upcomillas es

upcomillas es

8-Management of large data volumes

Advanced Computing Tools for Applied Research

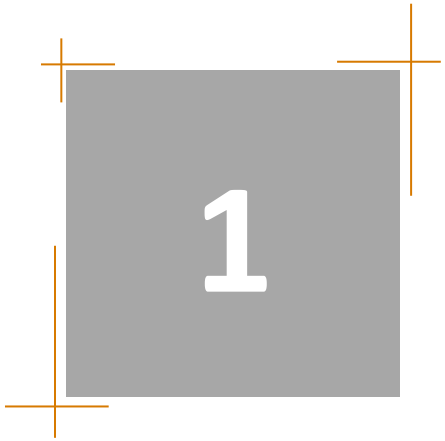
(*Herramientas Computacionales Avanzadas para la Investigación Aplicada*)

Rafael Palacios, Jaime Boal

Advanced Computing Tools for Applied Research
Contents

Storage systems

1. Introduction
2. Database description
3. Database management systems
4. SQL
5. Storage engines



Introduction

Introduction

- The maximum amount of memory that a program can handle efficiently is limited by the amount of RAM memory.
- When a program allocates too much memory, the operating system may begin to use virtual memory. Virtual memory decreases the performance dramatically.

Introduction

- Applications intended to work with large amounts of memory should be designed to load from the hard drive just the subset of information that they need every time.
- Distributed computing techniques could be applied
- Data could be stored in the hard disk in different ways:
  - Text files: more compatible, less efficient (access and storage)
  - Binary files: less compatible, more efficient
  - Databases: Best option for large datasets

2

Database description

Database description

- Database: Organized mechanism for storing, managing and retrieving information

- Databases use tables to store information

id	First Name	Last Name	Phone Number
1	Fernando	de Cuadra	6113
2	Rafael	Palacios	2758

Database description

- Database: Organized mechanism for storing, managing and retrieving information

- Databases use tables to store information

id	First Name	Last Name	Phone Number
1	Fernando	de Cuadra	6113
2	Rafael	Palacios	2758

Each row is a record

Each column is an attribute

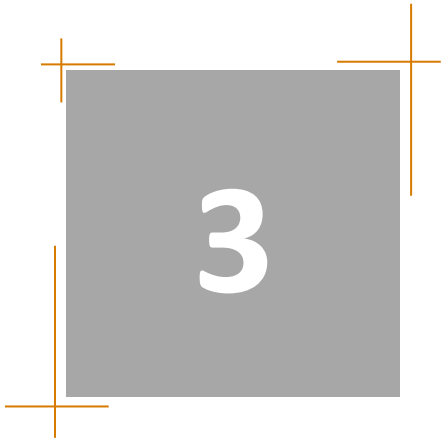
Database vs Spreadsheet

- A database that uses just one table is the same as a spreadsheet.
- The power of databases comes when several tables are used and they are linked together.
- Example:

Model	HP	CO2	Color	# units
Focus 1.6	100	159	Blue	2
Focus 1.6	100	159	Red	1
Focus 1.8D	115	137	White	1
Focus 1.8D	115	137	Red	2
Focus 1.8D	115	137	Blue	1

