MSDS 420 Section 58
Database Systems
Course Syllabus

Course Description

This course introduces data management and data preparation with a focus on applications in large-scale analytics projects utilizing relational, document, and graph database systems. Students learn about the relational model, the normalization process, and structured query language. They learn about data cleaning and integration, and database programming for extract, transform, and load operations. Students work with unstructured data, indexing and scoring documents for effective and relevant responses to user queries. They learn about graph data models and query processing. Students write programs for data preparation and extraction using various data sources and file formats. Recommended prior programming experience or 430-DL Python for Data Science. Prerequisites: None.

There are four language-focused courses across the MSDS program:

- MSDS 401-DL: Applied Statistics with **R**
- MSDS 420-DL: Database Systems [**SQL**]
- MSDS 430-DL: **Python** for Data Science
- MSDS 431-DL: Data Engineering with **Go**

This course introduces four database systems:

- Relational databases with **PostgreSQL**
- Document databases with **Elasticsearch**
- Graph databases with **Neo4j**
- Object data model with **EdgeDB**

Course Objectives

By the end of this course, you will be able to

- Compare and contrast Relational, Document-Oriented, and Graph database systems
- Analyze and interpret the entity relationship diagram (ERD)
- Apply the normalization process to create normalized relations
- Create and query a Relational database application
- Use Python to collect data from different database systems for exploratory data analysis (EDA)
- Create and query a Document-Oriented database application for information retrieval and indexing
- Develop SQL and NoSQL database applications from real-world datasets
- Use Python and SQL to execute Geospatial queries for spatial data analysis.
- Apply the object-oriented modeling techniques to create the schema for the object data model of the Knowledge graph
- Create and query a Graph database application for the information network (Graph)

Recommended Prior Programming Experience or MSDS 430-DL

MSDS 430-DL Python for Data Science or prior programming experience in one of the following high-level programming languages: Python, Go, R, C, C++, C#, or Java.

Required and Optional Readings and Resources

Required Textbook

DeBarros, Anthony. 2022. *Practical SQL: A Beginner's Guide to Storytelling with Data* (second edition). San Francisco: No Starch Press. [ISBN-13: 978-1-7185-0106-5]
Electronic copy available through Northwestern library, Safari Online:
https://learning.oreilly.com/library/view/practical-sql-2nd/9781098129866/
Code is in the GitHub repository: https://github.com/anthonydb/practical-sql-2/

Reference Books

Beazley, David M. 2022 *Python Distilled*. Boston: Addison-Wesley. [ISBN: 978-0-13-417327-6]

Chen, Daniel Y. 2023. *Pandas for Everyone* (second edition). Boston.: Addison-Wesley. [ISBN-13: 978-0137891153]

Connolly, Thomas M., and Carolyn E. Begg. 2015. *Database Systems: A Practical Approach to Design, Implementation, and Management* (sixth edition.). Upper Saddle River, N.J.: Pearson. [ISBN-13: 978-0132943260] Selected chapters on Course Reserves.

Donovan, Alan A., and Brian W. Kernighan. 2016. *The Go Programming Language.* Boston: Addison Wesley. [ISBN-13: 978-0-13-419044-0]

Gheorghe, Radu, Mathew Lee Hinman, and Roy Russo. 2016. *Elasticsearch in Action*. Shelter Island, NY: Manning. [ISBN-13: 978-1617291623]

Hellmann, Doug. 2017. *The Python 3 Standard Library by Example.* Boston: Addison-Weslen. [ISBN-13: 978-0-13-429105-5]

Kline, Kevin, Regina Obe, and Leo S. Hsu. 2022. *SQL in a Nutshell: A Desktop Quick Reference* (fourth edition). Sebastopol, CA: O'Reilly. [ISBN-13: 978-1492088868]

Kreibich, Jay A. 2010. *Using SQLite*. Sebastopol, CA: O;Reilly. [ISBN-13: 978-0-596-52118-9]

Obe, Regina O. and Leo S. Hsu. 2021. *PostGIS in Action* (third edition). Shelter Island, NY: Manning. [ISBN-13: 978-1617296697]

Robinson, Ian, Jim Webber, and Emil Eifrem. 2015. *Graph Databases: New Opportunities for Connected Data* (second edition). Sebastopol, CA: O'Reilly. [ISBN-13: 978-1-491-93089-2]
Wickham, Hadley, and Garrett Grolemund. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* Sebastopol, CA: O'Reilly. [ISBN-13: 978-1-491-91039-9]

Zhao, Alice. 2021. *SQL Pocket Guide: A Guide to SQL Usage* (fourth edition). Sebastopol, CA: O'Reilly. [ISBN-13: 978-1492090403]

Electronic copies of O'Reilly and Manning books are available from Safari Online.

