

# Introduction to Big Data

1



# Evolution of Big Data

2

- **The Model of Generating/Consuming Data has Changed**

**Old Model:** Few companies are generating data, all others are consuming data



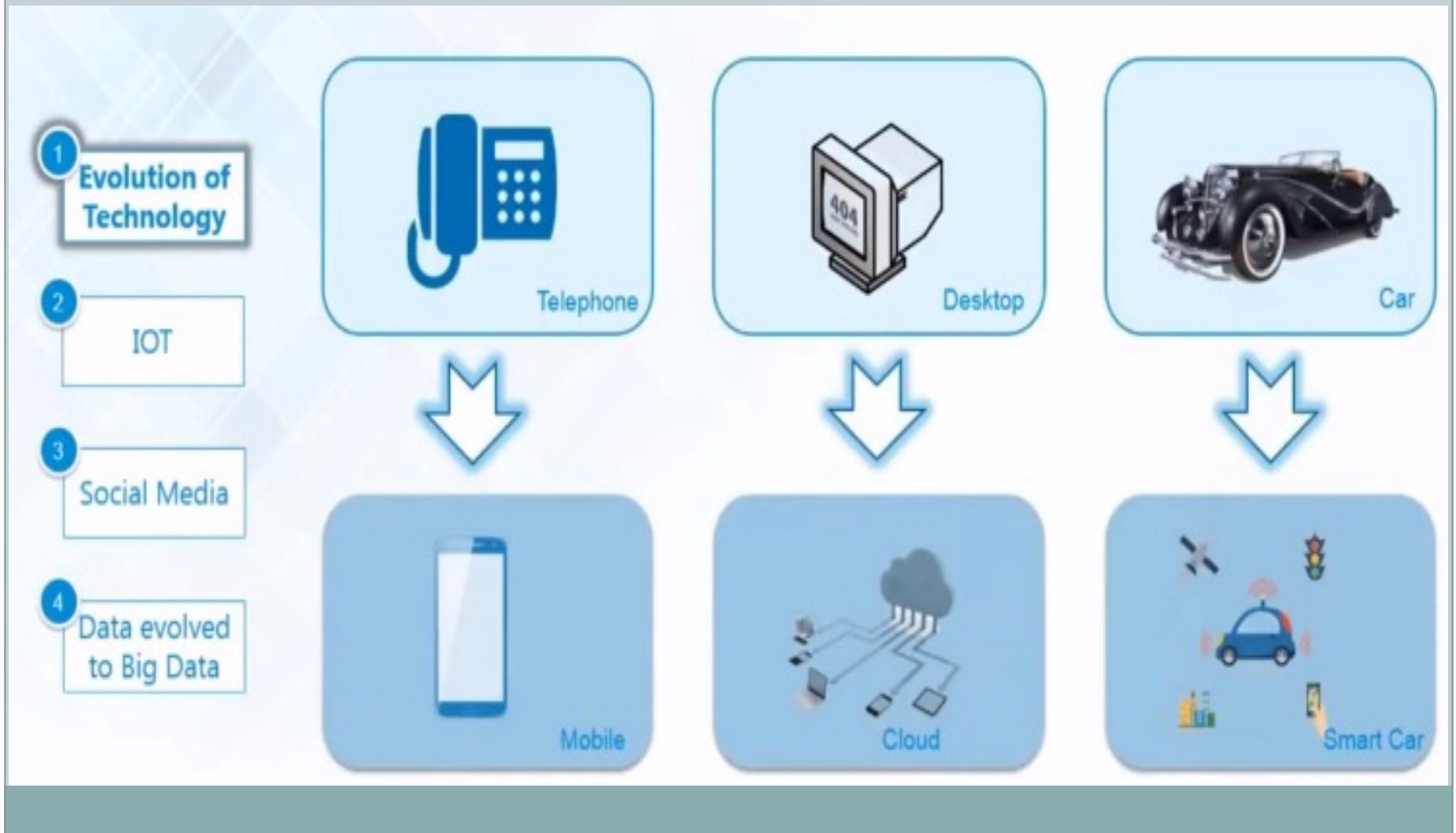
**New Model:** all of us are generating data, and all of us are consuming data



