NYU SCHOOL OF PROFESSIONAL STUDIES

Data Mining and Data Warehousing

MASY1-GC 3510 | 101 | Spring 2024 | 01/26/2024 - 05/03/2024 | 3 Credit
Modality: In-person
Course Site URL: https://brightspace.nyu.edu/

General Course Information

Name/Title: Sam Sultan, Adjunct Assistant Professor Mr.
NYU Email: sam.sultan@nyu.edu
Class Meeting Schedule: 01/26/2024 - 05/03/2024 / Friday | 02:00pm - 04:35pm
Class Location: Bldg:MIDC Room 1013
Office Hours: Fridays 4:45pn -5:45pm EST, onsite – Email to request an appointment.

Description

In an increasingly competitive information age, data mining and data warehousing are essential in business decision-making. This course teaches students concepts, methods and skills for working with data warehouses and mining data from these warehouses to optimize competitive business strategy. In this course, students develop analytical thinking skills required to identify effective data warehousing strategies such as when to use outsource or in-source data services. Students also learn to Extract, Transform and Load data into data warehouses (the ETL process) and use the CRISP approach to data mining to extract vital information for data warehouses. The course also teaches students how to secure data and covers the ethical issues associated with the uses of data and data models for business decisions.

Prerequisites

1210 - Quantitative Models for Decision Makers

Learning Outcomes

At the conclusion of this course, students will be able to:

- Translate business requirements into a well-constructed, normalized conceptual and logical data models
- Apply logical database design and the relational model
- Apply the CRISP model to conduct successful data mining
- Establish a successful ETL process to load a data warehouse
- Write basic SQL statements including some advanced SQL features
- Employ appropriate data governance principles to assure data quality and security

Communication Methods

Be sure to turn on your NYU Brightspace notifications and frequently check the "Announcements" section of the course site. This will be the primary method I use to communicate information critical to your success in the course. To contact me, send me an email. I will respond within 24 hours (48 during weekends).

Credit students must use their NYU email to communicate. Non-degree students do not have NYU email addresses. Brightspace course mail supports student privacy and FERPA guidelines. The instructor will use the NYU email address to communicate with students. All email inquiries will be answered within 24 hours. (48 during weekends)

Structure | Method | Modality
There are 14 sessions during this course. The session topics are organized into 1) Concepts, 2) Learning Principles, and 3) Hands on Design and Implementation.

During this course, there will be assignments, midterm exam, final exam, and a final team project. For the final team project, students will be divided into groups of 2 to 4 students. Course sessions are **in-person onsite**. Student engagement during class presentations is a must and will result in better overall grade. Important announcements, and all assignments and exam submissions will be done through the course site in NYU Brightspace.

This course is **onsite** and meets once a week on **Fridays**.

Expectations

Learning Environment
As graduate students, you are expected to conduct themselves in a professional manner and engage and collaborate with your classmates. Classrooms are diverse and include students who range in age, culture, learning styles, and levels of professional experience. To maintain an inclusive environment that ensures all students can equally participate with and learn from each other, as well as receive feedback and instruction from faculty during group discussions in the classroom, all course-based discussions and group projects should occur in a language that is professional and shared among all participants.

Participation
To receive full credit for class participation, you should attend all classes since much of the learning and engagement occurs during class lecture, presentation and class discussions. You must contribute and engage in class discussions/dialogue during every class session of the course. Please contact the instructor if you anticipate missing any part of the class. Please arrive on time so as not to disturb the flow of the lecture. Excessive lateness's may result in lower overall grade.
Please contact the instructor if you anticipate missing any part of the class. Participation grades will be based on:

- Involvement in class discussions, dialogues, and activities during each session
- Participation which demonstrates integration of reading, class work, and relevance application.
- Willingness to learn by accepting feedback, trying new skills and approaches, etc. Quality/quantity of providing effective and balanced feedback. You must ask at least one question and response to professor or other student inquiry at least one or multiple times per each session.

