Unit-9 Fundamentals of Database

Prashant Gautam

M.Sc. CSIT

We Live in a World of Data

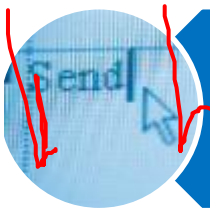
- Nearly 500 Exabytes per day are generated by the Large Hadron Collider experiments (not all recorded!)
- 2.9 million emails are sent every second
- 20 hours of video are uploaded to YouTube every minute
- 24 PBs of data are processed by Google every day
- 50 million tweets are generated per day
- 700 billion total minutes are spent on Facebook each month
- 72.9 items are ordered on Amazon every second

Data and *Big* Data

- The value of data as an organizational asset is widely recognized
- Data is literally exploding and is occurring along three main dimensions
  - "Volume" or the amount of data
  - "Velocity" or the speed of data
  - "Variety" or the range of data types and sources
- What is Big Data?
  - It is the proliferation of data that floods organizations on a daily basis
  - It is *high* volume, *high* velocity, and/or *high* variety information assets
  - It requires new forms of processing to enable *fast* mining, enhanced decision-making, insight discovery and process optimization

What Do We Do With Data and Big Data?

Store



Share

Query

Mine

Encrypt

…. and more!

We want to do these *seamlessly* and *fast*...

Using Diverse Interfaces & Devices

Computers

Mobile Devices

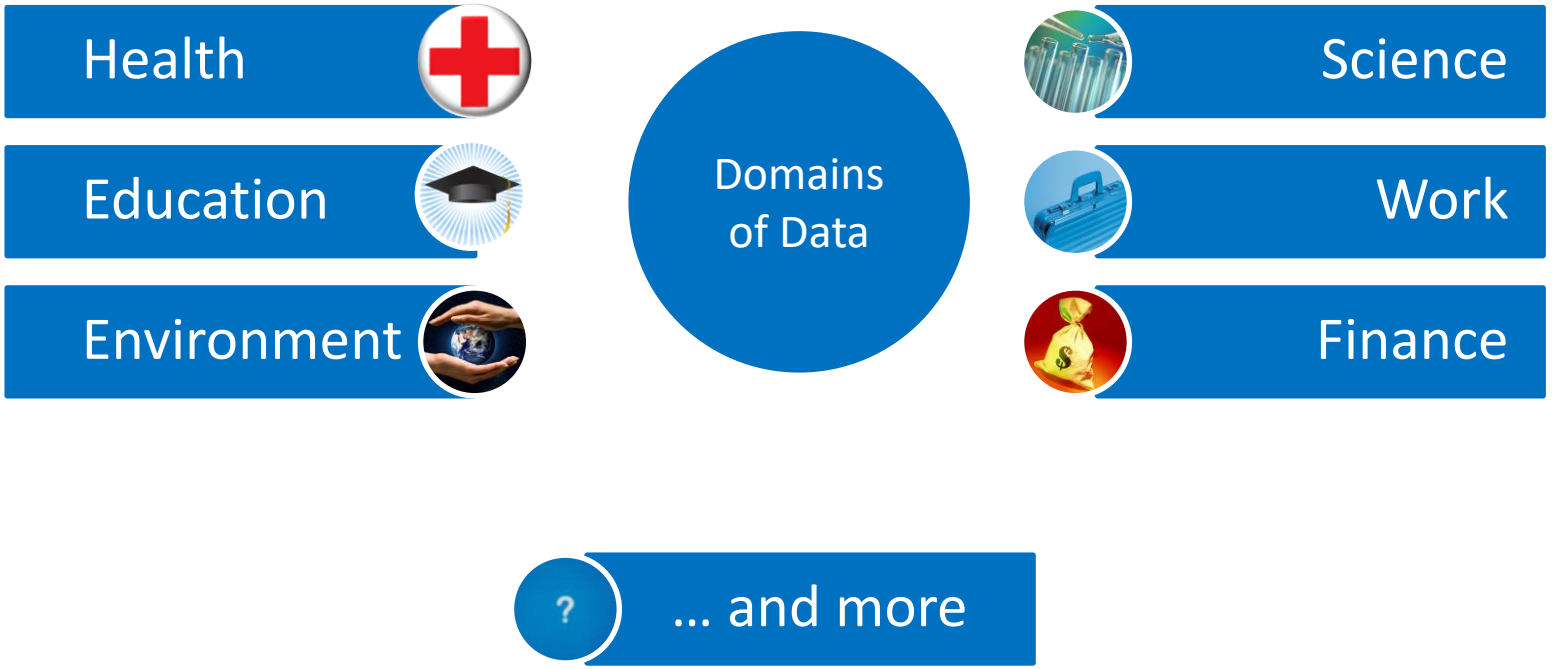
Consumer Electronics

Personal Monitors and Sensors

…and even appliances

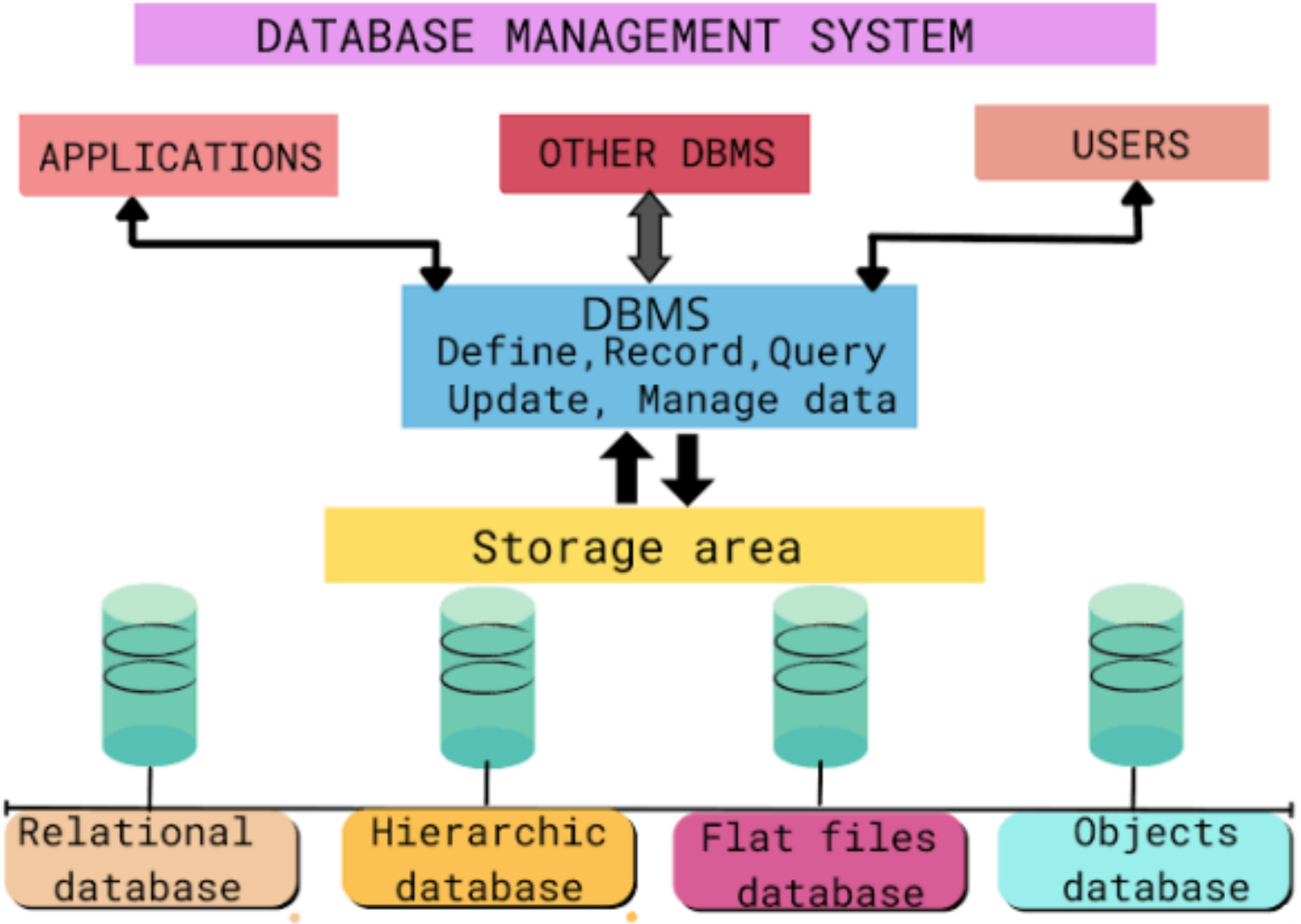
We also want to access, share and process our data from all of our devices, **anytime, anywhere**!

