

Database Management Systems

Vijaya Saradhi

IIT Guwahati

Mon, 06th Jan 2020

Twenty-twenty

2020

Wishing you all a very happy new year 2020

Outline

Theory

- About CS245
- Text books
- Pre-requisites
- Evaluation model
- Grading policy
- Re-evaluation policy

Outline

Databases Lab

- About CS246
- Pre-requisites
- Lab organization
- Lab material
- Lab teaching assistants (TAs)
- Evaluation model
- Grading policy

About CS245

Syllabus

Using DBMS as a black box

- ER Model
- Relational model and algebras,
- SQL
- Normalization

About CS245

Syllabus

Internals of relational DBMS

- File organizations
- Indexing (tree, hash, and bitmap)
- Implementation of relational operators

About CS245

Syllabus

Transaction management

- ACID properties
- Concurrency control
- Crash recovery

About CS245

Syllabus

Non-relational DBMS

- Consistency and availability trade-offs
- NoSQL DBMS (key-value, document, and graph)

CS245: Books

Text Books

- R. Ramakrishnan and J. Gehrke Database Management Systems, McGraw Hill, 2014
- P. Sadalage and M. Fowler NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence, Addison Wesley, 2012
- H. Garcia-Molina, J. Ullman, J. Widom, Database System: The Complete Book, 2nd Edition, Pearson, 2013

CS245: Pre-requisites

Course Pre-requisites

Have credited [CS203](#) Algorithms and Data Structures

CS245: Evaluation Model

- You must verify the exam timings with the Academic Section intranet webpage
- Exams Mode: closed text, closed notes

S. No.	Description	Weightage (%)	Exam Date
1	Mid Semester Exam	40	28-Feb-2019
2	End Semester Exam	60	03-May-2019

CS245: Grading and Re-evaluation Policy

- Relative grading is followed
- Answer script checking dates

Description	Start Date	End Date	Time
Mid Sem	09-Mar	13-Mar	10:00 - 13:00
End Sem	11-May	13-May	10:00 - 13:00

- Check your answers and compare with the published answers
- In case you cannot come to check answer script, nominate any person in writing to the course instructor.
- The nominated person produce the application and can check your answer script and request for re-evaluation if any.

CS245: Grading and Re-evaluation Policy

- In case of any issue with the evaluation and published answers do point out the same in writing on your [answer sheet](#).
- Re-evaluated marks will be published within [one week](#) after the deadline period of evaluation.
- Re-evaluation [concludes](#) once the published timeline for answer script checking is over for the appropriate evaluation.
- Constraints of any sort are hard to taken into account for relaxation of the published dates.

About CS246 Database Management Systems Lab

Syllabus

- **Using a relational DBMS** Writing SQL queries, accessing a DBMS from an external application
- **Implementing of parts of DBMS** various file organizations, indexing methods (Tree/ Hash/ Bitmap), external sorting algorithms, concurrency control schemes, and crash recovery schemes
- **Non-relational DBMS**
- **Performance comparison** non-relational DBMS with a relational DBMS for an application

About CS246 Database Management Systems Lab

Material/Text Books

- H. Garcia-Molina, J. Ullman, J. Widom, **Database System Implementation**, 2nd Edition, Pearson, 2002
- J. Groff and P. Weinberg, **SQL Complete Reference**, McGraw Hill, 3rd Edition, 2017.
- P. Sadalage and M. Fowler, **NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence**, Addison Wesley, 2012

CS246: Pre-requisites

Course Pre-requisites

- Have credited [CS203](#) Algorithms and Data Structures
- Have credited [CS204](#) Algorithms and Data Structures **Lab**

CS246: Labs Organization

Organization

- Every lab will have an assignment
- Material required to perform the lab is the contents of the previous week's CS245 theory classes
- Or material is explicitly shared with you so that you are equipped to perform the lab session
- Complete the assignment within the lab hours

CS246: Labs Organization

Evaluation and Grading

- The solution will be evaluated by the TAs
- Marks will be awarded as per your implemented solution
- Re-evaluation request must be made before the weekend (Saturday)
- Head TA: Swarup Ranjan Behera (b.swarup)
- Relative grading is followed

Databases

Introduction

- Integral part of our day-to-day life
- We are not aware that we are using one
- Examples

Supermarket

Database access

- Barcode reader scans every item that is purchased
- This reader is linked to a database application
- Finds out the price of the item from the product database
- Performs stock checking
- Update the stock entry
- Produces the bill

Purchase using credit card

Database access

- Purchasing goods using credit card
- Check for sufficient credit left to make the purchase
- This check is performed using credit card number
- Check for the total price of the goods
- $(\text{Money already spent in the given month}) + \text{above sum} < \text{credit limit}$ — $>$ confirm purchase
- Complex tasks
 - Credit card is not stolen one
 - Not in the list of lost cards

Complex examples

In terms of use cases

- Withdraw funds from a bank
- Booking an airline reservation
- Booking a railway ticket
- Booking a hotel room
- Purchase an item from online

Complex examples

In terms of volume

- Telecommunication companies
- Number of calls per second is 70000 (in 2007)
- Maintain call data record of the form