Unit of Data size	Exact size	Approximate Size	Examples	
KB (kilobyte )	$2^{10}$ or 1024 bytes	( $10^3$ or one thousand) bytes	A typical joke =1KB	
MB(megabyte )	$2^{20}$ bytes	( $10^6$ or one million) bytes	Complete work of Shakespeare =5MB	
GB (gigabyte )	$2^{30}$ bytes	( $10^9$ or one billion) bytes	Ten yards of books on a shelf = 1GB	
TB (terabyte)	$2^{40}$ bytes	( $10^{12}$ or one trillion) bytes	All the X-rays for a large hospital =1TB Tweets; created daily =121TB;	
PB (peta byte)	$2^{50}$ bytes	( $10^{15}$ or one quadrillion) bytes	All U.S. academic research libraries = 2PB Data processed in a day by Google =24PB	<b>B I G  D A T A</b>
EB (exa byte)	$2^{60}$ bytes	( $10^{18}$ or one Quintillion) bytes	Total global data created in 2006 = 161EB	
ZB (zetta byte)	$2^{70}$ bytes	( $10^{21}$ or one Sextillion) bytes	Total amount of global data created in 2012 = 2.7 ZB and expected 44 ZB by 2020	
YB (yotta byte)	$2^{80}$ bytes	( $10^{24}$ or one Septillion) bytes		

