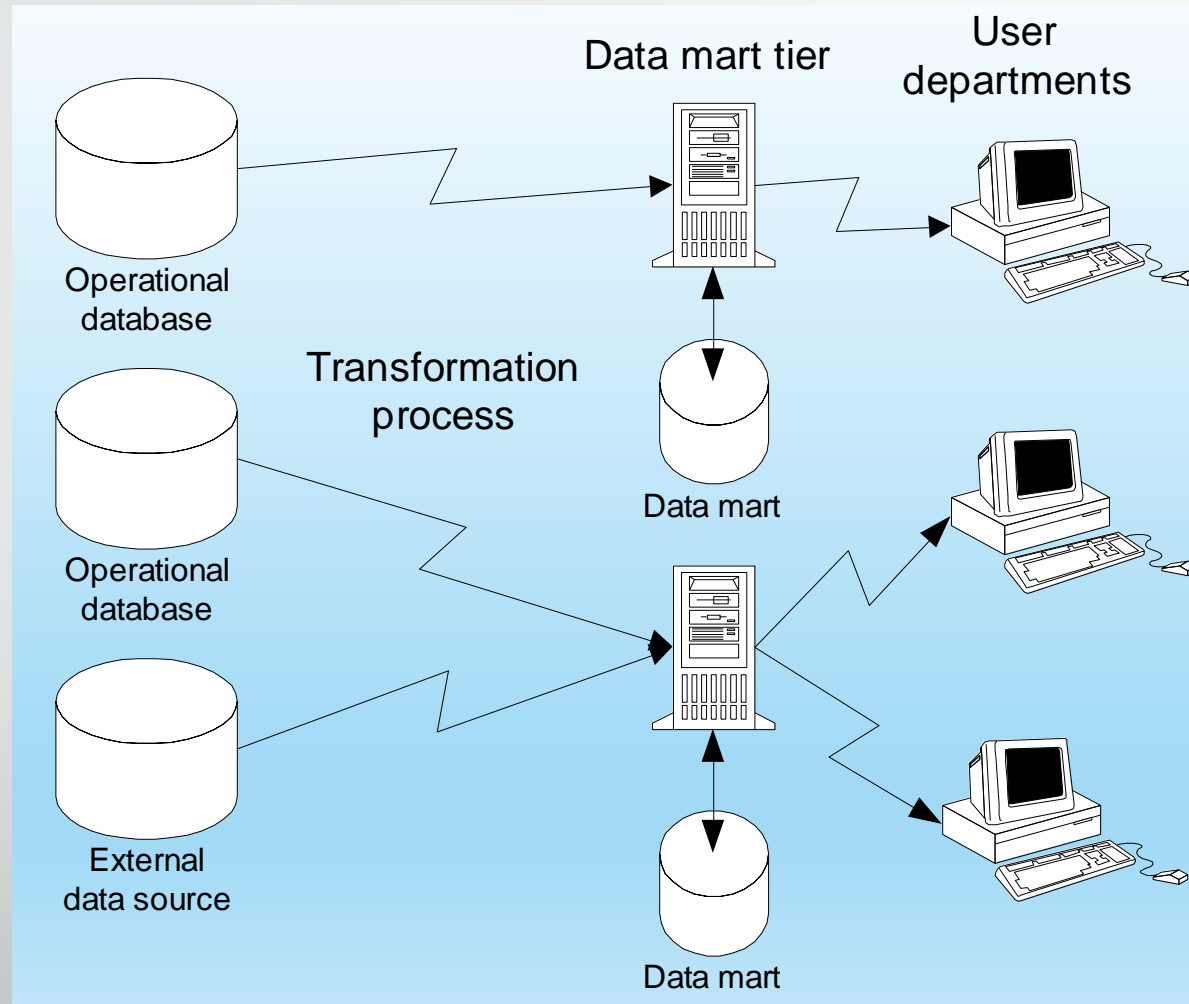
- Independent data marts
- Lower integration levels
- Logically decentralized
- Smaller project scope