Online Developer Resources, Documentation, and Tutorials:

PostgreSQL: https://www.postgresql.org/docs/
Elasticsearch: https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html
Neo4j: https://neo4j.com/docs/
EdgeDB: https://www.edgedb.com/docs/intro/quickstart

Course Reserves

Some readings will be available through the Course Reserves in the left navigation menu. The Syllabus and weekly roadmap will note which readings are to be accessed through Course Reserves. For assistance with Course Reserves, e-mail e-reserve@northwestern.edu. To ask a librarian for assistance, visit Northwestern's Ask A Librarian page.

Optional Readings and Resources

Optional readings will be listed on the syllabus in the Course Schedule.

Required and Optional Software

The primary database system used in this course is PostgreSQL, which is hosted on the Data Science Computing Cluster (DSCC). The primary computing language is Python. Faculty in some sections of the course may recommend additional database systems and programming languages.

Students will need to:

- Use the Anaconda/Python 3 tool downloaded from here
- Install PostgreSQL locally on their personal computers

- Utilize a virtual private network (VPN) connection to Northwestern's Data Science Computing Cluster (DSCC)
-

Sync Sessions

Students are expected to attend sync sessions or to watch the recordings. Students are responsible for material discussed in sync sessions.

Assignment Overview and Grading Breakdown

Grading and feedback turnaround will be one week from the due date. You will be notified if turnaround will be longer than one week. The discussion forums, assignments, and the final exam will be graded based on the specific criteria listed on the rubrics, which are available in the course. Grading components are shown in the following list:

- Discussion forums. 10 weeks, 10 points per week, 100 points total
- Assignments: 6 100-point assignments, 600 points total
- Term project (written research report): 100 points
- Final exam: 200 points

Grading Scale

Grade	Percentage	Total Points (out of 1000)
A	93%–100%	930– 1000 points
A-	90%–92%	900 – 929 points
B+	87%–89%	870 – 899 points
B	83%–86%	830 – 869 points
B-	80%–82%	800 – 829 points
C+	77%–79%	770 – 799 points
C	73%–76%	730 – 769 points
C-	70%–72%	700 – 729 points
F	0%–69%	0 – 699 points

Assignments utilize databases installed locally on the students' computers and on the remote Data Science Computing Cluster (DSCC), which is a Red Hat Linux environment.

Late Work Policy

Unless otherwise noted, every assigned task is due by the end of the week, Sunday by 11:55 pm CST (central time). This includes assignments, participation in discussions, and the final exam (due at the end of the term). Late is accepted with the instructor's permission only. Try not to fall behind in this course. We cover a lot of material. Falling behind is the primary reason students have difficulty with this course. Contact your instructor if you begin to fall behind or encounter an unanticipated event that may interfere with your coursework.

Term Paper

The term paper takes the structure of a formal research report. Research reports should answer questions as follows:

- Abstract. What is this research about and what did you learn? (Executive summary)
- Introduction. **Why** did you engage in this research?
- Literature Review. **Who** else has conducted research like this?
- Methods. **How** did you conduct the research?
- Results. **What** did you learn from the research?
- Conclusions. **So, what** will the research mean to management?

There is an audio overview regarding the structure of research papers. <https://northwestern.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=15434363-7b7e-43c7-9ab2-ac0100f4fd4d>

We use chapter 15 of the seventeenth edition of The Chicago Manual of Style (2019) as the standard for author/date citations and references. This manual also covers English grammar and punctuation. It is convenient to use Microsoft Word with Grammarly as a check on spelling, punctuation, and grammar.

The Writing Place is Northwestern's center for peer writing consultations. Consultations are free and available to anyone in the Northwestern community: undergraduates, graduate students, faculty, or staff. To book an appointment, go to The Writing Place website. <add the link>

In his essay "Politics and the English Language," George Orwell (1946) provided six rules of clear writing:

- Never use a metaphor, simile, or other figure of speech which you are used to seeing in print.
- Never use a long word where a short one will do.
- If it is possible to cut a word out, always cut it out.
- Never use the passive where you can use the active.
- Never use a foreign phrase, a scientific word, or a jargon word if you can think of an everyday English equivalent.
- Break any of these rules sooner than say anything outright barbarous.

Zinzer (2012) provides exceptional advice about nonfiction writing.

