Principles & Architecture

EPITA Bachelor of Science

Principles and Architecture of Information Systems

Chapter #4

Database Systems

Olivier BERTHET

Principles & Architecture

Evaluation

Description	Percentage	Marks
Participation	20.0	4.0
Exercise and Wooclaps	30.0	6.0
Quiz	30.0	6.0
Case study	20.0	4.0

Principles & Architecture

Structure

- Chapter 1 : Introduction and Organisations
- Chapter 2 : Hardware
- Chapter 3 : Software
- Chapter 4 : Database Systems
- Chapter 5 : Network
- Chapter 6 : Internet and E-Commerce
- Chapter 7 : Major Information Systems
- Chapter 8 : Systems Development
- Chapter 9 : Security,Privaca and Ethical issues

Principles & Architecture

Principles

- **Data management and modeling are key aspects of organizing data and information.**
- **A well-designed and well-managed database is an extremely valuable tool in supporting decision making.**
- **The number and types of database applications will continue to evolve and yield real business benefits.**

Principles & Architecture

Data Management

- **Without data and the ability to process it:**
  - **An organization could not successfully complete most business activities**
- **Data consists of raw facts**
- **To transform data into useful information:**
  - **It must first be organized in a meaningful way**

The Hierarchy of Data

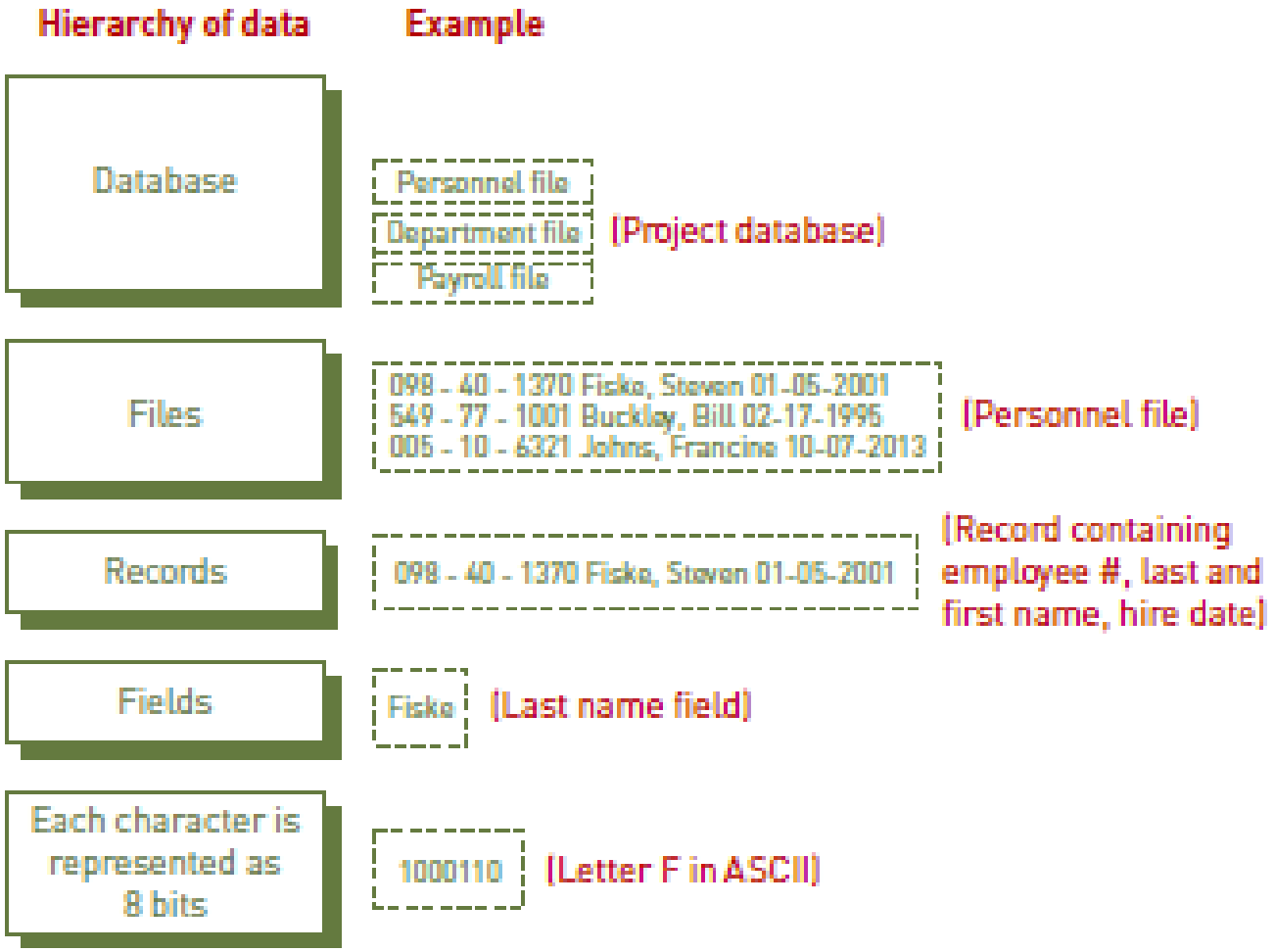
- **Bit (a binary digit)**
  - **Circuit that is either on or off**
- **Byte**
  - **Typically made up of eight bits**
- **Character**
  - **Basic building block of information**
- **Field**
  - **Name, number, or combination of characters that describes an aspect of a business object or activity**

The Hierarchy of Data

- **Record:**
  - Collection of related data fields
- **File:**
  - Collection of related records
- **Database:**
  - Collection of integrated and related files
- **Hierarchy of data:**
  - Bits, characters, fields, records, files, and databases

