BAX-421 – Data Management

TERM: Fall 2018

INSTRUCTOR: Mehul Rangwala 916.399.3271 mrangwala@ucdavis.edu

OFFICE HOURS:
Please feel free to contact the instructor any time if you have questions. I will stay back and be available after each class meeting as long as the students need. I encourage you to use this time to ask questions or review any part of the material that you are having difficulty with. If the time after class is not convenient, then a separate appointment can be arranged for a meeting either in-person or over Zoom. You can also ask me questions anytime without appointment via email, text, or phone. In short, my office hours are almost when you need me. It is critical that you clearly understand the concepts covered in the course. Getting your questions answered and helping you understand the material, exercises, and homework are my topmost priorities.

COURSE DESCRIPTION:
We live in an era of data deluge. There has been a dramatic increase in the amount of data, both structured and unstructured, obtained from plethora of sources including websites and devices. Steady decrease in the cost of storage will make this proliferation even more profound. "Big data" is not only about volume of data but also its velocity and variety. The success of an organization largely depends on how reliable and well managed its data is. This data not only fuels the business processes but also provides the organization the intelligence needed to support strategic decisions, measure success, and maintain competitive advantage. According to Data Management Body of Knowledge (DMBOK), *Data Management* is the development, execution, and supervision of plans, policies, programs, and practices that deliver, control, protect, and enhance the value of data and information assets, throughout their life cycle. This course covers two segments:

1. Understanding of data governance and other foundational activities such as data quality, data privacy, data security, and data quality; and
2. Extracting this data to provide the intelligence.

The topics in the second segment include a brief introduction to database design fundamentals to assist in effective assembly, storage, and organization of data and then using SQL and NoSQL to extract this data to drive business intelligence.

COURSE OBJECTIVES:

1. Understand the fundamentals of database systems and their use in analytics.
2. Understand the principles and best practices for data management and how adhering to these practices can help organizations get more value from the data.
3. Learn the concepts of Structured Query Language (SQL) and evaluate how it can be used to extract and transform data from relational databases.
4. Learn the concepts of NoSQL and evaluate how it can be used to store and extract data from non-relational databases.

TOPICS TO BE COVERED:

1. Types of Data
2. Foundations of Database Design
3. Data Governance
4. Data Ethics
5. Data Protection, Privacy, and Security
6. Data Quality
7. Introduction to General Data Protection Rights (GDPR)
8. Extracting data using SQL
9. Introduction to NoSQL for Business Analytics/Data Science

CLASS FORMAT:

Each session is two hours in length. We will be using the first hour of most weeks discussing one of the foundational activities in data management (data quality, data security, data quality, data governance). During this hour, we will also have a brief discussion on the Coursepack readings assigned for the week. The second hour will be entail working on SQL using BigQuery and MySQL when we cover relational databases and using MongoDB when we cover NoSQL databases.

TEXTBOOKS AND RESOURCES:

1. *Navigating the Labyrinth – An Executive Guide to Data Management* by Laura Sebastian-Coleman. (REQUIRED)
   Technics Publication.
   ISBN-10: 1634623754 | ISBN-13: 978-1634623759

2. *The Data Warehouse ETL Toolkit – Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data* by Ralph Kimball and Joe Caserta (OPTIONAL FOR YOUR READING ONLY)
   Wiley Publication.
   ISBN-10: 8126505540 | ISBN-13: 978-0764567575

3. A collection of the following Harvard Business Review articles (REQUIRED)

   The list is available on the next page.

	Article Name	Authors	Publication Date
1.	What's Your Data Strategy?	Leandro DalleMule, Thomas H. Davenport	Apr 30, 2017
2.	Data's Credibility Problem	Thomas C. Redman	Nov 30, 2013
3.	Assess Whether You Have a Data Quality Problem	Thomas C Redman	Jul 27, 2016
4.	If Your Data Is Bad, Your Machine Learning Tools Are Useless	Thomas C Redman	Apr 1, 2018
5.	The U.S. Needs a New Paradigm for Data Governance	Maya Uppaluru	Apr 15, 2018
6.	With Big Data Comes Big Responsibility	Alex "Sandy" Pentland, Scott Berinato	Oct 31, 2014
7.	The Enemies of Data Security: Convenience and Collaboration	Carl S Young	Feb 10, 2015
8.	If Data Is Money, Why Don't Businesses Keep It Secure?	Kuangyi Wei, Ryan LaSalle, Tim Cooper	Feb 9, 2015
9.	Protecting Customers' Privacy Requires More than Anonymizing Their Data	Matthew Schneider, Sachin Gupta	May 31, 2018
10.	How GDPR Will Transform Digital Marketing	Dipayan Ghosh	May 20, 2018
11.	GDPR and the End of the Internet's Grand Bargain	Larry Downes	Apr 8, 2018

These articles are available on Harvard Business Publishing Education website. You can purchase the articles in this Coursepack from https://hbsp.harvard.edu/import/540263.

4. Any book on SQL mentioned on the BAX-400 syllabus. There are plenty of resources available online on SQL and NoSQL. You can use any of these books or resources online for readings on SQL.

5. NoSQL for Mere Mortals (OPTIONAL)
   Publisher: Addison-Wesley Professional
   ISBN-10: 0134023218 | ISBN-13: 978-0134023212

6. MongoDB website: https://www.mongodb.com/what-is-mongodb

NOTES AND HANDOUTS:
I will upload notes and other handouts on Canvas every week before each class meeting.

