Database and Analytics

Tozammel Hossain
Oct 20, 2020

Data Science & Analytics
University of Missouri

Agenda

- Course Logistics
- Expectations
- Introduction to DBMS
- Module 1 overview

Course Logistics

- 3 credit hours, 8 weeks
- Flipped classroom strategy
- Striking a balance
  - Theory vs Training
- Evaluation
  - Discussion (10%)
  - Exercises (70%)
  - Final project (20%)

Course Logistics

- Canvas
  - Announcements
  - Grades
  - Slides and other docs
- Communication
  - Microsoft teams
  - Emails
    - Rajani (rypbc@mail.missouri.edu)
    - Instructor (hossaink@missouri.edu)

Course Objective

- Structured and unstructured data
- Database schema and ERD models
- Query plans and indexing, write SQL that involves multiple tables, and aggregation
- Nested queries and common table expressions
- Procedures and triggers

Topics to be Covered

- M1: Databases and SQL Introduction
- M2: Database Design
- M3: Advanced SQL-I
- M4: Advanced SQL-II
- M5: Advanced SQL-III
- M6: Data Engineering
- M7: Alternative Data Systems & Final Project Preparation
- Module 8: Final Project DB & Analytics

What Is a DBMS?

- A very large, integrated collection of data
- Models real-world *enterprise*
  - Entities (e.g., students, courses)
  - Relationships
    - Bernard Ebbers is taking ACC101
    - Martha Stewart is taking Ethics in Stock Trading
- A ***Database Management System*** *(DBMS)* is a software package designed to store and manage databases

Why Use a DBMS?

- Data independence and efficient access
- Reduced application development time
- Data integrity and security
- Uniform data administration
- Concurrent access, recovery from crashes

Any reason not to use DBMS?

- Simple application
  - A scripting program will do. Less overhead
- Large-Scaled Applications (NoSQL)
  - Cloud, Big Data, e.g. Google – Big Table, Amazon–Dynamo, Microsoft Azure Cosmos DB
- Specialized application
  - Existing query methods are not adequate
  - e.g. Show me information related to this face, show me protein structures similar to this new one

Why Study Databases?

- Shift from computation to information
  - at the “low end”: scramble to all data sources
  - at the “high end”: real-time decision supports
- Datasets increasing in diversity and volume
  - Digital libraries, interactive video, genome projects, satellite images, social media, etc
- DBMS encompasses most of CS
  - OS, languages, theory, AI, pattern recognition, machine learning, knowledge discovery, data mining, multimedia, logic, etc.

Data Models

- A *data model* is a collection of concepts for describing data
  - Relational data model
- A *schema* is a description of a particular collection of data, using a given data model
  - ***Students*** (*sid*: string, *name*: string, *major*: string, *age*: integer)

DBMS types

- Hierarchical database
  - Tree-like model; inflexible as one parent/many children
- Network Database
  - Network like structure; nodes and edges
- Graph database
  - Improvement over network database
  - A type of NoSQL language
  - Explicit connection
  - Neo4J (Cypher)

DBMS types

- Relational database
  - Most popular; table-oriented
  - Entity as 2D table
  - Implicit relationship
  - Focus of this course
- Object-oriented database
  - Coupling between programming lang + database
    - Both use the same representation

DBMS types

- Document database
  - Main categories of NoSQL database
  - Becoming popular
  - key-value store
  - flexible and scalable on the fly
    - Compare against the relational database

Database vs Data warehouse

- Database provides transactional support
  - OLTP (online transaction processing)
  - Simple queries
- Datawarehouse provides analytical support
  - Builds upon one or more OLTP database
  - Provides a layer optimized for and dedicated to analytics
  - Complex queries

Relational Database

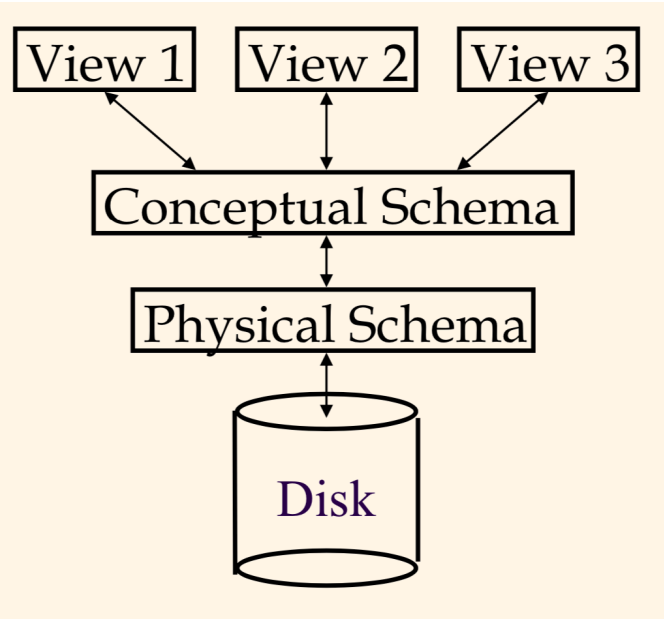
- The *relational model of data* is the most widely used model today
- Main concept: *relation (table)*, basically a table with rows and columns
- Every relation has a *schema (skeleton of a table)*
  - describes the columns, or fields
    - E.g., name, data type of each column or field

