Introduction to Data Structure (Data Management)

Felipe P. Vista IV

Course Outline

- Introduction, Data Models, SQL Basics
- SQL Aggregates, Grouping, Subqueries
- Wrapping-up SQL, Relational Algebra (RA), Datalog
- NoSQL, JSON
- JSON, SQL++
- SQL++, RA Part II, Query Evaluation
- Storage, Indexing Basics
- Basics of Query Optimization, Parallel Databases
- Map Reduce, Spark
- E/R Diagrams, Constraints
- Design Theory
- Transactions
- DB Techniques for Machine Learning

Introduction

- Class Administrative Matters
- Introduction
- Database Management Systems

Data Management

CLASS ADMIN MATTERS

Class Information

- Class Schedule
  - Wed: 15:00(3pm) – 16:00(4pm); Fri: 14:00(2pm) – 16:00(4pm)
- Mode of instruction
  - Online lecture via ZOOM
- Assignments
  - Given during lecture or posted at IELMS (http://ieilms_old.jbnu.ac.kr/)
- MidTerms and Finals
  - Most probably online
- Textbook
  - *"Database Systems: The Complete Book, 2nd Edition"*, Hector Garcia-Molina, Jeffrey Ullman, Jennifer Widom.

Grading

- Midterms : 20% 20
- Finals : 30% 30
- Attendance/participation : 20% 14
- Assignment : 30% 30

94 => 1

Grading

- Mid Terms (20%) and Finals (30%)
  - Enough time will be given
  - It is ok to discuss with classmates but **submit your own solution!**
  - Discussing is ok, cheating is "no-no"→ candidate for automatic "F"
  - Late submission = less points, maximum 90%-95% per item/number
    - In case we have to do tests online
  - With answer, minimum score is 75% per item/number
  - No answer is automatic 70% per item/number
  - Non submission is automatic 50% per item/number

Grading

2hr
W = 1 class
F = 2 classes
2hrs

- Attendance (8)
  - more than 15 mins late = *absent*, and 3 late = 1 *absent*
  - more than 3 absences = *problem (very biiig)*
  - Everybody start with 8 points for attendance
    - Become less if too much absences, ex: 70% of 8 = 5.6 points
  - Absences of more than ¼ total number of hours = F
    - Total is 15 wks x 3hrs/wk = 45; 45/4= 11.25 ≈12hrs → F

Grading

- Attendance (8) 10 – 5
  - more than 15 mins late = *absent*, and 3 late = 1 *absent*
  - more than 3 absences = *problem (very biiig)*
  - Everybody start with 8 points for attendance
    - Become less if too much absences, ex: 70% of 8 = 5.6 points
  - Absences of more than ¼ total number of hours = F
    - Total is 15 wks x 3hrs/wk = 45; 45/4= 11.25 ≈12hrs
- Participation(12) – 10 – 15
  - answer/raise questions during lecture to get points
  - everybody starts with 12 points for participation
    - Become less if you have less than 6 class participation, ex: 70% of 12 = 8.4

Grading

- Assignment : 30%
  - It will take some time
  - Mostly practical, to help learn
  - It is ok to discuss with classmates but **do it yourself!**
  - Assignments usually due one week after posting,

  - Late submission = less points, maximum 90%-95% per item/number
  - Submission with answer, min is 75% per item/number
  - Submission but no answer is automatic 70% per item/number
  - Non submission of assignment is automatic 50% per item/number

Grading

- Midterms : 20%
- Finals : 30%

MIDTERMS

1	2	3	4	5	6	7	8	9	10	Overall Score
10	5	10	10	10	10	9	8	5	5	82.00
7	9	10	10	10	10	7	9	10	10	92.00
5	5	10	10	10	8.5	5	8	9	8.5	79.00
										0.00
10	9	8.5	5	9	10	5	5	10	5	76.50
5	5	10	10	10	10	8.5	5	8.5	8.5	80.50
5	5	10	10	10	8.5	8.5	9.25	10	7	83.25
10	10	10	10	10	10	7	9	9	10	95.00
10	10	10	10	10	10	10	9.75	10	10	99.75
10	7	10	10	10	9	8	10	10	10	94.00