Optional Canvas course modules provide information relevant to possible term paper topics, including the following:

- Data Preparation and Feature Engineering
- Data Warehousing and Business Intelligence
- SQLite
- JavaScript Object Notation (JSON)
- CockroachDB Distributed Database
- Database System Performance Benchmarks
- Backend as a Service (BaaS) Systems
- Object-Relational Mappers
- Using Go with Databases
- Application Programming Interfaces (APIs)

References on Writing

Merriam-Webster's Collegiate Dictionary (eleventh ed.), 2008. Springfield, Mass.: Merriam-Webster.

Orwell, George., 1946, April. "Politics and the English Language." *Horizon.* Available online at http://www.orwell.ru/library/essays/politics/english/e_polit

Rodale, J. I, Laurence Urdang, and Nancy LaRoche, 1978. *The Synonym Finder.* Emmaus, Pa.: Rodale Press.

The Chicago Manual of Style (seventeenth ed.), 2017. Chicago: University of Chicago Press. Online information available at http://www.chicagomanualofstyle.org/home.html

Zinsser, William. 2012. *On Writing Well: An Informal Guide to Writing Nonfiction* (Thirtieth Anniversary Edition). New York: Harper Perennial.

Online Communication and Interaction Expectations

Discussion Forums

Discussion board participation is an essential and important part of this class and is designed to allow free exchange of ideas in a respectful and open environment. How often you post is less important than the contents of your contribution, although a minimum level of engagement is expected. You are encouraged to post actively and frequently, but please try not to clutter the board with irrelevant or insignificant material, which could work against you. Stay on topic, keep your language professional (abbreviated texting language is not appropriate), and try always to offer something new when you post (a "me too" type post doesn't count). When relevant, please remember to cite all sources, and avoid plagiarism.

You need to post at least one response to each discussion thread each week. Make at least one of those posts an original contribution to the discussion that includes references and citations, while being careful not to plagiarize or violate copyright. Also demonstrate engagement in the discussion forum by responding to others' posts. It may be helpful to read this guide to netiquette.

Participation and Attendance

This course will not meet at a particular time each week. All course goals, session learning objectives, and assessments are supported through classroom elements that can be accessed at any time. To measure class participation (or attendance), your participation in threaded discussion boards is required, graded, and paramount to your success in this course.

Student Support Services

AccessibleNU

This course is designed to be welcoming to, accessible to, and usable by everyone, including students who are English-language learners, have a variety of learning styles, have disabilities, or are new to online learning. Be sure to let me know immediately if you encounter a required element or resource in the course that is not accessible to you. Also, let me know of changes I can make to the course so that it is more welcoming to, accessible to, or usable by students who take this course in the future. Northwestern University and AccessibleNU are committed to providing a supportive and challenging environment for all undergraduate, graduate, professional school, and professional studies students with disabilities who attend the University. Additionally, the University and AccessibleNU work to provide students with disabilities and other conditions requiring accommodation a learning and community environment that affords them full participation, equal access, and reasonable accommodation. The majority of accommodations, services, and auxiliary aids provided to eligible students are coordinated by AccessibleNU, which is part of the Dean of Students Office. Please make sure you email your instructor if you have special needs the very first week of the quarter such that your instructor will ensure and plan for the accommodations of your special needs.

SPS Student Services

The Department of Student Services supports the academic and professional growth of SPS students. The Student Services team guides students through academic planning, policies, and administrative procedures, and promotes a supportive environment to foster student success. Students are encouraged to actively make use of the resources and staff available to assist them: Academic and Career Advisers, Counseling and Health Services, Student Affairs, Legal Services, Financial Aid and Student Accounts, among other services.

For a comprehensive overview of course and program processes and policies and helpful student resources, please refer to your SPS Student Handbook.

Academic Support Services

Northwestern University Library

As one of the leading private research libraries in the United States, Northwestern University

Library serves the educational and information needs of its students and faculty as well as scholars around the world. Visit the Library About page for more information or contact Distance Learning Librarian Tracy Coyne at 312-503-6617 or tracy-coyne@northwestern.edu.

