EECS 484: Database Management Systems

Instructor:

Atul Prakash (Email: aprakash@umich.edu)

Course Policies

Course Outline – EECS 484

- GOAL: Basic introduction to database management systems.
- Two perspectives:
  - External (*Database user*)
    - Data models, ER model, relational model, SQL, database design …
    - Java/JDBC Project: Common platform for building database applications
  - Internal (*Database implementer*)
    - File organizations, access methods, sorting, concurrency control, recovery, …
    - Minirel Project: Build components of a Relational Database System
- Textbook "Database Management Systems", by Raghu Ramakrishnan & Johannes Gehrke. 3rd ed.

Java

- Databases are most often accessed via SQL.
- But SQL is usually embedded in, and called from, a traditional programming language.
- Java is common choice, and so you will be using that in a project.

Groups

- Total of four projects and 5-6 homeworks
- Highly recommended: group of size 2
- Single-person project and homework submission is allowed, but discouraged
- Start looking for partners now!

Project Grading

- Mostly autograder, some human.
- Limited number of submissions, even for autograded portion.
  - Make sure to test extensively.
- Both partners expected to contribute and be familiar with all aspects of the project

Discussion Sections

- Not optional!
- Project and homework discussion
- Sometimes, discussions may run ahead of lectures or cover additional relevant topics

Honor Code – Course Policies

- CoE Honor Code for all students
- Key principle: No unfair advantage
- Your work must be original – no peeking at old solutions, sharing of code, or discussing the projects beyond your group
- **No public posting of solutions**, e.g., even after the course.
- **Private repos** to share with your partner or a potential employer are OK
- Posting questions on Piazza or using office hours for help is fine
- Questions? – Just ask.
- Also see Canvas for link to CoE Honor Code

Course Policies

- Projects
  - Due by **6 PM** on due date (Thursdays)
  - 4 free late days total across all projects for any reason. Free late days used up first
  - 10% penalty for each late day (or part thereof) used beyond the free day
  - Max of 5 late days, including grace days, allowed per project
- Homework assignments
  - Due by **6 PM** on due date (Thursdays)
  - No late submissions accepted, but lowest score dropped

Course Grading

First Exam	25%
Second Exam	25%
Homework assignments	10%
Four projects [each worth 10%]	40%

Note: Additional 0-2% may be allocated for participating in course evaluations or other bonus point opportunities, announced in advance. The above percentages may thus be marginally prorated.

Exams

- Two exams:
- One midterm and one final
- Non-cumulative.
- Closed book but a one 8.5”x11” page, two-sided, handwritten cheat sheet is permitted.

This week

- No discussion.
- No office hours.

- Regular schedule starts next week for discussion and office hours.

Lectures

- Notes are posted on canvas
- Sometimes updated after lecture.
  - To fix errors
  - To add clarifications

Overview of DBMS and Topics

What Is a DBMS?

- DBMS = Database Management System
- Database: Large, integrated collection of data.
- Models some real-world *enterprise*
  - Entities (e.g., students, courses)
  - Relationships (e.g., Lisa Simpson is taking EECS 484)
- DBMS: a software package designed to store and manage databases