Assignments and Deadlines
Students are expected to participate in each class session by offering their understanding of the subject, sharing ideas or discussing/commenting on other student's comments. In addition, students must complete and submit all assigned homework on time. Assignments are typically due within one week of assigned date (unless specifically mentioned to the contrary). Late submission of homework will either not be accepted, or will result in a lower grade. Students are also expected to develop with and present a team project with other students, as well as take and pass a midterm exam and a final exam.

Course Technology Use
We will utilize multiple technologies to achieve the course goals. I expect you to use technology in ways that enhance the learning environment for all students. All class sessions will require the use either the lab room desktop computers or your own personal laptop computer. You can use either a PC or a MAC based computer.

Feedback and Viewing Grades
I will provide timely meaningful feedback on all your work via our course site in NYU Brightspace. You can access your grades on the course site Gradebook.

Attendance
I expect you to attend all class sessions. Attendance will be taken into consideration when determining your final grade.

Excused absences are granted in cases of documented serious illness, family emergency, religious observance, or civic obligation. In the case of religious observance or civic obligation, this should be reported in advance. Unexcused absences from sessions may have a negative impact on a student's final grade. Students are responsible for assignments given during any absence.
Each unexcused absence or being late may result in a student's grade being lowered by a fraction of a grade. A student who has three unexcused absences may earn a Fail grade.

Refer to the SPS Policies and Procedures page for additional information about attendance.

Textbooks and Course Materials
Required

- Business Analytics – Communicating with Numbers (1st or 2nd Edition)
  - **Authors** - Sanjiv Jaggia, Alison Kelly, Kevin Lertwachara, Leida Che
  - **Publisher** – McGraw Hill, 2020 or 2022
  - **ISBN** - 978-1260785005
- The Data Warehouse Lifecycle Toolkit (2nd Edition) –Available through Amazon
  - **Authors** - Kimball, Ross, Thornthwaite, Mundy & Becker
  - **Publisher** – Wiley, 2008
  - **ISBN** - 978-0470149775
- The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd Edition) Available through Amazon
  - **Authors** - Ralph Kimball, Margy Ross

  - **Publisher** – Wiley, 2013
  - **ISBN** - 978-1118530801
- Instructor may also provide session by session content, which will be posted online.

Recommended Material

- Building the Data Warehouse (4th Edition)
  - **Author –** W. H. Inmon
  - **Publisher** – Wiley, 2005
  - **ISBN** - 978-0764599446

- The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence – Remastered Collection (2nd Edition)
  - **Authors** - Ralph Kimball, Margy Ross
  - **Publisher** – Wiley, 2016
  - **ISBN:** 978-1119216315

Grading | Assessment

Your grade in this course is based on your performance on multiple activities and assignments. Since all graded assignments are related directly to course objectives and learning outcomes, failure to complete any assignment will result in an unsatisfactory course grade.

All written assignments are to be completed with proper grammar, punctuation, and spelling. Please carefully proofread your written assignments before submitting them for a grade. Coding assignments should include both properly indented code and screenshot of the output. Grades and review comments will be posted by professor typically within 1 week of assignment submission or exam submission.

Midterm Exam: There will be a midterm exam. The exam will be an open book, open internet style exam. The exam will test the student's acquisition of topics, concepts and competencies learned in this class up to mid-term.

Final Exam: There will be a final exam. The exam will be an open book, open internet style exam. The exam will test the student's acquisition of topics, concepts and competencies learned in this class. The final exam will only cover materials covered after the midterm.

Please Note: Professor will not provide a “redo” or an opportunity for grade improvement for any assignment or exam for which a student received a low grade. It is the student responsibility to prepare for exams and to submit correct and most accurate assignments.

DESCRIPTION	PERCENTAGE
Homework	20%
Participation	10%
Midterm Exam	25%
Final Exam	25%
Final Team Project	20%

TOTAL POSSIBLE 100%

See the "Grades" section of Academic Policies for the complete grading policy, including the letter grade conversion, and the criteria for a grade of incomplete, taking a course on a pass/fail basis, and withdrawing from a course.

