



# Database Architecture

---

## Database Design

Department of Computer Engineering

Sharif University of Technology

Maryam Ramezani [maryam.ramezani@sharif.edu](mailto:maryam.ramezani@sharif.edu)



## ❑ Three characterizations:

- **Structured Data**
  - It is represented in a strict format.
  - The DBMS then checks to ensure that all data follows the structures and constraints specified in the schema.
- **Semi-Structured Data**
  - In some applications, data is collected in an ad-hoc manner before it is known how it will be stored and managed.
  - This data may have a certain structure, but not all the information collected will have identical structure.
  - The schema information is mixed in with the data values, since each data object can have different attributes that are not known in advance.
- **Unstructured Data**
  - There is very limited indication of the type of data.

# Structured Data



Structured data is highly organized and easily understood by machine language. Those working within relational databases can input, search, and manipulate structured data relatively quickly using a relational database management system (RDBMS).

id	name	age
1	Jim	28
2	Pam	26
3	Michael	42

id	subject	Teacher
1	Languages	John Jones
2	Track	Wally West
3	Swimming	Arthur Curry
4	Computers	Victor Stone

student_id	subject_id	grade
2	1	98
1	2	100
1	4	75
3	3	60
2	4	76
3	2	88



## ❑ JavaScript Object Notation (JSON) format.

```
## Document 1 ##
{
  "customerID": "103248",
  "name":
  {
    "first": "AAA",
    "last": "BBB"
  },
  "address":
  {
    "street": "Main Street",
    "number": "101",
    "city": "Acity",
    "state": "NY"
  },
  "ccOnFile": "yes",
  "firstOrder": "02/28/2003"
}
```

# Semi-structured Data



- ❑ Extensible Markup Language (XML) format.

```
<?xml version="1.0" standalone="yes"?>  
<projects>
```

```
<project>  
  <Name>ProductX</Name>  
  <Number>1</Number>  
  <Location>Bellaire</Location>  
  <DeptNo>5</DeptNo>  
  <Worker>  
    <SSN>123456789</SSN>  
    <LastName>Smith</LastName>  
    <hours>32.5</hours>  
  </Worker>  
  <Worker>  
    <SSN>453453453</SSN>  
    <FirstName>Joyce</FirstName>  
    <hours>20.0</hours>  
  </Worker>  
</project>  
<project>  
  <Name>ProductY</Name>  
  <Number>2</Number>  
  <Location>Sugarland</Location>  
  <DeptNo>5</DeptNo>  
  <Worker>  
    <SSN>123456789</SSN>  
    <hours>7.5</hours>  
  </Worker>  
  <Worker>  
    <SSN>453453453</SSN>  
    <hours>20.0</hours>  
  </Worker>  
  <Worker>  
    <SSN>333445555</SSN>  
    <hours>10.0</hours>  
  </Worker>  
</project>
```