Old-time Solution: Sorted Student Folders

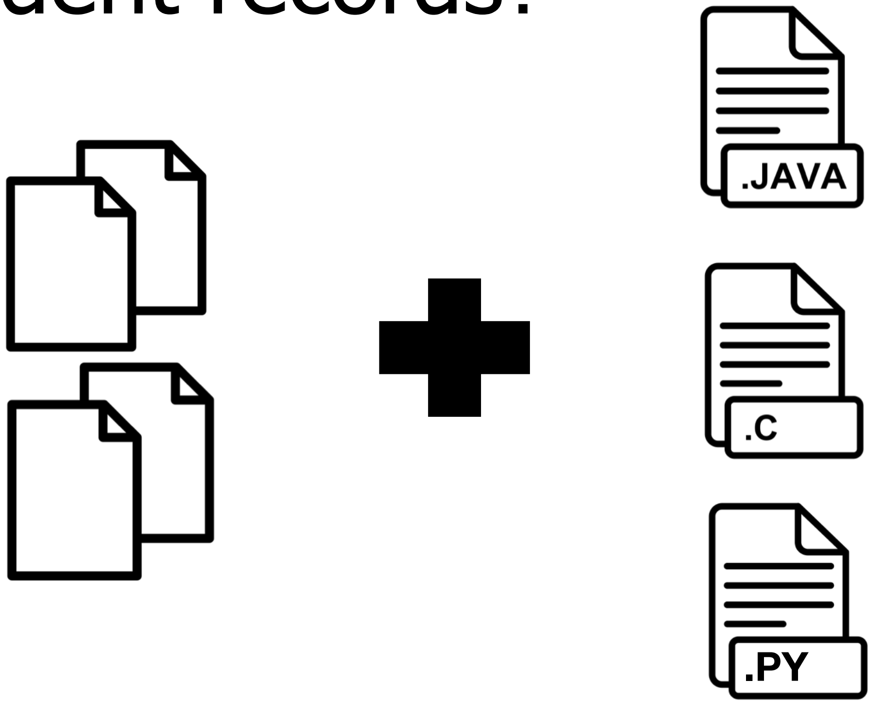
- Advantages?
- Disadvantages?

Old-time Solution: Sorted Student Folders

- Advantages?
  - cheap
- Disadvantages?
  - Large physical footprint
  - No sharing
  - No ad-hoc queries

Other Solution: Flat Files

- Access?
  - using programs in C, Java, Python etc.
- Layout for the student records?

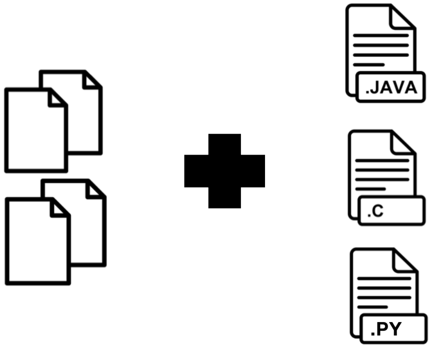


Other Solution: Flat Files

- Access?
  - using programs in C, Java, etc.
- Layout for the student records?

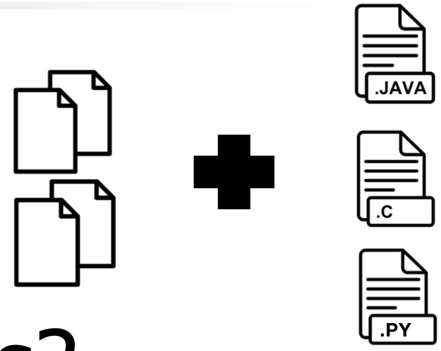
CSV:

Brown, Lisa, lbrown, db, A, os, B
Smith, Bart, bsmith
Tompson, Mary, mtom, vis, B+, db, A-
…
…



Other Solution: Flat Files

- Access?
  - using programs in C, Java, etc.
- Layout for the student records?



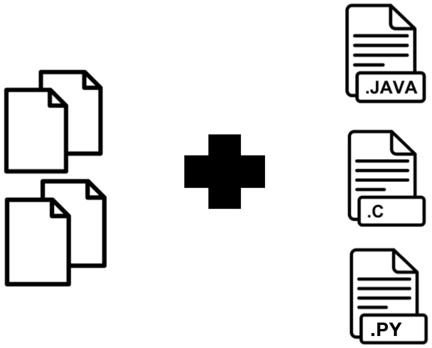
Multiple files:

Brown, Lisa, lbrown
Smith, Bart, bsmith
Tompson, Mary, mtom
…
…

lbrown, db, A
lbrown, os, B
mtom, vis, B+
mtom, db, A-
…
…

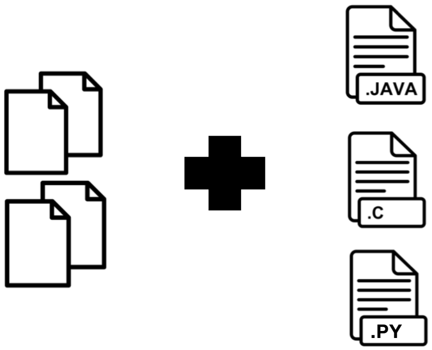
Other Solution: Flat Files

- Problems?



