INFSCI 1022
Database Management Systems

Name: Dmitriy Babichenko

Title: Clinical Associate Professor

Primary Appointment: School of Computing & Information

Secondary Appointment: School of Pharmacy

Today's Evil Plan

- Introduction to course and syllabus overview
- What is data and why do we care?
  - Big Data
  - Data Mining
- Database Management Systems overview
- Tools overview

Introduction to course and syllabus overview

What is Data and Why Do We Care?

Database Management Systems overview

Tools overview

Course Information

- **Term**: Spring 2020
- **Time**: Thursdays 6:00 PM – 8:30 PM
- **Location**: Information Science Building, Room 405

Course Instructor

- **Instructor**: Dmitriy Babichenko
- **Email**: dmb72@pitt.edu
- **Office**: Room 721, IS Building
- **Office Hours**:
  - TBD
  - By appointment

Textbook

- Relational Database Design and Implementation, 4th Edition.
- Author: Jan L. Harrington
- This book is available for free through the University of Pittsburgh library system (http://library.pitt.edu/).
- Direct link is http://proquest.safaribooksonline.com.pitt.idm.oclc.org/9780128499023

You are also responsible for any information/materials

- assigned readings from the textbook
- presented during the lecture by instructor or guest speakers
- linked to from the lecture slides (any links included in the slides are fair game on the quizzes and exams)

Objectives

- Develop solid understanding of database management systems
- Understand how to design and implement relational databases
- Learn how to ask the right questions about data and how to receive (hopefully) correct answers

Objectives

- Become proficient at data modeling and writing SQL queries
- Manage administrative tasks required in a database management environment
- Learn to import and export unstructured data using Python programming language (if we have time)

Assignments

- All of the homework assignments will be individual
- All assignments must be typed (handwritten submissions will not be accepted).
- All assignments must be submitted via Canvas.

Assignments

- If submitting multiple files, they must be zipped into a single file using standard .ZIP format. The final zipped file must be titled with the last names of the author, number of the assignment and course number separated by underscores. For example, if your last name is Smith, and you are submitting assignment 2, your final file should be named **Smith_Assignment2_INFSCI1024.zip**.
- You will lose 2 points for every submission that does not follow this naming convention.

Assignments

- Relational Model
- SQL (4)
- Entity-Relationship Model (MySQL)
- Final project

Late Submissions

Projects/assignments submitted after due date will be accepted, but your overall grade for that project/assignment will be reduced by 10% of the grade for every business day after the submission deadline. For example, if you will submit your work one week late, you will lose 50% of the grade.

Grading Policy

- Homework Assignments: 30%
- In-class labs: 10%
- Midterm Exam: 20%
- Final Exam: 20%
- Final Project: 20%

Grading Scale

- 93 <= A < 100
- 90 <= A- < 93
- 88 <= B+ < 90
- 82 <= B < 88
- 80 <= B- < 82
- 78 <= C+ < 80
- 72 <= C < 78
- 70 <= C- < 72
- 60 <= D < 70
- F < 60

Collaboration vs. Cheating

Collaboration on homework is permitted to an extent. Specifically, students are allowed to discuss the possible solutions to a problem and help each other with logic errors. However, handing your work to someone so that they may see a copy of your solution, or dictating code to a person on line-by-line basis is not within the spirit of the collaboration policy or the honor code of the university.

Academic Integrity Statement

Cheating/plagiarism will not be tolerated. All work must be your own, unless collaboration is specifically and explicitly permitted as in the course group project. Any unauthorized collaboration or copying will at minimum result in no credit for the affected assignment and may be subject to further action under the University Guidelines for Academic Integrity (http://www.provost.pitt.edu/info/ai1.html). You may incorporate excerpts from publications by other authors, but they must be clearly marked as quotations and properly attributed. You may discuss your ideas with others, but all substantive writing and ideas must be your own, or else be explicitly attributed to another, using a citation sufficiently detailed for someone else to easily locate your source.

Disability

If you have a disability for which you are or may be requesting an accommodation, you are encouraged to contact the Instructor and Disability Resources and Services, 216 William Pitt Union, (412) 648-7890 / (412) 383-7355 (TTY), as early as possible in the term. Disability Resources and Services reviews documentation related to a student's disability, provides verification of the disability, and recommends reasonable accommodations for specific courses.