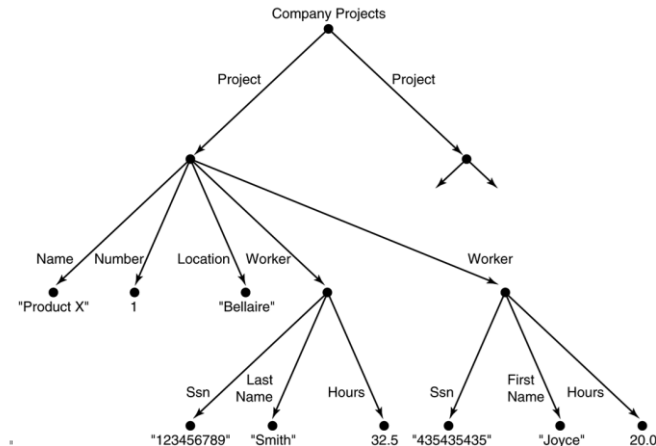
...

```
</projects>
```



## ❑ May be displayed as a directed graph...

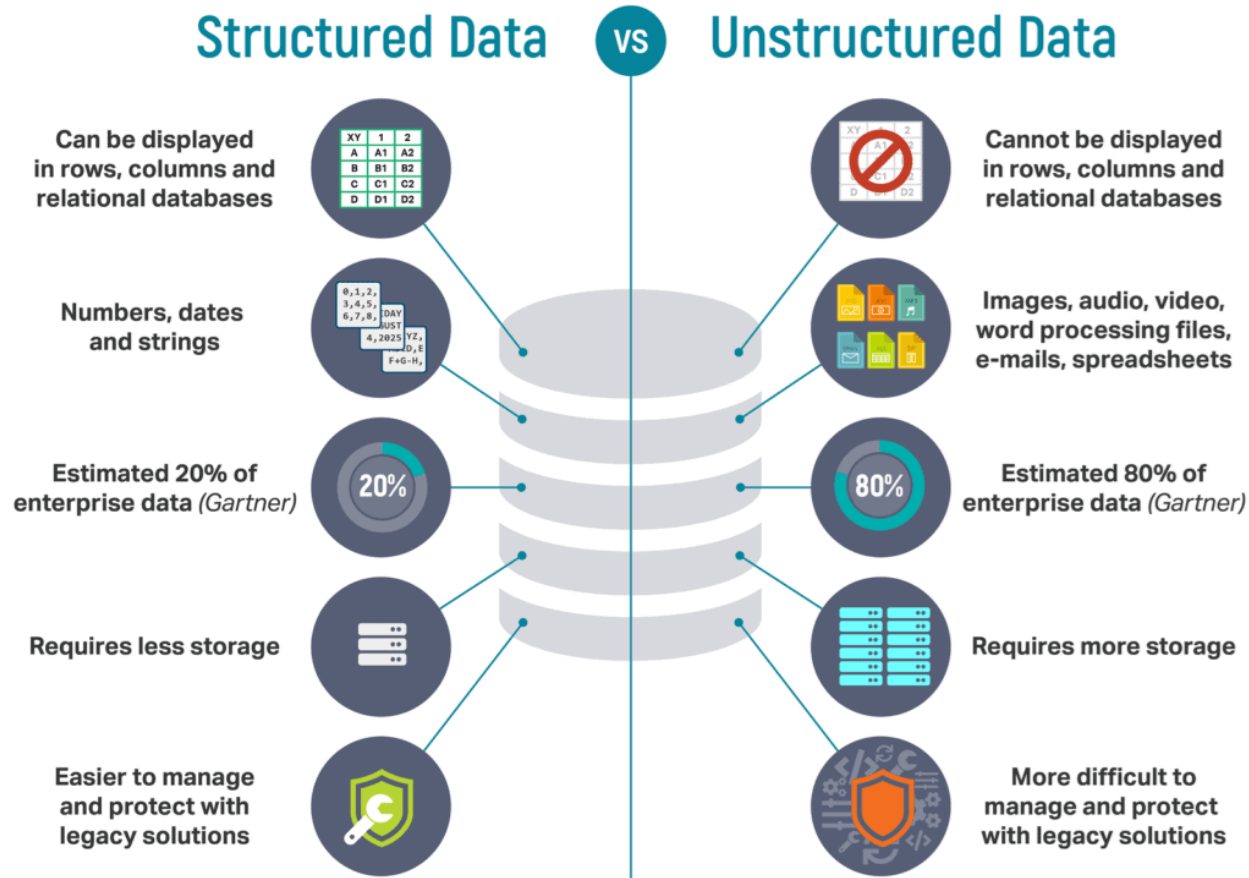
- The labels or tags on the directed edges represent the schema names—the names of attributes, object types (or entity types or classes), and relationships.
- The internal nodes represent individual objects or composite attributes.
- The leaf nodes represent actual data values of simple (atomic) attributes.





- ❑ Here are a few examples where unstructured data is being used in analytics today.
  - **Classifying image and sound.** Using deep learning, a system can be trained to recognize images and sounds. The systems learn from labeled examples in order to accurately classify new images or sounds.
  - **As input to predictive models.** Text analytics — using natural language processing (NLP) or machine learning — is being used to structure unstructured text.
  - **Chatbots in customer experience.** Chatbots have been in the market for a number of years, but the newer ones have a better understanding of language and are more interactive.

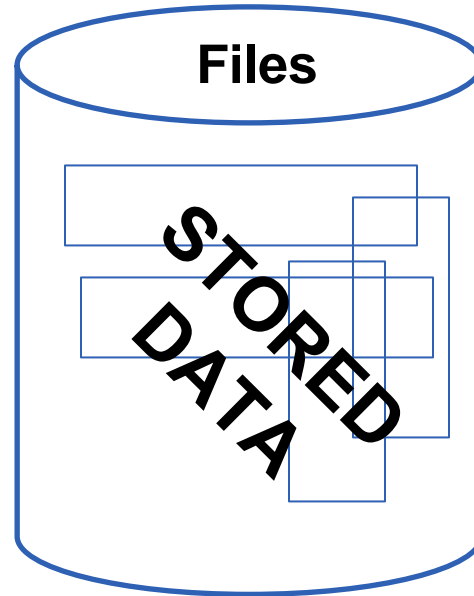
# Structured, Semi Structured and Unstructured Data