Other Solution: Flat Files

- Problems?
  - Inconvenient access to data
    - requires programming experience and knowledge of file layout
  - Data redundancy
  - Integrity problems
  - Atomicity problems (concurrent access issues)
  - Security problems



Why use a DBMS?

- It solves ALL these problems!
  - Data independence
    - Apps need a view of the data, not info about internal representation and storage
  - Efficient storage and access
  - Centralized data administration
  - Data integrity and security
  - Concurrent access, recovery from crashes
  - Reduced application dev time

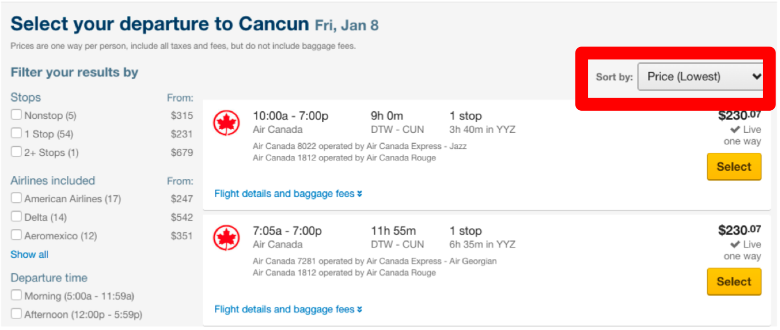
Who uses a DBMS?

Who uses a DBMS?

- Everyone!
  - Your bank
  - Your university
  - Your coffee shop
  - Your favorite hotel
  - Your favorite website
  - Your phone
  - Your government
- How many databases have you used so far today?

Why Study Databases?

- Data is useless without the tools to extract information (queries)
  - “Optimal” pricing of an airline ticket



Select your departure to Cancun Fri, Jan 8

Prices are one way per person, include all taxes and fees, but do not include baggage fees.

Filter your results by

Stops	From:
□ Nonstop (5)	$315
□ 1 Stop (54)	$231
□ 2+ Stops (1)	$679

Airlines included	From:
□ American Airlines (17)	$247
□ Delta (14)	$542
□ Aeromexico (12)	$351

Show all

Departure time

□ Morning (5:00a - 11:59a)

□ Afternoon (12:00p - 5:59p)

Sort by: Price (Lowest)

10:00a - 7:00p Air Canada | 9h 0m DTW - CUN | 1 stop 3h 40m in YYZ | $230.07 ✓ Live one way

Air Canada 8022 operated by Air Canada Express - Jazz
Air Canada 1812 operated by Air Canada Rouge

Select

Flight details and baggage fees

7:05a - 7:00p Air Canada | 11h 55m DTW - CUN | 1 stop 6h 35m in YYZ | $230.07 ✓ Live one way

Air Canada 7281 operated by Air Canada Express - Air Georgian
Air Canada 1812 operated by Air Canada Rouge

Select

Flight details and baggage fees

- Datasets increasing in diversity and volume
  - Websites, digital libraries, interactive video, human genome project, mobile applications

- Databases touch most of CS
  - OS, languages, theory, AI, multimedia, logic, …

Data Models

- Data model: a collection of concepts for describing data.
- Schema: a description of a particular collection of data, using a given data model.
- Relational model: the most widely-used model today.
- Entity-Relationship (ER) model: A “semantic” data model, i.e. a higher-level more user-intuitive model
  - A (relational) DBMS only understands the relational model
    ➔ Must translate an ER schema to a relational schema

Relational and Other Data Models

- **DBMS using the relational DM** ('70s-'80s)
  - IBM DB2
  - Informix
  - Oracle
  - Sybase
  - Microsoft Access
  - Tandem
  - Teradata
  - …

- **Other data models**
  - Hierarchical (mid '60s-'70s)
    - IBM IMS
  - Network ('70s)
    - IDMS, IDS
  - Object-oriented (~'90s)
    - ObjectStore
  - Object-relational (relational model + object DB concepts)
    - Oracle
  - …

Relational (Data) Model

- The most widely-used model today
  - A collection of relations
  - Relation = set of records – think of it as a table with rows and columns