COMPUTER PACKAGES/SOFTWARE:
1. Google's BigQuery https://cloud.google.com/bigquery/
2. MySQL
3. MongoDB for NoSQL

You should have Google's BigQuery account setup and MySQL installed prior to the first class session. You will be using Google's BigQuery in BAX-400 – Foundations of Analytics course, so it should already be available before this course begins.

The most important part of learning SQL is to get comfortable with the concepts and logic. Once you know the concept, it should not be very difficult to adapt to a software-specific syntax.

ASSESSMENT:
Your learning will be assessed and the final course grade will be determined from the following:

Homework (Individual)	30%
Midterm (Take-home)	30%
Final Exam (Take-home)	40%

ACADEMIC HONOR CODE:
All students are expected to adhere to the University of California, Davis' Code of Conduct as noted here: http://sja.ucdavis.edu/files/cac.pdf.

ADDITIONAL POINTS AND SUGGESTIONS:

1. Our class meetings will be interactive and involve a blend of lectures where we will cover the concepts and then practice writing queries to reinforce the concepts.

2. Becoming proficient with writing SQL queries takes practice and time. The idea behind the homework assignments and the exams is to help you develop the proficiency. It is an important skill to gain and quite a few of our alumni from our inaugural program write SQL queries as a part of their work assignments.

3. For writing complex queries, you will realize that it is best and less intimidating if you build components of the query separately and then bring all the components together.

4. The midterm and final exams will be computer-based and take-home. The formats of the midterm and final exams may be varied. Please note that the purpose of the exams is to assess your understanding of the concepts and your ability to apply concepts discussed in the class.

5. If you have difficulty with any material, please do not hesitate to contact me. My topmost priority is to ensure that you are successful in understanding of the material and prepare you for the other courses in the program and more importantly, for real world.

Schedule (Tentative): This is a tentative schedule. It may be adjusted according to the pace of the class.

	Date	Assignments Due	Topics Covered	Readings
1	Friday, 9/28/18 6:10 – 8:00 PM		• Types of Data • The Importance of Data Management • Fundamentals of Database Management System • Fundamentals of Database Design	Navigating The Labyrinth – Chapter 1
2	Friday, 10/5/18 6:10 – 8:00 PM		• Data Management Challenges • Review of SQL Basics o SELECT statements o DISTINCT, WHERE, ORDER BY clauses o UNION o UNION ALL	• Navigating The Labyrinth – Chapter 2 • For reading on the SQL basics, refer to any resource online or in *SQL: The Complete Reference* book.
3	Friday, 10/12/18 6:10 – 8:00 PM	Homework 1	• Data Ethics • Review of SQL Joins o Simple Joins o Inner Joins o Outer joins o Self Joins o Cross Joins	• Navigating The Labyrinth – Chapter 4 • For reading on the SQL basics, refer to any resource online or in *SQL: The Complete Reference* book.
4	Friday, 10/19/18 6:10 – 8:00 PM	Homework 2	• Data Governance • SQL Aggregate Functions – Summarizing, Grouping, and Filtering Grouped data	• Navigating The Labyrinth – Chapter 5 • Coursepack Readings – o The U.S. Needs a New Paradigm for Data Governance o What's Your Data Strategy? • For reading on the SQL basics, refer to any resource online or in *SQL: The Complete Reference* book.

	Date	**Assignments Due**	**Topics Covered**	**Readings**
5	Friday, 10/26/18 6:10 – 8:00 PM	Homework 3	• Data Protection, Privacy, and Security • Subqueries – Subqueries and Joins, Nested Subqueries, Correlated Subqueries, Temporary tables	• Navigating The Labyrinth – Chapter 9 • Coursepack Readings – o Protecting Customers' Privacy Requires More than Anonymizing Their Data o If Data Is Money, Why Don't Businesses Keep It Secure? o The Enemies of Data Security: Convenience and Collaboration o With Big Data Comes Big Responsibility • For reading on the SQL basics, refer to any resource online or in *SQL: The Complete Reference* book.
6	Friday, 11/02/18 6:10 – 8:00 PM	Midterm Exam (Take-Home) (Due) - Will be posted after the class on 10/26 or 10/27.	• Data Quality • Data Wrangling – Using String functions to clean data	• Navigating The Labyrinth – Chapter 11 • Coursepack Readings – o Data's Credibility Problem o Assess Whether You Have a Data Quality Problem o If Your Data Is Bad, Your Machine Learning Tools Are Useless
7	Friday, 11/09/18 6:10 – 8:00 PM	Homework 4	• GDPR • SQL Windows Functions	• Coursepack Readings – o How GDPR Will Transform Digital Marketing o GDPR and the End of the Internet's Grand Bargain
8	Friday, 11/16/18		• Performance Tuning SQL Queries • Pivoting Data using SQL	Readings/Notes will be posted on Canvas

	Date	**Assignments Due**	**Topics Covered**	**Readings**
	6:10 – 8:00 PM		• Introduction to Relational Algebra (if time allows)	
9	Friday, 11/30/18 6:10 – 8:00 PM	Homework 5	• Introduction to NoSQL for Analytics/Data Science • Limitations of Relational Databases • Comparison between NoSQL and Relational Databases • Types of NoSQL databases • Advantages of NoSQL databases • Document databases	Readings/Notes will be posted on Canvas
10	Friday, 12/07/18 6:10 – 8:00 PM		• Column Family databases • Graph databases	Readings/Notes will be posted on Canvas
11	Friday, 12/14/18	Final Exam (Take Home)	The exam will be posted on 12/8 and will be due on 12/14 11:59 PM.	