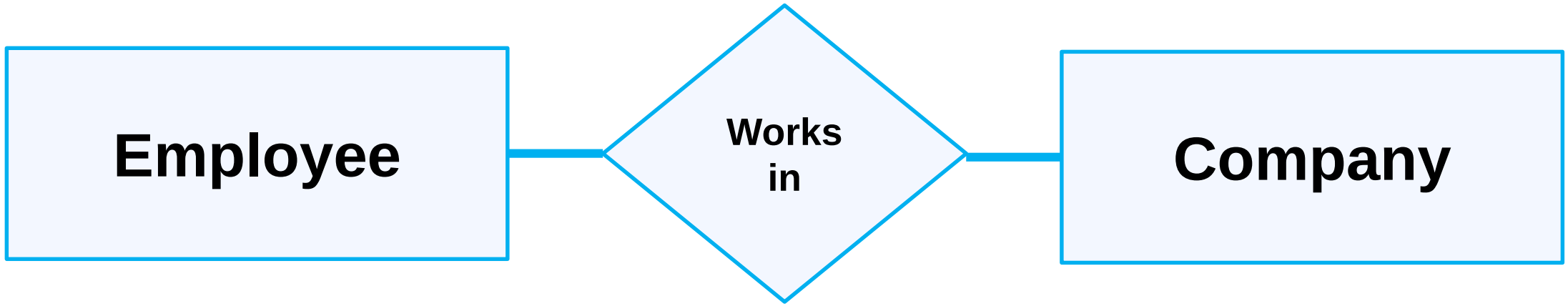
Principles & Architecture

The Hierarchy of Data



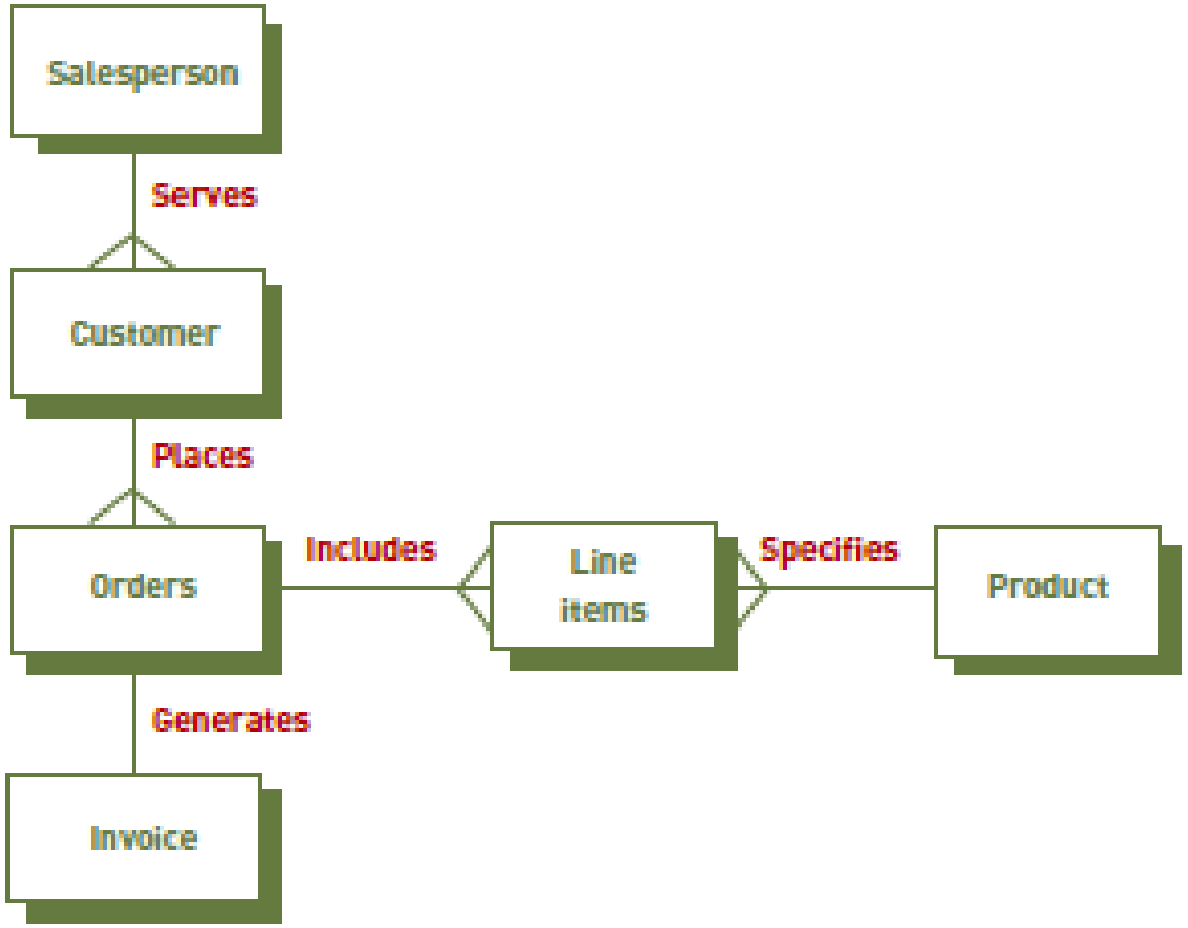
Entity-relationship (ER) diagram

Data models that use basic graphical symbols to show the organization of and relationships between data



Principles & Architecture

Entity Relationship model



Data Entities, Attributes, and Keys

- **Entity:**
  - Generalized class of people, places, or things (objects) for which data is collected, stored, and maintained
- **Attribute:**
  - Characteristic of an entity
- **Data item:**
  - Specific value of an attribute