Program-Specific Library Guides

- Data Science

Additional Library Resources

- Connectivity: Campus Wireless and Off-Campus Access to Electronic Resources
- Reserve a Library Study Room
- Sign up for an in-person or online Research Consultation Appointment
- Getting Available Items: Delivery to Long-Distance Patrons
- Social Science Data Resources
- Resources for Data Analysis

Learning Studios

Learning Studios are self-paced, self-directed, and individualized online tutorials to support SPS students and assist in student success. These Studios are optional, non-credit, and zero-tuition courses housed in Canvas, with no registration requirements in Caesar. Enrollment in such Studios will not be reflected on the student's transcript. While other students will be completing the studio, there are no required discussions or group activities. However, there will be an optional, web conference conducted weekly by an instructor for any students who have questions about the material. SPS is currently offering five Learning Studios for the current term: Programming in R, Programming in Python, Programming in Go, Microeconomics, and Academic Integrity Learning. Programming in Python and Programming in Go are especially relevant to this course. Students can self-enroll free by visiting the SPS Academic Services page.

The Writing Place

The Writing Place is Northwestern's center for peer writing consultations. Consultations are free and available to anyone in the Northwestern community: undergraduates, graduate students, faculty, or staff. To book an appointment, go to The Writing Place website.

The Math Place

The Math Place is a free tutorial service provided to students currently enrolled in Northwestern University's School of Professional Studies courses or in other Northwestern University courses. Students of all levels can benefit from the individual tutoring provided from this service, whether they are taking undergraduate or graduate level courses. To book an appointment, go to the Scheduling Location and select an available opening. We ask that students schedule up to one appointment per week. Appointments are currently offered over Zoom. For questions concerning appointments or additional information regarding tutoring services, please email spsmathplace@u.northwestern.edu.

Academic Integrity at Northwestern

Students are required to comply with University regulations regarding academic integrity. If you are in doubt about what constitutes academic dishonesty, speak with your instructor or graduate coordinator before the assignment is due and/or examine the University Web site. Academic dishonesty includes, but is not limited to, cheating on an exam, obtaining an unfair advantage, and plagiarism (e.g., using material from readings without citing or copying another student's paper). Failure to maintain academic integrity will result in a grade sanction, possibly as severe as failing and being required to retake the course, and could lead to a suspension or expulsion from the program. Further penalties may apply. For more information, visit The Office of the Provost's Academic Integrity page. Some assignments in SPS courses may be required to be submitted through Turnitin, a plagiarism detection and education tool. You can find an explanation of the tool here.

Course Technology

This course will involve a number of different types of interactions. These interactions will take place primarily through the Canvas system. Please take the time to navigate through the course and become familiar with the course syllabus, structure, and content and review the list of resources below.

Network Connection to DSCC

The DSCC serves as a research and training facility for graduate students in the Master of Science in Data Science program. User accounts are not associated with individual courses or instructors. They are for student and faculty use only and remain available as long as users maintain valid Northwestern NetIDs. Users should not share user account NetIDs and passwords with others. Each account is tied with its Northwestern University network identity For DSCC technical support related to server connection/login issues, email or call :

sps-it@northwestern.edu and phone number 312-503-3333

Using VPN to connect Northwestern Network

If you are a remote user (not on Northwestern University campus/network), you need to connect to the university network through VPN in order to connect to DSCC. Please follow the instructions from this link for NU VPN.

Canvas

The Canvas Student Center includes information on communicating in Canvas, navigating a Canvas course, grades, additional help, and more. The Canvas at Northwestern website provides information of getting to know Canvas at Northwestern and getting Canvas support.

The Canvas Student Guide provides tutorials on all the features of Canvas. For additional Canvas help and support, you can always click the Help icon in the lower left corner to begin a live chat with Canvas support or contact the Canvas Support Hotline. The Canvas Accessibility Statement and Canvas Privacy Policy are also available.

Zoom

We will use Zoom for synchronous meetings. The Zoom support page provides additional guidance for using Zoom, and the Zoom for Students in Canvas page has guidance specifically for students The Zoom Privacy Policy and the Accessibility Features on Zoom are also available. These synchronous sessions will be recorded, so you will be able to review the session afterward. Students are expected to attend the sync sessions or to watch the recordings. Students are responsible for material discussed in the sync sessions.

Panopto

Videos in this course may be hosted in Panopto. If you have not used Panopto in the past, you may be prompted to login to Panopto for the first time and authorize Panopto to access your Canvas account. You can learn more about using Panopto and login to Panopto directly by visiting the Panopto guide on the Northwestern IT Resource Hub. Depending on the assignment requirements of this course, you may be asked to create videos using Panopto in addition to viewing content that your instructor has provided through Panopto.
Watch this Tutorial on how to create Video using Panopto and here is another Tutorial on how to share a video The Panopto Privacy Policy and the Accessibility Features on Panopto are also available.