ID	Model	HP	CO2
F100	Focus 1.6	100	159
F110	Focus 1.8D	110	137

ID	Color	# units
F100	Blue	2
F100	Red	1
F110	White	1
F110	Red	2
F110	Blue	1



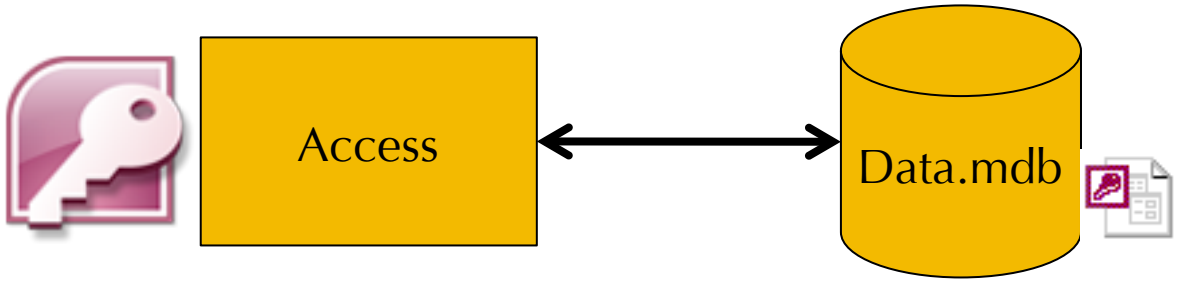
Database management systems

Database management systems

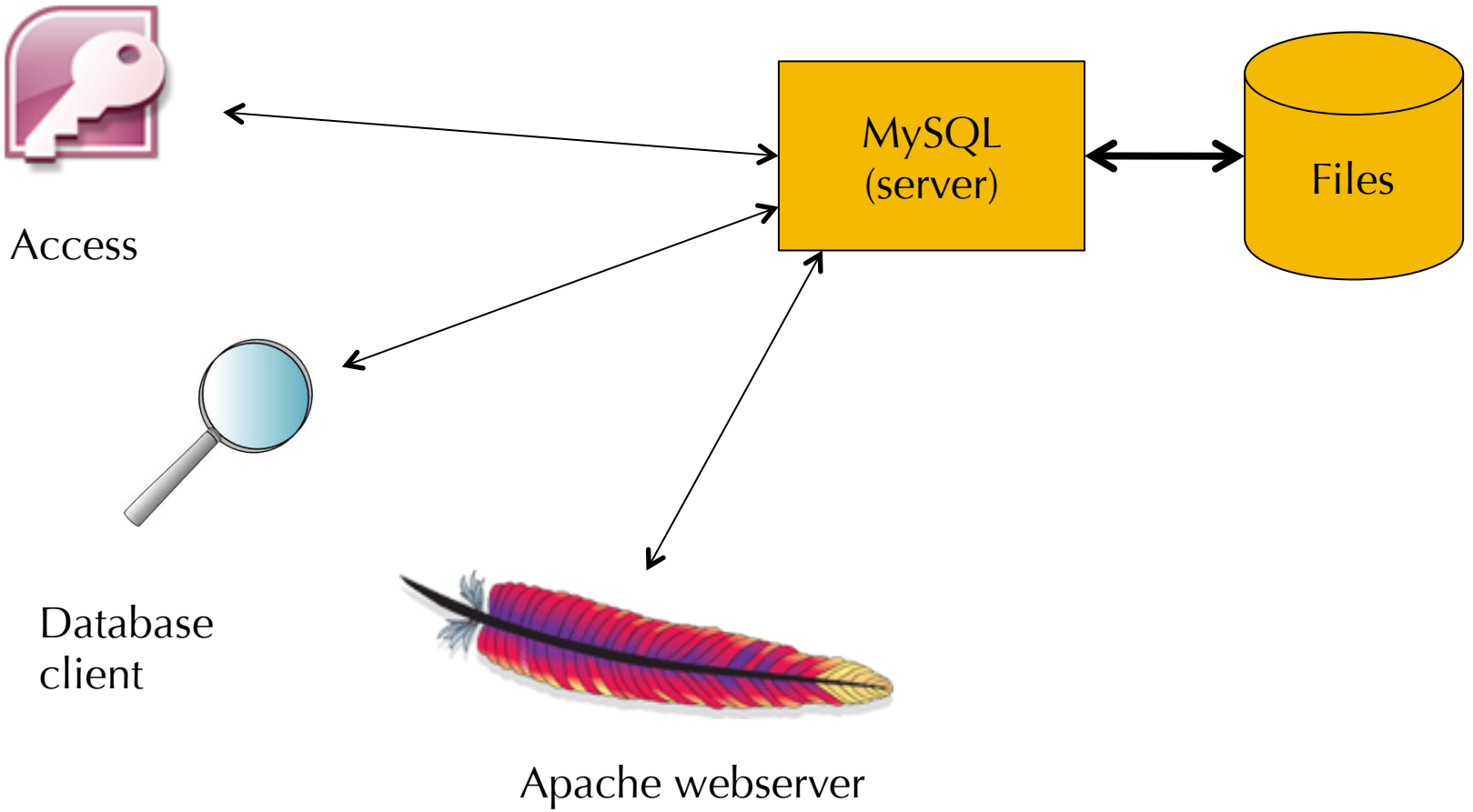
- Desktop databases
  - Single user
  - Ex. Microsoft Access, File Maker

- Server databases
  - Multiuser applications
  - Ex. MySQL, MS SQL server, Oracle, DB2…
  - Databases for web servers
  - Important characteristics: Performance, Scalability