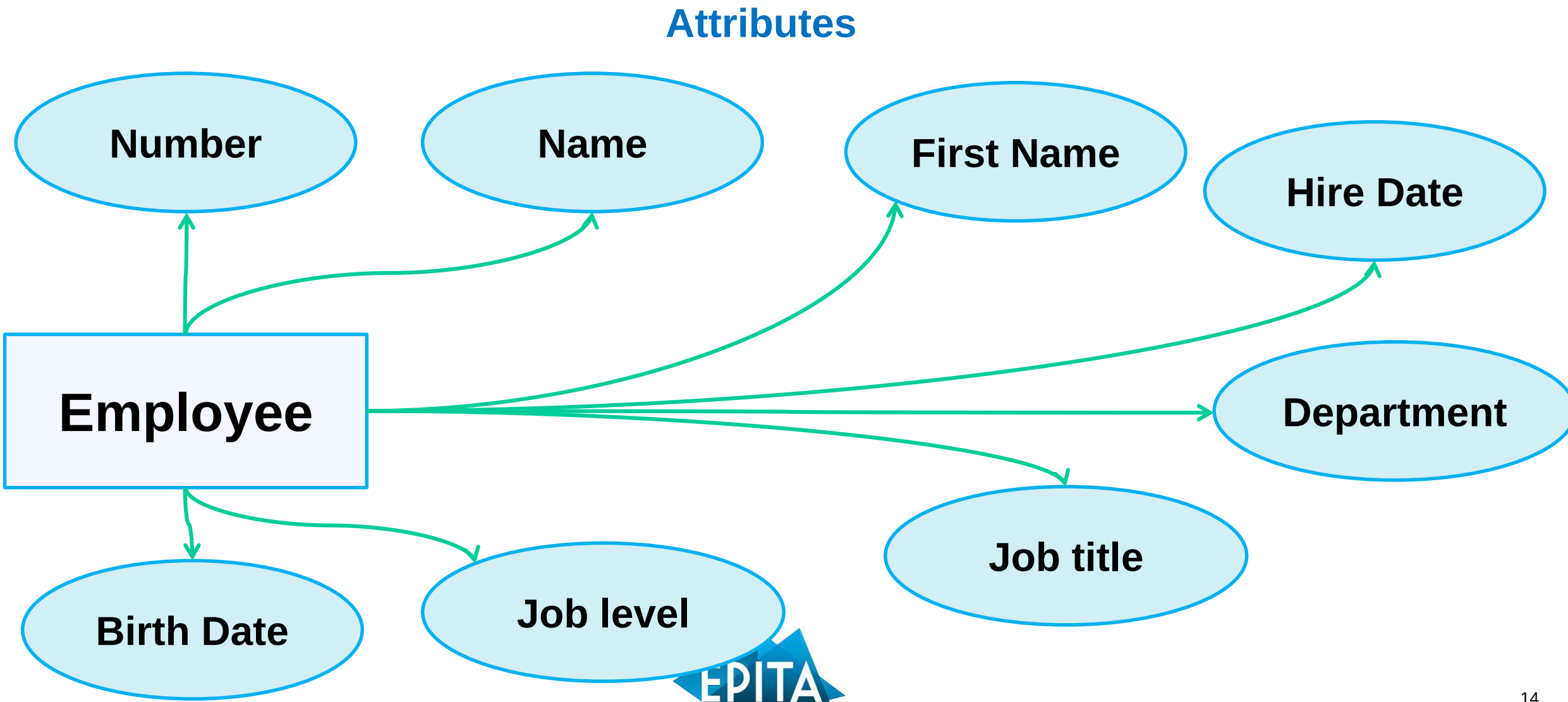
Entity : one example

Employee

What are the attributes of the Entity Employee ?

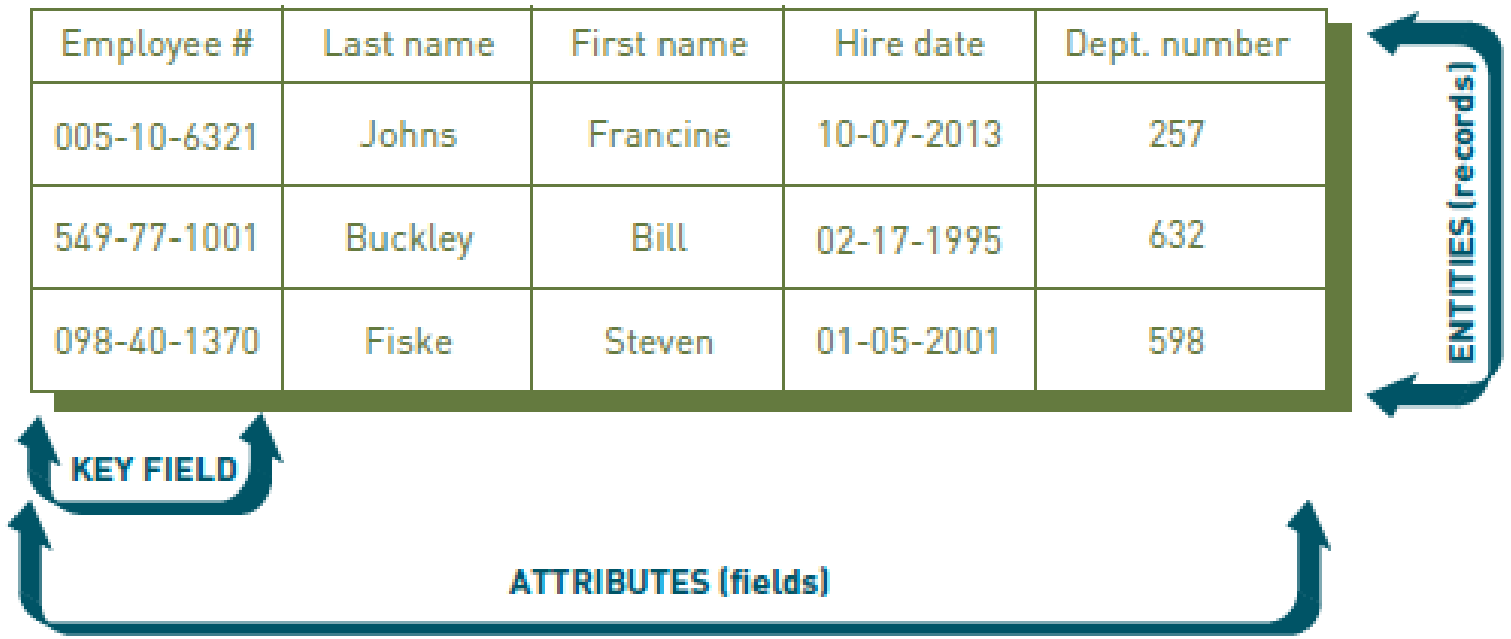
Employee

Principles & Architecture



Principles & Architecture

Data Entities, Attributes, and Keys



Principles & Architecture

Another exercise with entity Student

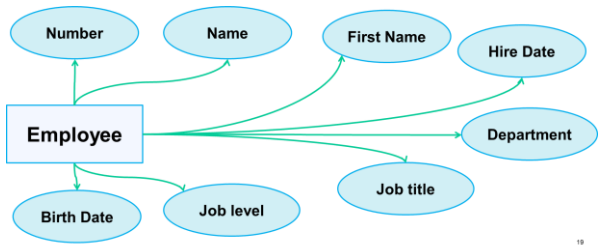
Student

Principles & Architecture

Conceptual, Logical and Physical Design

Conceptual

Logical



Physical

Employee #	Last name	First name	Hire date	Dept. number
005-10-6321	Johns	Francine	10-07-2013	257
549-77-1001	Buckley	Bill	02-17-1995	632
098-40-1370	Fiske	Steven	01-05-2001	598