Minimum Required Technical Skills

Students in an online program should be able to do the following:

- Communicate via email and Canvas discussion forums.
- Use web browsers and navigate the World Wide Web.
- Use the learning management system Canvas.
- Use integrated Canvas tools (e.g., Zoom, Panopto, Course Reserves).
- Use applications to create documents and presentations (e.g., Microsoft Word, PowerPoint).
- Use applications to share files (e.g., Box, Google Drive).

Systems Requirements for Distance Learning

Students and faculty enrolled in SPS online master's degree programs should have access to a computer with the Minimum System Requirements.

Technical Help and Support

The SPS Help Desk is available for Faculty, Students and Staff to support their daily IT needs.

For additional technical support, contact the Northwestern IT Support Center.

Course Week-by-Week Schedule

Module	Topic	Activities
1	Introduction to Database Systems and PostgreSQL	▪ Introduce Yourself ▪ Complete Required Readings/Media ▪ Complete Class Discussions
2	Database Design, Entity-Relationship Diagrams, Normalization, and SQL	▪ Complete Required Readings/Media ▪ Complete Class Discussions ▪ Complete Assignment 1
3	Languages and Tools for Working with SQL	▪ Complete Required Readings/Media ▪ Complete Class Discussions ▪ Complete Assignment 2
4	SQL: Data Definition, Manipulation, and Aggregation;	▪ Complete Required Readings/Media ▪ Complete Class Discussions ▪ Complete Assignment 3 ▪ Term Project Checkpoint 1
5	SQL: Geographical Information Systems and PostGIS	▪ Complete Required Readings/Media ▪ Complete Class Discussions ▪ Complete Assignment 4
6	Document Databases and Indexing	▪ Complete Required Readings/Media ▪ Complete Class Discussions
7	Document Query Relevance and Information Retrieval	▪ Complete Required Readings/Media ▪ Complete Class Discussions ▪ Complete Assignment 5 ▪ Term Project Checkpoint 2
8	Graph Databases and Query Languages	▪ Complete Required Readings/Media ▪ Complete Class Discussions ▪ Register for Self-Proctored Final Exam using Panopto
9	Object Data Modeling for Graph Databases	▪ Complete Required Readings/Media ▪ Complete Class Discussions ▪ Complete Assignment 6 ▪ Review Practice Document for Final Exam
10	Database Systems Review	▪ Complete Required Readings/Media ▪ Complete Class Discussions ▪ Complete Self-Proctored Final Exam using Panopto ▪ Complete Term Project

Course Schedule

Module 1. Introduction to Database Systems and PostgreSQL

Learning Objectives

After this week the student will be able to

- Differentiate between structured, unstructured, and semistructured data.
- Compare and contrast Relation (SQL), Document-Oriented (NoSQL), and Graph based (Cypher) databases.
- Reflect on the skills and competencies they would like to obtain from taking this course.
- Discuss a professional experience with a database engine and the types of processing you have undertaken.
- Explain current trends for database management systems and database technologies.
- Explain how modern databases evolved from file processing systems.

In addition, you will need to be able to do the following:

- Compare and contrast file-processing and database management systems.
- Define the main functions of a database management system (DBMS).
- Illustrate the process to extract data to generate reports that support the business decision making process.
- Explain the concepts of data warehouses, business intelligence, and big data analytics.

Required Readings

DeBarros, Anthony. 2022. *Practical SQL: A Beginner's Guide to Storytelling with Data* (second edition). San Francisco: No Starch Press. [ISBN-13: 978-1-7185-0106-5]
Electronic copy available through Northwestern library, Safari Online:
https://learning.oreilly.com/library/view/practical-sql-2nd/9781098129866/
Code is in the GitHub repository: https://github.com/anthonydb/practical-sql-2/
Chapters 1–3 (pages 1–40) and passim.

Assignments

Discussion board participation.

Sync Session

Tuesday, January 3, 7 pm Central Time. Recorded session, Students are expected to attend the sync session or to watch the recording. Students are responsible for material discussed in the sync session.

Module 2. Database Design, Entity-Relationship Diagrams, Normalization, and SQL

Learning Objectives

After this week the student will be able to:

- Describe the importance of the Entity Relationship Diagram (ERD) to the design of relational database application.
- Explain the differences between the Crows Foot, Chen, and UML notations.
- Evaluate the usefulness of the\ Crows Foot, Chen, and UML notations.
- Use the conceptual data model and its core data-modeling blocks: attributes, entities, relationships, and cardinalities in order to generate and explain the business rules for the given data model.
- Create a database using SQL language.