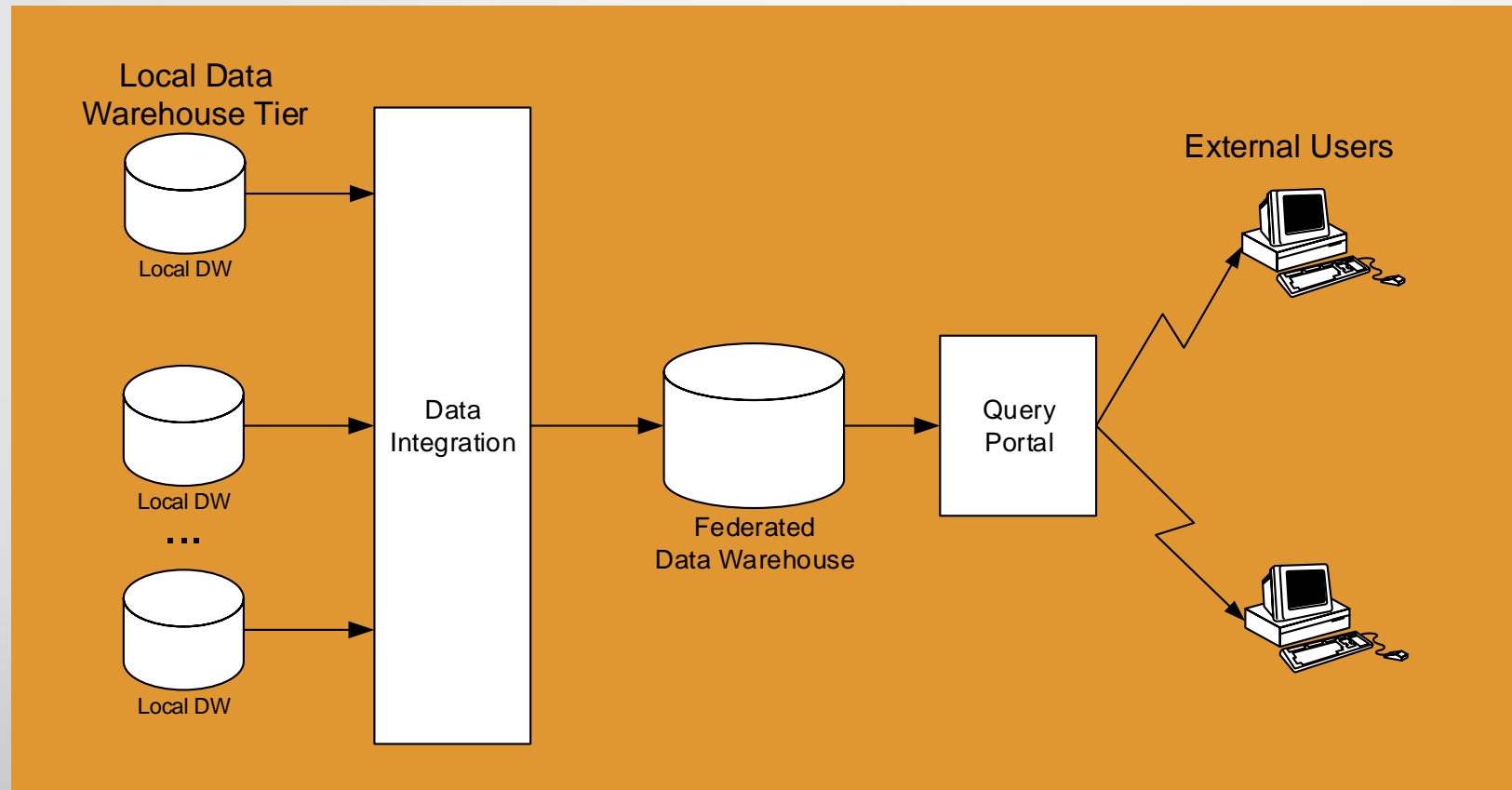
Top-Down Architecture



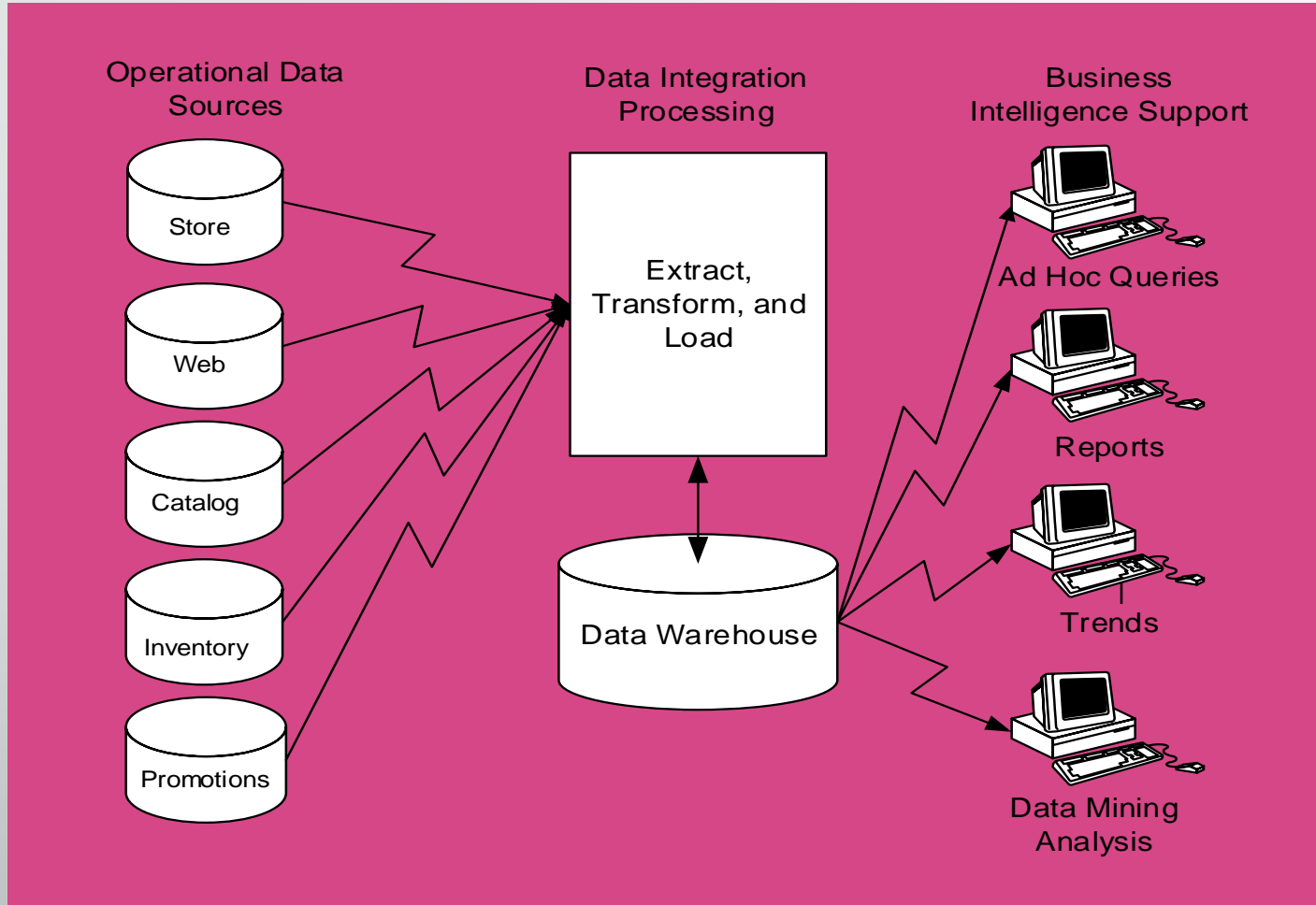
Bottom-up Architecture



Federated Architecture



TPC-DS Data Warehouse



Architecture Selection Factors

- Learning effects
 - Project risk
 - Intangible business value
- Strategic view of information technology
 - Level of sponsorship
 - Information independence
 - Task routineness

Summary

- Characteristics of business architectures
- Maturity model to guide investment decisions and data warehouse development over time



Exercise 2



15 min

Break

Data Visualization

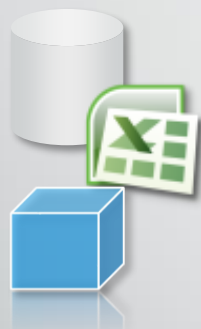
- Companies and individuals increasingly rely on data to make good decisions.
 - Because data is so voluminous, there is a need for visual tools that help people understand it.
- Data or information visualization
 - is the use of visual representations to explore, make sense of, and communicate data.
- What is portrayed in data visualizations
 - is the information (aggregations, summarizations, and contextualization) and not the data.

Visual Data Discovery Empowers Business People to Conduct Analysis without IT Help

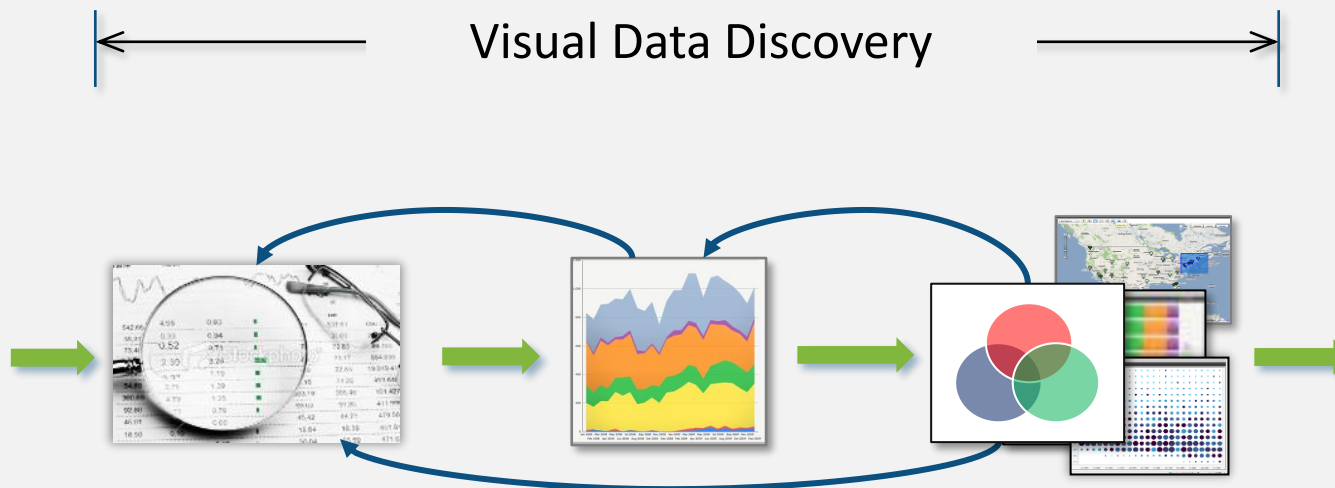
Start

Visual Data Discovery

Finish



Data



Familiarize

Visualize

Analyze &
Explore

Insight


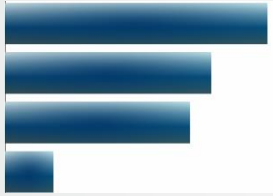
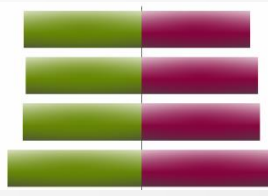
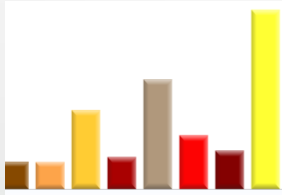
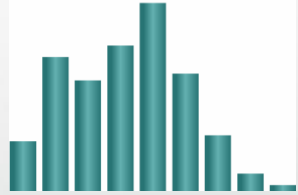
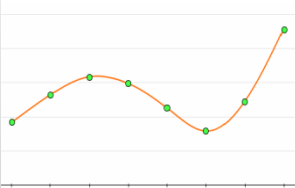
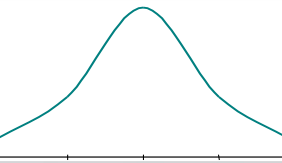
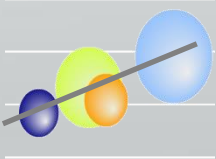
Types of Data