Keys

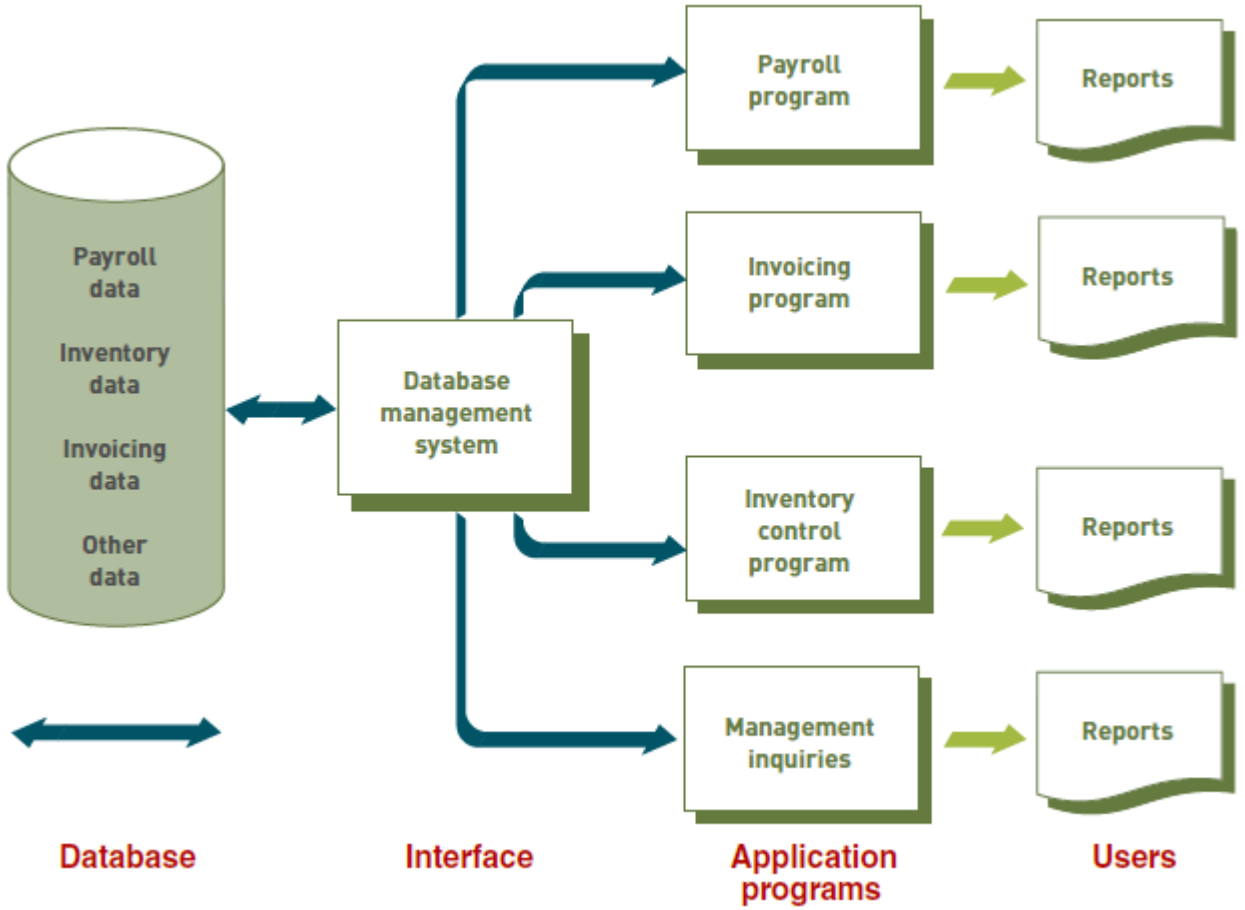
- **Key:**
  - **Field or set of fields in a record that is used to identify the record**
- **Primary key:**
  - **Field or set of fields that uniquely identifies the record**

The Database Approach

- **Traditional approach to data management:**
  - **Each distinct operational system used data files dedicated to that system**
- **Database approach to data management:**
  - **Pool of related data is shared by multiple application programs**

Principles & Architecture

Database approach to data management



The Database Approach

- **When building a database, an organization must consider:**
  - **Content: What data should be collected and at what cost?**
  - **Access: What data should be provided to which users and when?**
  - **Logical structure: How should data be arranged so that it makes sense to a given user?**
  - **Physical organization: Where should data be physically located?**

Advantages of the Database approach

Advantages	Explanation
Improved strategic use of corporate data	Accurate, complete, up-to-date data can be made available to decision makers where, when, and in the form they need it. The database approach can also give greater visibility to the organization's data resources.
Reduced data redundancy	Data is organized by the DBMS and stored in only one location. This results in a more efficient use of system storage space.
Improved data integrity	With the traditional approach, some changes to data were not reflected in all copies of the data. The database approach prevents this problem because no separate files are maintained.
Easier modification and updating	The DBMS coordinates data modifications and updates. Programmers and users do not have to know where the data is physically stored. Data is stored and modified once. Modification and updating is also easier because the data is commonly stored in only one location.
Data and program independence	The DBMS organizes the data independently of the application program, so the application program is not affected by the location or type of data. Introduction of new data types not relevant to a particular application does not require rewriting that application to maintain compatibility with the data file.
Better access to data and information	Most DBMSs have software that makes it easy to access and retrieve data from a database. In most cases, users give simple commands to get important information. Relationships between records can be more easily investigated and exploited, and applications can be more easily combined.
Standardization of data access	A standardized, uniform approach to database access means that all application programs use the same overall procedures to retrieve data and information.
A framework for program development	Standardized database access procedures can mean more standardization of program development. Because programs go through the DBMS to gain access to data in the database, standardized database access can provide a consistent framework for program development. In addition, each application program need address only the DBMS, not the actual data files, reducing application development time.
Better overall protection of the data	Accessing and using centrally located data is easier to monitor and control. Security codes and passwords can ensure that only authorized people have access to particular data and information in the database, thus ensuring privacy.
Shared data and information resources	The cost of hardware, software, and personnel can be spread over many applications and users. This is a primary feature of a DBMS.