Introduction to course and syllabus overview

What is Data and Why Do We Care?

Database Management Systems overview

Tools overview

What is Data?

What is data?

1. Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation
2. Information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful
3. Information in numerical form that can be digitally transmitted or processed

https://www.merriam-webster.com/dictionary/data

What is data?

- Data is a set of values of **qualitative** or **quantitative** variables.
- Pieces of data are individual pieces of information.
- While the concept of data is commonly associated with scientific research, data is collected by a huge range of organizations and institutions, including businesses, governments, etc...

https://en.wikipedia.org/wiki/Data

What is Big Data?

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

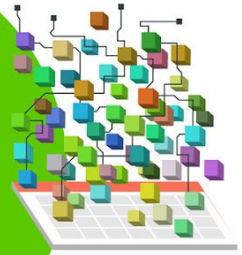
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day



6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

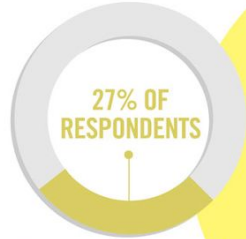
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions



in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Introduction to course and syllabus overview

What is Data and Why Do We Care?

Database Management Systems overview

Tools overview

What is a server?

What is the difference between a spreadsheet and a database?

Databases are safer. Excel, for example, does everything in memory, so that any unsaved data may be lost if your system crashes. Databases write data to the hard drive immediately.

http://www.pcmag.com/article2/0,2817,1435148,00.asp

Databases can handle more data. Sure, Excel can technically handle more than 65,000 rows of data, but doing so will likely bog down even the fastest PC.

http://www.pcmag.com/article2/0,2817,1435148,00.asp

Databases can easily link tables of related data together, such as customers and orders or musical groups and albums (as well as the songs on each album). This is where the words *relational* and *database* come together. Storing related data together in a single table or spreadsheet can be unwieldy and invite errors.

http://www.pcmag.com/article2/0,2817,1435148,00.asp

Most importantly, databases can be used to answer complex questions.

Patient	Date	Symptom	Country
1	07/12/2014	cough, fever	Guinea
2	07/12/2014	cough with blood production, diarrhea, fever	Liberia
3	07/13/2014	reddened eyes, joint and muscle pain, fever	Liberia
4	07/13/2014	fever, fatigue, weakness, reddened eyes	Liberia
5	07/13/2014	joint and muscle pain, headache, nausea and vomiting	Sierra Leone
6	07/13/2014	fever, fatigue, malaise, and weakness, reddened eyes, joint and muscle pain, headache, nausea and vomiting	Guinea
7	07/13/2014	fever, fatigue, weakness, reddened eyes	Sierra Leone
8	07/14/2014	joint and muscle pain, headache, nausea and vomiting	Guinea
9	07/14/2014	cough with blood production, diarrhea, fever	Liberia
10	07/14/2014	fever, fatigue, malaise, and weakness, reddened eyes, joint and muscle pain, headache, nausea and vomiting	Liberia

Types/Brands of Relational DBMS

- Microsoft SQL Server (a.k.a MSSQL)
- Oracle
- MySql
- Microsoft Access
- PostgreSQL
- MariaDB

Microsoft SQL Server

http://www.microsoft.com/en-us/server-cloud/products/sql-server-editions/sql-server-express.aspx

Oracle

http://www.oracle.com/technetwork/database/enterprise-edition/downloads/index.html

MySql

http://www.mysql.com/

Microsoft Access

http://office.microsoft.com/en-us/access/

PostgreSQL

http://www.postgresql.org/

Database Management System (DBMS)

- DBMS contains information about a particular **enterprise**
  - Collection of interrelated data
  - Set of programs to access the data
  - An environment that is both *convenient* and *efficient* to use

Database Applications

- Banking: all transactions
- Airlines: reservations, schedules
- Universities: registration, grades
- Sales: customers, products, purchases
- Online retailers: order tracking, customized recommendations
- Manufacturing: production, inventory, orders, supply chain
- Human resources: employee records, salaries, tax deductions

