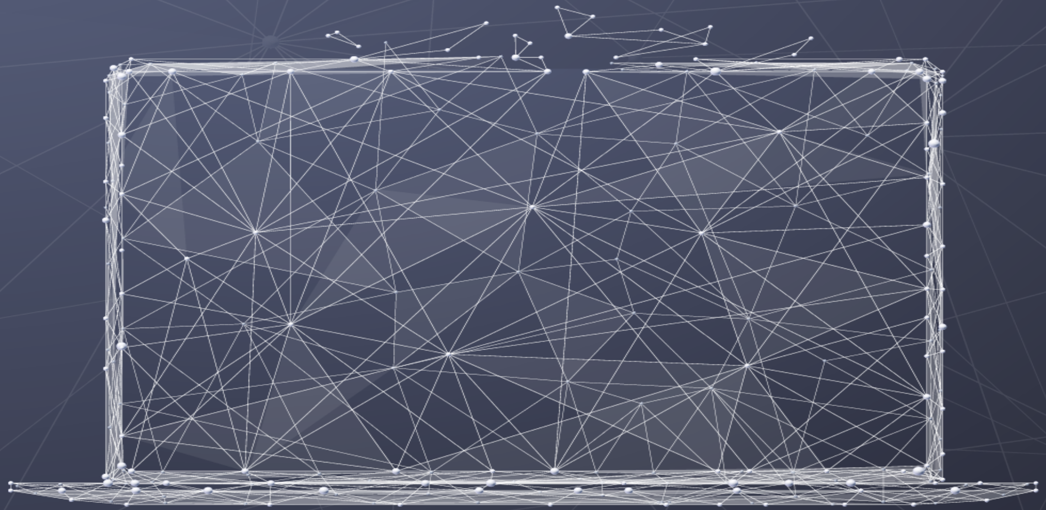
Data Science
“Data Engineering II”

Data Engineering II



PURDUE UNIVERSITY®

College of Science

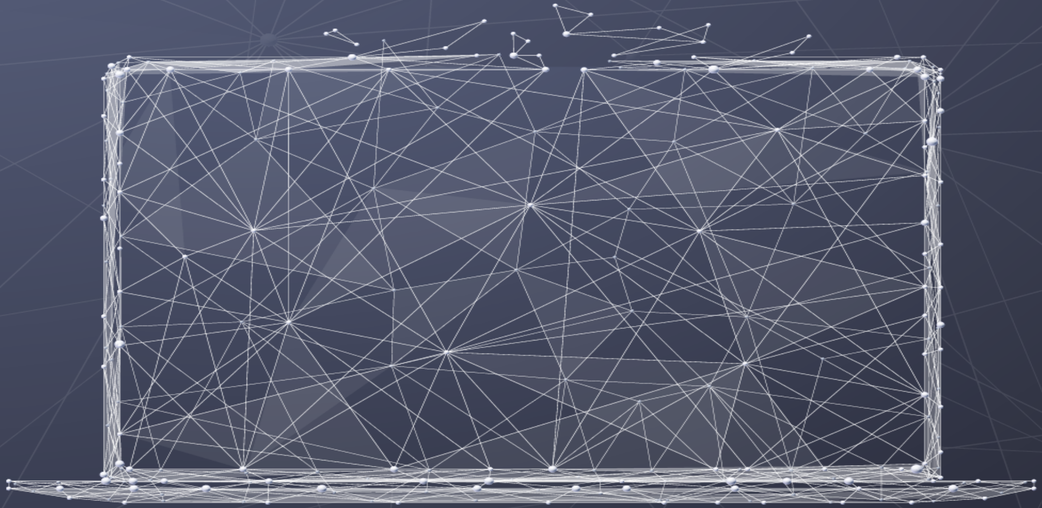
Course Outline

Summary and Topics

- This module is a short introduction to the use of database management systems from a user's perspective.
- We will cover
  - Modeling an enterprise using Entity-Relationship Diagrams
  - Transforming the model into a Relational or NoSQL database
  - Querying and updating the database using SQL
  - An introduction to query processing and indexing
- This module will not cover the design and implementation of the database management system itself.