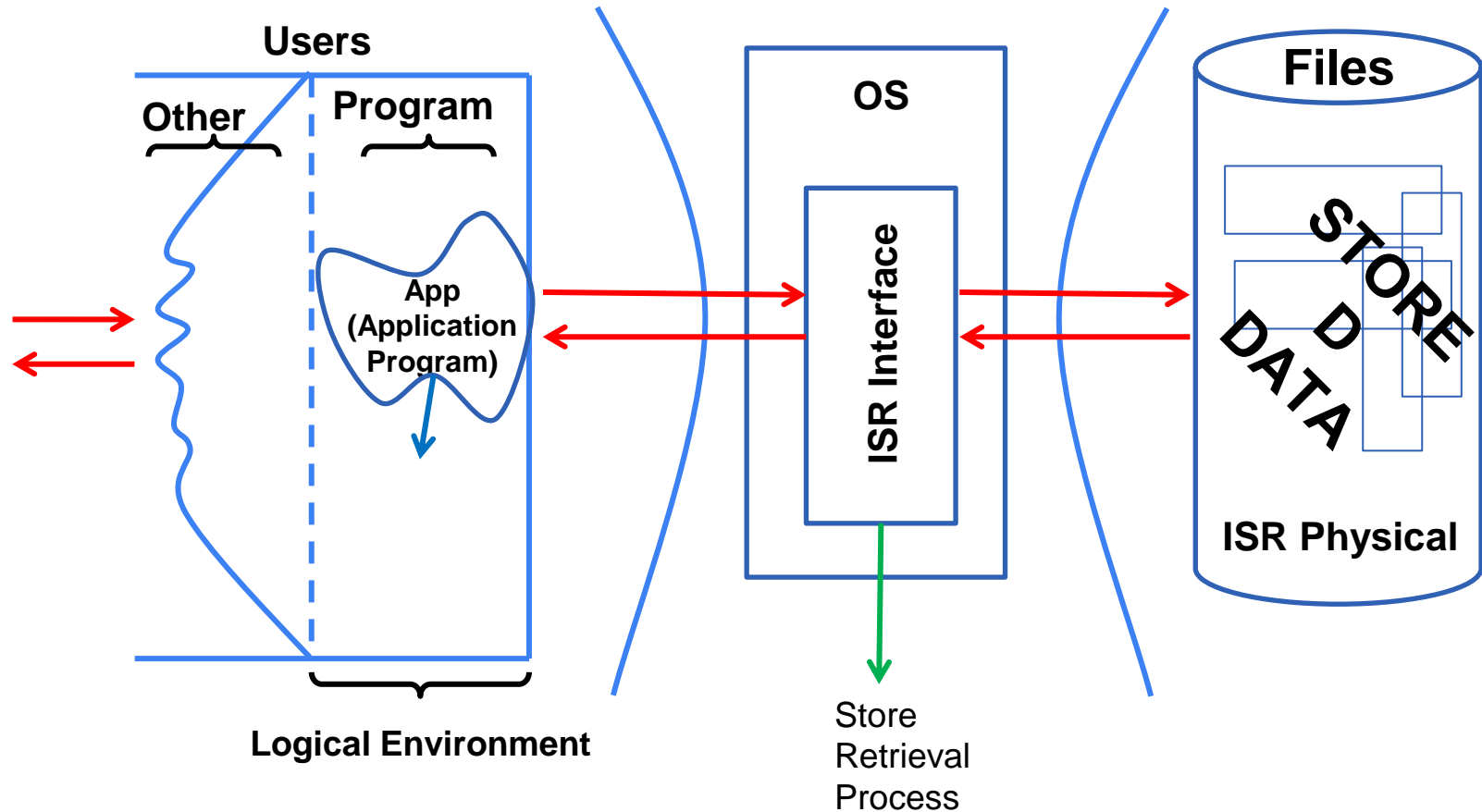




- ❑ We need an interface system for creating, managing, and using the ISR.



# ISR Interface





- ❑ In its “raw” form, data has little meaning. In this case it simply looks like a couple of lists of integer numbers. There is no context on which to base the data.

**Data:**

0	11,500
5	12,300
10	12,800
15	10,455
20	12,200
25	13,900
30	14,220



- ❑ By “processing” the data we have transformed it into something with more meaning. In this example, the processing consisted primarily of placing the data in context (which is usually done by adding more data! Although this additional data is really *metadata* (see page 14)). Now the data begins to take on more meaning.

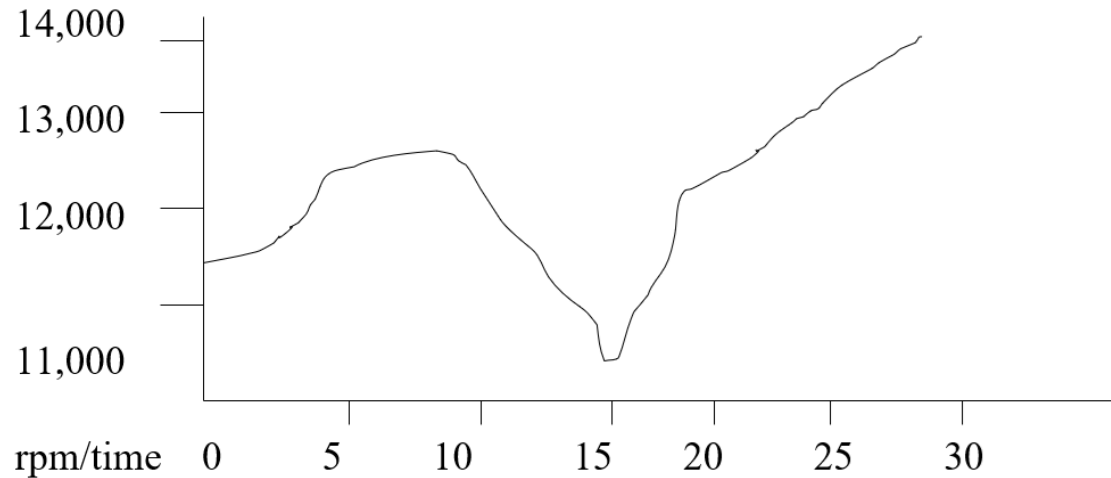
**Information:** Engine RPM Data: Roebling Road 10/22/2009 – Yamaha Heavy