<p>Nominal Attributes Data that be counted, but not ordered or aggregated.</p> <p>Examples:</p> <ul style="list-style-type: none">•Products – Books, Movies, Music•Gender – Male, Female•State – Virginia, Nevada, California	<p>Ordinal Attributes Data that can be counted and ordered, but not aggregated.</p> <p>Examples:</p> <ul style="list-style-type: none">•Date – 1/1/2014, 1/2/2014...•Grades – A, B, C...•Ranks – Like, Neutral, Dislike
<p>Metrics Quantitative data that can be counted, ordered, and aggregated.</p> <p>Examples:</p> <ul style="list-style-type: none">•Revenue, Cost, Profit•Number of Customers•Temperature•Time	<p>Ordinal Attributes and Metrics Some data can be used as either attributes or metrics. Their classification is dependent on usage.</p> <p>Examples:</p> <ul style="list-style-type: none">•Age•Scores

This Chart-Comparison Matrix Identifies the Best Chart Type to Maximize Data Comprehension

Comparison Type

Basic Chart Form

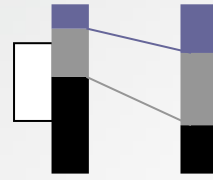
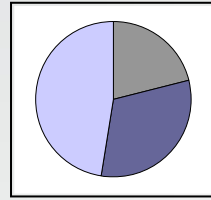
	COMPONENT	ITEM	TIME SERIES	FREQUENCY	CORRELATION
PIE					
BAR					
COLUMN					
LINE					
DOT					

Adopted from Zelazny, G. *Saying It With Charts: The executive's Guide to Visual Communication* McGraw Hill Professional, March 15, 2001

Source: Adapted from Gene Zelazny (2001). *Say It With Charts*

In Most Cases, One of Five Basic Chart Types Provides the Most Effective Data Presentation

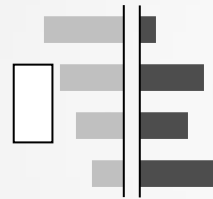
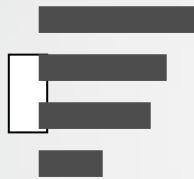
Pie Chart



5%

Recommended usage frequency

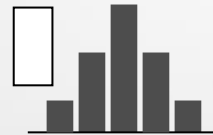
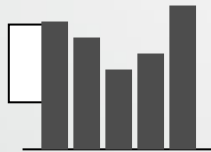
Bar Chart



25%

Recommended usage frequency

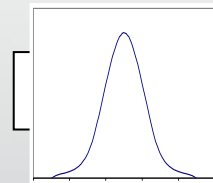
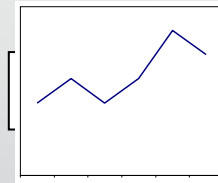
Column Chart



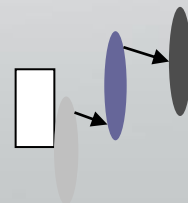
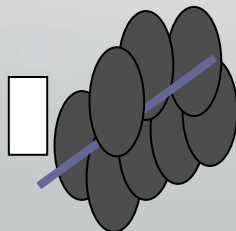
50%

Recommended usage frequency (combined)

Line Chart



Dot Chart



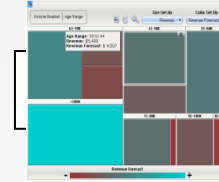
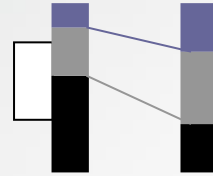
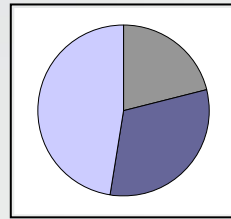
10%

Recommended usage frequency

Source: Adapted from Gene Zelazny (2001). *Say It With Charts*

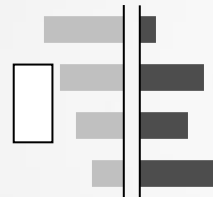
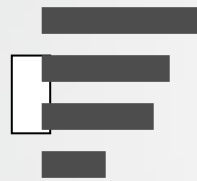
Composite Charts Convey More Business Dimensions and Metrics into a Single Display

Pie Chart



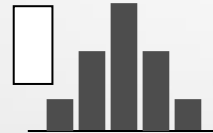
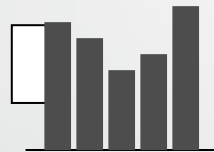
Heat Map

Bar Chart



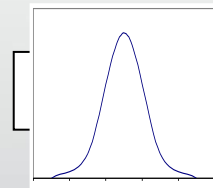
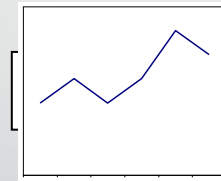
Bullet Graph

Column Chart



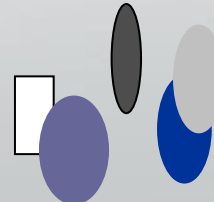
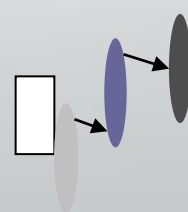
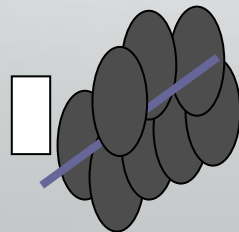
Micro Chart