Students

sid	name	login	age
13	Lisa	lsimp	40
41	Bart	bart	20

Courses

cid	cname	cred.
E-484	EECS484	4
E-584	EECS584	3

Enrolled

sid	cid	grade
41	E-484	A-
13	E-584	A+

Relational (Data) Model

- **Schema** = a description of data in terms of a data model
  - Every relation has a schema
  - Specifies the name of the relation, the name and type of the columns (or fields or attributes)
  - Each row also called a tuple or a record

Students(sid:`string`, name:`string`, login:`string`, age:`integer`)
Courses(cid:`string`, cname:`string`, credits:`integer`)
Enrolled(sid:`string`, cid:`string`, grade:`string`)

Students

sid	name	login	age
13	Lisa	lsimp	40
41	Bart	bart	20

Courses

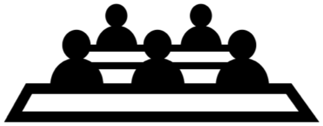
cid	cname	cred.
E-484	EECS484	4
E-584	EECS584	3

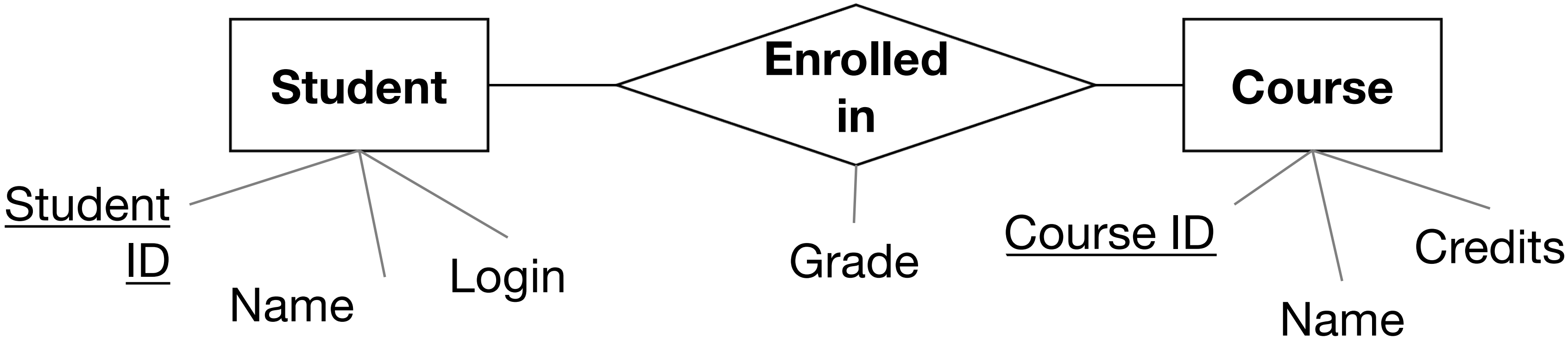
Enrolled

sid	cid	grade
41	E-484	A-
13	E-584	A+

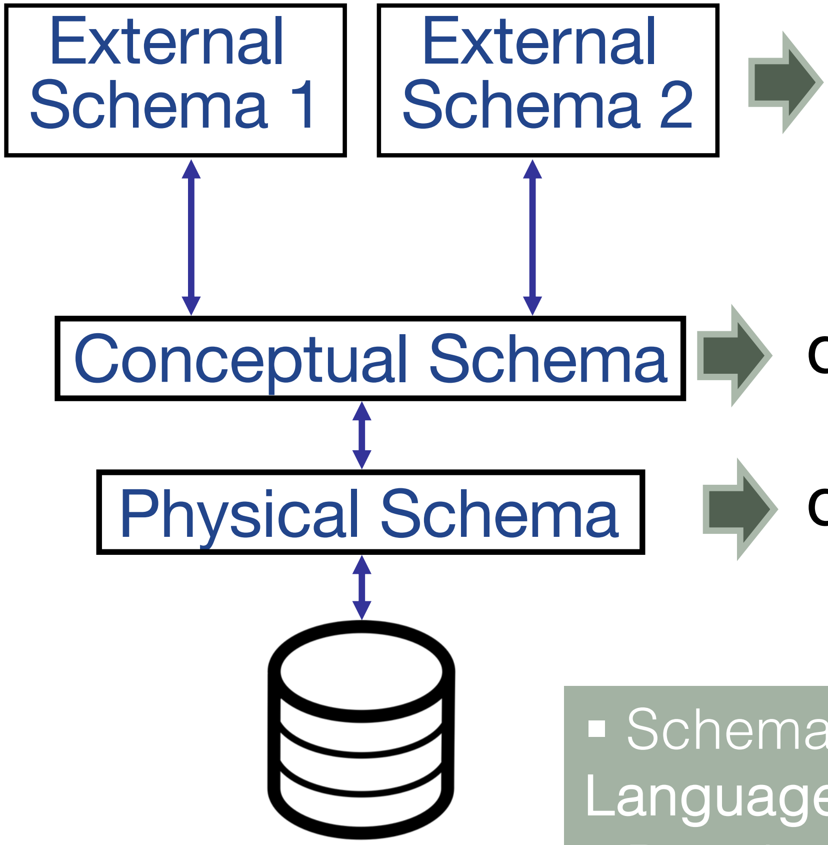
Entity-Relationship (ER) Model

- A “semantic” data model
  - a higher-level, more user-intuitive model
- Entity-Relationship diagram:
  - Entities: Student, Course
  - Relationship: Enrolled_in





Levels of Abstraction



External Schema 1

External Schema 2

Conceptual Schema

Physical Schema