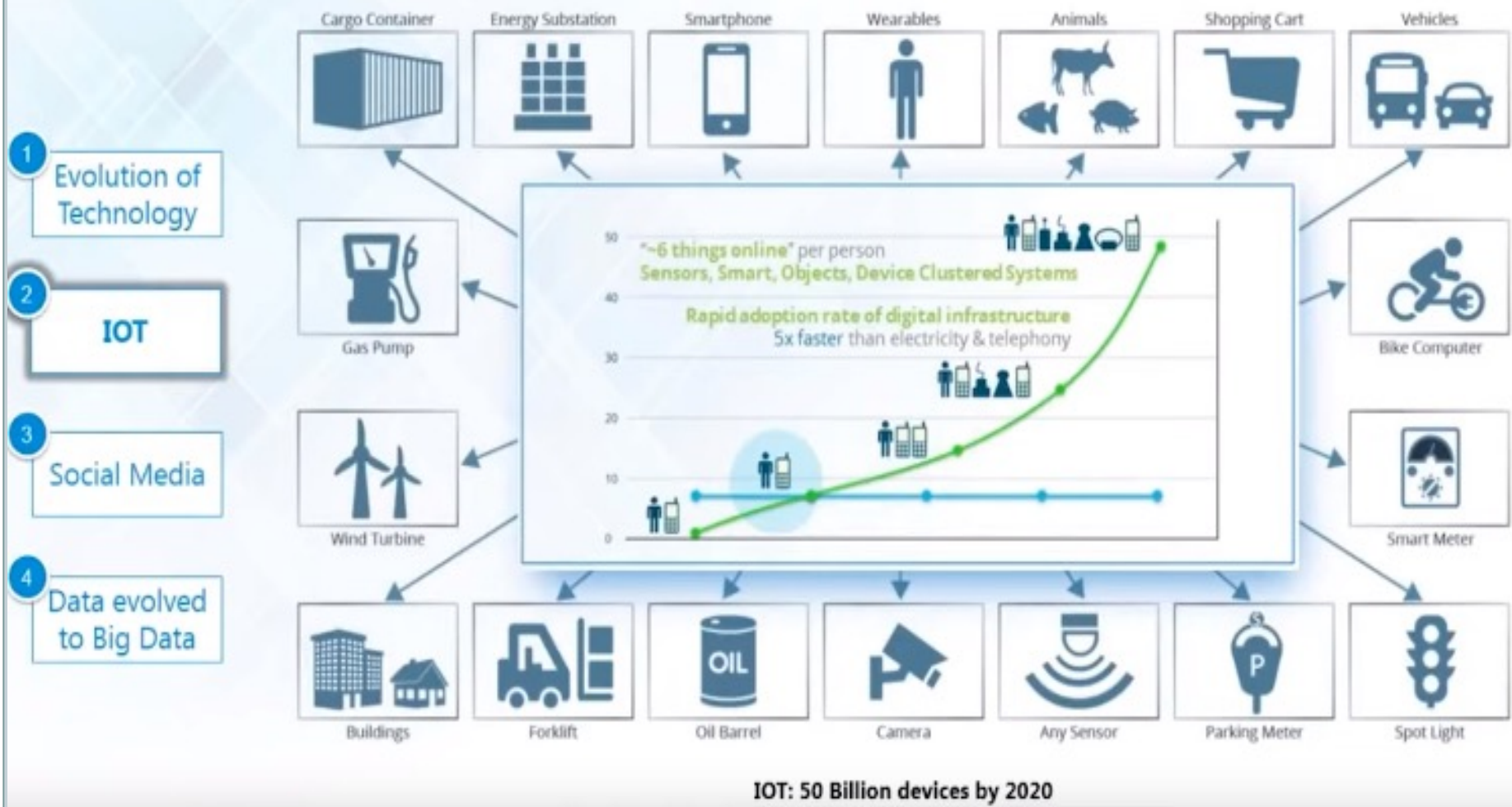
# Evolution of Big Data by technology

4



# Evolution of Big Data by Internet Of Things

5

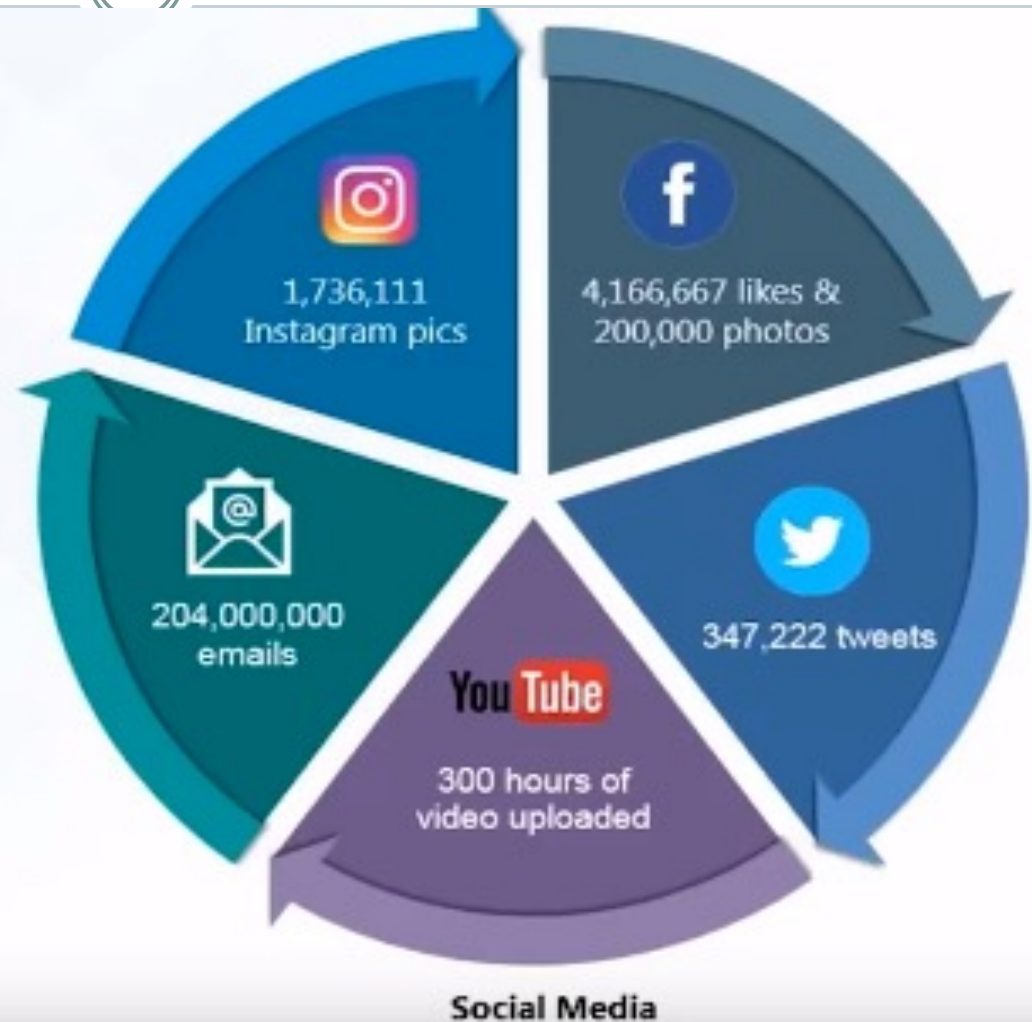




# Evolution of Big Data by Social Media

6

- 1 Evolution of Technology
- 2 IOT
- 3 **Social Media**
- 4 Data evolved to Big Data



# Evolution of Big Data by other factors

7



# Big Data sources

8

- **Human Generated Data**

- is emails, documents, photos and tweets. We are generating this data faster than ever. Just imagine the number of videos uploaded to You Tube and tweets swirling around. This data can be Big Data too.