Desktop databases



Server Databases



4

SQL

SQL

- SQL is the standard interface with all databases

- SQL = Structured Query Language

- Standard: ANSI SQL
- Specific
  - Oracle **PL/SQL**
  - Microsoft SQL Server **Transact-SQL**

- SQL is divided into two modules:
  - DDL Data definition language. To create tables, keys, etc…
  - DML Data Manipulation Languate. To insert, delete, update and retrieve data

SQL Data definition example

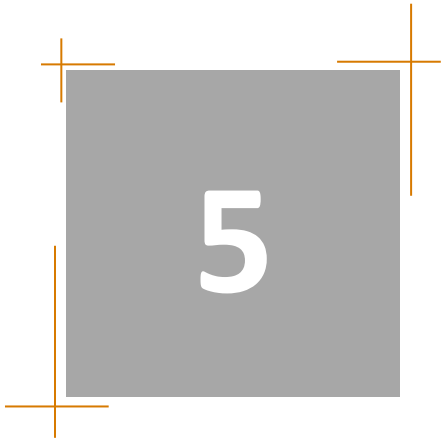
```
CREATE TABLE mytable (
  col1 INT,
  col2 CHAR(5),
  col3 DATE
);
```

```
CREATE TABLE `usuarios` (
  `uid` int(11) NOT NULL DEFAULT '0',
  `nombre` varchar(30) DEFAULT NULL,
  `apellidos` varchar(50) DEFAULT NULL,
  PRIMARY KEY (`uid`)
) ENGINE=MyISAM DEFAULT CHARSET=latin1;
```

SQL Data manipulation example

```
SELECT * from employees where salary>50000;
```

```
SELECT employees.first_name, employees.last_name
FROM employees,department
WHERE
   employees.id_department=department.id_department
   AND department.school='ICAI'
   AND employees.salary>50000;
```



Storage engines

Types storage engines

- Transaction (Ex. InnoDB)
  - Avoids inconsistencies
    - Some difficulties for changing data structures
    - Deleting one row may delete several records in linked tables
  - Fine grain privileges
  - Safer. Easier to recover data from server crashes
  - Better concurrency for tables with many updates and reads
  - Disable autocommit for better performance:
    - Several statements can be combined then accepted (commit) at the same time
    - Rollback can be used to ignore changes

InnoDB

- MySQL InnoDB characteristics
  (dev.mysql.com/doc)

Table 13.4. InnoDB Storage Engine Features

Storage limits	64TB	*Transactions*	Yes	*Locking granularity*	Row
MVCC	Yes	*Geospatial data type support*	Yes	*Geospatial indexing support*	No
B-tree indexes	Yes	*Hash indexes*	No	*Full-text search indexes*	No
Clustered indexes	Yes	*Data caches*	Yes	*Index caches*	Yes
Compressed data	Yes [a]	*Encrypted data* [b]	Yes	*Cluster database support*	No
Replication support [c]	Yes	*Foreign key support*	Yes	*Backup / point-in-time recovery* [d]	Yes
Query cache support	Yes	*Update statistics for data dictionary*	Yes		

Types storage engines

- Non-Transaction (Ex. MyISAM)
  - Much faster
  - Lower disk space requirements
  - Less memory required to perform updates
  - May have inconsistencies
    - Easier to change data structures
    - Unexpected results due to inconsistencies

MyISAM

- MySQL MyISAM characteristics (dev.mysql.com/doc)

Table 13.2. `MyISAM` Storage Engine Features

Storage limits	256TB	*Transactions*	No	*Locking granularity*	Table
MVCC	No	*Geospatial data type support*	Yes	*Geospatial indexing support*	Yes
B-tree indexes	Yes	*Hash indexes*	No	*Full-text search indexes*	Yes
Clustered indexes	No	*Data caches*	No	*Index caches*	Yes
Compressed data	Yes [a]	*Encrypted data* [b]	Yes	*Cluster database support*	No
Replication support [c]	Yes	*Foreign key support*	No	*Backup / point-in-time recovery* [d]	Yes
Query cache support	Yes	*Update statistics for data dictionary*	Yes		

Instituto de Investigación Tecnológica
C/ Santa Cruz de Marcenado, nº 26
28015 Madrid
Tel +34 91 542 28 00
Fax + 34 91 542 31 76
info@iit.upcomillas.es

www.iit.upcomillas.es

UNIVERSIDAD PONTIFICIA ICAI ICADE COMILLAS MADRID