Relational Database

SID	*Name*	*Major*	*Age*
000-321	John Smith	CS	20
000-890	Jim Beering	ECE	21
001-788	Seth Norton	CS	38
000-021	Mary Harper	MUII	42

Levels of Abstraction

- Many *views*, single *conceptual (logical) schema* and *physical schema*
  - Views describe how users see the data
  - Conceptual schema defines logical structure
  - Physical schema describes the files and indexes used
- Structured Query Language (SQL)



DDL vs DML

- *Schemas are defined using **DDL (Data Definition Language)***
  - ***Create table …***
- *Data is modified/queried using **DML (Data Manipulation Language)***
  - ***select, update, delete***

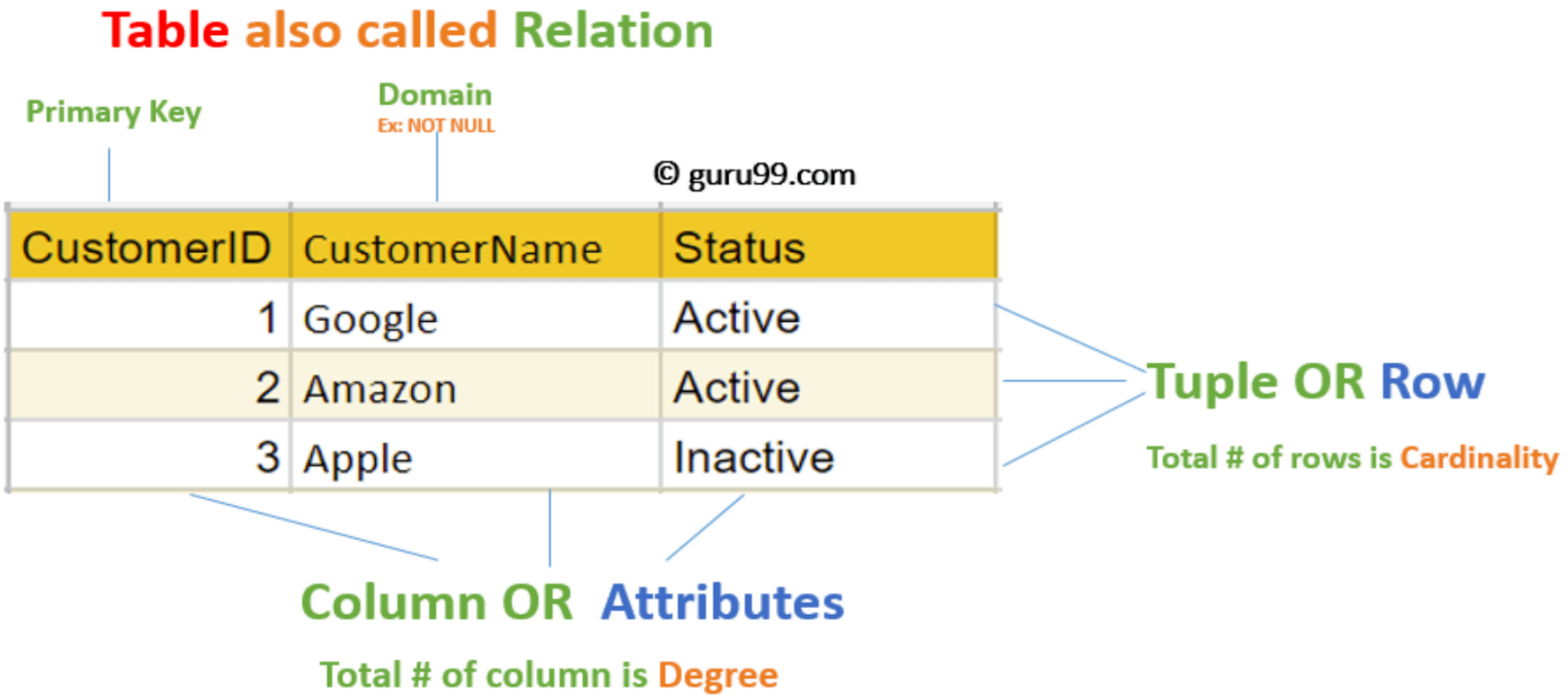
Example: University Database

- Conceptual schema
  - *Students(sid:string, name: string, login: string, age: integer, gpa: real)*
  - *Courses(cid: string, cname: string, credits: integer)*
  - *Enrolled(sid:string, cid:string, grade:string)*
- Physical schema
  - Relations stored as unordered files
  - A B+-tree index on first column of Students