Entities

- A concept in the business or user environment about which the organization wishes to maintain data
- Person, place, object, event, or concept

Entities Examples

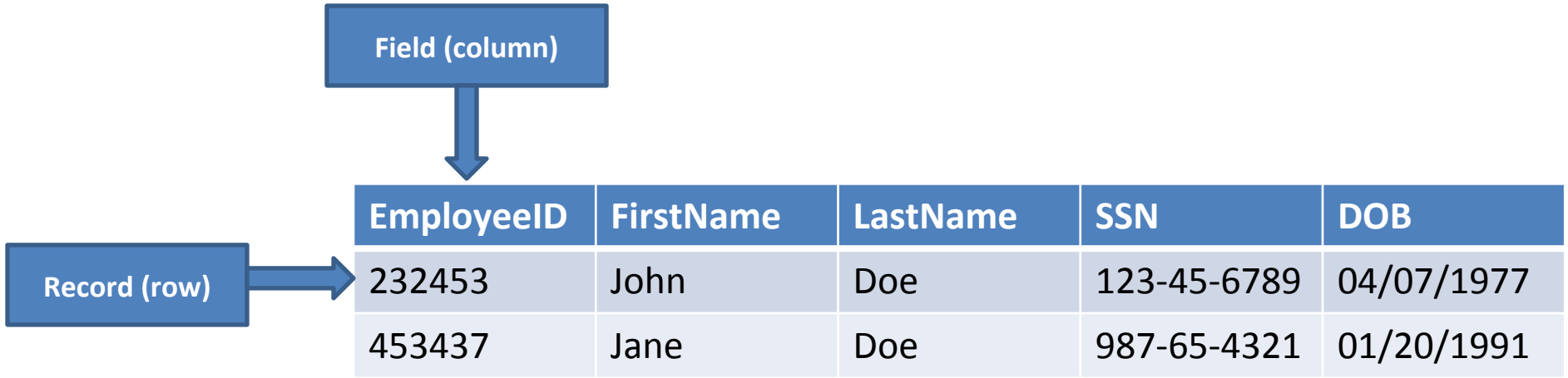
- Person: Employee, Student, Patient
- Place: Store, Warehouse, State
- Object: Machine, Building, Automobile, Product
- Event: Sale, Registration
- Concept: Account, Course

Attributes

- An attribute is a property or characteristic of an entity
- Example: Model, make, year, color are attributes of **Car** entity.