Disadvantages of the Database approach

Disadvantages	Explanation
More complexity	DBMSs can be difficult to set up and operate. Many decisions must be made correctly for the DBMS to work effectively. In addition, users have to learn new procedures to take full advantage of a DBMS.
More difficult to recover from a failure	With the traditional approach to file management, a failure of a file affects only a single program. With a DBMS, a failure can shut down the entire database.
More expensive	DBMSs can be more expensive to purchase and operate than traditional file management. The expense includes the cost of the database and specialized personnel, such as a database administrator, who is needed to design and operate the database. Additional hardware might also be required.

The Relational Database Model

- **The relational database model is a simple but highly useful way to organize data into collections of two-dimensional tables called relations. Each row in the table represents an entity, and each column represents an attribute of that entity**

Data Table 1: Project Table

Project	Description	Dept. number
155	Payroll	257
498	Widgets	632
226	Sales manual	598

Data Table 2: Department Table

Dept.	Dept. name	Manager SSN
257	Accounting	005-10-6321
632	Manufacturing	549-77-1001
598	Marketing	098-40-1370

Data Table 3: Manager Table

SSN	Last name	First name	Hire date	Dept. number
005-10-6321	Johns	Francine	10-07-1997	257
549-77-1001	Buckley	Bill	02-17-1979	632
098-40-1370	Fiske	Steven	01-05-1985	598

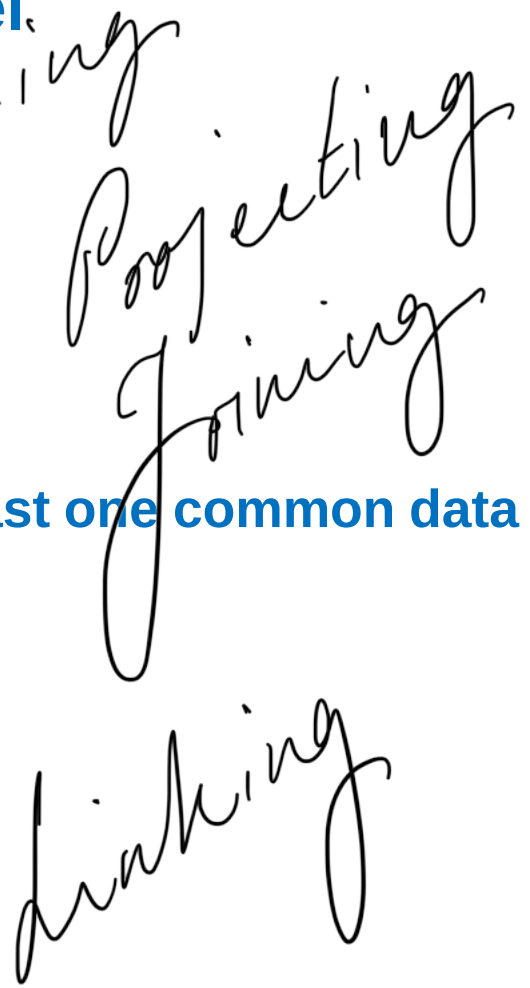
The Relational Database Model

- **Relational model:**
  - Describes data using a standard tabular format
  - Each row of a table represents a data entity (record)
  - Columns of the table represent attributes (fields)
  - Domain: Allowable values for data attributes

Principles & Architecture

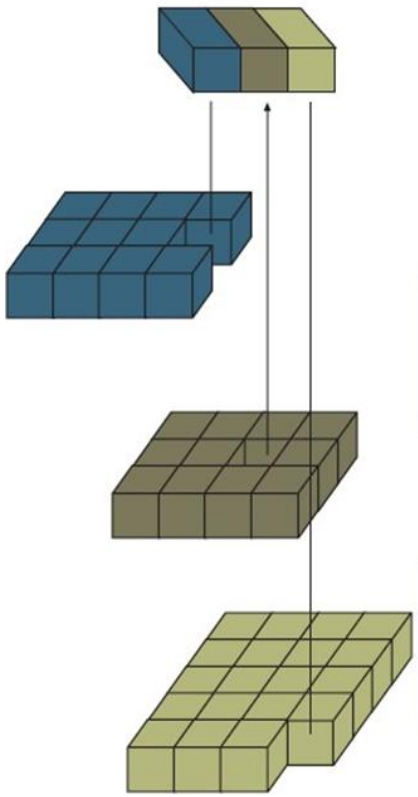
The Relational Database Model

- **Manipulating data:**
  - **Selecting: Eliminates rows according to certain criteria**
  - **Projecting: Eliminates columns in a table**
  - **Joining: Combines two or more tables**
  - **Linking: Manipulating two or more tables that share at least one common data attribute**



Principles & Architecture

Linking



Data Table 1: Project Table

Project number	Description	Dept. number
155	Payroll	257
498	Widgets	632
226	Sales manual	598

Data Table 2: Department Table

Dept. number	Dept. name	Manager SSN
257	Accounting	005-10-6321
632	Manufacturing	549-77-1001
598	Marketing	098-40-1370

Data Table 3: Manager Table

SSN	Last name	First name	Hire date	Dept. number
005-10-6321	Johns	Francine	10-07-1997	257
549-77-1001	Buckley	Bill	02-17-1979	632
098-40-1370	Fiske	Steven	01-05-1985	598