Example: University Database

- External Schema (View)
  - *Course_info(cid:string,enrollment:integer)*

Cheat Sheet of a Table

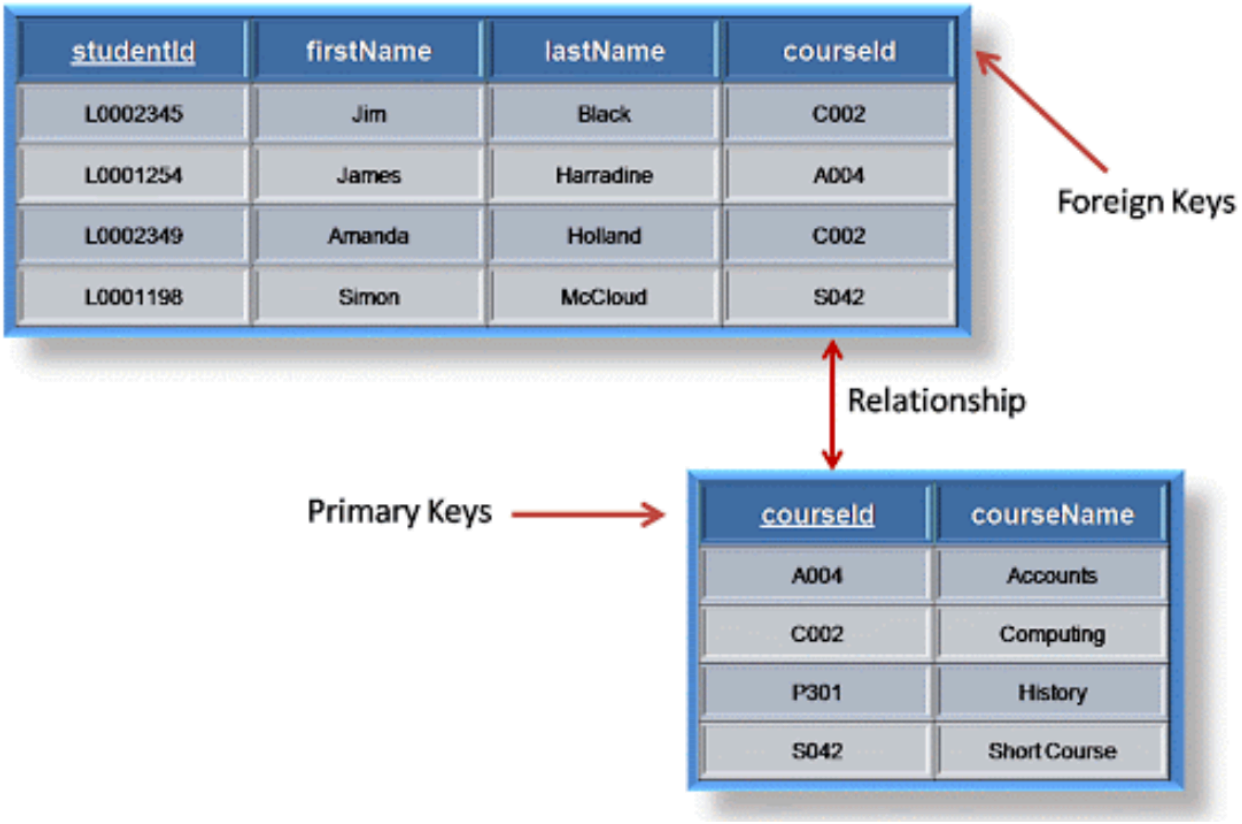


- Relational model is based upon set theory

Quiz

- Why a relational database is called a “relational model”?

Answer


studentId
firstName
lastName
courseId
L0002345
Jim
Black
C002
L0001254
James
Harradine
A004
L0002349
Amanda
Holland
C002
L0001198
Simon
McCloud
S042
Foreign Keys
Relationship
Primary Keys
courseId
courseName
A004
Accounts
C002
Computing
P301
History
S042
Short Course

Summary

- DBMS is used to maintain and query large datasets
- Benefits include recovery from system crashes, concurrent access, quick application development, data integrity and security
- Levels of abstraction give data independence

Summary

- A DBMS typically has a layered architecture
- DBAs hold responsible jobs and are *well-paid*
- DBMS and Big Data Analytics is one of ***the broadest, most exciting areas*** in CS
- Data Science/Machine Learning is getting more and more attention