Line Chart



Graph Matrix

Dot Chart



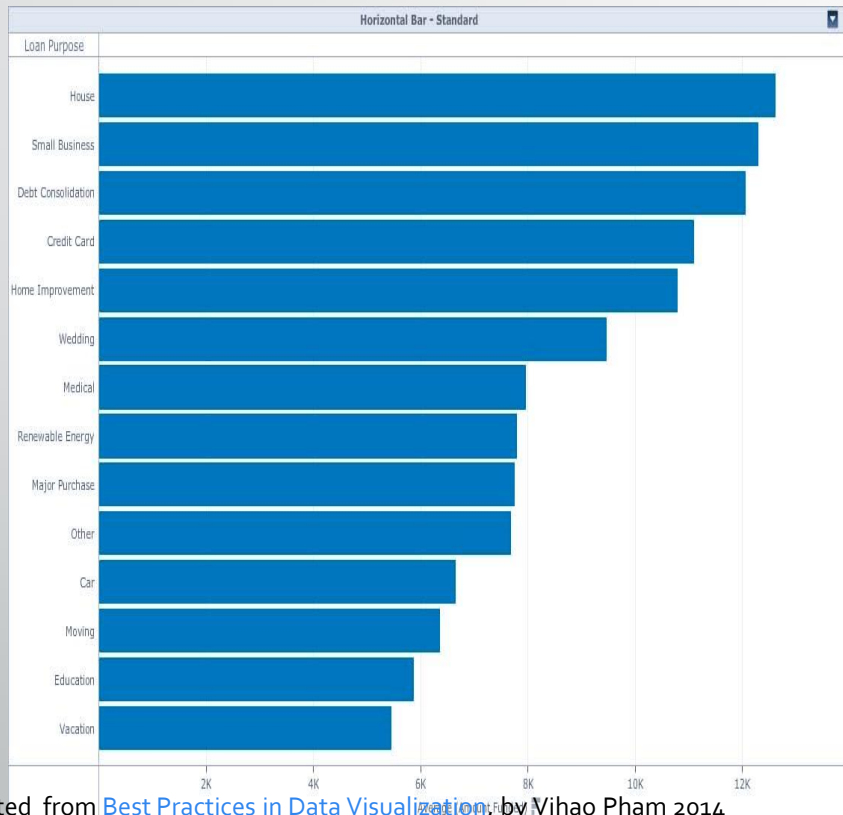
Bubble Chart

Source: Adapted from Gene Zelazny (2001). *Say It With Charts*

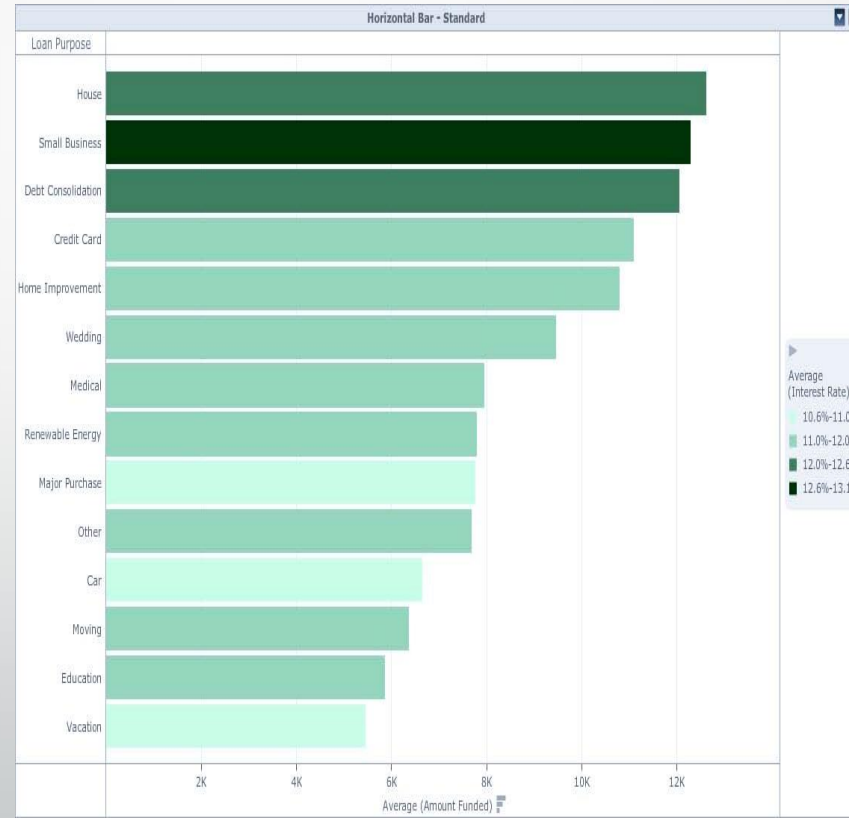
Visualizations

Attribute (nominal) and Metric

Comparative Analysis - Bar Chart
Sorted



Comparative analysis- Bar Chart
with Color to highlight Metric Patterns

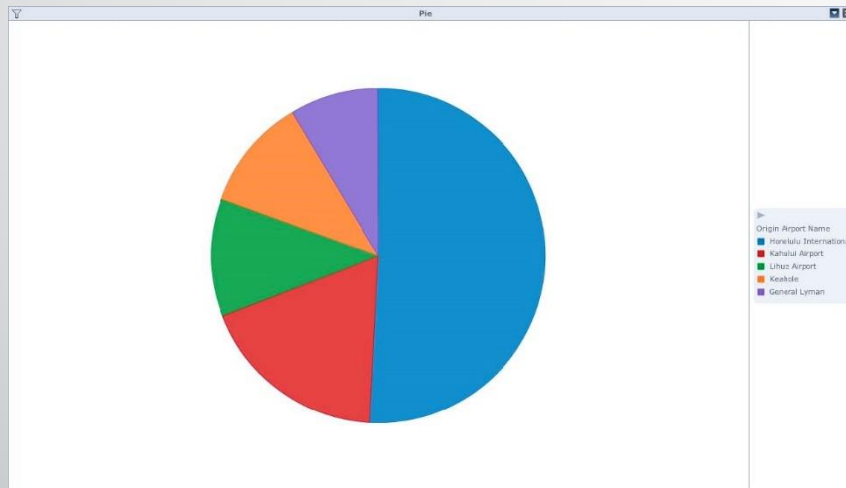


Adapted from [Best Practices in Data Visualization](#) by Mihao Pham 2014.