In addition, you will need to be able to do the following:

- Illustrate the Database Life Cycle (DBLC) process.
- Define database design terminologies, including entities, fields, records, files, tables, candidate keys, and primary keys.
- Explain the difference between conceptual design, logical design, and physical design.
- Create entity relationship diagrams (ERD) using notations of Crows Foot, Chen, and UML notations.

Required Readings

DeBarros, Anthony. 2022. *Practical SQL: A Beginner's Guide to Storytelling with Data* (second edition). San Francisco: No Starch Press. [ISBN-13: 978-1-7185-0106-5] Electronic copy available through Northwestern library, Safari Online: https://learning.oreilly.com/library/view/practical-sql-2nd/9781098129866/
Code is in the GitHub repository: https://github.com/anthonydb/practical-sql-2/
Chapters 4–7 (pages 41–115) and passim.

Assignments

- Discussion board participation
- Assignment 1

Sync Session

Tuesday, January 10, 7 pm Central Time. Recorded session, Students are expected to attend the sync session or to watch the recording. Students are responsible for material discussed in the sync session.

Module 3. Languages and Tools for Working with SQL

Learning Objectives

After this week the student will be able to:

- Explain the rationale behind the normalization process.
- Define the normal forms 1NF, 2NF, and 3NF.
- Discuss the basic concepts of functional dependencies: partial and transitive.
- Define the relational model's core element: relations.
- Discuss cases when the relationship between entities is represented by the relation.
- Discuss how the conceptual data model main components: Entities, Attributes, Relationships in ERD are mapped into relations that are composed of the logical constructs: rows (tuples) and columns (attributes).

In addition, you will need to be able to do the following:

- Explain how the relational database model offers a logical view of data.
- Discuss how relations are implemented as tables in a relational DBMS.
- Explain how data redundancy is handled in the relational database model.
- Discuss the basic concepts of data anomalies and data redundancy.
- Transform E-R diagrams to relations.
- Describe how normal forms can be transformed from lower normal forms to higher normal forms.
- Use normalization to decompose anomalous relations to well-structured relations.

Required Readings

DeBarros, Anthony. 2022. *Practical SQL: A Beginner's Guide to Storytelling with Data* (second edition). San Francisco: No Starch Press. [ISBN-13: 978-1-7185-0106-5]
Electronic copy available through Northwestern library, Safari Online:
https://learning.oreilly.com/library/view/practical-sql-2nd/9781098129866/
Code is in the GitHub repository: https://github.com/anthonydb/practical-sql-2/
Chapters 8–10 (pages 116–181) and passim.

Assignments

- Discussion board participation
- Assignment 2

No Sync Session

Module 4. SQL: Data Definition, Manipulation, and Aggregation

Learning Objectives

After this week the student will be able to:

- Describe the strengths and weaknesses of using SQL for data collection.
- Explain why general purpose programming languages are used along with SQL for certain aspects of data collection and preparation.
- Perform data analysis tasks on data read from a CSV file and loaded into a DataFrame object.
- Use SQL in the relational database implementation and utilize different SQL statements to access and query the constructed database.

In addition, you will need to be able to do the following:

- Discuss the role of SQL in the implementation of relational database applications
- Define the categories of SQL statements: Data definition language (DDL) and Data manipulation language (DML).
- Describe how to enforce referential integrity using SQL.
- Discuss the basic commands and functions of SQL.
- Choose SQL - DDL to create tables and indexes.
- Choose SQL - DML for to add, modify, delete, and retrieve data.
- Explain how to use SQL to query a database to extract data.
- Discuss data indexing, selection, and filtering.

Required Readings

DeBarros, Anthony. 2022. *Practical SQL: A Beginner's Guide to Storytelling with Data* (second edition). San Francisco: No Starch Press. [ISBN-13: 978-1-7185-0106-5]
Electronic copy available through Northwestern library, Safari Online:
https://learning.oreilly.com/library/view/practical-sql-2nd/9781098129866/
Code is in the GitHub repository: https://github.com/anthonydb/practical-sql-2/
Chapters 11–14 (pages 183–273) and passim.

Assignments

- Discussion board participation
- Term Paper Checkpoint 1
- Assignment 3

Sync Session

Tuesday, January 24, 7 pm Central Time. Recorded session, Students are expected to attend the sync session or to watch the recording. Students are responsible for material discussed in the sync session.