FINAL EXAMS

1	2	3	4	5	6	7	8	9	10	Overall Score
10	9	10	9	9	10	10	7	10	10	94.00
7	7	10	7	7	10	7	7	7	10	79.00
10	10	10	5	8	7	5	7	5	5	72.00
										0.00
10	7	10	8	5	5	5	5	5	5	65.00
5	5	5	5	5	5	5	5	5	5	50.00
10	10	10	10	9	9	10	9	10	5	92.00
10	8	10	9	5	5	5	5	10	9	76.00
10	10	10	7	10	10	10	10	10	10	97.00
10	10	10	10	10	10	10	9	10	10	99.00

Grading

- Attendance/participation : 20%

	Week14		Week15 (Finals)			Raw Score	Grade Equivalent
16-Jun	16-Jun	18-Jun	23-Jun	23-Jun	25-Jun		
						0.00	100
						3.00	70
						1.00	100
1	1	1	1	1	1	20.00	0
						1.00	100
						3.00	70
		1				2.00	100
		1				3.00	70
						0.00	100
						2.00	100

					CLASS PARTICIPATION					
Week14		Week15 (Finals)			No of Times participated	Improtant ones missed	Grades I	Regular Conversation	Addl Points based on Regular Conversation	Grade I + Addl Points
16-Jun	18-Jun	23-Jun	23-Jun	25-Jun		(Total of 5)			(0.5 points per)	
	1	1		1	7.00	0	100	2	1	101
					4.00	3	97	2	1	98
					1.00	5	95	1	0.5	95.5
					0.00	5	95	0	0	0
					2.00	5	95	2	1	96
					2.00	4	96	1	0.5	96.5
					0.00	5	95	0	0	95
					2.00	5	95	2	1	96
	1				11.00	0	100	6	3	103
					4.00	5	95	4	2	97

Grading

- Assignment : 30%

ASSIGNMENTS										
1	2	3	4	5	6	7	8	9	10	Overall Score
	90	80	86	95	95	95			95	**90.86**
	100	95	97	100	100	70			100	**94.57**
	90	85	84	85	85	70			70	**81.29**
										0.00
	90	90	94	90	90	88			100	**91.71**
	70	70	70	70	70	70			70	**70.00**
	90	85	84	85	80	70			70	**80.57**
	89	90	96	90	90	98			100	**93.29**
	100	100	98	90	100	90			100	**96.86**
	100	100	97	100	100	100			100	**99.57**

20 30 20 30

Midterms (30%)	Finals (35%)	Attendance/Participation (20%)		Assignment (15%)	Overall Score	Equivalent Score
		Attendance (5%)	Participation (15%)			
82.00	94.00	100.00	101.00	90.86	91.28	Ao
92.00	79.00	70.00	98.00	94.57	87.64	B+
79.00	72.00	100.00	95.50	81.29	80.42	Bo
0.00	0.00	0.00	0.00	0.00	0.00	F
76.50	65.00	100.00	96.00	91.71	78.86	C+
80.50	50.00	70.00	96.50	70.00	70.13	Co
83.25	92.00	100.00	95.00	80.57	88.51	B+
95.[illegible]	76.00	70.00	96.00	93.29	86.99	B+
99.75	97.00	100.00	103.00	96.86	98.85	A+
94.00	99.00	100.00	97.00	99.57	97.34	A+

Student Responsibilities

- Download/Install ZOOM app for online lecture
  - Zoom profile must be your name similar to OASIS
- Regularly login and check on-line learning system for updates, notifications
  - http://ielms_old.jbnu.ac.kr
  - Presentations & lecture videos will be uploaded after class
  - Assignments will be posted online
- Regularly check Kakao Group Chat
  - Everybody must have a Kakao talk account
  - Search & add account "botjok" then you will be added to the group chat

Basic Requirements for Course

- Laptop/Notebook/Computer
  - Necessary for the practical exercises/assignments
- Operating System
  - Linux/ MS Windows/ Mac OS
- Internet Connectivity

Who me?

- Faculty member GFC – School of International Eng'g & Science
  - Network System Control Lab – Electronic Eng'g Dept., JBNU
  - Advanced Electronic Research Information Center, JBNU
- PhD in Electronic Engineering, JBNU
- Worked at Industry & Government of Philippines
  - Industry: Drivers license, NBI CHD (ORACLE); PNRC (Sybase)
  - Government: Electronic Governance (LAMP)
- Research Interests:
  - Systems Design, Software Development, Fuzzy Logic, Sensor Fusion, Embedded Systems, Navigation systems, Marine Information System, Signal Processing & Augmented Reality.