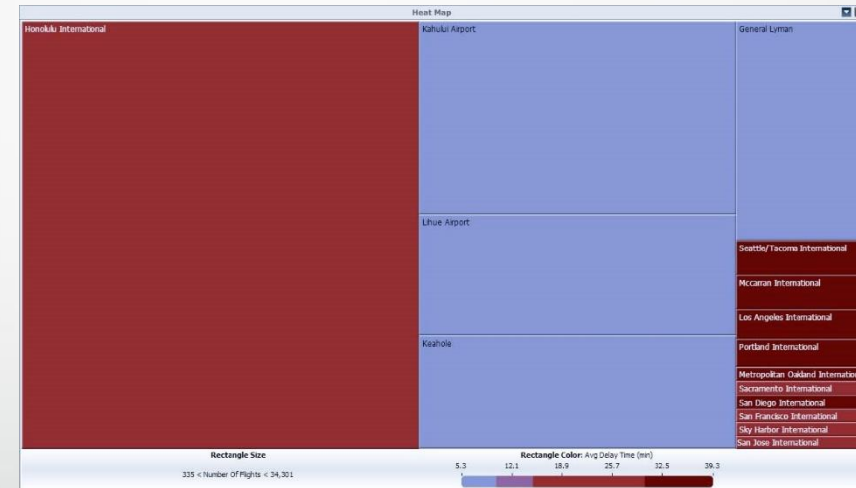
Visualizations

Attribute(nominal) and Metric

Contribution analysis-few elements- Pie Chart



Contribution Analysis- Many Elements- Heat Map

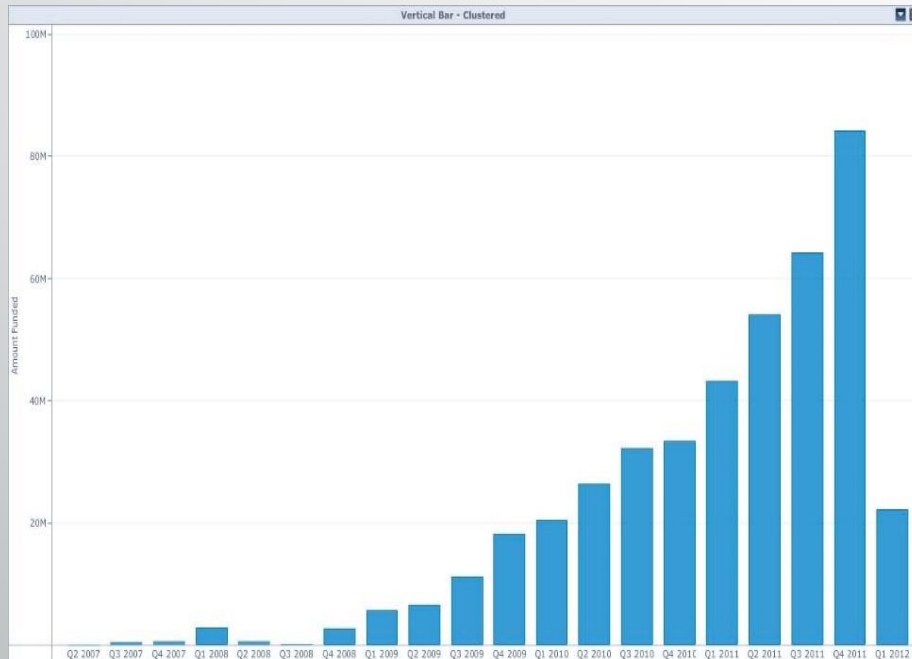


Adapted from [Best Practices in Data Visualization](#), by Vihao Pham 2014

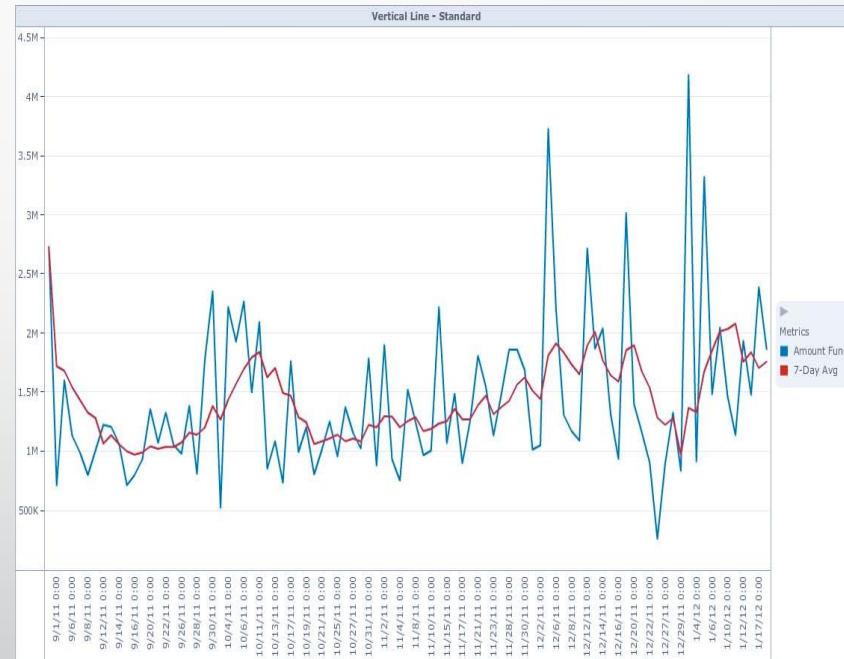
Visualizations

Attribute (ordinal) and Metric

Time-series analysis- Few elements- Column Chart



Time-series Analysis- Many Elements- Line chart

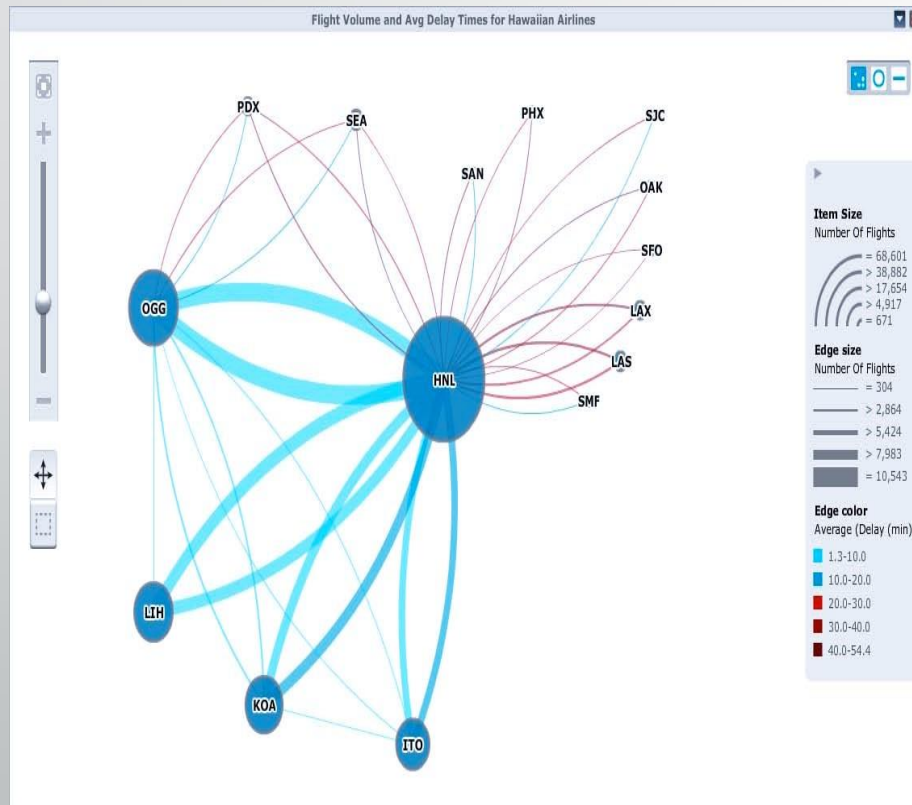


Adapted From [Best Practices in Data Visualization](#), by Vihao Pham 2014

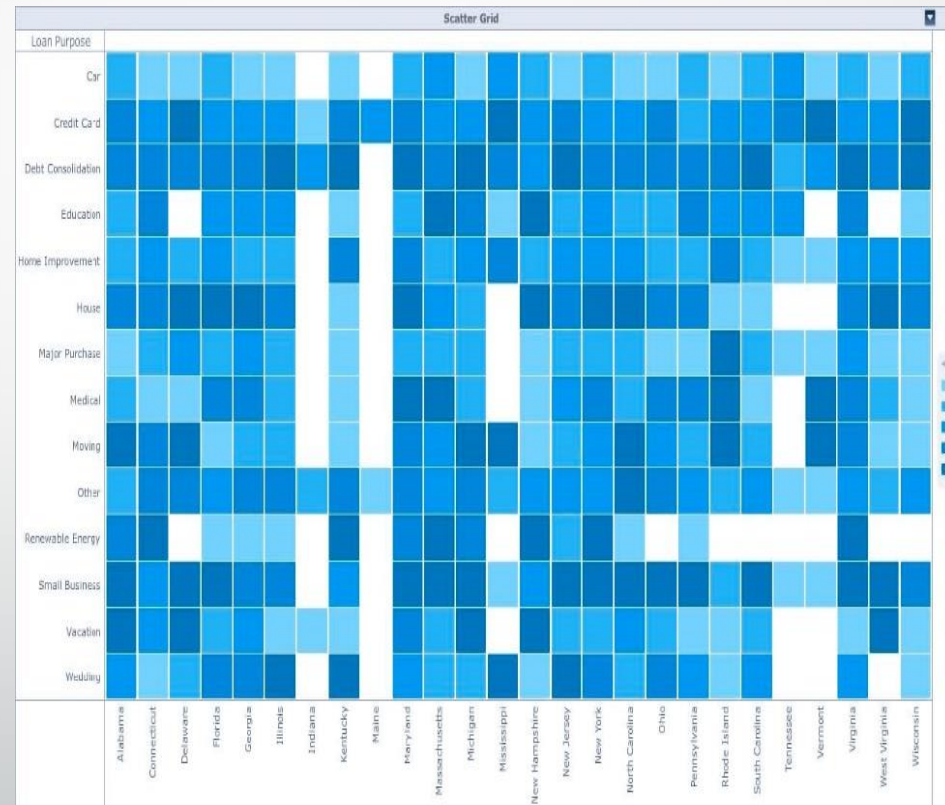
Visualizations

Attribute (Nominal) and Attribute (Nominal)

Market Basket or Network Analysis- Network Visualization



Market Basket or Network Analysis-
Avoid Scatter Grid- Implied ordinality

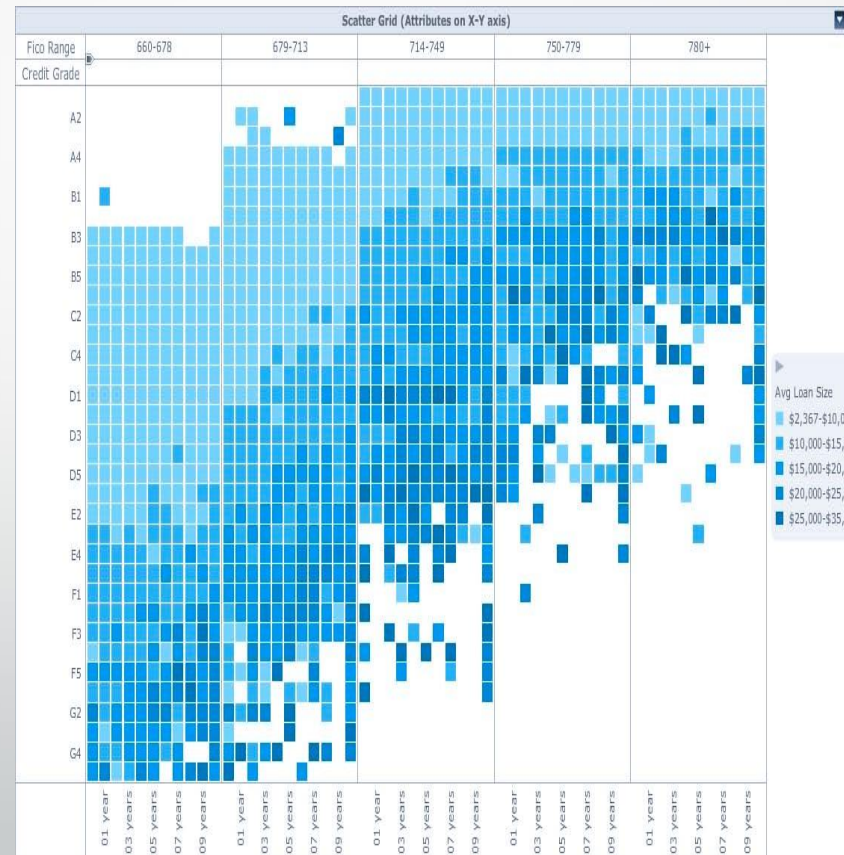


Adapted from [Best Practices in Data Visualization](#), by Vihao Pham 2014

Visualizations

Attribute (nominal) and Attribute (ordinal)