Module 5. SQL: Geographical Information Systems and PostGIS

Learning Objectives

After this week the student will be able to

- Compare the group by functions in Python/Dataframe and SQL.
- Evaluate the scenarios that are appropriate for using SQL vs. Python tools for aggregating data.
- Use PostGIS/PosgreSQL to execute location-based SQL queries.
- Execute geospatial queries to provide and plot descriptive statistics on Choropleth map.

In addition, you will need to be able to do the following:

- Explain the different types of subqueries and correlated queries.
- Discuss advanced SQL JOIN operator syntax.
- Illustrate how to use SQL functions to manipulate dates, strings, and other data.
- Discuss relational set operators UNION, UNION ALL, INTERSECT, MINUS.
- Create and aggregate data using SQL group by and having operators.
- Discuss pivoting and hierarchical indexing of data.
- Discuss PostGIS/PosgreSQL spatial operators, spatial functions, spatial data types, and spatial indexing that are used to implement Geospatial queries.

Required Readings

DeBarros, Anthony. 2022. *Practical SQL: A Beginner's Guide to Storytelling with Data* (second edition). San Francisco: No Starch Press. [ISBN-13: 978-1-7185-0106-5]
Chapters 15–16 (pages 275–335) and passim.

Obe, Regina O. and Leo S. Hsu. 2021. *PostGIS in Action* (third edition). Shelter Island, NY: Manning. [ISBN-13: 978-1617296697]
Chapters 1–2 (pages 1–65) and passim.

Assignments

- Discussion board participation
- Assignment 4

Sync Session

Tuesday, January 31, 7 pm Central Time. Recorded session, Students are expected to attend the sync session or to watch the recording. Students are responsible for material discussed in the sync session.

Module 6. Document Databases and Indexing

Learning Objectives

After this week the student will be able to:

- Explain the role of “precision” and “recall” in the assessment of information retrieval systems.
- Explain the structure of inverted indexes.
- Construct hierarchical indexes.
- Select and group data to create pivot tables.
- Interact with a NoSQL (document-oriented) database engine, ElasticSearch.
- Experient with different NoSQL queries and evaluate the output to fine-tune results for better precision/accuracy/relevance.
- Create and run NoSQL queries required for this assignment requirements.
- Use NoSQL to retrieve and plot geospatial data on HeatMaps
-

In addition, you will need to be able to do the following:

- Define the information retrieval process.
- Explain the concept of finding relevant documents.
- Discuss the architecture of search engines for information retrieval.
- List the Information retrieval tasks to store unstructured data collection.
- List the tasks to retrieve the relevant documents from a document collection.
- Discuss data structures used for Indexing.
- Examine basic information retrieval evaluation metrics: recall and precision.
- Illustrate how to parse text data with regular expressions in general programming language.

Required Readings

The following chapters are available from Safari Books online for free through Northwestern Library: https://www.library.northwestern.edu/ . Use your netid and email handle to access these videos.

Gheorghe, Radu, Mathew Lee Hinman, and Roy Russo. 2016. *Elasticsearch in Action*.
Shelter Island, NY: Manning. [ISBN-13: 978-1617291623]
Chapters 1–2 (pages 1–52) and passim.

Assignments

- Discussion board participation

No Sync Session

Module 7. Document Query Relevance and Information Retrieval

Learning Objectives

After this week the student will be able to:

- Evaluate the type of similarity model that is most appropriate for searches of data from the Internet.
- Differentiate between the various types of similarity models.
- Reflect on the skills and competencies learned in the course.
- Interact with a NoSQL (document-oriented) database engine, ElasticSearch.
- Experiment with different NoSQL queries and evaluate the output to fine-tune results for better precision/accuracy/relevance.

In addition, you will need to be able to do the following:

- Discuss commonly used models for information retrieval.
- Describe similarity and matching strategies.
- Explain how to rank-order the documents matching a query.
- Discuss natural language processing and text analytics for document collection.
- Compare and contrast Boolean, vector, and probabilistic similarities.

Required Readings

The following chapters are available from Safari Books online for free through Northwestern Library: https://www.library.northwestern.edu/ . Use your netid and email handle to access these videos.

Gheorghe, Radu, Mathew Lee Hinman, and Roy Russo. 2016. *Elasticsearch in Action*. Shelter Island, NY: Manning. [ISBN-13: 978-1617291623]
Chapters 3–4 (pages 53–117) and passim.

Assignments

- Discussion board participation
- Term Project Checkpoint 2
- Assignment 5

Sync Session

Tuesday, February 14, 7 pm Central Time. Recorded session, Students are expected to attend the sync session or to watch the recording. Students are responsible for material discussed in the sync session.