Data Management

INTRODUCTION

Data.. Data… Lotsa' Data…

Per Minute

- Users watch 4,146,600 videos on Youtube
- 456,000 tweets are sent on Twitter
- Instagram users post 46,740 photos
- 16 million text messages sent
- 156 million emails are sent
- 15,000 GIFs are sent via Facebook messenger
- 154,200 calls on Skype

Per second

- Google processes more than 40,000 searches

https://www.bernardmarr.com/default.asp?contentID=1438

Data.. Data… Lotsa' Data…

shutterstock.com • 628802006

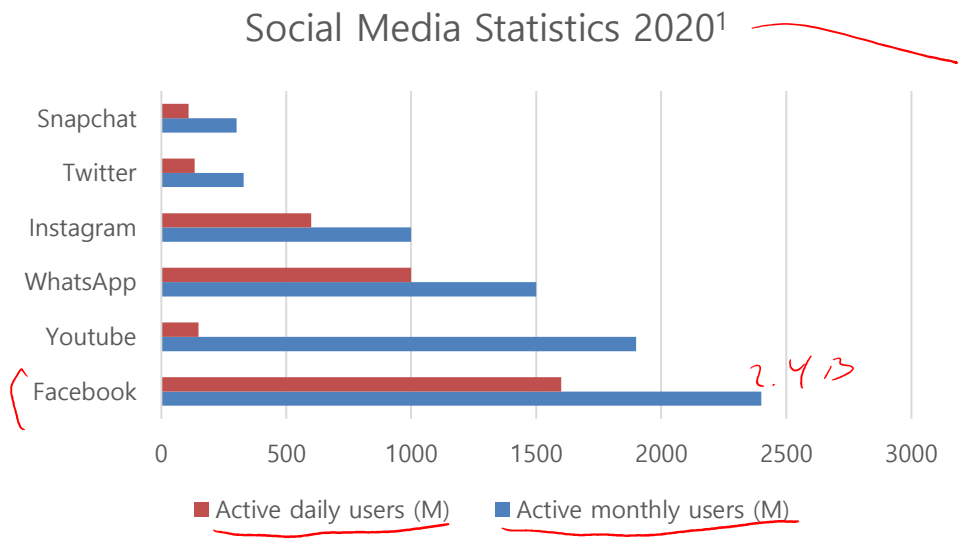
- Twitter: 12 TB of data per day[1]
  - 1 TB = 1024 GB
- Facebook: 4 petabytes of data per day[2]
  - 1 PB = 1024 TB = 1M GB
- Instagram: 95M photos & videos shared daily[3]
  - Bonus Participation point for who can give average data per day

[1] https://bigdatashowcase.com/how-much-big-data-companies-make-on-internet/

[2] https://research.fb.com/blog/2014/10/facebook-s-top-open-data-problems/

[3] https://business.instagram.com/blog/500000-advertisers/

Data.. Data… Lotsa' Data…



[1] https://dustinstout.com/social-media-statistics/

[2] https://www.pingdom.com/blog/webpages-are-getting-larger-every-year-and-heres-why-it-matters/

[3] https://www.littledata.io/average/pages-per-session-(all-devices)

[4] https://www.spinutech.com/digital-marketing/analytics/analysis/7-website-analytics-that-matter-most/

- Assume
  - 2.07MB bytes per page/site[2], 2.8 pages per session[3], 2-3 mins/session[4]
- If analyze trend for 3 months of data:
  - FB(2.4B/mo): 1.02×10^{20} bytes → **0.102 Zettabyte**, 1 ZB = 1×10^{21} bytes

Data Management is Universal

- Managing data is critical for most apps/services/ programs
  - old or new systems
  - small or large amount of data

Data Management is Universal

- Managing data is critical for most apps/services/ programs
  - old or new systems
  - small or large amount of data

- Even small amount of data can bring tough problems

Data Management is Universal

- Managing data is critical for most apps/services/ programs
  - old or new systems
  - small or large amount of data
- Even small amount of data can bring tough problems
- Managing data properly makes everything else easier

Motivation

- The world is drowning in data

Motivation

- The world is drowning in data
- Professionals are needed to help manage
  - help scientists discover new things
  - help companies offer better service
  - help government more efficient

Motivation

- The world is drowning in data
- Professionals are needed to help manage
  - help scientists discover new things
  - help companies offer better service
  - help government more efficient
- This course
  - will cover both theories & tools