- **Machine Generated Data**

- is a new breed of data. This category consists of sensor data, and logs generated by 'machines'
- such as email logs, click stream logs, etc. Machine generated data is orders of magnitude larger than Human Generated Data.



# Big Data sources

9

- **Web Data**

- **Social media data** : Sites like Facebook, Twitter, LinkedIn generate a large amount of data
- **Click stream data** : when users navigate a website, the clicks are logged for further analysis (like navigation patterns). Click stream data is important in on line advertising and E-Commerce

12+ TBs of tweet data every day



25+ TBs of  
log data every day



? TBs of data every day

# Big Data sources

10

**sensor data** : sensors embedded in roads to monitor traffic and misc.

*30 billion* RFID tags today  
(1.3B in 2005)

*4.6 billion*  
camera phones  
world wide

*100s of millions*  
of GPS enabled  
devices sold  
annually

*2+ billion*  
people on the Web  
by end 2011

*76 million* smart meters in 2009...  
200M by 2014



# What is Big Data?

11

## **Big data**

is the term for a collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.

## **Real world examples of Big Data**

- Facebook : has 40 PB of data and captures 100 TB / day
- Yahoo : 60 PB of data
- Twitter : 8 TB / day
- EBay : 40 PB of data, captures 50TB/ day



# Introduction to Big data and Analytics

12

ANY QUESTIONS /  
DOUBTS

???

# Databases and Data Warehouse

13





# Learning Objectives

14

Upon successful completion of this chapter, you will be able to:

- Describe the differences between data, information, and technology
- Define the term *database* and identify the steps to creating one
- Describe the role of a database management system
- Describe the characteristics of a data warehouse
- Define data mining and describe its role in an organization

# Data, Information, and Knowledge

15



- Data is raw bits and pieces of information
  - Quantitative – numeric
  - Qualitative – descriptive
  - Alone is not useful
- Information is when data is given context and more specific
- Knowledge is developed when information has been aggregated and analyzed to make decisions, set policies, and spark innovation
- Wisdom is the combination of knowledge and experience
  - May take years to develop

