

WIP - DSC 100 Course Notes

UC San Diego

Kyle Shannon

July 1, 2023

Contents

Lecture 1: Introduction and Overview	2
Lecture 2: The Relational Data Model	3

Lecture 1: Introduction and Overview

In the first lecture, we will cover the basics and the importance of databases in Society, Data Science, AI, and ML. Students will understand the role and significance of databases in various applications, and learn about some of the history of databases. We will cover course expectations and onboarding.

- Course introduction & onboarding
- Importance of Databases in analytical and nonanalytical roles
- Data, Databases, and DBMS
- Brief on Databases in DS, AI, and ML Applications with some examples

Relevant Resource: [TBD](#)

Lecture 2: The Relational Data Model

The second lecture focuses on the key differences between a relational data model and other data models students are likely to have used (e.g. csv, excel, and data frames). Students will learn the basic concepts of data models and the fundamentals of relational databases through an interactive demo. They will also learn about SQLite and Database Management Systems.

- Basic concepts of data models
- Introduction & Terminology to Relational Databases
- Database Design for DS Projects
- SQLite Demo

Relevant Resource: [TBD](#)

Lecture: Introduction to SQL

The third lecture introduces SQL, a vital language for managing and manipulating databases. Students will learn about basic commands, and how SQL is used in practice. There will be several in-class demos.

- Importance of SQL
- Basic SQL commands (SELECT, FROM, WHERE)
- - DDL (CREATE, ALTER, DROP)
- - DML (INSERT, UPDATE, DELETE)
- Writing SQL in analytical uses vs. end user applications (e.g. a bank interface)

Relevant Resource: [TBD](#)

Lecture: SQL Joins I

In this lecture, students will gain understanding of how to use SQL joins and aggregates. The knowledge of these techniques will be crucial in retrieving data for complex analysis tasks.

- SQL JOIN operations Inner, Outer, & Self

Relevant Resource: [TBD](#)

Lecture: SQL Joins II

In this lecture, We will continue to focus on joins, but use them in slightly more complex manners.

- SQL JOIN operations Inner, Outer, & Self
- Using SQL joins for data analysis

Relevant Resource: [TBD](#)

Lecture: Aggregates & Groupings

In this lecture, We will continue to focus on aggregates, & groupings to solve more complex problems.

- SQL aggregate functions SUM, COUNT, AVG, MAX, MIN
- Using SQL joins, aggregates, & groupings for data analysis

Relevant Resource: [TBD](#)

Lecture: Subqueries and Advanced Queries I

This lecture will delve deeper into SQL by explaining subqueries, which are instrumental for complex data manipulations and analysis.

- Introduction to SQL subqueries
- Using subqueries and complex queries for advanced data analysis

Relevant Resource: [TBD](#)

Lecture: Subqueries and Advanced Queries II

This lecture will delve deeper into SQL by explaining subqueries, which are instrumental for complex data manipulations and analysis.

- Introduction to SQL subqueries
- Using subqueries and complex queries for advanced data analysis

Relevant Resource: [TBD](#)

Lecture: Set Operations & NULL

This lecture will delve deeper into SQL by explaining subqueries, which are instrumental for complex data manipulations and analysis.

- Set operations (UNION, INTERSECT, EXCEPT)
- Null values and their treatment in SQL
- Case Study: SQL for Data Cleaning in DS

Relevant Resource: [TBD](#)

Lecture: Relational Algebra

This lecture will delve deeper into SQL by explaining subqueries, which are instrumental for complex data manipulations and analysis.

- Relational Algebra's connection to SQL and DataFrames

Relevant Resource: [TBD](#)

Midterm

Midterm exam in person and will cover material up to Lecture: Subqueries and Advanced Queries

Lecture: SQL for Data Cleaning & Data Analysis

In this lecture, students will learn the importance of data cleaning in data science and how SQL can be used for this purpose.

- Importance of data cleaning in data science
- Using SQL for data cleaning
- Advanced SQL queries for data analysis
- Window functions (ROW_NUMBER, RANK, DENSE_RANK, etc.)
- PIVOT, UNPIVOT

Relevant Resource: [TBD](#)

Lecture: Storage, Indexing, & Size Estimation

This lecture introduces the concept of indexing, which is crucial for improving database query performance.

- Introduction to indexing
- How indexing works (B-tree, hash index)
- How data is stored on disk & accessed
- Size estimation techniques and sharding
- Impact of indexing on data retrieval in data science

Relevant Resource: [TBD](#)

Lecture: Database Design Principles & ER Diagrams I

In this lecture, students will learn the principles of database design and the concept of E-R diagrams, which are crucial for effectively representing data relationships.

- Introduction to database design principles
- Understanding E-R diagrams
- Use of E-R diagrams in database design for data science projects
- Understand schema design

Relevant Resource: [TBD](#)

Lecture: Database Design Principles & ER Diagrams II

In this lecture, students will learn the principles of database design and the concept of E-R diagrams, which are crucial for effectively representing data relationships.

- Introduction to database design principles
- Understanding E-R diagrams
- Use of E-R diagrams in database design for data science projects
- Understand schema design
- Views, Stored Procedures, Triggers
- Transactions

Relevant Resource: [TBD](#)

Lecture: Relational Data Modelling and Introduction to Normalization

Students will learn about relational data modeling and be introduced to normalization, a process to eliminate redundancy and dependency.

- Introduction to relational data modeling
- Introduction to normalization
- Use of normalization in data science

Relevant Resource: [TBD](#)

Lecture: Advanced Normalization

This lecture dives deeper into the concept of normalization, covering advanced normalization techniques. Students will learn how to optimize database designs to ensure data integrity and efficiency.

- Advanced normalization techniques (3NF, BCNF)
- Normalization in database design for data science projects

Relevant Resource: [TBD](#)

Lecture: Database Systems in AI and ML Applications

Students will learn how databases are used in AI and ML applications, with focus on data storage, retrieval and manipulation for model training and prediction.

- Role of databases in AI and ML
- Importance of database design for AI and ML applications
- Case studies of databases in AI and ML

Relevant Resource: [TBD](#)

Lecture: Concurrency & ACID/BASE Properties of SQL & NOSQL

This lecture dives deeper into the concept of normalization, covering advanced normalization techniques. Students will learn how to optimize database designs to ensure data integrity and efficiency.

- ACID vs BASE
- Concurrency
- Types of NoSQL databases (Document, Key-value, Column, Graph)

Relevant Resource: [TBD](#)

Lecture: Security, 'The Cloud', & Warehousing

The lecture discusses the importance of database security and how to implement security measures to protect sensitive data. And students will also learn about cloud-based database systems, which are becoming increasingly popular in the data science industry.

- Importance of database security
- SQL for database security (GRANT, REVOKE)
- Introduction to cloud databases
- Advantages and challenges of cloud databases
- Cloud databases in data science
- Data Ingestion and Preparation in DS

Relevant Resource: [TBD](#)

Final Exam or Project

The final exam will cover materials from all the lectures.