Lap 12: time rpm

0	11,500
5	12,300
10	12,800
15	10,455
20	12,200
25	13,900
30	14,220



- ❑ Considering the same data as was presented in the previous slide, consider the following processing of that data.



Graph: Partial Lap 12 – Roebling Road 10/22/2009 – Yamaha Heavy



- ❑ **Data**: Unorganized and unprocessed facts; static; a set of discrete facts about events
- ❑ **Information**: Aggregation of data that makes decision making easier
- ❑ **Knowledge** is derived from information in the same way information is derived from data; it is a person's range of information



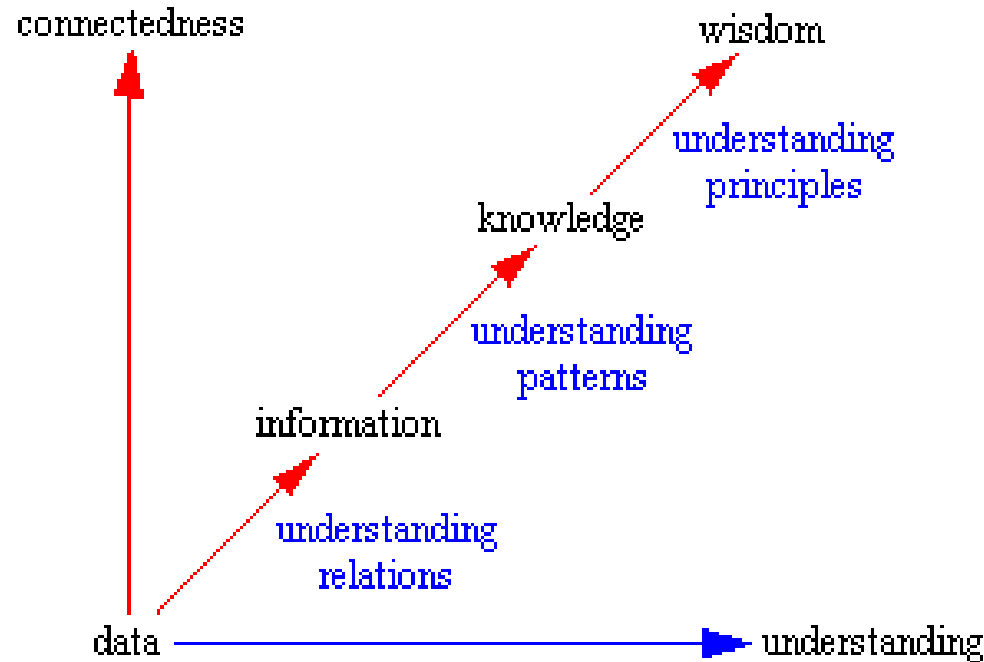
- ❑ Data represents a fact or statement of event without relation to other things.
  - Ex: It is raining.
  
- ❑ Information embodies the understanding of a relationship of some sort, possibly cause and effect.
  - Ex: The temperature dropped 15 degrees and then it started raining.
  
- ❑ Knowledge represents a pattern that connects and generally provides a high level of predictability as to what is described or what will happen next.
  - Ex: If the humidity is very high and the temperature drops substantially the atmosphere is often unlikely to be able to hold the moisture so it rains.
  
- ❑ Wisdom embodies more of an understanding of fundamental principles embodied within the knowledge that are essentially the basis for the knowledge being what it is. Wisdom is essentially systemic.
  - Ex: It rains. And this encompasses an understanding of all the interactions that happen between raining, evaporation, air currents, temperature gradients, changes, and raining.

# The DIKW Pyramid





# A Sequential Process of Knowing



Understanding supports the transition from one stage to the next, it is not a separate level in its own right



- ❑ What is this (note the point when you realize what it is but do not say)
  - I have a box.
  - The box is 3' wide, 3' deep, and 6' high.
  - The box is very heavy.
  - When you move this box you usually find lots of dirt underneath it.
  - Junk has a real habit of collecting on top of this box.
  - The box has a door on the front of it.
  - When you open the door the light comes on.
  - You usually find the box in the kitchen.
  - It is colder inside the box than it is outside.
  - There is a smaller compartment inside the box with ice in it.
  - When I open the box it has food in it.