S. No	Date	Time	To-Phone	Duration	Amount
-------	------	------	----------	----------	--------

- Business rules for billing
 - Within the same network
 - Across networks
 - Taking care of increased pricing
- Have to manage several exabytes of data (of the order of 1000000 TB)

Complex examples

In terms of data types

- Text databases
- Image databases
- Video databases
- Audio databases
- Source Code databases
- Geographical information databases
- Graph databases
- Streaming databases
- and many more

Definitions

Database

Collection of [related data](#)

Definitions

Database

Collection of **related data**

Data

- Known facts that can be recorded and that have **implicit meaning**
- Example: names, telephone numbers, and addresses of people you know
- This information is stored on a hard drive using a PC and software such as Microsoft Access or Excel
- This collection of **related facts** is a database

Definitions

Database

Collection of **related data**

Data

- Known facts that can be recorded and that have **implicit meaning**
- Example: names, telephone numbers, and addresses of people you know
- This information is stored on a hard drive using a PC and software such as Microsoft Access or Excel
- This collection of **related facts** is a database

Implicit Properties

- A **mini-world**: Should represent some aspect of real world called Universe of Discourse (UoD).
- A database is a logically coherent collection of data with inherent

File Systems

Limitations - setting the context

- Assume that our institute has large collection of data around 500 GB
- About employees, departments, students, registration, marks, attendance, grades, etc.
- Questions about this data to be answered quickly
 - List the toppers of every department
 - List which department students got silver medal the most
 - List which department gave highest number of AA grades
 - Which department got least number of F grades due to attendance

File Systems

Limitations

- We do not have 500 GB of main memory (earlier days)
- 32-bit computer system can refer directly to only 4GB of data
- Write special purpose programs for every question
- Protect data from inconsistent changes
- Protect data from unauthorized access
- Protect data from failures
- Restore data in case of disk failures (what is the mechanism?)
- Issues in making copies of data
- Issues in updating data

File Systems

Limitations

- Write special purpose programs for every question
 - Lengthy development time
 - Difficulty in getting the answers quickly (some one has to develop these programs)
 - Maintaining data AND programs becomes complex task (earlier we have only data)
 - Flexible formatting is not possible

Advantages of DBMS

Advantages

- **Data independence** Application programs should not be exposed to details of data representation
- **Efficient data access** **effective storage** and **efficient retrieval**
- **Data integrity and security** always accessed through DBMS which enforces constraints
- **Data administration** centralizing the administration of data leads to significant improvements
- **Concurrent access**
- **Crash Recovery**
- **Reduce application development time**

DBMS

What it contains?

- A DBMS is a **collection of programs** that enables users to **create and maintain** a database
- DBMS is a **general-purpose software system**
- It facilitates the process of
 - Defining
 - Constructing
 - Manipulating and
 - Sharing database among various users and applications

Defining

Elements

Defining a database involves specifying

- the **data types**
- **data structures**
- **constraints** on the data to be stored in the database
- Database descriptive information is also stored by the DBMS
- The description is in the form of database catalog which is the **meta-data**

Constructing, Manipulating and Sharing

Elements

- Constructing the database is the process of storing the data on some **storage medium** that is controlled by the DBMS
- **Manipulating** a database includes functions such as
 - Querying the database to retrieve specific data
 - Updating the database to reflect changes in the mini-world
 - Generating reports from data
- **Sharing** a database allows multiple **users** and **programs** to access the database **simultaneously**

Size of Database - Specific Example - Approximate computation

- A train has 72 berths, 26 bogies
- Assume uniformity in bogies; reservation policies
- Berth reservation has **associated data**
 - User name (50), Age (4), Gender (1), Identity (12)
 - From (50) and To (50) information
 - Where the passenger is boarding (50)
 - Date of travel (10)
 - Reservation can be made 120 days before date of travel
- Berth assignment has **associated data**
 - Train number (5)
 - Train Date (Note this date is different from user entered travel date) (10)
 - Bogie number (4)
 - Birth number (4)

Size of Database - Specific Example - Approximate computation

- Total memory is $(50+4+1+12+50+50+50+10) + (5 + 10+ 4+4) = 246$
- Information storage per train per day:

$$\frac{(250 * \overbrace{72}^{\text{berths}} * \overbrace{26}^{\text{bogies}} * \overbrace{120}^{\text{before}})}{(\underbrace{1024}_{\text{kilo}} * \underbrace{1024}_{\text{mega}} * \underbrace{1024}_{\text{giga}})} = 0.0523 \text{ GB}$$

- Two distinct trains play between stations
- Assume 1000 trains runs across
- $0.0523 * 2 * 1000 = 104.606 \text{ GB/per day}$

Database - Queries

- List trains between GHY and VSKP
- List trains playing on 10-Oct-2018
- List passengers who name is "abcd" and travelling between GHY and VSKP
- List trains having only second AC

Database - Queries

- Reserve a berth on 10-Oct-2018 for train no. 12345 from GHY to VSKP
- Cancel a reservation
- Update date of travel
- Update passenger name

Database - Queries

- Include a new train
- Delete a train
- Modify train information
- Cancel a train
- Schedule a train

Users of the Database

- Passengers
- Agents
- Tellers
- Engineerings
- Managers
- Administrators

Database - User Interfaces

- Each user has different roles to perform on the stored data
- Web forms
- Client connections
- Application Front Ends
- SQL interfaces