Course Outline
Start/End Dates: 01/26/2024 - 05/03/2024 / Friday
Time: 02:00pm - 04:35pm
No Class Date(s): Friday – 03/22/2024
Special Notes: Spring Break 03/18/24 - 03/24/24

Course Outline

Session 1, 01/26/24 - Introduction Data Warehouse?

- *Data Warehousing ROI*
- *DSS - Decision Support Systems*
- *Operational vs. Analytical Systems*
- *Evolution of DSS and Data Warehousing*
- *OLTP - Online Transaction Processing*
- *Characteristics of a Data Warehouse*
- *What is a Data Mart? Creating a Data Mart*
- *Data Comparison Chart*
- *OLAP - Online Analytical Processing*

- ***Assignments (due one week from today)***
  - ***Reading:*** *Chapter 1 (from both Data Warehouse Lifecycle Toolkit, and Building the Data Warehouse). Skim thru glossary (of Data Warehouse Lifecycle Toolkit)*

Session 2, 02/02/24 - Planning and Building the Data Warehouse

- *Planning & Building the Data Warehouse*
- *Sponsorship and Cost Justification*
- *Project Prerequisites*
- *Barriers, Challenges and Risks*
- *Preparing for Implementation*
- *Developing the Data Warehouse*
- *SDLC Methodologies - Waterfall vs. RUP Approach*
- *Planning & Project Management*
- *Analysis*
- *Logical & Physical Design*
- *Implementation and Deployment*
- *Operations*

- ***Assignments (due one week from today):***

- ***Reading:*** *Chapter 1, 2 (The Data Warehouse Lifecycle Toolkit)*

Session 3, 02/09/24 - Data Warehouse Design

- *Data Warehouse Design*
- *Drivers for Multi-Dimensional Analysis*
- *Limitations of Relational Models*
- *The Data Cube*
- *What is dimensional modeling?*
- *Advantages of Dimensional Models*
- *Logical and Physical Design*
- *Data Normalization*
- *Benefits and Drawbacks of Data Normalization*
- *De-Normalizing of Data*
- *Characteristics of a Data Warehouse*
- *Subject Oriented, Integrated, Time Variant, Non-Volatile*
- *The Star Schema*

- ***Assignments (due one week from today):***
  - ***Reading:*** *Chapter 6 (The Data Warehouse Lifecycle Toolkit)*

Session 4, 02/16/24 - Data Warehouse Schemas

- *Data Warehouse Schemas*
- *Dimensions and Dimension Tables*
- *Facts and Fact Tables*
- *The Star Schema*
- *The Snowflake Schema*
- *Degenerate and Junk Dimensions*
- *The Data Warehouse Bus Architecture*
- *Conformed Dimensions and Standard Facts*
- *Data Granularity*
- *Changing Dimensions*

- ***Assignments (due one week from today):***
  - ***Reading:*** *Chapter 6 (The Data Warehouse Lifecycle Toolkit)*
  - ***Assignments:*** *Create a logical database model using data normalization rules*

Session 5, 02/23/24 - Components of a Data Warehouse

- *Components of a Data Warehouse*
- *Source Systems, Staging Area, Presentation, Access Tools*
- *Building the Data Matrix*
- *The Four Steps Process*
- *Multiple Fact Tables in a single Data Mart*
- *Chain, Heterogeneous, Transaction/Snapshot & Aggregate Facts*
- *Fact and Dimension Table Detail*
- *Identifying Source for each Fact & Dimension*

- *Mapping from Source to Target*

- ***Assignments (due one week from today):***
  - ***Reading:*** *Chapter 7, 4 (The Data Warehouse Lifecycle Toolkit)*
  - ***Assignments:*** *Create SELECT statements that perform various table joins from our database*

Session 6, 03/01/24 - The ETL Process

- *The ETL Process*
- *Extracting the Data into the Staging Area*
- *The Challenge of Extracting from Disparate Platforms*
- *Full vs. Incremental Extracts*
- *Detecting Changes to Data*
- *Transforming the Data*
- *Complexity of Data Integration*
- *Dealing with Missing & Dirty Data*
- *Data Transformation Tasks*
- *Loading the Data*
- *Timing and Job Control of Data Loads*