Overview of Database Types

- **Flat file**
  - Simple database program whose records have no relationship to one another
- **Single user**
  - Only one person can use the database at a time
  - Examples: Access, FileMaker Pro, InfoPath
- **Multiple users**
  - Allow dozens or hundreds of people to access the same database system at the same time
  - Examples: Oracle, MS SQL Server , PostgreSQL, IBM DB2, MySQL

Principles & Architecture

Some definitions

- **DBMS:**
  - Database Management System
- **Data definition language (DDL):**
  - Collection of instructions and commands used to define and describe data and relationships in a specific database
  - Allows database's creator to describe data and relationships that are to be contained in the schema
- **Data dictionary:**
  - Detailed description of all the data used in the database

Some definitions

- **Data manipulation language (DML):**
  - **Commands that manipulate the data in a database**
- **Structured query language (SQL):**
  - **Adopted by the American National Standards Institute (ANSI) as the standard query language for relational databases**
- **Once a database has been set up and loaded with data:**
  - **It can produce reports, documents, and other outputs**

Database administration

- **DBA:**
  - **Works with users to decide the content of the database**
  - **Works with programmers as they build applications to ensure that their programs comply with database management system standards and conventions**
- **Data administrator:**
  - **Responsible for defining and implementing consistent principles for a variety of data issues**

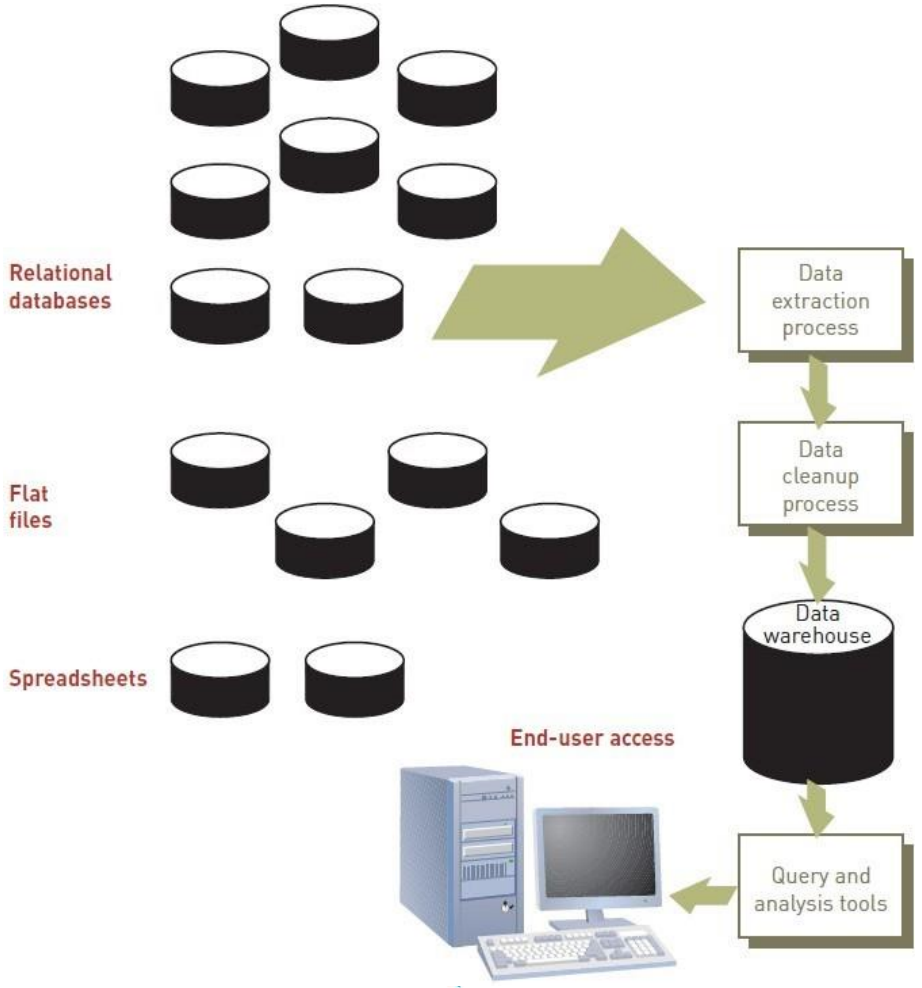
Principles & Architecture

Data Warehouses, Data Marts, and Data Mining

- **Data warehouse**
  - **Database that holds business information from many sources in the enterprise**
- **Data mart**
  - **Subset of a data warehouse**
- **Data mining**
  - **Information-analysis tool that involves the automated discovery of patterns and relationships in a data warehouse**

Principles & Architecture

Data Warehouses, Data Marts, and Data Mining



Predictive Analysis

- **Form of data mining that combines historical data with assumptions about future conditions to predict outcomes of events**
- **Used by retailers to upgrade occasional customers into frequent purchasers**
- **Software can be used to analyze a company's customer list and a year's worth of sales data to find new market segments**

Principles & Architecture

Data Warehouses, Data Marts, and Data Mining

Application	Description
Branding and positioning of products and services	Enable the strategist to visualize the different positions of competitors in a given market using performance (or other) data on dozens of key features of the product and then to condense all that data into a perceptual map of only two or three dimensions.
Customer churn	Predict current customers who are likely to switch to a competitor.
Direct marketing	Identify prospects most likely to respond to a direct marketing campaign (such as a direct mailing).
Fraud detection	Highlight transactions most likely to be deceptive or illegal.
Market basket analysis	Identify products and services that are most commonly purchased at the same time (e.g., nail polish and lipstick).
Market segmentation	Group customers based on who they are or on what they prefer.
Trend analysis	Analyze how key variables (e.g., sales, spending, promotions) vary over time.