Cluster or Heat map Analysis
Scatter Grid



Appropriate Visualizations

	Metric	Attribute (Nominal)	Attribute (Ordinal)
Attribute (Nominal)	Bar Heatmap	Network	Line w/ Break-By Bar w/ Break-By
Attribute (Ordinal)	Column Line		Scatter Grid
Metric	Scatter/Bubble		

Adopted from [Best Practices in Data Visualization](#), by Vihao Pham 2014

General Rules for Charts

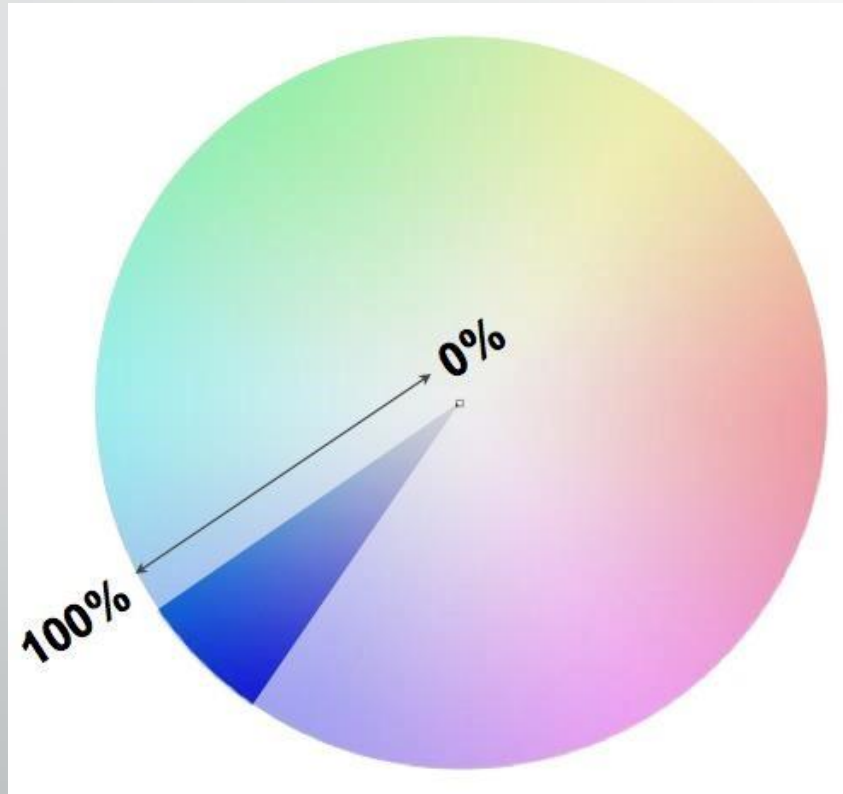
- Think about which increments are best to use in charts.
- Avoid combining unrelated charts into one.
- Consider the importance of direction when presenting data.
- Avoid cramming too much information into each chart.
- Make sure that any data labels or data legends are legible.
- Think about the order and direction of the slices in a pie chart.
- Avoid using too many effects to differentiate each slice.
- Think about the size of sparklines.
- Ensure they are placed in relevant sections.

Source: Adopted from [Free E-Book: How to Create Compelling Business Dashboards](#): Everything you need to know to design best-practice dashboards and data visualizations. Matillion Business Intelligence.

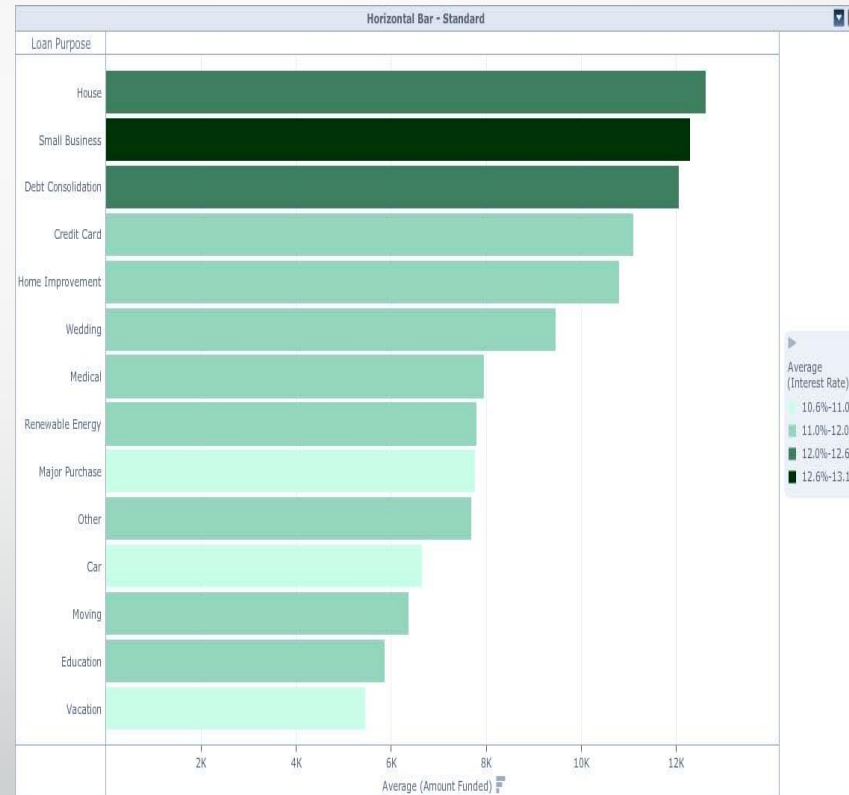
Visualization Considerations

Use Color Saturation Correctly

Less Saturation: Small Values



More Saturation: Greater Values



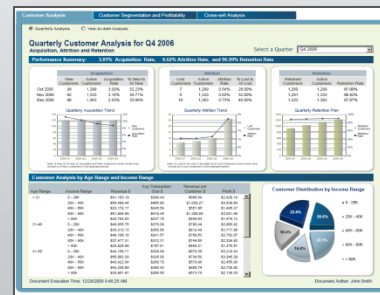
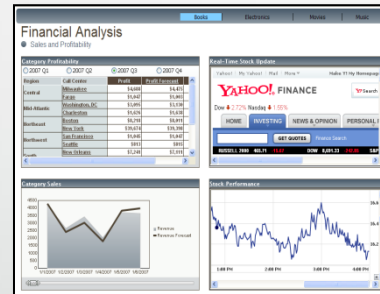
Adapted from [Best Practices in Data Visualization](#), by Vihao Pham 2014

Dashboards Combine Many Reports into a Single, Easy to Understand, Dashboard Application

Dozens of Reports
Frequently Used Together



Business Dashboards
Consolidates Dozens of Reports



Multi-layout Dashboard Book
Consolidates Multiple Independent Dashboard Designs



Dashboard Types

Scope	<ul style="list-style-type: none">• Broad: Displaying information about the entire organization• Specific: Focusing on a specific function, process, product, etc.
Business role	<ul style="list-style-type: none">• Strategic: Provides a high-level, broad, and long-term view of performance• Tactical: measure the business's progress according to related trends, in accordance with each strategic initiative• Operational: Provides a focused ,near-term, and operational and business processes view of performance
Time horizon	<ul style="list-style-type: none">• Historical: Looking backwards to track trends• Snapshot: Showing performance at a single point in time• Real-time: Monitoring activity as it happens• Predictive: Using past performance to predict future performance

Source; Adopted from: [A Guide to Creating Dashboards People love to use, translating Delicious Data into a Beautiful Design](#)
Version 2.0. May 2015

Action items to consider before you start

- Define the type of data you are working with.
- Consider timeliness of this information, and how frequently the dashboard itself will be updated.
- Find out about users.
- Evaluate the suitability of the BI platform for the design and deployment of the dashboards

What to include on your dashboard

- Define your dashboard functionality
- Don't sacrifice substance over style
- Know your users' requirements
- Validate their information requirements
- Select a right metric
- Select the right visual representation

From: How to Create Compelling Business
Dashboards - Complete Guide-

What Makes a perfect metric?