Data is Becoming Critical to Our Lives



What is data, database, DBMS

- Data: Known facts that can be recorded and have an implicit meaning; raw
- Database: a highly organized, interrelated, and structured set of data about a particular enterprise
  - Controlled by a database management system (DBMS)
- DBMS
  - Set of programs to access the data
  - An environment that is both *convenient* and *efficient* to use
- Database systems are used to manage collections of data that are:
  - Highly valuable
  - Relatively large
  - Accessed by multiple users and applications, often at the same time.
- A modern database system is a complex software system whose task is to manage a large, complex collection of data.


DATABASE MANAGEMENT SYSTEM
APPLICATIONS
OTHER DBMS
USERS
DBMS
Define, Record, Query
Update, Manage data
Storage area
Relational database
Hierarchic database
Flat files database
Objects database

What is Database

The database is a collection of inter-related data which is used to retrieve, insert and delete the data efficiently. It is also used to organize the data in the form of a table, schema, views, and reports, etc.

For example: The college Database organizes the data about the admin, staff, students and faculty etc.

Using the database, you can easily retrieve, insert, and delete the information.

Database Management System

- Database management system is a software which is used to manage the database. For example: MySQL, Oracle, etc are a very popular commercial database which is used in different applications.
- DBMS provides an interface to perform various operations like database

DBMS allows users the following tasks:

- **Data Definition:** It is used for creation, modification, and removal of definition that defines the organization of data in the database.
- **Data Updation:** It is used for the insertion, modification, and deletion of the actual data in the database.
- **Data Retrieval:** It is used to retrieve the data from the database which can be used by applications for various purposes.

User Administration: It is used for registering and monitoring users, maintain data integrity, enforcing data security, dealing with concurrency control, monitoring performance and recovering information corrupted by unexpected failure.

Characteristics of DBMS

- It uses a digital repository established on a server to store and manage the information.
- It can provide a clear and logical view of the process that manipulates data.
- DBMS contains automatic backup and recovery procedures.
- It contains ACID properties which maintain data in a healthy state in case of failure.
- It can reduce the complex relationship between data.
- It is used to support manipulation and processing of data.
- It is used to provide security of data.
- It can view the database from different viewpoints according to the requirements of the user.

Advantages of DBMS

- **Controls database redundancy:** It can control data redundancy because it stores all the data in one single database file and that recorded data is placed in the database.
- **Data sharing:** In DBMS, the authorized users of an organization can share the data among multiple users.
- **Easily Maintenance:** It can be easily maintainable due to the centralized nature of the database system.
- **Reduce time:** It reduces development time and maintenance need.
- **Backup:** It provides backup and recovery subsystems which create automatic backup of data from hardware and software failures and restores the data if required.
- **multiple user interface:** It provides different types of user interfaces like graphical user interfaces, application program interfaces

Disadvantages of DBMS

- **Cost of Hardware and Software:** It requires a high speed of data processor and large memory size to run DBMS software.
- **Size:** It occupies a large space of disks and large memory to run them efficiently.
- **Complexity:** Database system creates additional complexity and requirements.
- **Higher impact of failure:** Failure is highly impacted the database because in most of the organization, all the data stored in a single database and if the database is damaged due to electric failure or database corruption then the data may be lost forever.

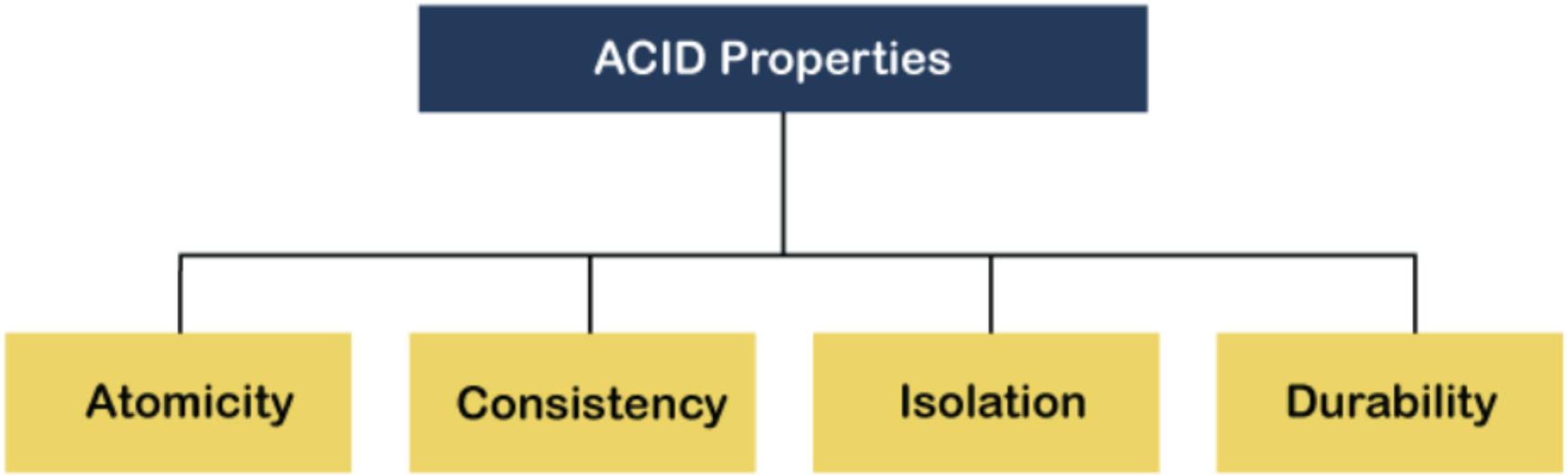
Advantages of DBMS

1. Control of data redundancy
2. Data Consistency
3. Sharing of data
4. Improved security
5. Enforcement of standards
6. Economy of scale
7. Balance of conflicting requirements
8. Improved data accessibility and responsiveness.
9. Increased productivity
10. Increased maintenance through data independence.
11. Increased concurrency.
12. Improved backup and recovery services.