# Rate of Motion towards Knowledge












- ☐ It was a refrigerator
- ☐ At some point in the sequence you connected with the pattern and understood
- ☐ When the pattern connected the information became knowledge to you
- ☐ If presented in a different order you would still have achieved knowledge but perhaps at a different rate

## ❑ FS: File System

- Examples of file systems include NTFS, FAT32, ext4, and HFS+, commonly used in Windows, embedded systems, Linux, and macOS respectively. Basic file system operations include file creation, file reading, file writing, file deletion, and directory traversal, crucial for data management.

## ❑ DBMS: Database Management System

	MySQL	▼		Oracle	▼		Microsoft SQL Server	▼
	PostgreSQL	▼		MongoDB	▼		Relational model	▼
	SQLite	▼		IBM Db2	▼		Microsoft Access	▼

## KBMS: Knowledge Base Management System



Canva



Evernote



MediaWiki



Confluence



Google



Microsoft



Bloomfire



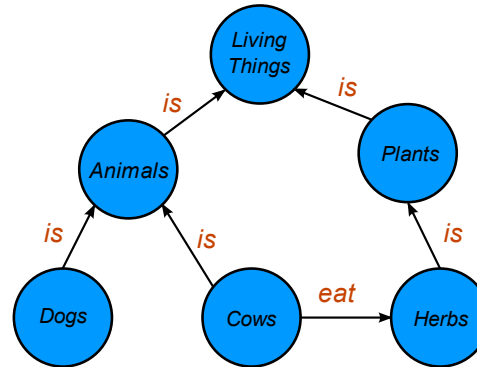
Knowledge bases



Notion



### Knowledge Graph



## ❑ DM: Data Mining System



KNIME



Oracle Data Mining



RapidMiner



IBM SPSS Modeler



Intrusion Detection



Shopping market analysis



WEKA



Bioinformatics



Criminal investigation



## ❑ KDS: Knowledge Discovery System

- Expert systems
- Decision Support Systems (DSS)
- Advisor systems
- Fault diagnosis (or troubleshooting) systems
- Help desk systems.



- ❑ It is a set of **stored** data, which is **persistent**, **integrated** and **interconnected**, even if possible **without redundancy**, (having its own architecture, based on a specific **data model**), using one or more users in an organization (in) an environment): **multiuser**, **shared** and **concurrent**.

# What Is a DBMS?



- ❑ A Database is a very large, integrated collection of data.
- ❑ Models real-world enterprise.
  - Entities (e.g., students, courses)
  - Relationships (e.g., John is taking ECE459)
- ❑ A Database Management System (DBMS) is a software package designed to store and manage databases.



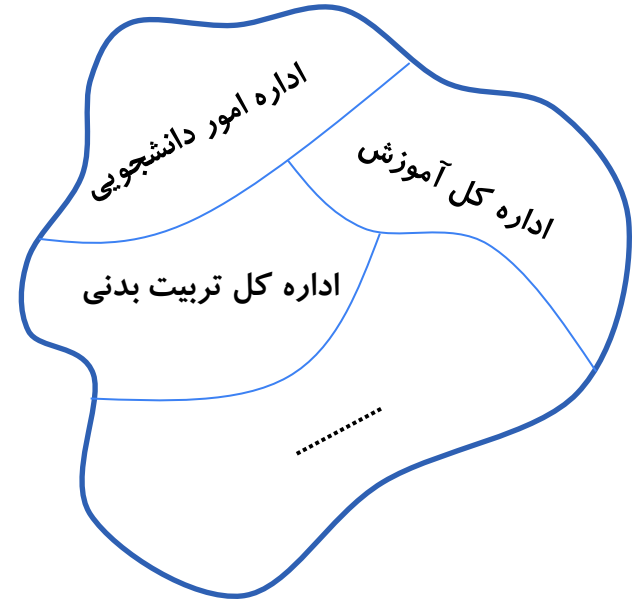


- ❑ Application must stage large datasets between main memory and secondary storage (e.g., buffering, page-oriented access, 32-bit addressing, etc.)
- ❑ Special code for different queries
- ❑ Must protect data from inconsistency due to multiple concurrent users
- ❑ Crash recovery
- ❑ Security and access control



- ❑ Data independence (abstract view of data) and efficient access.
- ❑ Reduced application development time.
- ❑ Data integrity (enforce constraints) and security.
- ❑ Uniform (central) data administration.
- ❑ Concurrent access, recovery from crashes.