- Actionable
 - Metric involves a specific and repeatable action that can be linked to the observed data
- Transparent
 - Metric involves relatively simple calculations, making it easy for users to follow them
- Accessible
 - Metric involves data which is easy accessible, and simple to maintain
- Recognizable
 - There is a clear, distinct, and consistent understanding of what the metric means throughout the whole dashboard

Source: Stephen Few, Information Dashboard Design 2013, [A Guide to Creating Dashboards People love to use](#), Juice, 2009-2010.

Example of Metrics

Category	Measures	Category	Measures
Sales	<ul style="list-style-type: none"> • Bookings • Billings • Pipeline (anticipated sales) • Number of orders • Order amounts • Selling prices 	Fulfilment	<ul style="list-style-type: none"> • Number of days to ship • Backlog • Inventory levels
Marketing	<ul style="list-style-type: none"> • Market share • Campaign success • Customer demographics 	Manufacturing	<ul style="list-style-type: none"> • Number of units manufactured • Manufacturing times • Number of defects
Finance	<ul style="list-style-type: none"> • Revenues • Expenses • Profits 	Human resources	<ul style="list-style-type: none"> • Employee satisfaction • Employee turnover • Count of open positions • Count of late performance reviews
Technical support	<ul style="list-style-type: none"> • Number of support calls • Resolved cases • Customer satisfaction • Call durations 		

Adopted from: Stephen Few, Information Dashboard Design 2013

Ensure Metrics are Comparable

- Time comparison
 - Allows for representing trends in data, and making comparisons against points in the past, or even against future forecasts.
- Cross comparison
 - Allows for analyzing certain variables in relation to one another, to see if there is any correlation between them.
- Goal comparison
 - Allows for charting progress against predetermined goals and targets.

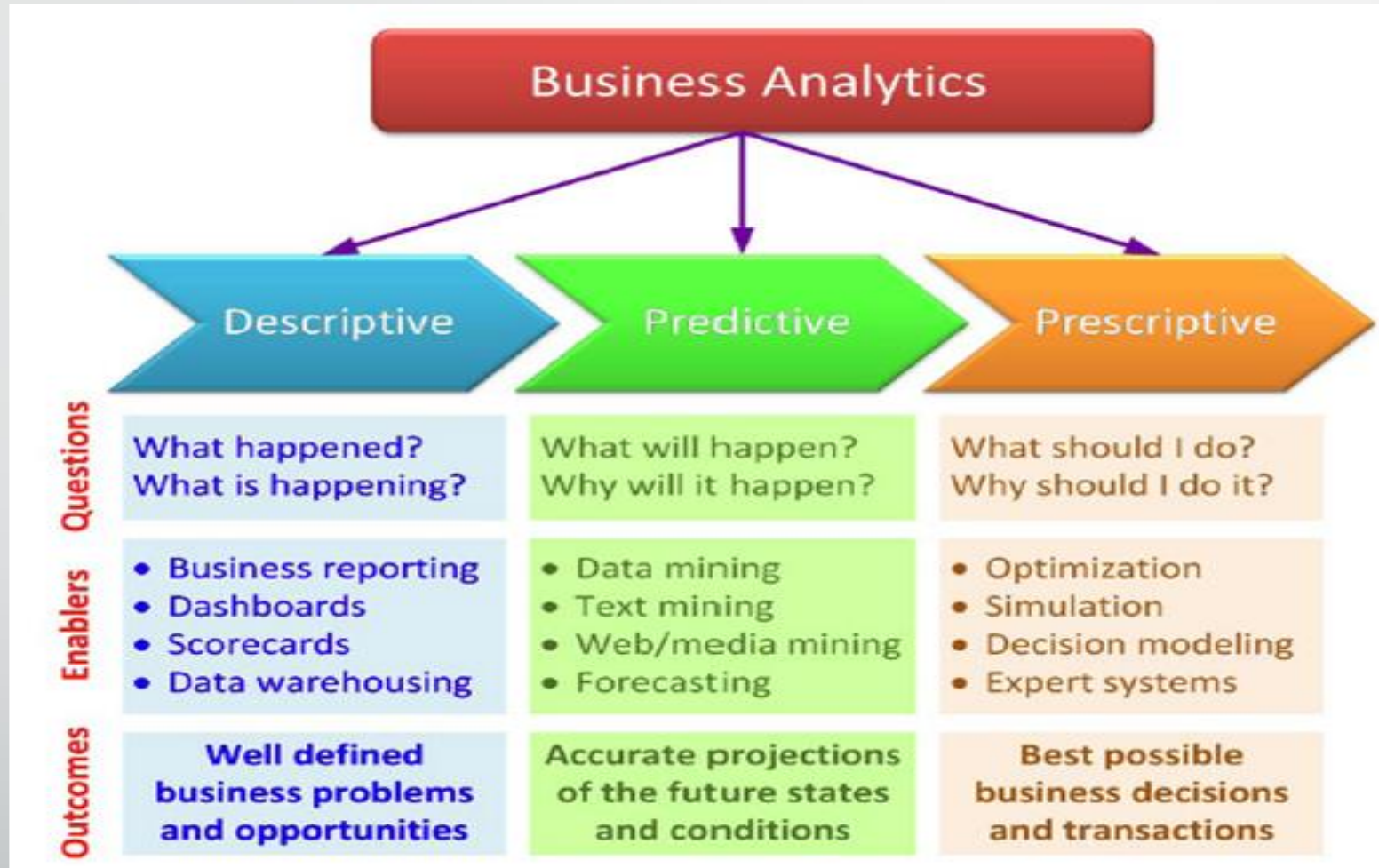
Adopted from **Stephen Few, Information Dashboard Design 2013**

Business Analytics

Visual analytics is the combination of visualization and predictive analytics.

- ***Information visualization*** is aimed at answering
 - “what happened” and “what is happening” and
 - is closely associated with business intelligence (routine reports, scorecards, and dashboards),
- ***Predictive analytics*** is aimed at answering
 - “why is it happening,” “what is more likely to happen,” and
 - is usually associated with business analytics (forecasting, segmentation, correlation analysis).

Business Analytics Overview



From SHARDA, RAMESH; DELEN, DURSUN; TURBAN, EFRAIM, BUSINESS INTELLIGENCE AND ANALYTICS: SYSTEMS FOR DECISION SUPPORT, 10th Edition, © 2015. Used by permission of Pearson Education, Inc., New York, NY. All Rights Reserved.

Descriptive Analytics

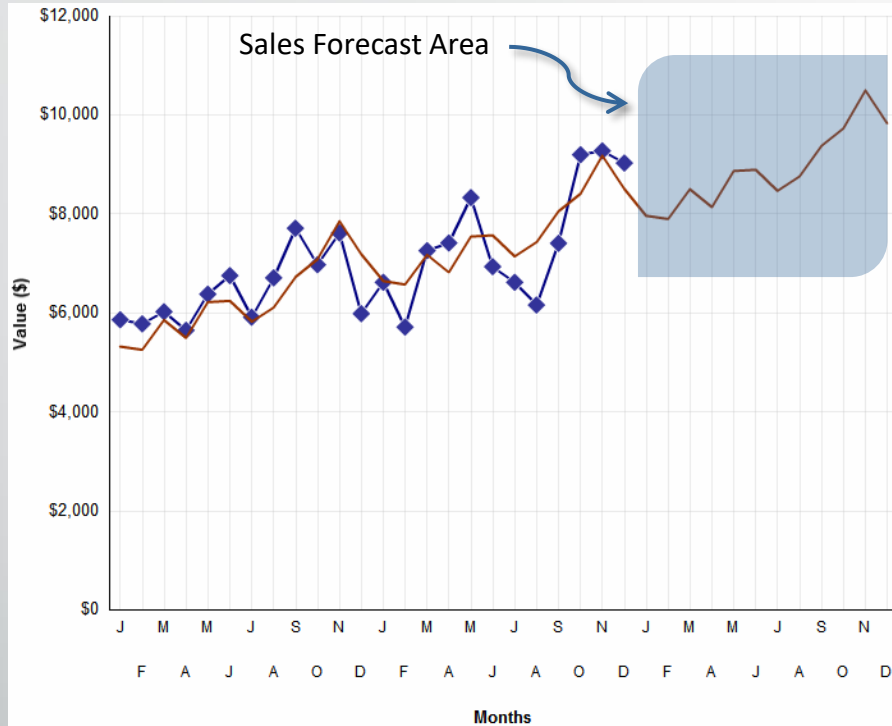
- Descriptive or reporting analytics refers to
 - knowing what is happening in the organization
 - understanding some underlying trends and causes of such occurrences
- It involves
 - consolidation of data sources and availability of all relevant data in a form that enables appropriate reporting and analysis.
 - usually development of this data infrastructure is part of data warehouses