Disadvantages of DBMS

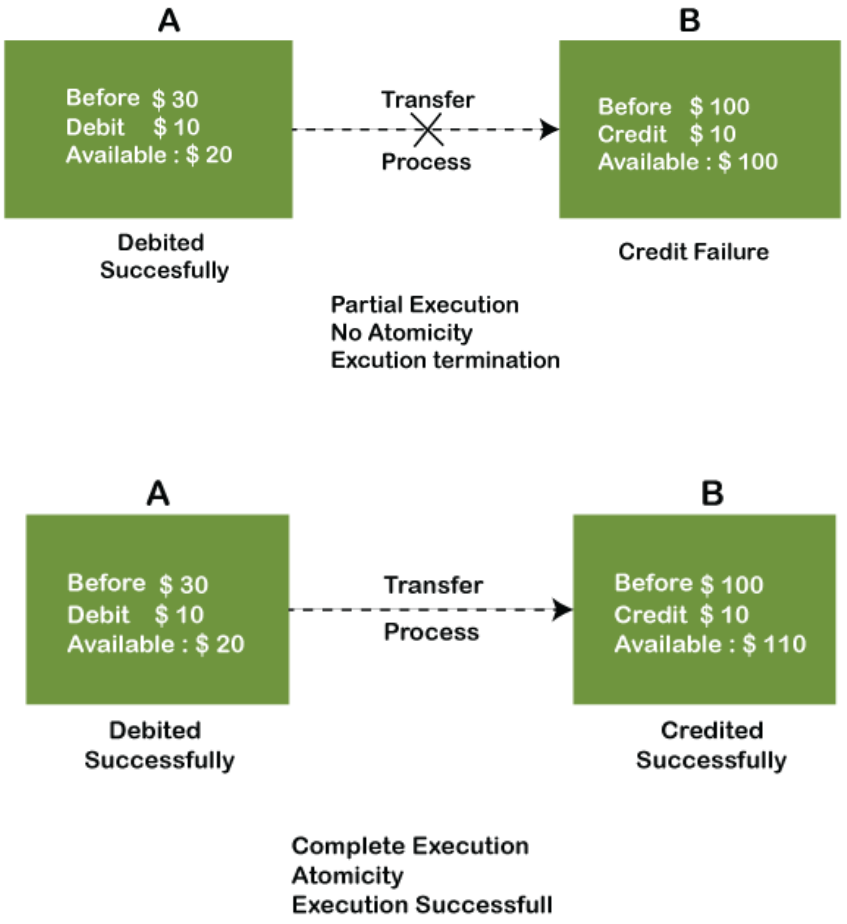
- Complexity
- Size
- Cost of DBMS
- Additional hardware cost
- Cost of conversion
- Performance
- Higher impact of failure.

ACID Properties in DBMS



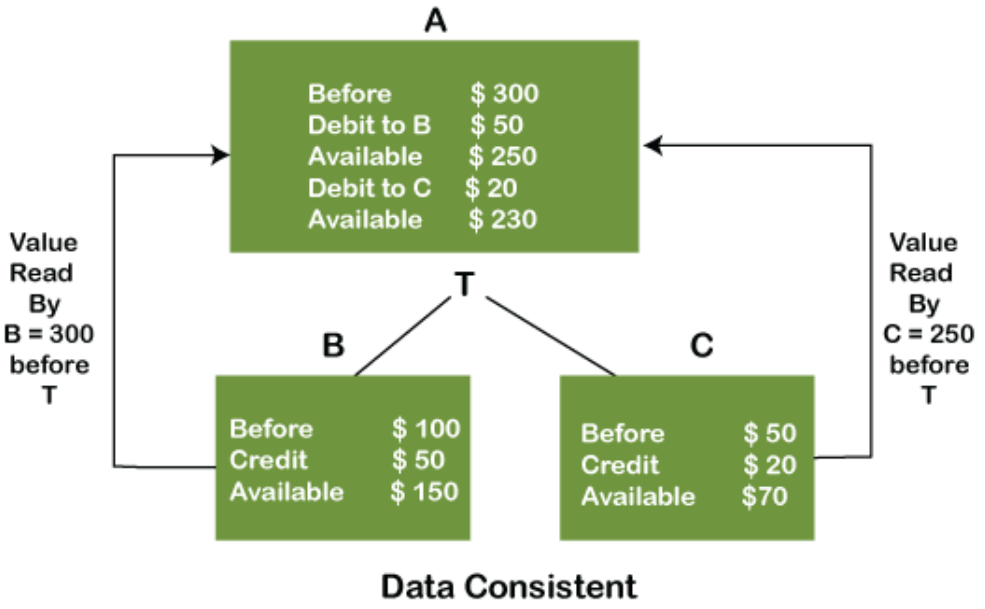
Atomicity

- It means if any operation is performed on the data, either it should be performed or executed completely or should not be executed at all.



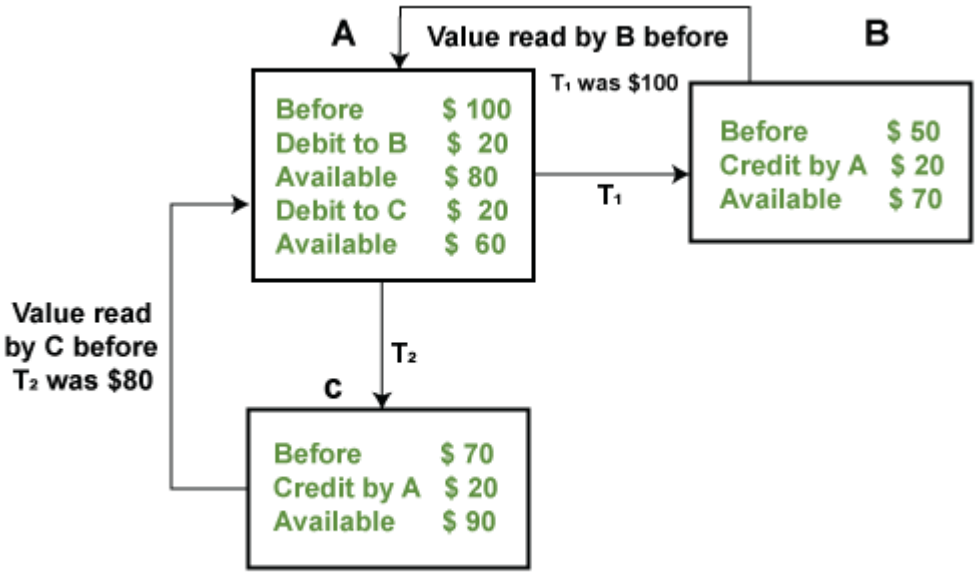
Consistency

- The integrity of the data should be maintained, which means if a change in the database is made, it should remain preserved always.