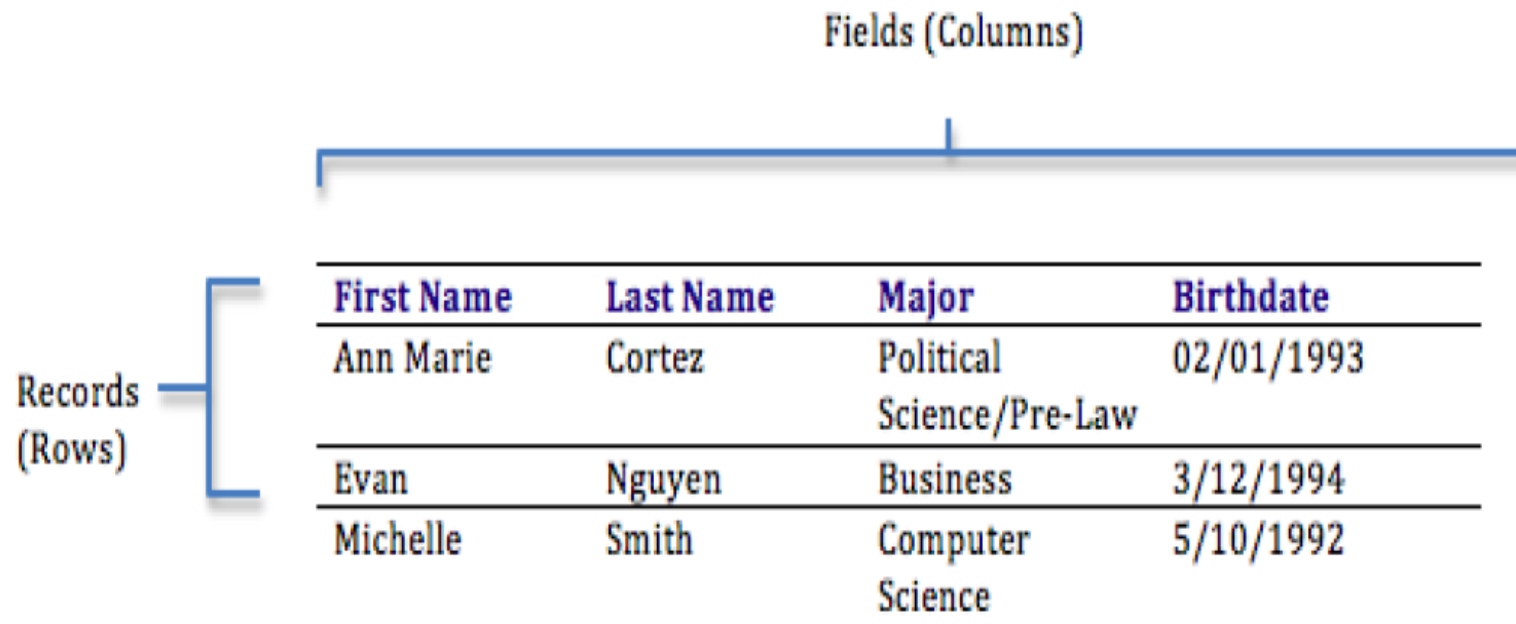
# Databases

16

- Organized collection of related information to generate knowledge for decision making purposes
  - For example, a university transcript database may contain information on students, classes taken, and grades received
- A separate university database would be created to maintain your financial information
- Relational databases (such as Microsoft Access) where data is organized into one or more tables
  - Tables are a collection of fields
    - ✦ E.g., Student ID, Course ID, Grade Earned
  - Record is an instance in the table
    - ✦ E.g., your specific information in the table

# Databases continued

17



The diagram illustrates a database table structure. A horizontal blue line with a vertical tick mark in the center is labeled "Fields (Columns)" above it. Below this line is a table with four columns: "First Name", "Last Name", "Major", and "Birthdate". To the left of the table, a vertical blue bracket is labeled "Records (Rows)". The table contains three rows of data. The first row has "Ann Marie", "Cortez", "Political Science/Pre-Law", and "02/01/1993". The second row has "Evan", "Nguyen", "Business", and "3/12/1994". The third row has "Michelle", "Smith", "Computer Science", and "5/10/1992".

First Name	Last Name	Major	Birthdate
Ann Marie	Cortez	Political Science/Pre-Law	02/01/1993
Evan	Nguyen	Business	3/12/1994
Michelle	Smith	Computer Science	5/10/1992

*Rows and columns in a table*

# Database Design

18

- Design is a critical first step in creating a database
  - Understand the goal of how the database will be used
  - Identify the data needed as part of accomplishing this goal
  - Identify how the data is related to each other
  - Identify tables and fields to organize the data
    - ✦ Each table needs a primary key of which field(s) is unique to each record and will not change
      - For example, our Bronco ID
    - ✦ Normalization is performed to eliminate duplicated data



# Database Design

19

- Design is a critical first step in creating a database
  - Understand the goal of how the database will be used
  - Identify the data needed as part of accomplishing this goal
  - Identify how the data is related to each other
  - Identify tables and fields to organize the data
    - ✦ Each table needs a primary key of which field(s) is unique to each record and will not change
      - For example, our Bronco ID
    - ✦ Normalization is performed to eliminate duplicated data

# Database Reports

20

- Structured Query Language (SQL) is a tool/language that helps extract information from the database for analysis purposes

Gender	Student ID	Last Name	First Name	Math grade
F	35965	Parker	Mary	76
	62242	Barker	Megan	67
	65784	Catbog	Nina	45
	78999	Catbog	Jasmine	83
	87900	Pass	Rosa	65
	98646	Lee	Lily	74
M	06754	Woloch	Gilbert	90
	24567	Wang	Ryan	84
	45679	Scott	Randy	98
	75768	Obama	barak	100
	76567	Li	Daniel	73
	76890	Huynh	Mathew	97
	98548	Barker	Josh	86
	98750	Parker	Franklin	56