Data Management

DATABASE MANAGEMENT SYSTEM

Database

- What is Database?
  - Set of files that store related data

Database

- What is Database?
  - Set of files that store related data

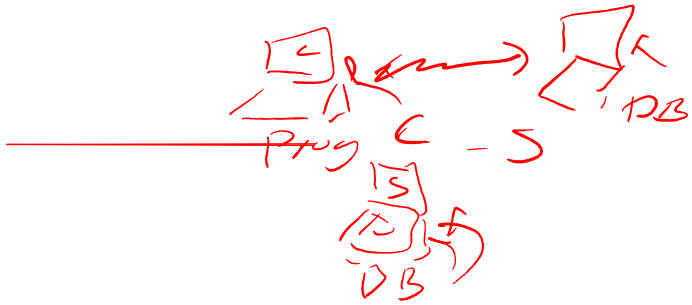
- Database examples
  - Business: Accounts (customers), payroll (salary)
  - Academe: JBNU students (grades, class) OASIS
  - E-Commerce: Coupang products
  - Transportation reservation: KORAIL, Express Bus

Database Management System

- What is DBMS?
  - Big program so that we can efficiently manage a large database and allow it to persist for long period of time → 「exist」

Database Management System

- What is DBMS?
  - Big program so that we can efficiently manage a large database and allow it to persist for long period of time
- DBMS examples
  - Oracle, Microsoft SQL Server, Sybase, Vertica, Teradata → Commercial
  - Open source: MySQL (Oracle/Sun), PostgreSQL FB → MySQL
  - Open source library: SQLite (not client-server but embedded in program)



Database Management System

- What is DBMS?
  - Big program so that we can efficiently manage a large database and allow it to persist for long period of time
- DBMS examples
  - Oracle, Microsoft SQL Server, Sybase, Vertica, Teradata
  - Open source: MySQL (Oracle/Sun), PostgreSQL
  - Open source library: SQLite (not client-server but embedded in program)
- Focus on relational DBMS

An Example: Online Concert Tickets Vendors

An Example: Online Concert Tickets Vendors

- What data do we need?
  - Data about artists, concerts, venues, reservation, hot artists, order histories, trends, preferences, etc.
  - Session data (searches, clicks, pages)
  - Note: data must be persistent (last longer than app)
    - *Data will be very large… cannot fit all in memory*

An Example: Online Concert Tickets Vendors

- What data do we need?
  - Data about artists, concerts, venues, reservation, hot artists, order histories, trends, preferences, etc.
  - Session data (searches, clicks, pages)
  - Note: data must be persistent (last longer than app)
    - *Data will be very large… cannot fit all in memory*
- What capabilities on the data we need?

An Example: Online Concert Tickets Vendors

- What data do we need?
  - Data about artists, concerts, venues, reservation, hot artists, order histories, trends, preferences, etc.
  - Session data (searches, clicks, pages)
  - Note: data must be persistent (last longer than app)
    - *Data will be very large… cannot fit all in memory*
- What capabilities on the data we need?
  - Add/insert concerts, find concerts by artist/date/etc., analyze past order history, recommended concerts/artists
  - Data must be accessible efficiently, by lots of users
  - Data must be safe from failures, malicious users, and bugs!

Multi-User Issues

Multi-User Issues

- Janin & Pat both have Code Num for gift certificates (credit) of ₩300K they got as birthday gifts
  - Janin using her phone bought a "MAROON 5" ticket concert: ₩180K
  - Pat used her laptop to buy "RHCP" ticket concert: ₩150K

Multi-User Issues

- Janin & Pat both have Code Num for gift certificates (credit) of ₩300K they got as birthday gifts
  - Janin using her phone bought a "MAROON 5" ticket concert: ₩180K
  - Pat used her laptop to buy "RHCP" ticket concert: ₩150K

- Questions:
  - What is credit left?
  - What if another concert, like BTS, costs ₩140?
  - What if the server crashes?
  - What if data center goes offline?