Isolation

- 'isolation' means separation.
- Isolation is the property of a database where no data should affect the other one and may occur concurrently.
- the operation on one database should begin when the operation on the first database gets complete.



Isolation - Independent execution of T_1 & T_2 by A

Durability

- Durability ensures the permanency of something.
- In DBMS, the term durability ensures that the data after the successful execution of the operation becomes permanent in the database.
- The durability of the data should be so perfect that even if the system fails or leads to a crash, the database still survives.
- However, if gets lost, it becomes the responsibility of the recovery manager for ensuring the durability of the database.
- For committing the values, the COMMIT command must be used every time we make changes.

Assignment

- Explain the ACID properties of DBMS.

Database Architecture in DBMS

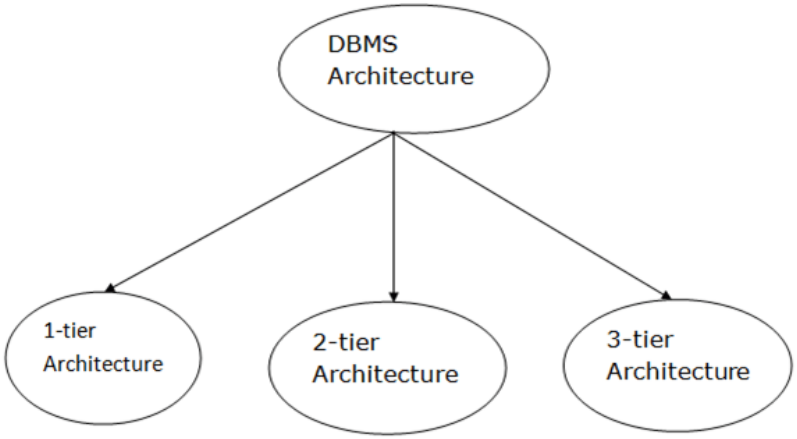
Introduction

- representation of DBMS design.
- helps to design, develop, implement, and maintain the database management system.
- allows dividing the database system into individual components that can be independently modified, changed, replaced, and altered.
- helps to understand the components of a database.

- A Database stores critical information and helps access data quickly and securely. Therefore, selecting the correct Architecture of DBMS helps in easy and efficient data management.

Types of DBMS Architecture

- One Tier Architecture (Single Tier Architecture)
- Two Tier Architecture
- Three Tier Architecture



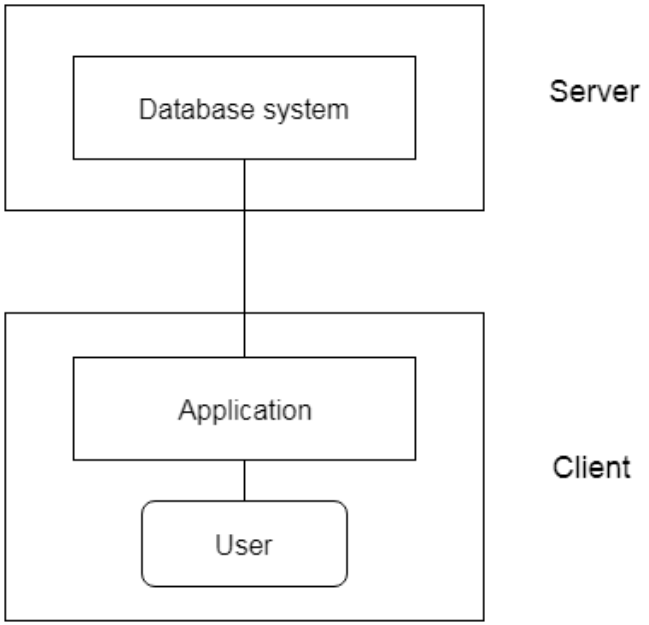
1 Tier Architecture

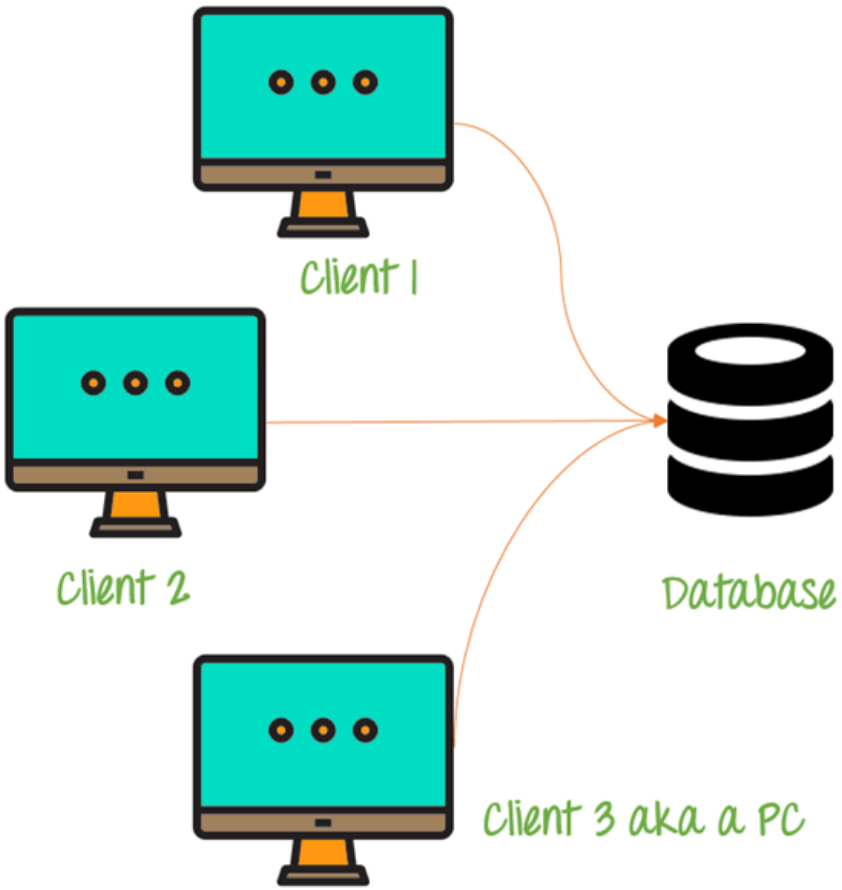
- the simplest architecture of Database in which the client, server, and Database all reside on the same machine.
- Eg. anytime you install a Database in your system and access it to practice SQL queries. But such architecture is rarely used in production.