Business Intelligence

- **Involves gathering enough of the right information:**
  - **In a timely manner and usable form and analyzing it to have a positive impact on business strategy, tactics, or operations**
- **Competitive intelligence:**
  - **Limited to information about competitors and the ways that knowledge affects strategy, tactics, and operations**

Business Intelligence

- **Counterintelligence:**
  - **Steps organization takes to protect information sought by “hostile” intelligence gatherers**
- **Data loss prevention (DLP):**
  - **Refers to systems designed to lock down data within an organization**
  - **Powerful tool for counterintelligence**
  - **A necessity in complying with government regulations that require companies to safeguard private customer data**

Distributed Databases

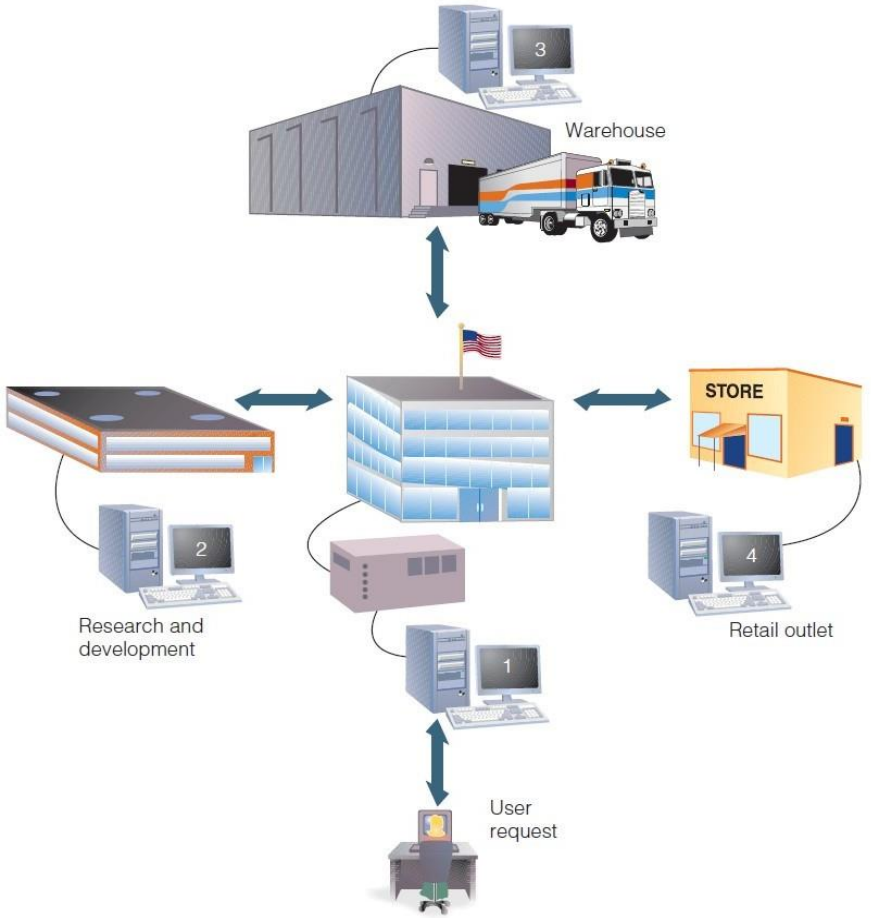
- **Distributed database:**
  - **Database in which the data may be spread across several smaller databases connected via telecommunications devices**
  - **Gives corporations more flexibility in how databases are organized and used**
- **Replicated database:**
  - **Holds a duplicate set of frequently used data**

Principles & Architecture

Distributed Databases

- **For a clothing manufacturer, computers might be located at corporate headquarters, in the research and development center, in the warehouse, and in a company owned retail store.**

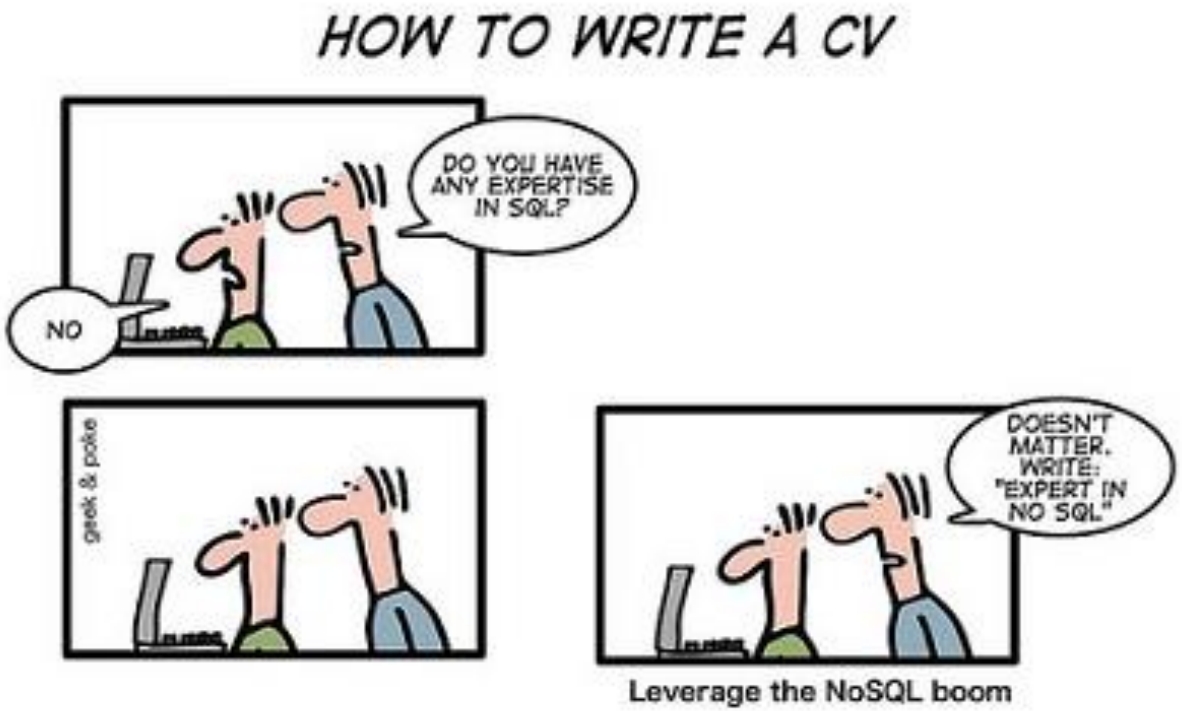
- **Telecommunications systems link the computers so that users at all locations can access the same distributed database no matter where the data is actually stored.**



Principles & Architecture

NoSQL

Two DBAs go into a NoSQL bar, but they leave right away because they can't find a table!