Data Science
“Data Engineering II”

Week1: Lecture 1



PURDUE
UNIVERSITY®

College of Science

Week 1, Lecture 1 Outcome

Topics

In this lecture, we will introduce data management challenges faced my most modern enterprises and discuss the advantages of using a database management system over a file based solution for managing enterprise data.

Databases are Everywhere

Modern Applications Require Management of Data

- Retail operations
- Banking applications
- Online services: Amazon, Application Stores, Yelp, AirBNB, …
- University records
- Governmental services: FDA, Social Security, Medicare, …
- ….

Data Management Needs

Typical requirements for an enterprise

- Complexity: lots of different types of data, processes, policies
- Scale: data collections are often very large in size
- Concurrent Access: multiple users need to access and update parts of the data at the same time
- Crash Recovery: data and updates should not be lost or tampered due to a power failure, or software bug, disk drive failure etc.
- Security: access to read and/or modify data should be limited

Sample Enterprise

University Grade Management System

- Assume that you are asked to develop a system to manage Purdue University's Grades.
  - Students — information about each student
    - ID, name, GPA, courses enrolled
  - Instructors — a record of each instructor
    - ID, name, office, email
  - Courses — information about each course
    - Instructor, room, schedule, department
  - Grades — a record of each student's grade for each course

Simplistic File Based Solution

Data Organization Decisions

- We could choose to store the data the following three files:
  - `students.txt` (information for each student per line)
  - `instructors.txt` (information for each instructor per line)
  - `courses.txt` (information for each course offered per line)
  - `grades.txt` (information for each grade per line)
- We need to decide the order of each piece of information, how to handle missing or multiple values, how to separate fields, etc.
- How to record who is enrolled in what course? How to record the grades?

File Based Solution

Data Access and Update

- We need to write code to parse these files to insert, extract, and update the necessary information, e.g.,
  - Add new courses, instructors, grades,
  - Find the details of a given student or course
  - Find the student with the highest GPA
  - Update the grade of a student for a given course
  - …

Drawbacks of File Based Solution - 1

Handling Changes to the Enterprise

- Suppose that we need to
  - Find instructors teachi
- Or, ensure certain constra
  - Ensure that no studen
- Or, add new features
  - Introduce multiple sections for courses

Each of these changes would require writing new code and potentially rewriting existing code. This require an expert programmer that understands all the details of each file, format, existing codebase, …

Drawbacks of File Based Solution - 2

Concurrent Users and Recovery

- Every piece of code tha
  ensure
  - Correctness despite
  - Resilience to failure
  - Only authorized users are allowed run code that updates or views certain information

Requires careful attention by the developer to EVERY potential interleaving of multiple users, and failures, while ensuring that ALL access honors the end user's security policies.

PURDUE UNIVERSITY | College of Science

Drawbacks of File Based Solution - 3

Efficiency

- Imagine that the size of c
- How do we ensure that c
  - Data reorganization r
  - Managing the use of
  - How to choose betw linear search)
    - Depends upon the data and the hardware.

Any changes to how answers are computed would require code rewrites, and be closely tied to the physical data layout and the architecture of the machine on which the program is run — not easy to dynamically adapt.

Drawbacks of File Based Solution - 4

Data Quality and Consistency

- How do we ensure
  - The grade for a c
  - A student is not e
  - A course has only
  - ….

Such tests would need to be consistently enforced throughout EVERY piece of code that modifies or inserts relevant data into a file.
It is up to the developers to ensure that these tests are applied consistently in each instance — and updated consistently when rules change.