Required Functionality for Db Management

- Describe real-world entities in terms of stored data
- Persistent storage of large datasets
- Efficient query and update
  - Can handle complex queries about data
  - Can handle sophisticated updates
  - Performance matters! (users can feel lag)
- Easily change structure (add attributes/characteristics)
- Allow simultaneous updates
- Recover when system crashes
- Security and integrity

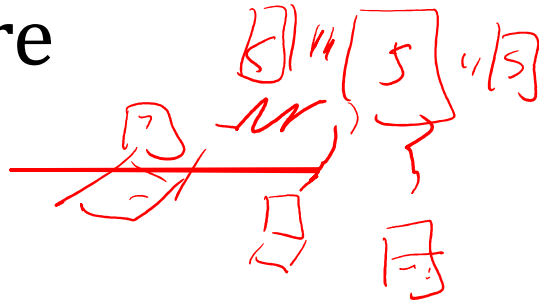
DataBase Management System (DBMS)

- Very difficult to implement all these features inside the application (correctly)
- DBMS provides these features (and more)
- DBMS simplifies application development

Client-Server Architecture

- One server that stores the database (DBMS):
  - Usually robust
  - Can also be a desktop … or a huge cluster running parallel DBMS

Client-Server Architecture



- One server that stores the database (DBMS):
  - Usually robust
  - Can also be a desktop … or a huge cluster running parallel DBMS

- Many clients run apps and connect to DBMS
  - Sybase Adaptive Server Enterprise; psql for PostgreSQL; MySQL server
  - Or some Java/C++ program (typical)

Client-Server Architecture

- One server that stores the database (DBMS):
  - Usually robust
  - Can also be a desktop … or a huge cluster running parallel DBMS
- Many clients run apps and connect to DBMS
  - Sybase Adaptive Server Enterprise; psql for PostgreSQL; MySQL server
  - Or some Java/C++ program (typical)
- Clients "talk" to server using JDBC protocol
  - Usually phone ←~→ web server ←~→ DBMS

* JDBC– Java DB Connectivity; connect java-based clients to different DBs

Key Persons

- DB application developer:
  - Write program that queries and modify data

Key Persons

- DB application developer:
  - Write program that queries and modify data
- DB designer
  - Establishes the schema

Key Persons

- DB application developer:
  - Write program that queries and modify data
- DB designer
  - Establishes the schema
- DB administrator
  - Load data, tune system, maintain whole system running

Key Persons

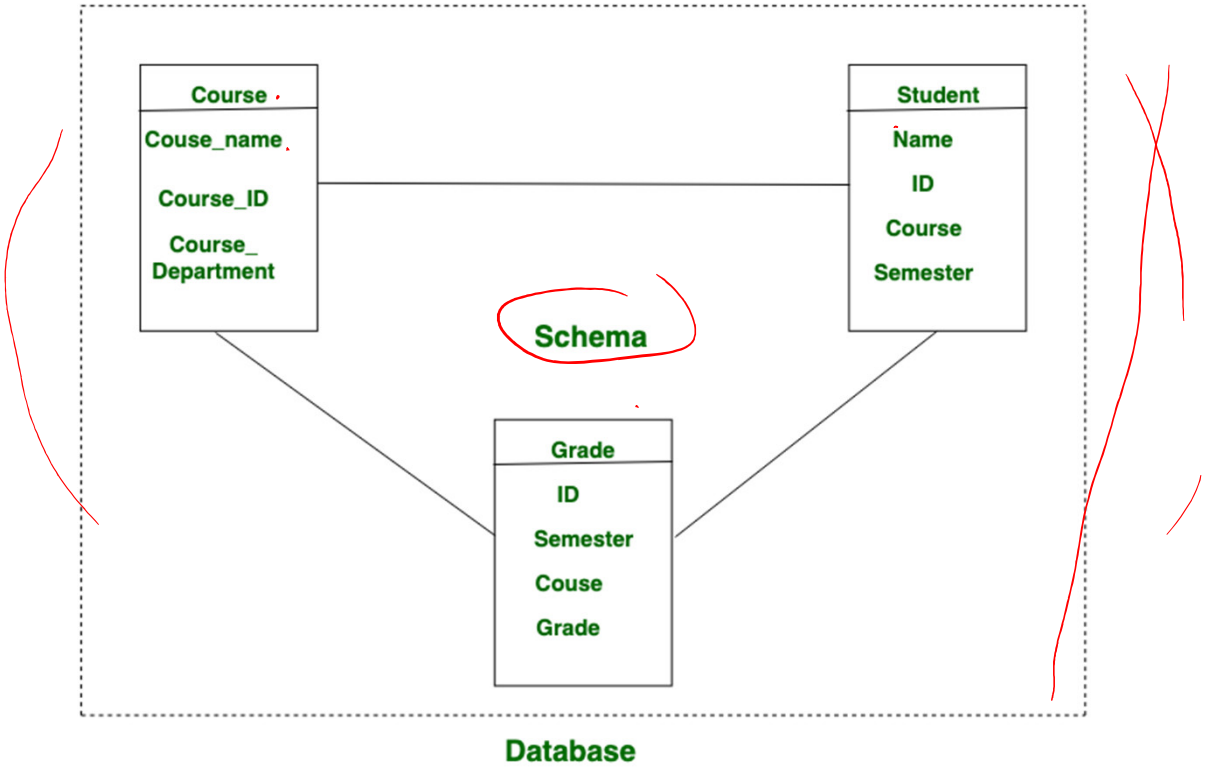
- DB application developer:
  - Write program that queries and modify data
- DB designer
  - Establishes the schema
- DB administrator
  - Load data, tune system, maintain whole system running
- DB analyst
  - data mining, data integration