Single Tier Architecture

2-Tier Architecture

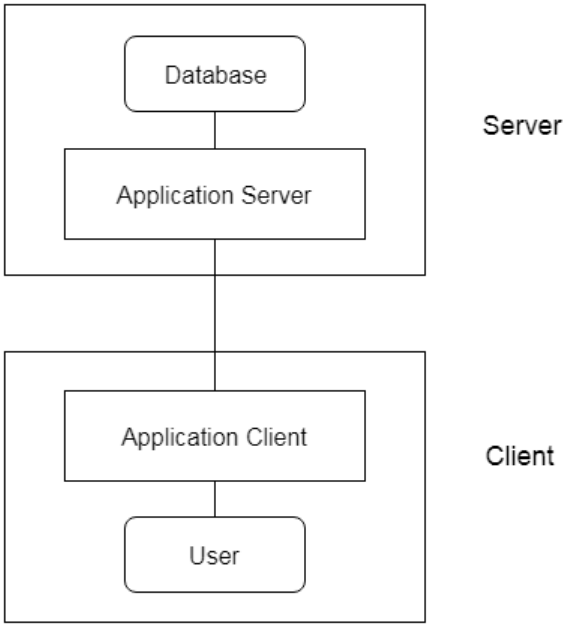
- A **2 Tier Architecture** in DBMS is a Database architecture where the presentation layer runs on a client (PC, Mobile, Tablet, etc.), and data is stored on a server called the second tier.
- Two tier architecture provides added security to the DBMS as it is not exposed to the end-user directly.
- It also provides direct and faster communication.
- Eg. A Contact Management System created using MS- Access.




Client 1
Client 2
Database
Client 3 aka a PC

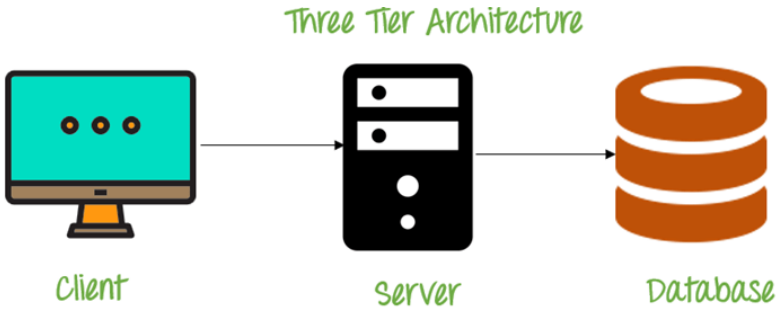
3 Tier Architecture

- most popular client server architecture in DBMS in which the development and maintenance of functional processes, logic, data access, data storage, and user interface is done independently as separate modules.
- Three Tier architecture contains a presentation layer, an application layer, and a database server.

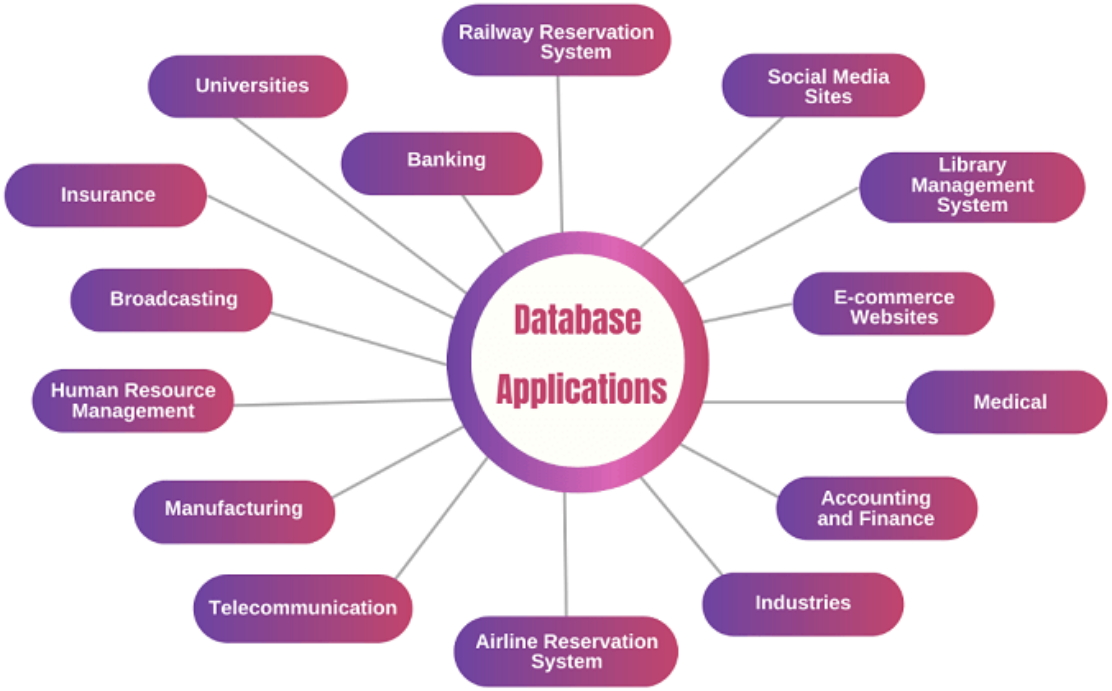


- 3-Tier database Architecture design is an extension of the 2-tier client-server architecture. A 3-tier architecture has the following layers:
- Presentation layer (your PC, Tablet, Mobile, etc.)
- Application layer (server)
- Database Server