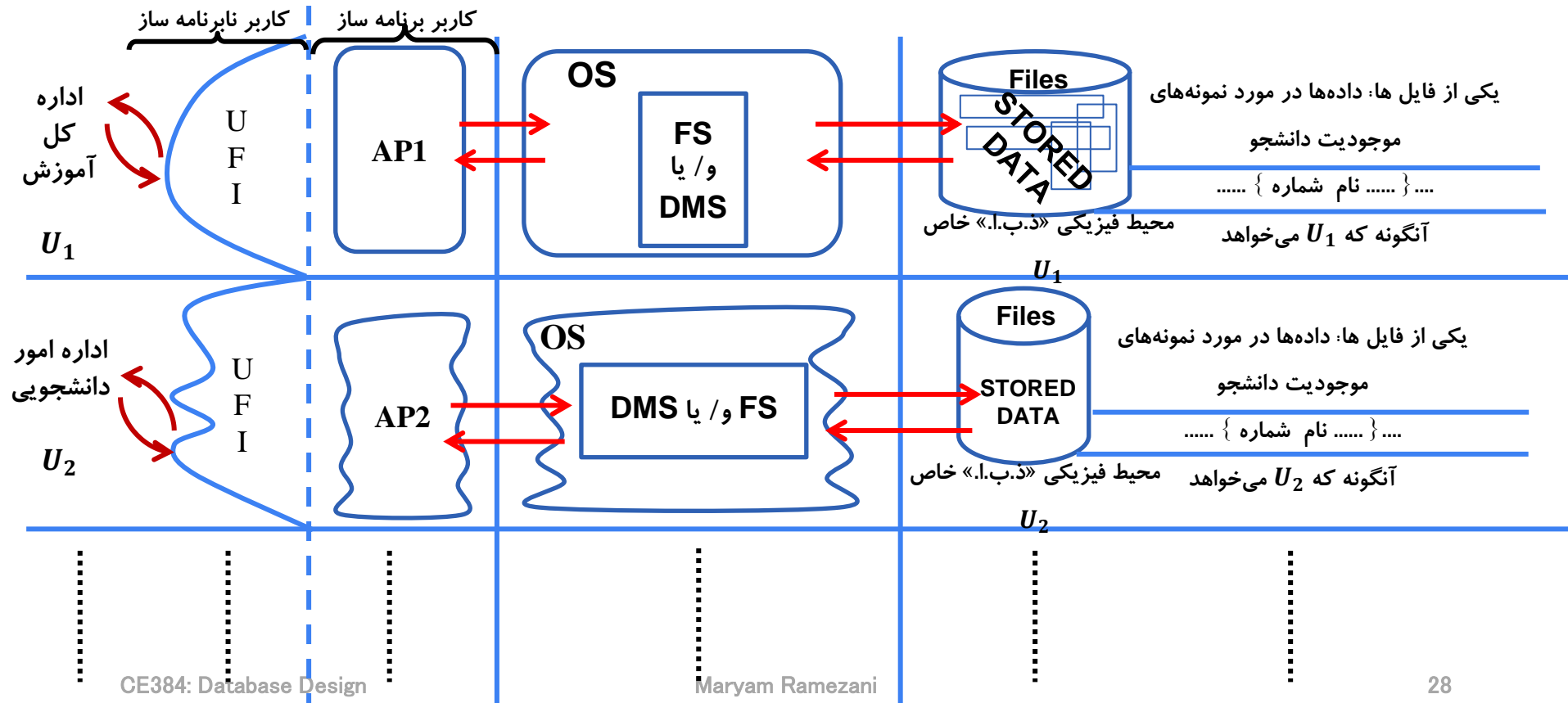
- ❑ Entity?
- ❑ Two approach for this environment:
  - File System
  - Database