Tables

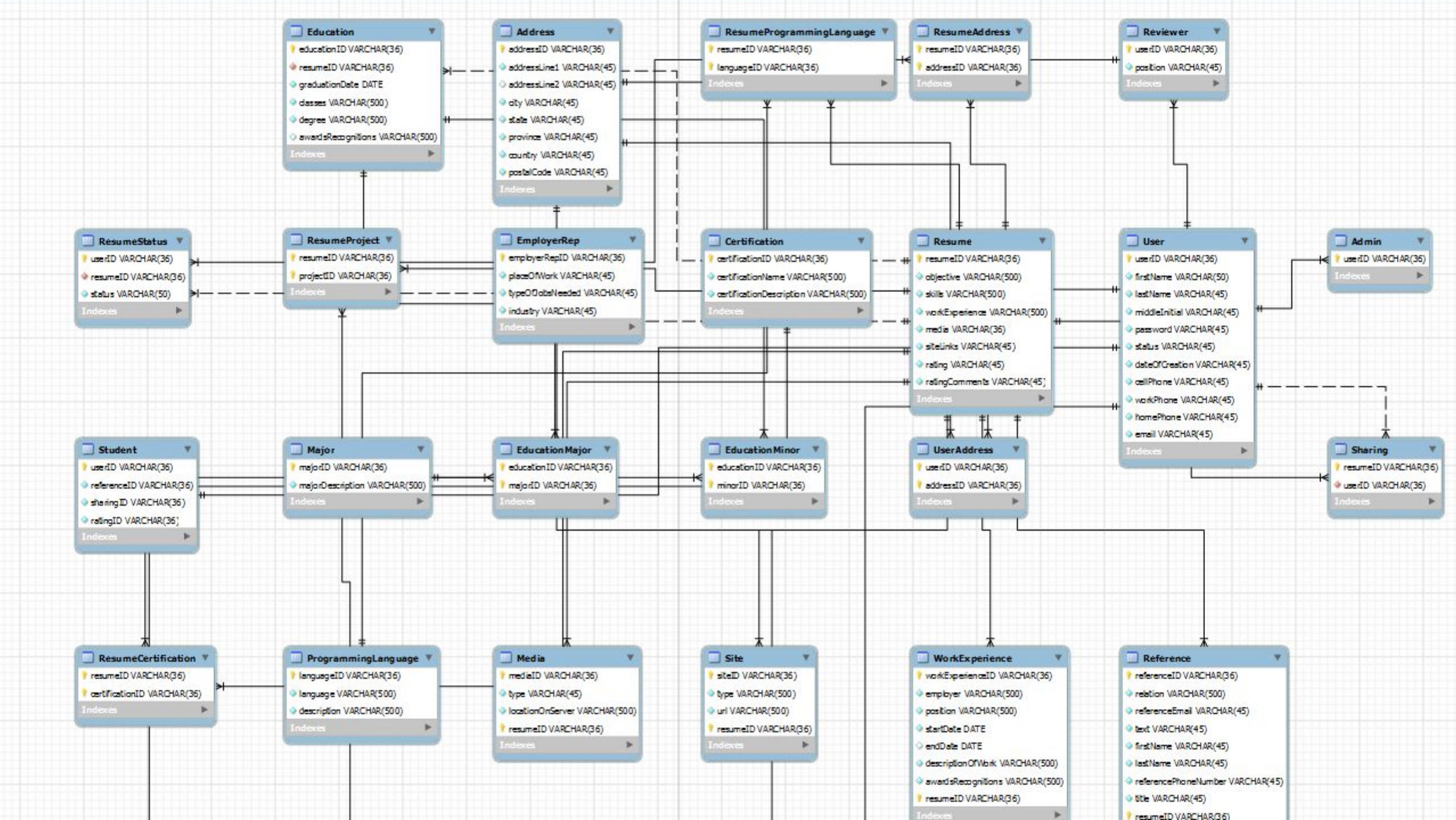
- A database is a collection of tables.
- Tables are referred to as “**entities**”.
- Each table contains records (or tuples) - horizontal rows in the table.
- Each record contains fields - vertical columns of the table.
- Columns are referred to as “**attributes**”



EmployeeID	FirstName	LastName	SSN	DOB
232453	John	Doe	123-45-6789	04/07/1977
453437	Jane	Doe	987-65-4321	01/20/1991

Schema

- A collection of related tables and relationships between those tables is called a **schema**.
- Some database management applications (such as Microsoft SQL Server) allow multiple schemas per database