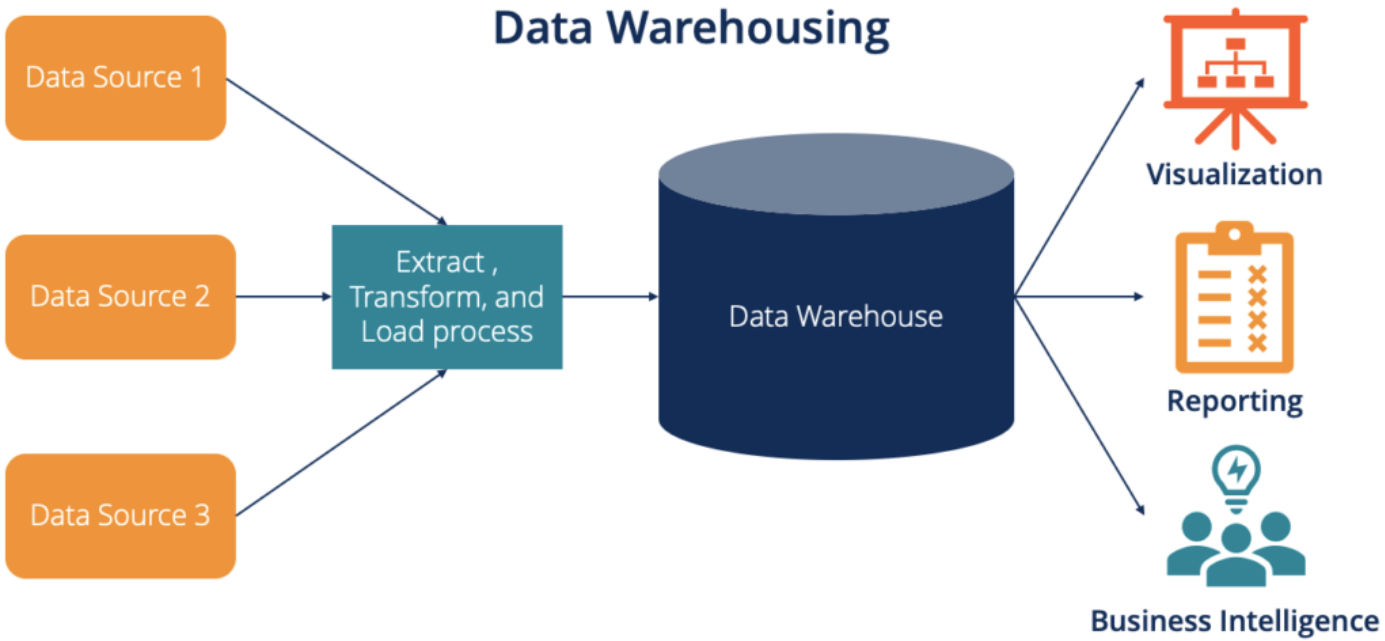
EG. Any large website on the internet



Database Applications



Data Warehousing



Data Warehousing

- Data
  - Raw piece of information that is capable of being moved and store.
- Database
  - An organized collection of such data in which data are managed in tabular form with relationship.
- Data Warehouse
  - System that organizes all the data available in an organization, makes it accessible & usable for the all kinds of data analysis and also allows to create a lots of reports by the use of mining tools.

Data Warehouse…

- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."

- Data warehousing:
  - The process of constructing and using data warehouses.
  - Is the process of extracting & transferring operational data into informational data & loading it into a central data store (warehouse)

Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

Data Warehouse—Subject-Oriented

- Organized around major subjects, such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain “time element”.

Data Warehouse—Non-Volatile

- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*.

Database vs. Datawarehouse

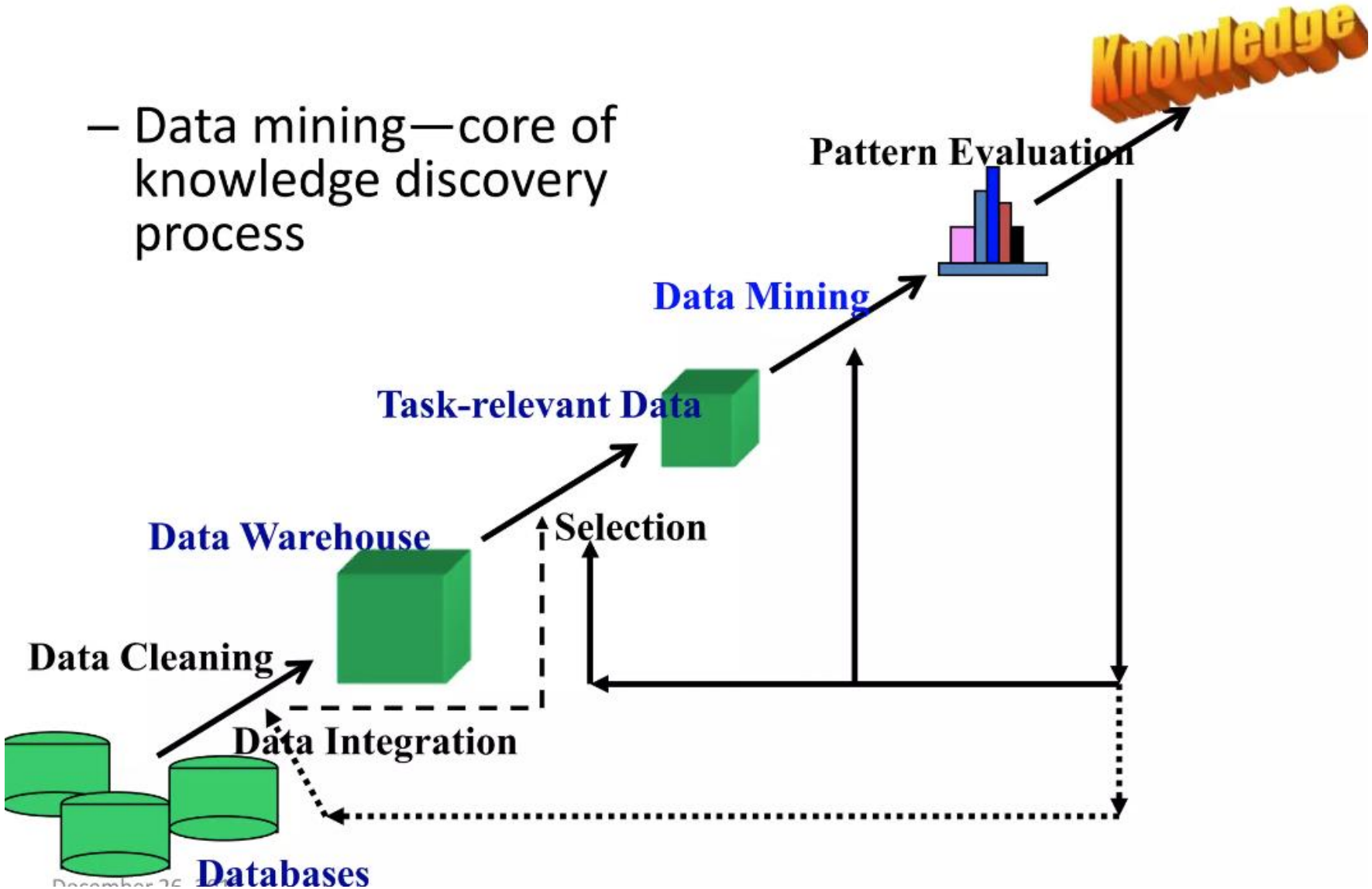
	Database	**Data warehouse**
Processing type	OLTP	OLAP
Optimized operations	CRUD transactions	Complex analytical queries
Data sources	Usually one	Usually multiple
Data model	Normalized	Denormalized
Data timelines	Daily to monthly	Historical
Data volume	Low to mid	Mid to large

What Is Data Mining?