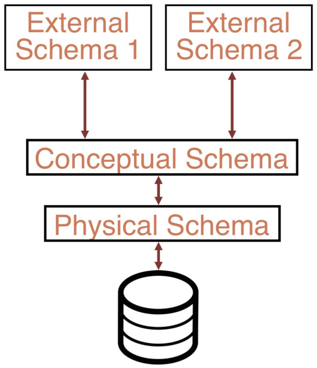
describe how users see the data, views

defines the logical structure (schema)

describes the files and indexes used

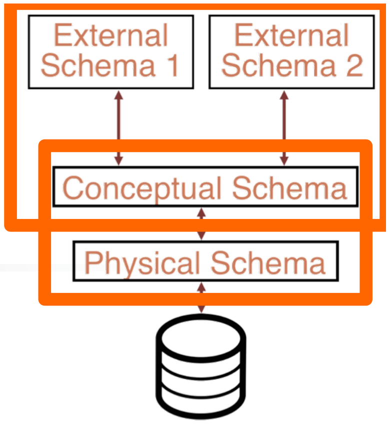
- Schemas are defined using **Data Definition Language (DDL)** – e.g., `CREATE TABLE X (…)`
- Data is modified/queried using **Data Manipulation Language (DML)** – e.g., `SELECT FROM X WHERE …`

Example



- Conceptual schema (1):
  - Students(sid:`string`, name:`string`, login:`string`, age:`integer`)
  - Courses(cid:`string`, cname:`string`, credits:`integer`)
  - Enrolled(sid:`string`, cid:`string`, grade:`string`)
- Physical schema (1):
  - Relations stored as unordered files.
  - Index on first column of Students.
- External Schema (≥ 1):
  - View: Course_info(cid:`string`, enrollment:`integer`)
  - View: Class_rank(sid:`string`, gpa:`real`, rank:`integer`)

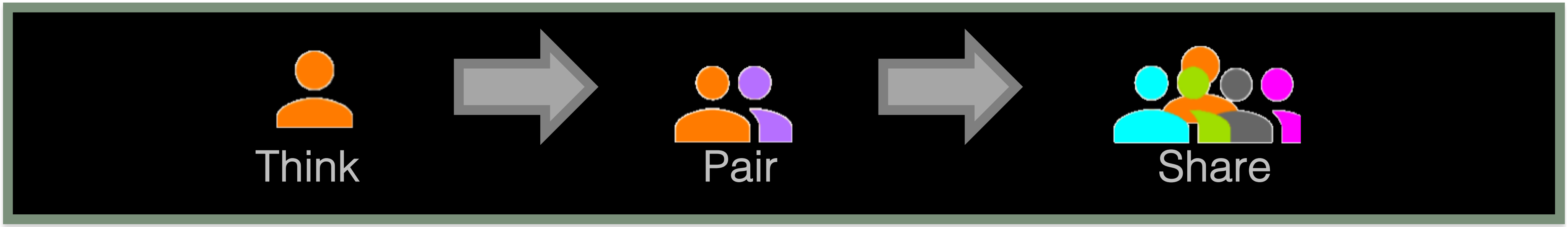
Data Independence



- Applications insulated from data format and storage details
- Logical data independence: Protection from changes in *logical* structure of data
  - External / Conceptual schemas
- Physical data independence: Protection from changes in *physical* structure of data
  - Conceptual / Physical schemas