Education
educationID VARCHAR(36)
resumeID VARCHAR(36)
graduationDate DATE
classes VARCHAR(500)
degree VARCHAR(500)
awardsRecognitions VARCHAR(500)
Indexes
Address
addressID VARCHAR(36)
addressLine1 VARCHAR(45)
addressLine2 VARCHAR(45)
city VARCHAR(45)
state VARCHAR(45)
province VARCHAR(45)
country VARCHAR(45)
postalCode VARCHAR(45)
Indexes
ResumeProgrammingLanguage
resumeID VARCHAR(36)
languageID VARCHAR(36)
Indexes
ResumeAddress
resumeID VARCHAR(36)
addressID VARCHAR(36)
Indexes
Reviewer
userID VARCHAR(36)
position VARCHAR(45)
Indexes
ResumeStatus
userID VARCHAR(36)
resumeID VARCHAR(36)
status VARCHAR(50)
Indexes
ResumeProject
resumeID VARCHAR(36)
projectID VARCHAR(36)
Indexes
EmployerRep
employerRepID VARCHAR(36)
placeOfWork VARCHAR(45)
typeOfJobsNeeded VARCHAR(45)
industry VARCHAR(45)
Indexes
Certification
certificationID VARCHAR(36)
certificationName VARCHAR(500)
certificationDescription VARCHAR(500)
Indexes
Resume
resumeID VARCHAR(36)
objective VARCHAR(500)
skills VARCHAR(500)
workExperience VARCHAR(500)
media VARCHAR(36)
siteLinks VARCHAR(45)
rating VARCHAR(45)
ratingComments VARCHAR(45)
Indexes
User
userID VARCHAR(36)
firstName VARCHAR(50)
lastName VARCHAR(45)
middleInitial VARCHAR(45)
password VARCHAR(45)
status VARCHAR(45)
dateOfCreation VARCHAR(45)
cellPhone VARCHAR(45)
workPhone VARCHAR(45)
homePhone VARCHAR(45)
email VARCHAR(45)
Indexes
Admin
userID VARCHAR(36)
Indexes
Student
userID VARCHAR(36)
referenceID VARCHAR(36)
sharingID VARCHAR(36)
ratingID VARCHAR(36)
Indexes
Major
majorID VARCHAR(36)
majorDescription VARCHAR(500)
Indexes
EducationMajor
educationID VARCHAR(36)
majorID VARCHAR(36)
Indexes
EducationMinor
educationID VARCHAR(36)
minorID VARCHAR(36)
Indexes
UserAddress
userID VARCHAR(36)
addressID VARCHAR(36)
Indexes
Sharing
resumeID VARCHAR(36)
userID VARCHAR(36)
Indexes
ResumeCertification
resumeID VARCHAR(36)
certificationID VARCHAR(36)
Indexes
ProgrammingLanguage
languageID VARCHAR(36)
language VARCHAR(500)
description VARCHAR(500)
Indexes
Media
mediaID VARCHAR(36)
type VARCHAR(45)
locationOnServer VARCHAR(500)
resumeID VARCHAR(36)
Indexes
Site
siteID VARCHAR(36)
type VARCHAR(500)
url VARCHAR(500)
resumeID VARCHAR(36)
Indexes
WorkExperience
workExperienceID VARCHAR(36)
employer VARCHAR(500)
position VARCHAR(500)
startDate DATE
endDate DATE
descriptionOfWork VARCHAR(500)
awardsRecognitions VARCHAR(500)
resumeID VARCHAR(36)
Indexes
Reference
referenceID VARCHAR(36)
relation VARCHAR(500)
referenceEmail VARCHAR(45)
text VARCHAR(45)
firstName VARCHAR(45)
lastName VARCHAR(45)
referencePhoneNumber VARCHAR(45)
title VARCHAR(45)
resumeID VARCHAR(36)

SQL

- SQL stands for Structured Query Language
- **Semi-standardized** language for querying relational databases
- SQL differs slightly between database brands
- **SQL != Database**

Query to select first 5 rows from a database table called Employees:

- Microsoft SQL Server:
  - *SELECT **TOP 5** * FROM Employees*
- MySQL
  - *SELECT * FROM Employees **LIMIT 5**;*

Database Design

- The process of designing the general structure of the database consists of:
  - Logical design
  - Physical design

Logical Design

- Logical Design – deciding on the database schema. Database design requires that we find a “good” collection of relation schemas.
  - **Business decision** – What attributes should we record in the database?
  - **Computer Science decision** – What relation schemas should we have and how should the attributes be distributed among the various relation schemas?

Physical Design

- Physical Design – Deciding on the physical layout of the database
  - File system
  - Indexes

Database Architecture

The architecture of a database systems is greatly influenced by the underlying computer system on which the database is running:

- Centralized *(our focus in this class)*
- Client-server
- Parallel (multi-processor)
- Distributed

Data Warehouses

- System used for reporting and data analysis
- Considered a core component of business intelligence
- Central repositories of integrated data from one or more disparate sources.
- Store current and historical data in one single place
- Used for creating analytical reports for knowledge workers throughout the enterprise

https://en.wikipedia.org/wiki/Data_warehouse

Data Warehouses

- The data stored in the warehouse is uploaded from the operational systems (such as marketing or sales).
- The data may pass through an operational data store and may require data cleansing for additional operations to ensure data quality before it is used in the DW for reporting.

https://en.wikipedia.org/wiki/Data_warehouse

Three SQL programmers walk into a NoSQL bar.

A little while later they walked out… because they couldn’t find a table

NoSQL

- **Document databases** pair each key with a complex data structure known as a document. Documents can contain many different key-value pairs, or key-array pairs, or even nested documents.
- **Graph stores** are used to store information about networks of data, such as social connections. Graph stores include Neo4J and Giraph.