# Case Study



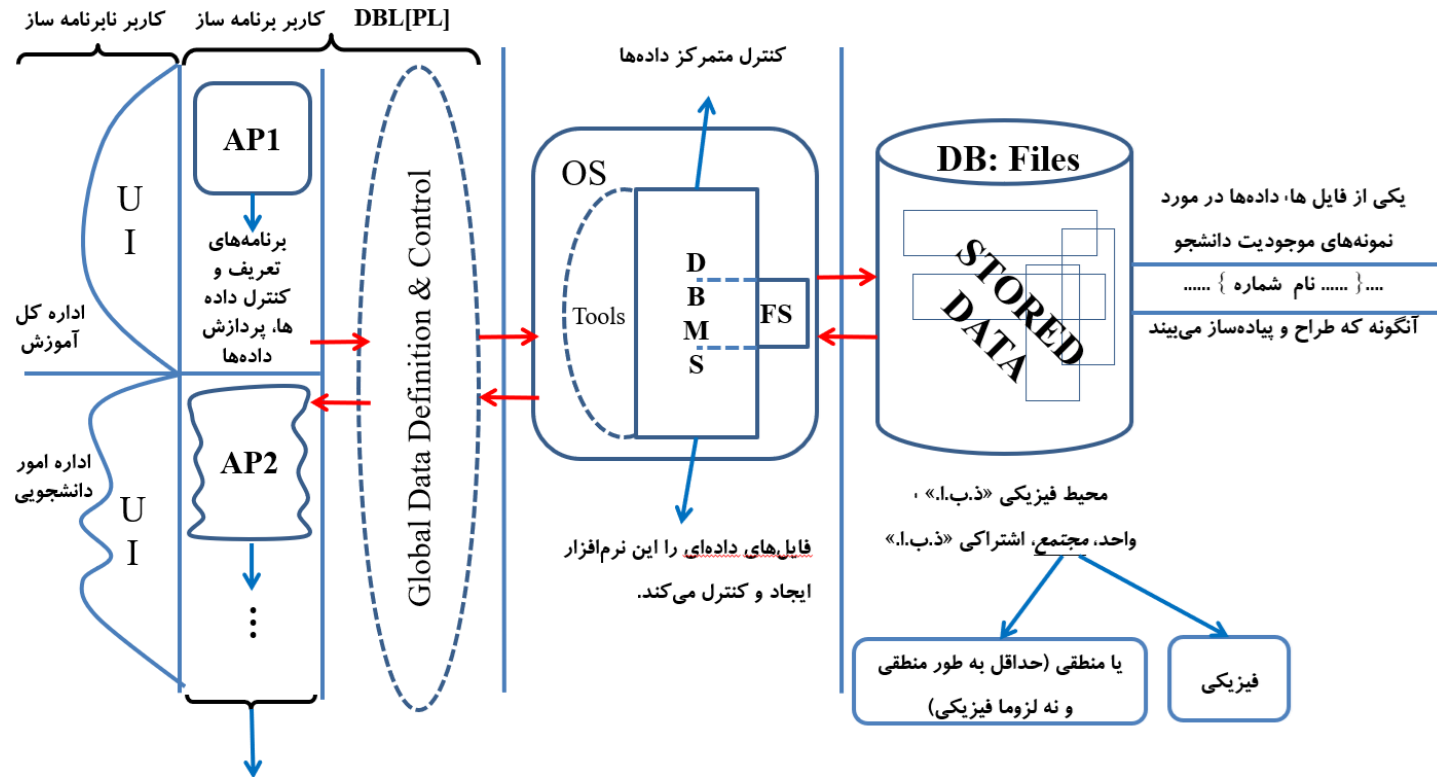
## File System Approach



# Case Study



## □ Database Approach



# Types of Redundancy





## افزونگی در معنای محدود ☐

☐ عبارت است از تکرار ذخیره سازی مقادیر (value) یک صفت یا بیش از یک صفت مربوط به یک نوع

موجودیت در یک فایل داده‌ای یا فایل کمکی آن.

☐ این نوع افزونگی گونه‌هایی دارد:

۱- **طبیعی**: ناشی از ماهیت داده‌های محیط (مثل صفت رشته دانشجو که برای دانشجویان مختلف می‌تواند یکسان و در نتیجه تکراری باشد)

▪ **کنجکاوی**: برای کاهش مصرف حافظه در حالت افزونگی طبیعی چه باید کرد؟

۲- **تکنیکی**: ناشی از استفاده از یک تکنیک معمولا برای افزایش سرعت (مثل نمایه سازی یا شاخص‌بندی -

(Indexing)

## افزونگی در معنای گسترده

□ عبارت است از تکرار ذخیره‌سازی داده‌ها در مورد نمونه‌های یک یا بیش از یک نوع موجودیت از یک محیط (تحت کنترل سیستم‌های مختلف).

- این نوع افزونگی بسته به مشی طراحی می‌تواند نه از نوع طبیعی و نه از نوع تکنیکی بلکه ناشی از رهیافت انتخاب شده برای طراحی و تولید سیستم‌های کاربردی باشد.
- به طور مثال تکرار اطلاعات دانشجویان در دو زیرسیستم اداره کل آموزش و زیرسیستم اداره امور دانشجویی.

□ نکته:

- افزونگی از نوع طبیعی و تکنیکی در پایگاه داده هم می‌تواند وجود داشته باشد.
- این نوع افزونگی با ملاحظات تکنیکی هم می‌تواند اتفاق بیفتد (مانند تکرار کل/بخشی از داده‌ها برای افزایش کارایی در پاسخگویی به پرس‌وجوها)