Module 8. Graph Databases and Query Languages

Learning Objectives

After this week the student will be able to:

- Explain the rationale and applications of graph databases.
- Compare and contrast graph database to relational database and document-oriented database.
- Use a graph query language
- Create and query a graph database application

Required Readings

The following chapters are available from Safari Books online for free through Northwestern Library: https://www.library.northwestern.edu/ . Use your netid and email handle to access these videos.

Robinson, I., Webber, J. and Eifrem, E. (2015) Graph Databases, (2nd Edition). O'Reilly Media, Inc. [ISBN: 9781491930892]
Free Access to Textbook from Safari Books online/Northwestern Library:
https://learning.oreilly.com/library/view/graph-databases-2nd/9781491930885
Chapters 1–3 (pages 1–64) and passim.

Optional Readings

- The Neo4j Cypher Manual v4.1
  https://neo4j.com/docs/cypher-manual/current/

Media

- Chapter 3. Graph Databases, Neo4j
  https://learning.oreilly.com/videos/learning-neo4j-graphs/9781787287358/9781787287358-video3_1

Assignments

- Discussion board participation
- Register for Self-Proctored Final Exam

Sync Session

Tuesday, February 28, 7 pm Central Time. Recorded session, Students are expected to attend the sync session or to watch the recording. Students are responsible for material discussed in the sync session.

Module 9. Object Data Modeling for Graph Databases

Learning Objectives

After this week the student will be able to:

- Identify alternative graph query languages, including Cypher, EdgeQL, and GraphQL.
- Describe what is meant by a graph-relational model and object-oriented databases (drawing on the example of EdgeDB.
- Create EdgeDB graph schema
- Use EdgeDB driver (blocking IO and asyncio) for Python to create and query graph-relational database applications
- Use GraphQL of the EdgeDB built-in graphql extension to create and query graph-relational database applications
- Compare and contrast GraphQL and REST API
- Work with an EdgeDB-based application.

Online Readings and Tutorials

Work with the Get Started and Quickstart materials on the EdgeDB site: https://www.edgedb.com/docs

Assignments

- Discussion board participation
- Assignment 6
- (Optional) Practice Final Exam Available

Sync Session

Tuesday, March 1, 7 pm Central Time. Recorded session, Students are expected to attend the sync session or to watch the recording. Students are responsible for material discussed in the sync session.

Module 10. Database Systems Review

Learning Objectives

After this week the student will be able to:

- Distinguish among primary types of database systems: relational, document, graph, and object-relational systems.
- List major database systems in use today and identify these systems by type.
- Distinguish among database query languages, including SQL, Cypher, and GraphQL.
- Contrast alternative methods for working with unstructured and semi-structured text across database systems, both relational and document-oriented systems.

No additional readings this week.

Assignments

- Discussion board participation
- Term Paper
- Final Exam

No Sync Session

Module A (Optional). Data Preparation with Python Pandas

Learning Objectives

After this week the student will be able to:

- Evaluate the various methods of cleaning and transforming data with Python Pandas DataFrame and explain why your chosen ones are most effective.
- Identify bad data problems.
- Convey how to screen data for potential problems, identifying outliers and miscoded data.
- Analyze the dataset in the given CSV file.
- Clean the given dataset.
- Load the dataset into sqlite database engine.
- Execute different SQL queries.

In addition, you will need to be able to do the following:

- Clean and update data items using a general programming language.
- Filter and detect missing and duplicate values.
- Address problems of missing data in surveys and databases

Note optional practice exercises are available to review data preparation with Python Pandas.

Module B (Optional). Document Warehousing and Business Intelligence

Learning Objectives

After this week the student will be able to:

- Explain how online transaction processing (OLTP) systems differ from online analytical processing (OLAP).
- Use SQL to answer ad-hoc queries for a data warehouse represented by a star-schema.

In addition, you will need to be able to do the following:

- Discuss the main concepts and benefits associated with data warehousing.
- Define the architecture and main components of a data warehouse.
- Discuss the concept of a data mart and the main reasons for implementing a data mart.
- Explain how business intelligence provides a comprehensive business decision support framework.
- Discuss the relationship and differences between operational data and decision support data.
- Explain how to prepare data for a data warehouse.
- Define what a star schema is and how it is constructed.

Note. Your instructor may review data warehouse and business intelligence concepts and systems as part of selected sync sessions and discussion boards.

Additional Modules

Module C (Optional). Backend as a Service (Baas)

Module D (Optional). The Go Programming Language