- ***Assignments (due one week from today):***
  - ***Reading:*** *Chapter 9 (The Data Warehouse Lifecycle Toolkit)*

Session 7, 03/08/24 – Midterm Exam

- ***Midterm Exam***

Session 8, 03/15/24 - Aggregating Data

- *Aggregating Data*
- *Goals and Risks of Data Aggregation*
- *Deciding What to Aggregate*
- *Data Sparsity*
- *Design Requirement for Aggregates*
- *The problem with Aggregates*
- *Aggregate Navigators*

- ***Assignments (due one week from today):***
  - ***Reading:*** *Chapter 8 p353-357 (The Data Warehouse Lifecycle Toolkit)*
  - ***Assignments:*** *Create SELECT statements that perform data analytics against the data warehouse using aggregate functions and the GROUP BY clause*

Session 9 (Self-Study), 03/29/24 - Selecting the Business Subject

- *Self-Study Topics*
- *Selecting the Business Subject*
- *Declaring the Grain*
- *Choosing the Dimension*

- *Identify the Fact*
- *Avoiding Null Keys*
- *Retail Market Basket Analysis*
- *Additive and Semi-Additive Facts*
- *The Value Chain Integrated Inventory Model*
- *Order Management Data Marts*
- *Date and Other Dimension Role Playing*
- *Allocation to Lower Level Facts*
- *Profit and Loss Data Marts*

- ***Assignments (due one week from today):***
  - ***Reading:*** *Chapter 2, 3, 5 (The Data Warehouse Toolkit)*

Session 9, (Self Study), 03/29/24 - CRM and Financial Data Warehouses

- *Self-Study Topics*
- *CRM Overview*
- *Customer Dimension*
- *Demographic Dimension Outriggers*
- *Date Dimension Outriggers*
- *Large Changing Customer Dimension*
- *Mini-Dimensions*
- *Commercial Customer Hierarchies*
- *Fixed vs. Variable Level Hierarchies*
- *General Ledger Accounting*
- *OLAP role in G/L and Chart of Accounts*
- *Time Stamped Employee Dimensions*

- ***Assignments (due one week from today):***
  - ***Reading:*** *Chapter 6, 7, 8 (The Data Warehouse Toolkit)*
  - ***Assignments:*** *Use the logical database model created in session 4 to create the physical tables in your database*

Session 10, 04/05/24 – Web Analytics and Mining

- *Web Data Warehouses, Mining and Analytics*
- *Overview of Web Based Interaction*
- *Challenges of Tracking Data*
- *Creating Persistent State on the Web*
- *Techniques for Tracking States*
- *Working with Cookies*
- *User Registration*
- *Web Server Log Files*
- *Online Advertising and Tracking*
- *Online Page Tracking and Analytics*
- *User Dimension and Page Hits Facts*

- ***Assignments (due one week from today):***

- ***Reading:*** *Chapter 15 (The Data Warehouse Toolkit)*

Session 11, 04/12/24 - Introduction to Data Mining

- *Data Mining and Concepts*
- *What is Data Mining Good For?*
- *Statistics, Artificial Intelligence & Machine Learning*
- *Data Mining Examples and Tools*
- *Connection between Data Mining and Data Warehousing*
- *Retrospective Reporting vs. Predictive*
- *Data Mining Applications*
- *Data Mining vs. Statistics vs. OLAP*
- *Data Mining Statistical Techniques (Sampling, Regression & Decision Trees)*
- *Clustering, Segmentation and Nearest Neighbor Techniques*
- *Keys to commercial success of Data Mining*

- ***Assignments (due one week from today):***
  - ***Reading:*** *Chapters 1,2,3,8 (Business Analytics – Communicating with Numbers)*

Session 12, 04/19/24 - Data Mining Techniques I

- *Data Mining Techniques – Part 1*
- *Terminology*
- *Bayesian Theorem and Probabilities*
- *Naive Bayes Classifier and Predictions*
- *Naive Bayes with multiple variables and multiple targets*
- *Linear Regression*
- *Linear Regression with multiple independent variables*
- *Other Data Mining techniques*
- *Examples using Weka*