دلایل بروز افزونگی در سیستم‌های ISR به ویژه سیستم‌های پایگاهی کدامند؟





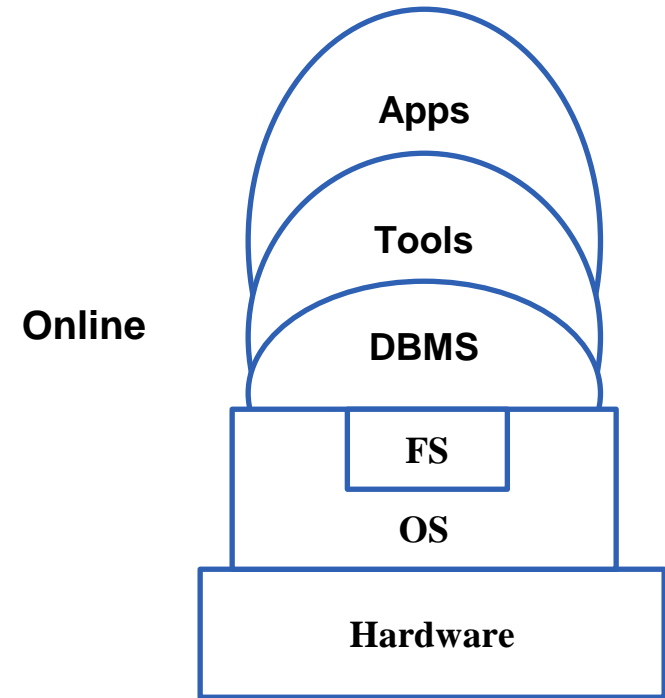
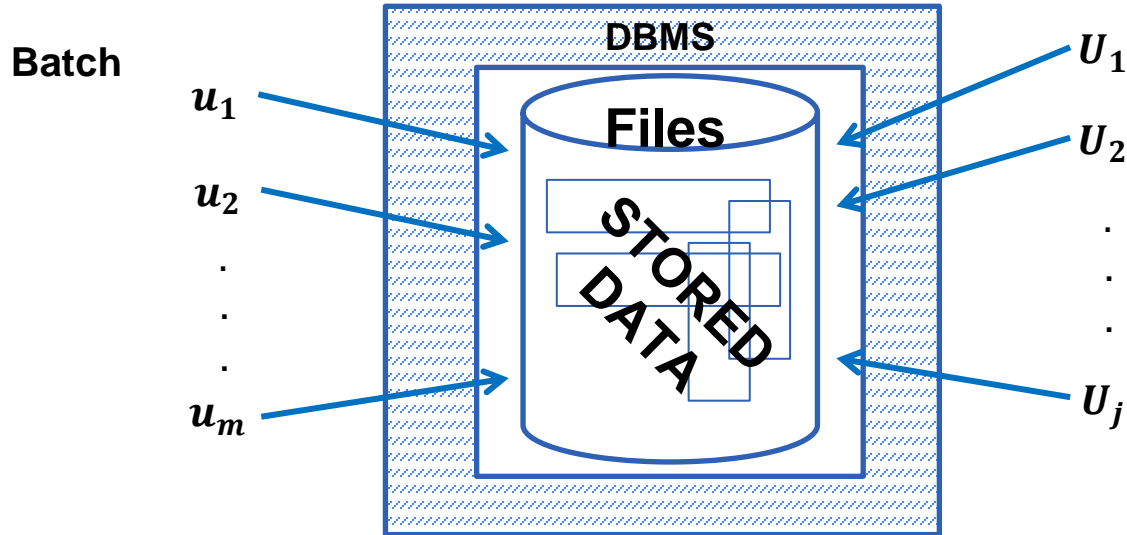


- ❑ Transaction Manager: There are certain guarantees that a DBMS must make when performing operations on a database. These guarantees are often referred to as the ACID properties.
  - **Atomicity**: all of a transaction is executed or none of it is executed.
  - **Consistency**: data cannot be in a inconsistent state.
  - **Isolation**: concurrent transactions must be isolated from each other both in effect and in visibility.
  - **Durability**: changes to the database caused by a transaction must not be lost even if the system fails immediately after the transaction completes.

# Components of Database Environment



- ☐ Hardware
- ☐ Software
- ☐ User
- ☐ Data
  - User
  - System



# Batch – Streaming – Interactive



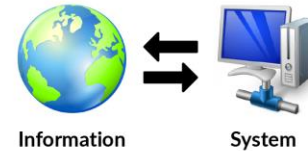
## Batch Processing

20 Min



## Stream Processing

Less Than 1 Sec





OLAP (Online Analytical Processing)	OLTP (Online Transaction Processing)
سیستم‌های پردازشی که امکان ترکیب و تجمیع داده‌ها را به منظور تحلیل آنها از ابعاد مختلف فراهم می‌کنند.	سیستم‌های پردازش تراکنش برخط، که نوعی سیستم‌های پردازش داده هستند که امکان اجرای تعداد زیادی تراکنش را توسط تعداد زیادی کاربر به صورت همروند فراهم می‌کنند.
تاکید روی زمان پاسخ پرس‌وجوهای پیچیده روی حجم زیادی از داده‌های تاریخچه‌ای و تجمیعی	تاکید روی سرعت و کارایی اجرای پرس‌وجوها و حفظ جامعیت داده‌ها در دسترسی‌های همروند
حاوی پرس‌وجوهای اغلب کم ولی پیچیده و تجمیعی و شامل حجم زیادی داده از نوع select	حاوی تعداد زیادی پرس‌وجوهای استاندارد و معمولاً ساده از نوع insert و delete و update و select
کاربران آن اغلب شامل مدیران و تحلیلگران کسب‌وکار	کاربران آن اغلب اپراتورهای برنامه‌های کاربردی برای انجام عملیات روزانه جهت ارائه خدمات
استفاده از سیستم‌های انبار داده (Data Warehouse)	استفاده از سیستم‌های مدیریت پایگاه داده (DBMS)