NoSQL Data

- Great for unstructured data:
  - Emails
  - Text files
  - Spreadsheets
  - Digital Images
  - Video
  - Audio
  - Social media posts
- List of NoSQL databases: http://nosql-database.org/

The Benefits of NoSQL

- More scalable than relational DBs
- Provide superior performance
- Their data model addresses large volumes of rapidly changing structured, semi-structured, and unstructured data
- Object-oriented programming that is easy to use and flexible
- Geographically distributed scale-out architecture instead of expensive, monolithic architecture

https://www.mongodb.com/nosql-explained

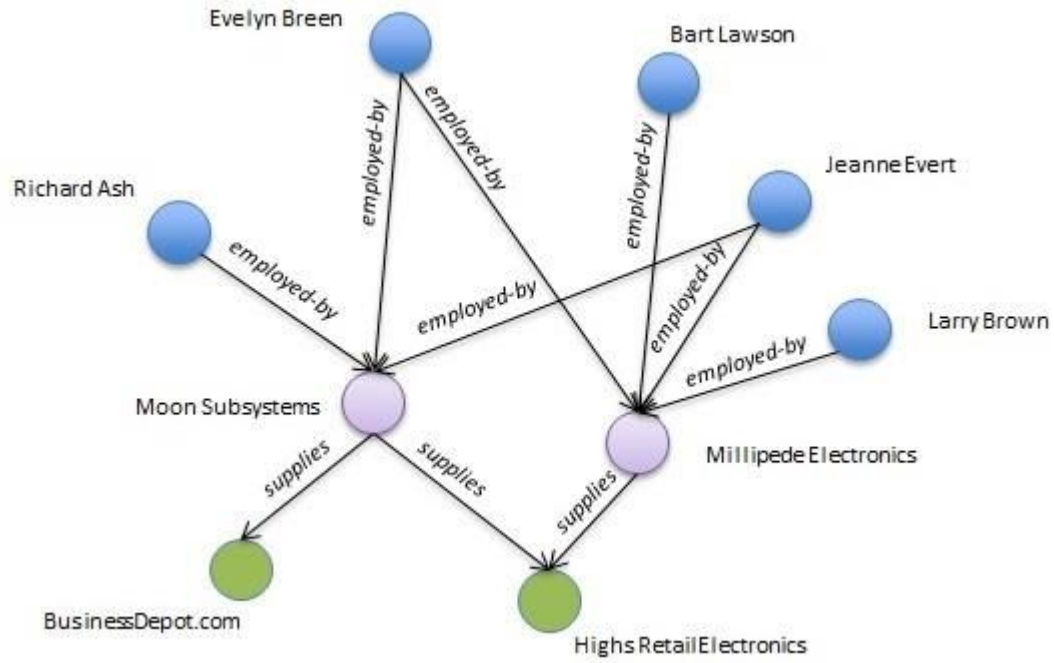
Document Databases

Designed for storing, retrieving, and managing document-oriented, or semi structured data.

https://docs.mongodb.com/manual/core/data-modeling-introduction/

Graph Databases

- Use graph structures for semantic queries with nodes, edges and properties to represent and store data.
- A key concept of the system is the graph where entities are **nodes** and relationships are **edges**.



https://upside.tdwi.org/articles/2016/07/25/graph-databases-for-analytics-1.aspx

Introduction to course and syllabus overview

What is Data and Why Do We Care?

Database Management Systems overview

Tools overview

Installing MySQL Workbench

MySQL Workbench is a visual database design and management tool created specifically for MySQL database management system. You can download MySQL Workbench 6 from https://dev.mysql.com/downloads/workbench/6.0.html

Connect MySQL Workbench to Server

- Connection name: INFSCI1022
- Hostname: sis-teach-01.sis.pitt.edu
- Port: 3306
- Username: is1022practice
- Password: 1s!O22Pra6t1c3

Draw.IO

- Login to Google Drive (https://drive.google.com)
- Go to New → More → Connect More Apps
- Find draw.IO app
- Connect it to your Google Drive

Questions?