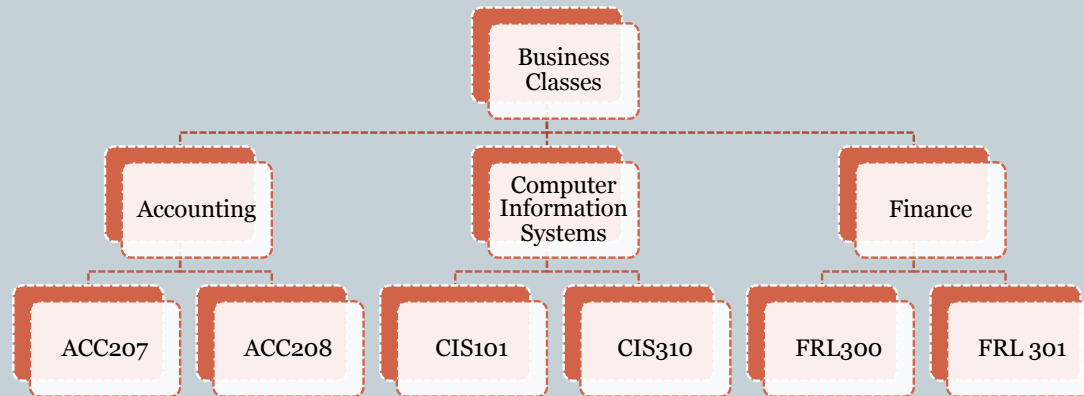
Monday, October 06, 2008

Page 1 of 1

# Other Database Types

21

- Hierarchical - parent/child relationship between data



- Document-centric – places data into documents that can be manipulated
- NoSQL – usually on multiple machines and, in some cases, machines in multiple data centers

# Database Management Systems

22

- Database Management Systems (DBMS) is an application that allows data to be:
  - Entered, Modified, and Deleted
  - Read
  - Reported
- Has a user friendly interface to design the database
- Relational databases use Microsoft Access installed on one machine with one user access at a time
- Enterprise Databases serve the entire organization

# Data Warehouse

23

- Consists of extracts from one or more of the organization's databases
- Allows the data to be copied and stored for analysis
  - Needs to be refreshed as the data changes
- Data is time-stamped when extracted
  - Allows comparisons between different time periods
- Data is standardized
  - All similar fields (e.g., calendar dates) are structured the same
    - ✦ Date is MM/DD/YYYY
- Data marts are smaller subsets of data warehouses for specific business problems



# Data Warehouse Benefits

24

- Forces organizations to better understand the data
- Centralized view of data to identify inconsistent data
- Once inconsistencies are resolved, higher quality data is used to make better business decisions
- Data can be analyzed over multiple time periods
- Tools are available to combine data and gain more insight into business operations

# Data Mining

25

- Automated process of analyzing data
  - To find previously unknown trends, patterns, and associations
  - To make better business decisions
- Starts with a hypothetical result in mind
- Privacy concerns
  - Easier to combine disparate sources of information and when aggregated tell you much more about the individual
  - Data brokers now to sell this information
- Business intelligence – collecting and analyzing information to increase their competitive advantage
- Business analytics – uses internal company data to improve business processes and practices

# Knowledge Management (KM)

26



- Companies and individuals accumulate knowledge
- Not consistently written down or saved
- If recorded, not consistently organized
- KM is the process of formalizing the capture, indexing, and storing of knowledge