A Better Solution: Database Management Systems

Advantages of a DBMS over Files

- Provides a uniform high-level logical view of the data
  - Independent of the data and physical storage
- Ensures correctness in presence of concurrent users and crashes
- Reduces data inconsistency problems
- Allows ad hoc queries — no need for programming each time
- Much faster development of the database application
- Automatic optimization of queries
- Ensures access follows a high-level specification of rules

A Better Solution: Database Management Systems

Advantages of a DBMS over Files

- Using a DBMS allows the user to describe and manipulate the data at a higher, logical level
  - Only need to specify what is particular to the given enterprise
- Application developers can create enterprise-specific transactions/operations built on top of the DBMS layer using a simple notion of transactions
  - They can ignore concurrent access, failures, access control, and efficiency — all of which the DBMS automatically handles

Basic Database Concepts

What is a Database?

- A very large, integrated collection of data.
- Models a real-world enterprise (e.g., university grades)
- The database is managed by a Database Management System (DBMS), which is a software package designed to store and manage databases.
  - Oracle, DB2 (IBM), MS SQL Server, MS Access
  - PostgreSQL, MySQL

Clarification

DBMS vs. File System

- We can choose to use a flat file based solution to manage our data or use a DBMS
- A DBMS stores data using the file system
  - Both for the data and additional information used by the DBMS
- The DBMS carefully manages the organization of the data in files and when it is moved to/from main memory
  - The DBMS shields the user from many of these aspects
  - Physical Independence!

Basic Database Concepts

Data Model

- A set of concepts to describe the structure of a database, the operations for manipulating these structures, and certain constraints that the data should obey.
- *Relational, Object-Oriented,* and *Object-Relational* are most common

Relational Data Model

- All data is organized in tables consisting of rows and columns
- Built upon the formal mathematical notion of a relation
- Each table, or relation, has a well-defined structure derived from the application

Example

Consider the following enterprise of a university:

- STUDENTs enroll in COURSES
- INSTRUCTORs teach COURSES
- COURSEs are offered by DEPARTMENTs

Example Schema

Database Schema

STUDENTS

Student ID	First Name	Last Name	GPA
Integer	String	String	Real

COURSES

CourseID	Department	Credits	Instructor
Integer	String	Real	Integer

INSTRUCTORS

Instructor ID	First Name	Last Name	Office
Integer	String	String	String

DEPARTMENTS

Name	Head	Office	College
String	Integer	String	String

Example

Database Instance

STUDENTS

Student ID	First Name	Last Name	GPA
12345	Jane	Doe	4.0
4988	John	Doe	3.58
877788	Nicole	Parker	3.75

COURSES

CourseID	Department	Credits	Instructor
CS54100	Computer Science	3.0	8778
CS54200	Computer Science	3.0	98748

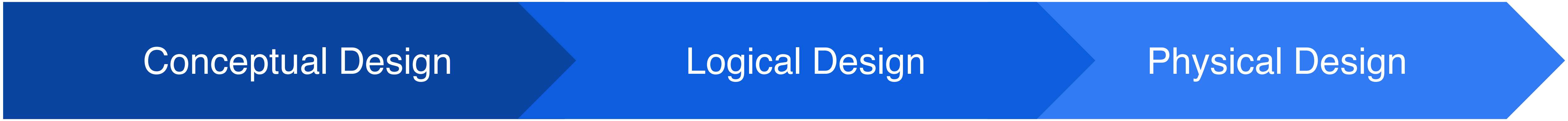
INSTRUCTORS

Instructor ID	First Name	Last Name	Office
8778	Sunil	Prabhakar	LWSN 2116K
98748	Walid	Aref	LWSN 2116J

DEPARTMENTS

Name	Head	Office	College
Computer Science	838348	Lawson 3144	Science

Database Design



Creating an Entity Relationship Diagram (ERD) which describe the real world enterprise entities, attributes and the relationships among them.

Transforming ERD to relational model: tables, keys (constraints), etc.

Creating the database and other supporting structures based on a specific DBMS. E.g. Mysql

Week 1, Lecture 1 Outcome

Topics

In this lecture, we will introduce data management challenges faced my most modern enterprises and discuss the advantages of using a database management system over a file based solution for managing enterprise data.