- ***Assignments (due one week from today):***
  - ***Reading:*** *Chapters 4-7 (Business Analytics – Communicating with Numbers)*

Session 13, 04/26/24 - Data Mining Techniques II

- *Data Mining Techniques – Part 2*
- *Supervised vs. Unsupervised Data Mining*
- *Classification and Segmentation*
- *Tree Inductions*
- *Entropy and Information Gain*
- *The ID3 and C4.5 classifier process*
- *Clustering*
- *Co-occurrence grouping*
- *Examples using Weka*

- ***Assignments (due one week from today):***

- ***Reading:*** *Chapters 9-11 (Business Analytics – Communicating with Numbers)*

Session 14, 05/03/24 - Weka

- *Weka Data Mining Tool*
- *Weka Examples*

- ***Final Exam***
- ***Data Warehouse Team Projects Due***

NOTES:

The syllabus may be modified to better meet the needs of students and to achieve the learning outcomes.

The School of Professional Studies (SPS) and its faculty celebrate and are committed to inclusion, diversity, belonging, equity, and accessibility (IDBEA), and seek to embody the IDBEA values. The School of Professional Studies (SPS), its faculty, staff, and students are committed to creating a mutually respectful and safe environment (*from the SPS IDBEA Committee*).

New York University School of Professional Studies Policies

1. Policies - You are responsible for reading, understanding, and complying with University Policies and Guidelines, NYU SPS Policies and Procedures, and Student Affairs and Reporting.

2. Learning/Academic Accommodations - New York University is committed to providing equal educational opportunity and participation for students who disclose their dis/ability to the Moses Center for Student Accessibility. If you are interested in applying for academic accommodations, contact the Moses Center as early as possible in the semester. If you already receive accommodations through the Moses Center, request your accommodation letters through the Moses Center Portal as soon as possible (mosescsa@nyu.edu | 212-998-4980).

3. Health and Wellness - To access the University's extensive health and mental health resources, contact the NYU Wellness Exchange. You can call its private hotline (212-443-9999), available 24 hours a day, seven days a week, to reach out to a professional who can help to address day-to-day challenges as well as other health-related concerns.

4. Student Support Resources - There are a range of resources at SPS and NYU to support your learning and professional growth. For a complete list of resources and services available to SPS students, visit the NYU SPS Office of Student Affairs site.

5. Religious Observance - As a nonsectarian, inclusive institution, NYU policy permits members of any religious group to absent themselves from classes without penalty when required for compliance with their religious obligations. Refer to the University Calendar Policy on Religious Holidays for the complete policy.

6. Academic Integrity and Plagiarism - You are expected to be honest and ethical in all academic work. Moreover, you are expected to demonstrate how what you have learned incorporates an understanding of the research and expertise of scholars and other appropriate experts; and thus recognizing others' published work or teachings—whether that of authors, lecturers, or one's peers—is a required practice in all academic projects.

Plagiarism involves borrowing or using information from other sources without proper and full credit. You are subject to disciplinary actions for the following offenses which include but are not limited to cheating, plagiarism, forgery or unauthorized use of documents, and false form of identification

Turnitin, an originality detection service in NYU Brightspace, may be used in this course to check your work for plagiarism.

Read more about academic integrity policies at the NYU School of Professional Studies on the Academic Policies for NYU SPS Students page.

7. Use of Third-Party Tools - During this class, you may be required to use non-NYU apps/platforms/software as a part of course studies, and thus, will be required to agree to the “Terms of Use” (TOU) associated with such apps/platforms/software.

These services may require you to create an account but you can use a pseudonym (which may not identify you to the public community, but which may still identify you by IP address to the company and companies with whom it shares data).

You should carefully read those terms of use regarding the impact on your privacy rights and intellectual property rights. If you have any questions regarding those terms of use or the impact on the class, you are encouraged to ask the instructor prior to the add/drop deadline.