- Data mining (knowledge discovery in databases):
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases.
- Alternative names and their "inside stories":
  - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- What is not data mining?
  - (Deductive) query processing.
  - Expert systems

Knowledge Discovery (KDD) Process

– Data mining—core of knowledge discovery process

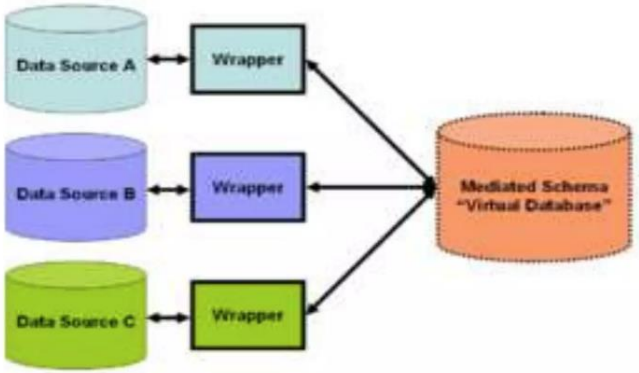


DATA CLEANING

- Remove Noise and Inconsistent Data

DATA INTEGRATION

- Where multiple data sources may be combined



Data Selection: where data relevant to analysis task are retrieved from the DB

DATA TRANSFORMATION

- Where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operation

Data Mining

- An essential Process where intelligent methods are applied to extract data patterns

PATTERN EVALUATION

- To identify the truly interesting patterns representing knowledge based on interestingness measures

KNOWLEDGE REPRESENTATION

- Where visualization and knowledge representation techniques are used to present mined knowledge to users

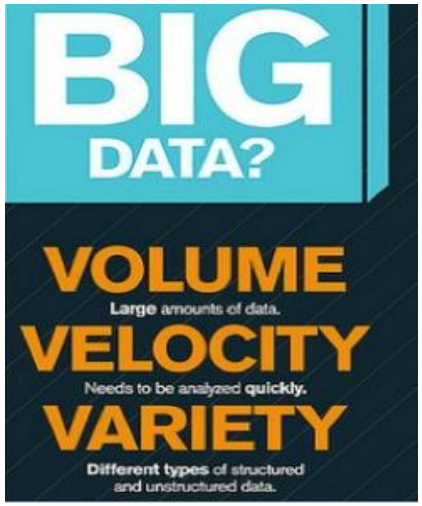
Introduction to Big Data

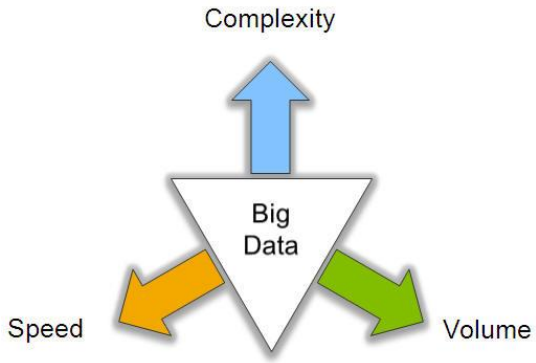
What's Big Data?

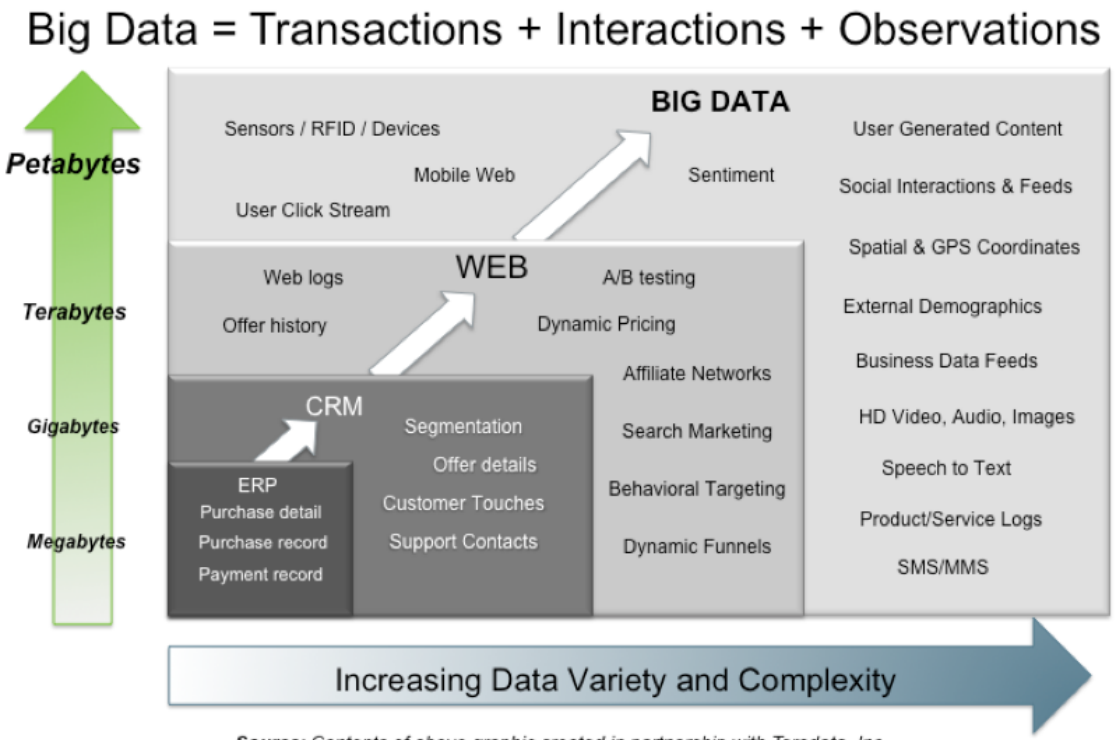
No single definition; here is from Wikipedia:

- **Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.
- The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."

Big Data: 3V's



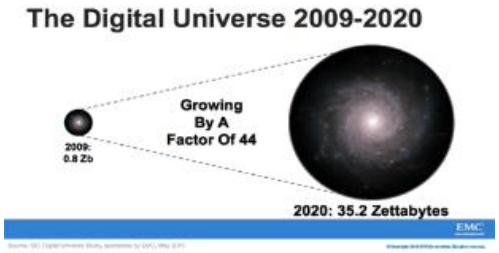


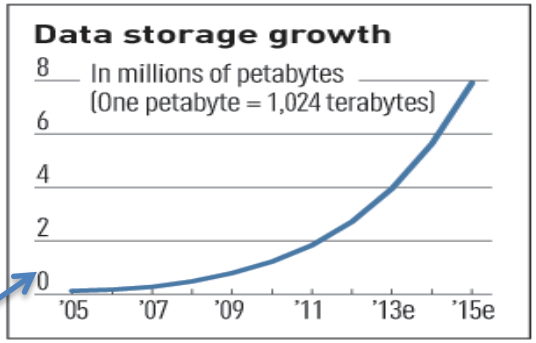


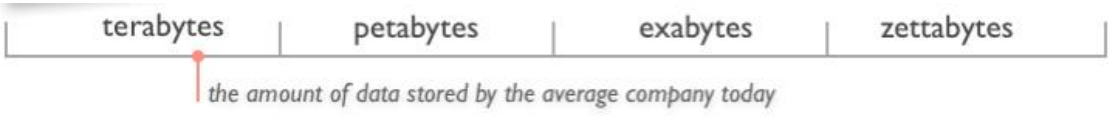
Source: Contents of above graphic created in partnership with Teradata, Inc.