Principles & Architecture

Where does NoSQL come from?

- **Non-relational DBMSs are not new**
- **But NoSQL represents a new incarnation**
  - **Due to massively scalable Internet applications**
  - **Based on distributed and parallel computing**
- **Development**
  - **Starts with Google**
  - **First research paper published in 2003**
  - **Continues also thanks to Lucene's developers/Apache (Hadoop) and Amazon (Dynamo)**
  - **Then a lot of products and interests came from Facebook, Netfix, Yahoo, eBay, Hulu, IBM, and many more**

NoSQL Not Only SQL

- **A NoSQL database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.**
- **NoSQL databases are increasingly used in big data and real-time web applications.**
- **NoSQL systems are also sometimes called Not only SQL to emphasize that they may support SQL-like query languages or sit alongside SQL databases**
- **The data structures used by NoSQL databases (e.g. key–value pair, wide column, graph, or document) are different from those used by default in relational databases, making some operations faster in NoSQL. The particular suitability of a given NoSQL database depends on the problem it must solve.**

NoSQL and Big Data

- **NoSQL comes from Internet, thus it is often related to the “big data” concept**
- **How much big are “big data”?**
  - **Over few terabytes Enough to start spanning multiple storage units**
- **Challenges**
  - **Efficiently storing and accessing large amounts of data is difficult, even more considering fault tolerance and backups**
  - **Manipulating large data sets involves running immensely parallel processes**
  - **Managing continuously *evolving schema* and metadata for *semi-structured and un-structured* data is difficult**

Principles & Architecture

How did we get here?

- **Explosion of social media sites (Facebook, Twitter) with large data needs**
- **Rise of cloud-based solutions such as Amazon S3 (simple storage solution)**
- **Just as moving to dynamically-typed languages (Python, Ruby, Groovy), a shift to dynamically-typed data with frequent schema changes**
- **Open-source community**

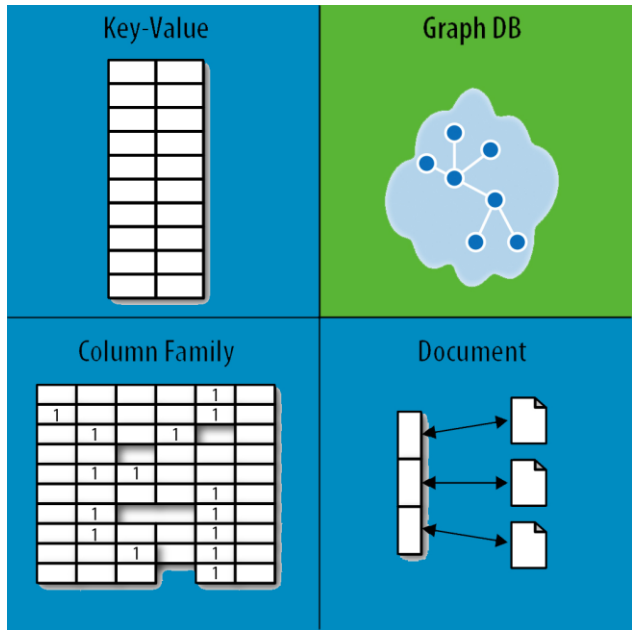
Principles & Architecture

Why are RDBMS not suitable for Big Data

- **The context is Internet**
- **RDBMSs assume that data are**
  - **Dense**
  - **Largely uniform (structured data)**
- **Data coming from Internet are**
  - **Massive and sparse**
  - **Semi-structured or unstructured**
- **With massive sparse data sets, the typical storage mechanisms and access methods get stretched**

Principles & Architecture

NoSQL Database Types



Large variety of types:

- **Sorted ordered Column Store**
  - –Optimized for queries over large datasets, and store columns of data together, instead of rows
- **Document databases:**
  - –pair each key with a complex data structure known as a document.
- **Key-Value Store :**
  - –are the simplest NoSQL databases. Every single item in the database is stored as an attribute name (or 'key'), together with its value.
- **Graph Databases :**
  - –are used to store information about networks of data, such as social connections.

Principles & Architecture

NoSQL Main vendors by Types

Key–Value	Document	Graph	Column
HyperDEX	Lotus Notes	Allegro	Accumulo
Couchbase Server	Couchbase Server	Neo4J	Cassandra
Oracle NoSQL Database	Oracle NoSQL Database	InfiniteGraph	Druid
OrientDB	OrientDB	OrientDB	Vertica
	MongoDB	Virtuoso	HBase

IMDB - In Memory Database

- **An in-memory database (IMDB) is a database management system that stores the entire database in random access memory (RAM).**
- **This approach provides access to data at rates much faster than storing data on some form of secondary storage (e.g., a hard drive or flash drive) as is done with traditional database management systems.**
- **IMDBs enable the analysis of big data and other challenging data-processing applications, and they have become feasible because of the increase in RAM capacities and a corresponding decrease in RAM costs.**
- **In-memory databases perform best on multiple multicore CPUs that can process parallel requests to the data, further speeding access to and processing of large amounts of data**

Principles & Architecture

IMDB - In Memory Database

Database Software Manufacturer	Product Name	Major Customers
Altibase	HDB	E*Trade, China Telecom
Oracle	Times Ten	Lockheed Martin, Verizon Wireless
SAP	High-Performance Analytic Appliance (HANA)	eBay, Colgate
Software AG	Terracotta Big Memory	AdJuggler

Principles & Architecture

Hadoop

- **Hadoop is an open-source software framework that includes several software modules that provide a means for storing and processing extremely large data sets,**
- **Hadoop has two primary components: a data processing component (a Java-based system called MapReduce and a distributed file system (Hadoop Distributed File System, HDFS) for data storage.**
- **Hadoop divides data into subsets and distributes the subsets onto different servers for processing.**
- **A Hadoop cluster may consist of thousands of servers. In a Hadoop cluster, a subset of the data within the HDFS and the MapReduce system are housed on every server in the cluster.**

Principles & Architecture

Hadoop