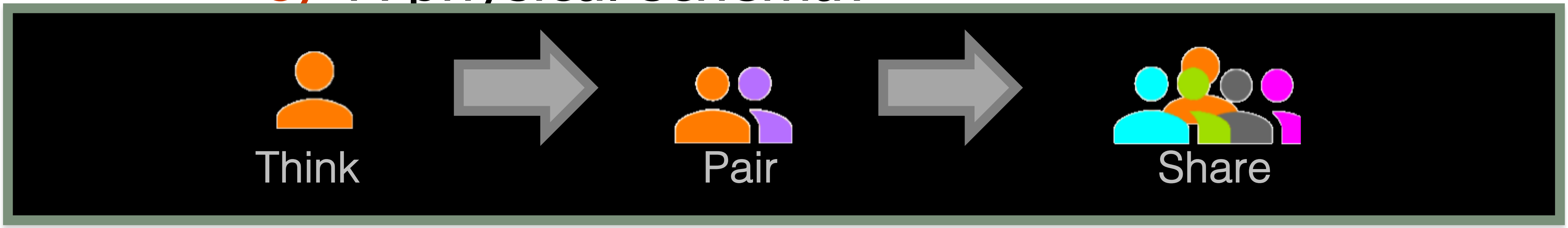
CYU

- Which of these are more suitable for storing in a DBMS rather than files in an OS?
  (a) Grades for students at the university
  (b) Source code for a program
  (c) Contents of a textbook



CYU

- Let's say UM provides you access to a relational table that gives just your grades in various courses. Does that relation represent:
  a) An external schema?
  b) A conceptual schema?
  c) A physical schema?



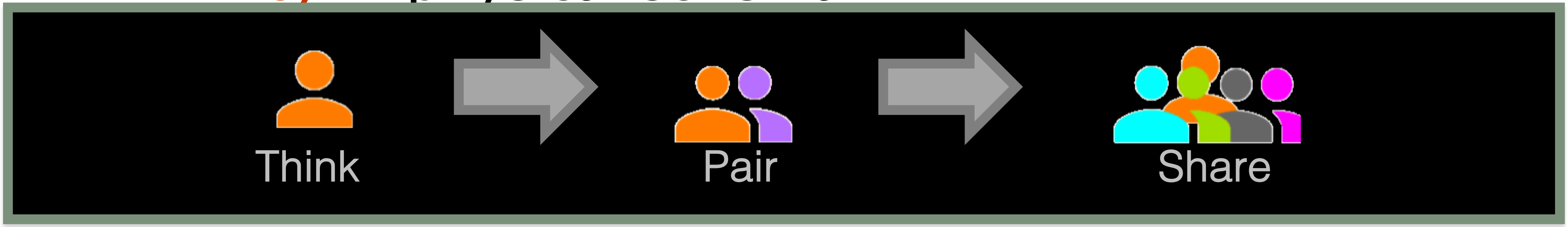
Think

Pair

Share

CYU

- The relational table with student grade information is very large and stored on multiple servers for performance. Does the storage scheme represent:
  a) An external schema?
  b) A conceptual schema?
  c) A physical schema?



Think

Pair

Share

Lots of People use DBMS …

- DBMS vendors
- DB application programmers
  - E.g. smart webmasters
- *Database administrator (DBA)*
  - Designs logical /physical schemas
  - Handles security and authorization
  - Data availability, crash recovery
  - Database tuning as needs evolve

Must understand how a DBMS works!

Summary

- DBMS used to maintain, query large datasets.
- Benefits include recovery from system crashes, concurrent access, quick application development, data integrity and security.
- Levels of abstraction give data independence.
- DBAs hold responsible jobs and are well-paid!
- DBMS R&D is one of the most exciting areas in CS.