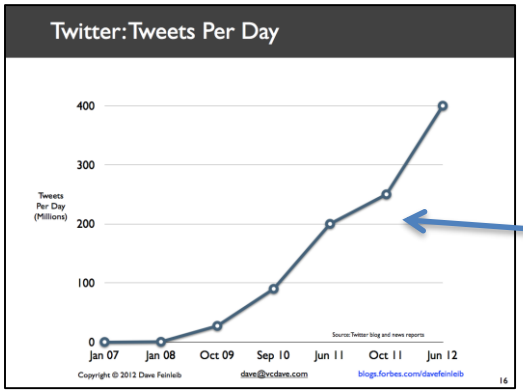
Volume (Scale)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially







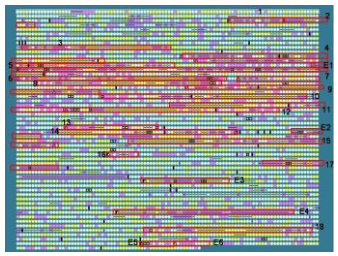


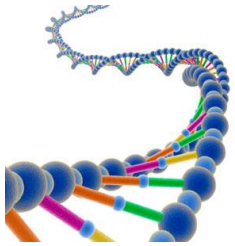
Exponential increase in collected/generated data

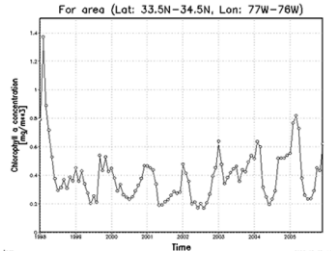
200 Petabytes of data per day is processed by Google Search alone.

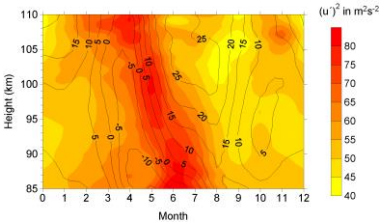
Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)



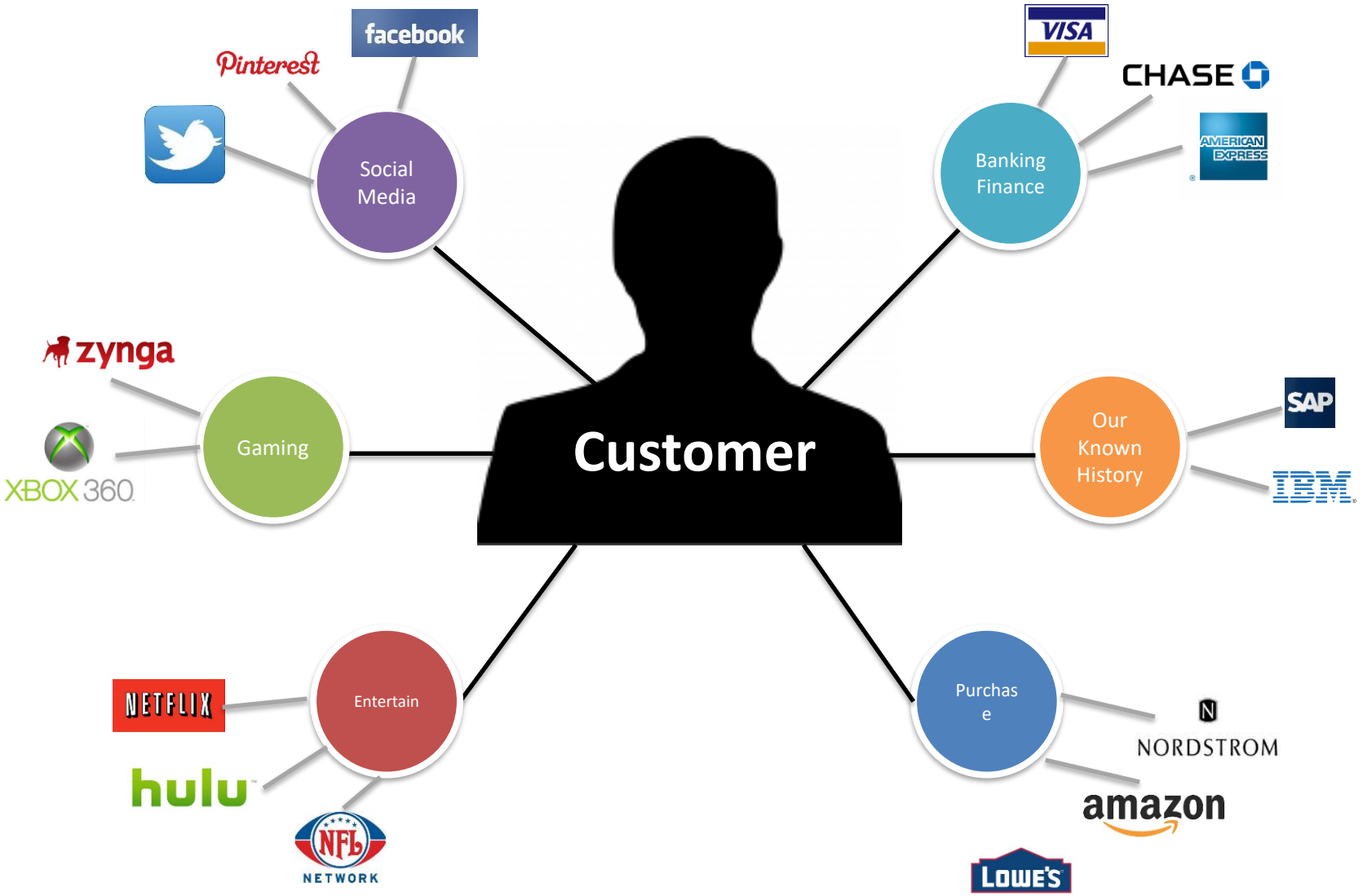






To extract knowledge➔ all these types of data need to linked together

A Single View to the Customer



Velocity (Speed)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

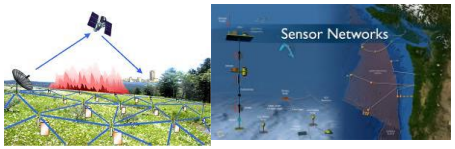
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Real-time/Fast Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

Some Make it 4V's