Predictive Analytics

- Predictive analytics aims to determine what is likely to happen in the future.
- Uses statistical and data mining techniques to predict if the customer is likely to
 - switch to a competitor (“churn”),
 - buy next and how much,
 - respond to promotion,
 - worth the risk

Bring Predictive Analysis into the Mainstream for Business Users

Typical Predictive Analyses Based on Regression Techniques



Powerful Predictive Analyses Based on Data Mining Techniques

DETERMINE WHO IS LIKELY TO ...

- Achieve Revenue
- Stay in Budget
- Respond
- Purchase
- Defraud
- Be Profitable
- Be On Time

Linear Regression
Logistic Regression
Tree Regression

Decision Tree
Clustering

Time Series
Association Rules

Neural Network
Rule Set

Support Vector Machines
Ensembles of Models

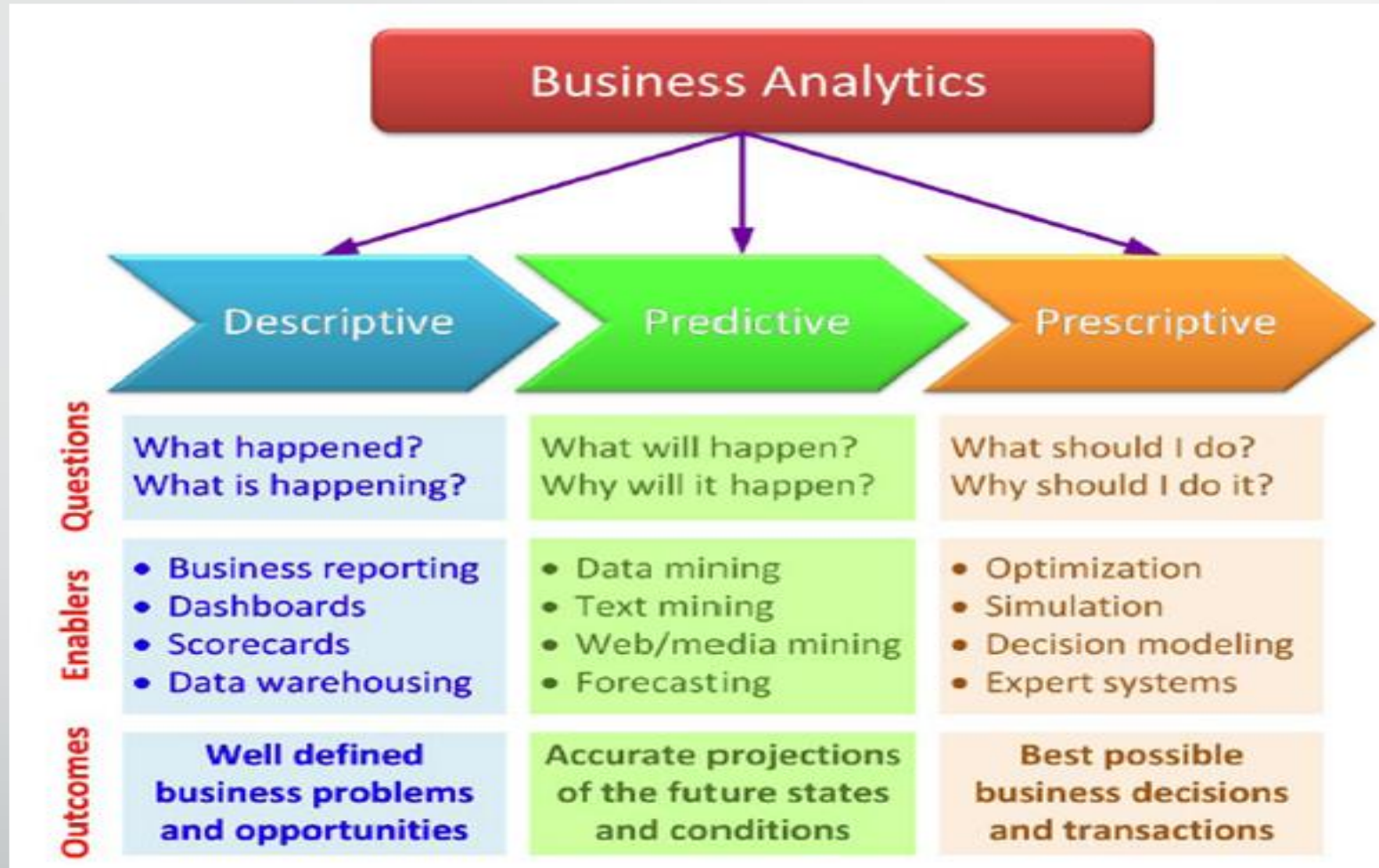
Prescriptive Analytics

Prescriptive analytics aim to recognize

- what is going on,
- the likely forecast, and
- the best performance possible using optimization techniques

From SHARDA, RAMESH; DELEN, DURSUN; TURBAN, EFRAIM, BUSINESS INTELLIGENCE AND ANALYTICS: SYSTEMS FOR DECISION SUPPORT, 10th Edition, © 2015. Used by permission of Pearson Education, Inc., New York, NY. All Rights Reserved.

Business Analytics Overview



From SHARDA, RAMESH; DELEN, DURSUN; TURBAN, EFRAIM, BUSINESS INTELLIGENCE AND ANALYTICS: SYSTEMS FOR DECISION SUPPORT, 10th Edition, © 2015. Used by permission of Pearson Education, Inc., New York, NY. All Rights Reserved.

Big Data Analytics

- Big Data? It refers to data that
 - cannot be stored in a single storage unit
 - is arriving in many different forms, be they structured, unstructured, or in a stream
 - are clickstreams from Web sites, postings on social media sites such as Facebook, or data from traffic, sensors, or weather
- Big Data analytics solution
 - instead of pushing data to a computing node, solution pushes computation to the data

From SHARDA, RAMESH; DELEN, DURSUN; TURBAN, EFRAIM, BUSINESS INTELLIGENCE AND ANALYTICS: SYSTEMS FOR DECISION SUPPORT, 10th Edition, © 2015. Used by permission of Pearson Education, Inc., New York, NY. All Rights Reserved.

Big Data

Extremely large sets of data that require advanced computational and analytical techniques to process and derive insights from.

The term "big" doesn't necessarily refer to the physical size of the data, but rather its complexity, variety, and velocity.

Big data can come from a variety of sources, such as social media, IoT devices, and business transactions.



Volume: Big data refers to data sets that are too large to be processed by traditional computing systems. They can range from terabytes to petabytes in size.



Variety: Big data is often heterogeneous and comes in a variety of formats, such as text, images, and videos. This presents a challenge in terms of data integration and processing.



Velocity: Big data is generated at an unprecedented speed, requiring real-time processing and analysis.



Veracity: Big data is often incomplete, inconsistent, and contains errors. It requires advanced techniques to ensure its accuracy and reliability.

Characteristics of Big Data

Applications of Big Data

Predictive analytics: Using big data to predict future outcomes and trends, such as customer behavior and market trends. Amazon

Personalized marketing: Using big data to create targeted marketing campaigns based on customer preferences and behavior.

Fraud detection: Using big data to identify and prevent fraudulent transactions and activities.

Healthcare: Using big data to improve patient outcomes and healthcare delivery.



Assignment 2



Questions