# Database and Data Warehouse

27

ANY QUESTIONS /  
DOUBTS

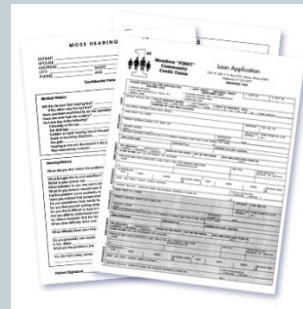
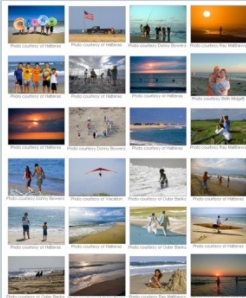
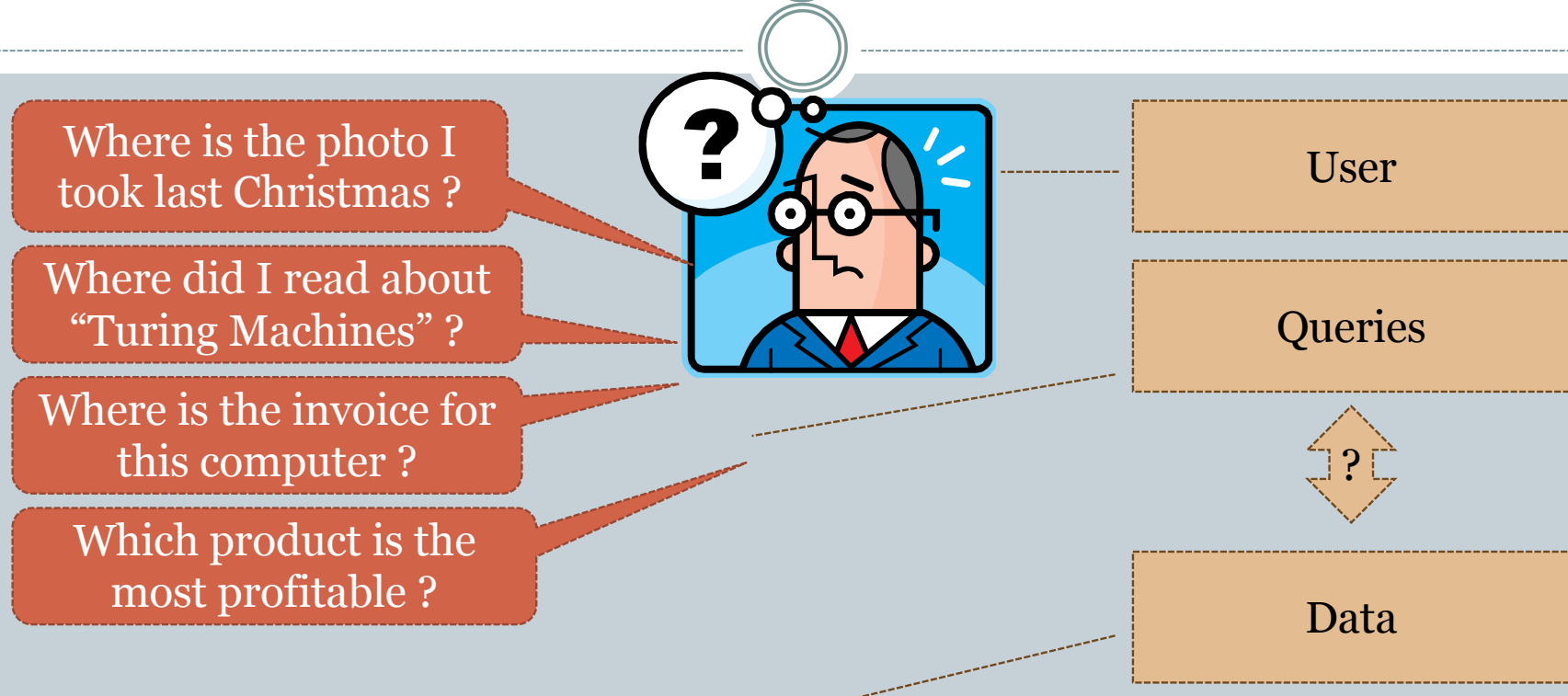
???

## 29





# The Data Management Problem



# What is ``data” ?

31

- **Data** are known facts that can be recorded and that have implicit meaning.
- Three broad categories of data
  - Structured data
  - Semi-structured data
  - Unstructured data
- ``**Structure**” of data refers to the organization within the data that is identifiable.

# What is a database ?

32

- A **database** : a collection of related data.
  - Represents some aspect of the real world (aka universe of discourse).
  - Logically coherent collection of data
  - Designed and built for specific purpose
- A **data model** is a collection of concepts for describing/organizing the data.
- A **schema** is a description of a particular collection of data, using the a given data model.

# Structured vs unstructured data

33

- Structured data tends to refer to information in “tables”

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

Typically allows numerical range and exact match (for text) queries, e.g.,

*Salary < 60000 AND Manager = Smith.*

# Unstructured data

34

- Typically refers to free text
- Allows
  - Keyword queries including operators
  - More sophisticated “concept” queries e.g.,
    - ✦ find all web pages dealing with *drug abuse*
- Classic model for searching text documents

# Semi-structured data

35

- In fact almost no data is “unstructured”
- E.g., this slide has distinctly identified zones such as the *Title* and *Bullets*
  - ✧ ... to say nothing of linguistic structure
- Facilitates “semi-structured” search such as
  - *Title* contains data AND *Bullets* contain search
- Or even
  - *Title* is about Object Oriented Programming AND *Author* something like stro\*rup
  - where \* is the wild-card operator

# Database and Data Warehouse

36

ANY QUESTIONS /  
DOUBTS

???