Key Persons

- DB application developer:
  - Write program that queries and modify data
- DB designer
  - Establishes the schema
- DB administrator
  - Load data, tune system, maintain whole system running
- DB analyst
  - data mining, data integration
- DBMS implementer
  - Builds the DBMS

DB vs Schema

- Schema
  - Blueprint of a database; organization & structure of a database



Key Concepts for the course

- Data models: how to describe real-world data
  - Relational? XML? JSon?

Key Concepts for the course

- Data models: how to describe real-world data
  - Relational? XML? JSon?
- Schema vs data
  - Establishes the schema

Key Concepts for the course

- Data models: how to describe real-world data
  - Relational? XML? JSon?
- Schema vs data
  - Establishes the schema
- Declarative query language
  - Say what we want, not how to get it

Key Concepts for the course

- Data models: how to describe real-world data
  - Relational? XML? JSon?
- Schema vs data
  - Establishes the schema
- Declarative query language
  - Say what we want, not how to get it
- Data independence
  - Physical: can change how data stored in disk w/o affecting apps
  - Logical: can change schema w/o affecting apps

Key Concepts for the course

- Data models: how to describe real-world data
  - Relational? XML? JSon?
- Schema vs data
  - Establishes the schema
- Declarative query language
  - Say what we want, not how to get it
- Data independence
  - Physical: can change how data stored in disk w/o affecting apps
  - Logical: can change schema w/o affecting apps
- Query optimizer & compiler

Key Concepts for the course

- Data models: how to describe real-world data
  - Relational? XML? JSon?
- Schema vs data
  - Establishes the schema
- Declarative query language
  - Say what we want, not how to get it
- Data independence
  - Physical: can change how data stored in disk w/o affecting apps
  - Logical: can change schema w/o affecting apps
- Query optimizer & compiler
- Transactions: isolation & atomicity

Key Concepts for the course

- Data models: how to describe real-world data
  - Relational? XML? JSon?
- Schema vs data
  - Establishes the schema
- Declarative query language
  - Say what we want, not how to get it
- Data independence
  - Physical: can change how data stored in disk w/o affecting apps
  - Logical: can change schema w/o affecting apps
- Query optimizer & compiler
- Transactions: isolation & atomicity

Course contents

- Focus on using DBMS

Course contents

- Focus on using DBMS
- Relational Data Model
  - SQL, Relational Algebra, Datalog

Course contents

- Focus on using DBMS
- Relational Data Model
  - SQL, Relational Algebra, Datalog
- Semistructured Data Model
  - JSon, NoSQL, AsterixDB

Course contents

- Focus on using DBMS
- Relational Data Model
  - SQL, Relational Algebra, Datalog
- Semistructured Data Model
  - JSon, NoSQL, AsterixDB
- Conceptual Design
  - E/R diagrams, Views, and DB normalization

Course contents

- Focus on using DBMS
- Relational Data Model
  - SQL, Relational Algebra, Datalog
- Semistructured Data Model
  - JSon, NoSQL, AsterixDB
- Conceptual Design
  - E/R diagrams, Views, and DB normalization
- Transactions

Course contents

- Focus on using DBMS
- Relational Data Model
  - SQL, Relational Algebra, Datalog
- Semistructured Data Model
  - JSon, NoSQL, AsterixDB
- Conceptual Design
  - E/R diagrams, Views, and DB normalization
- Transactions
- Parallel DBs, MapReduce, and Spark

Assignment

- Setup/Install SQLite on your PC/laptop
- Take screen capture/picture and upload to IELMS ☺

Thank you.