**Data Mining and Data Warehousing**

**MASY1-GC 3510 | 200 | Fall 2024 | 09/04/2024 - 12/04/2024 | 3 Credit Modality:** Online (Sy)

**Course Site URL:** <https://brightspace.nyu.edu/>

**General Course Information**

Name/Title: Farid Razzak

NYU Email: fr497@nyu.edu

Class Meeting Schedule: 09/04/2024 - 12/04/2024 | Wednesday | 07:00pm -- 09:35pm

Class Location: Online(Sy)

Office Hours: By appointment via NYU Zoom Please email farid.razzak@nyu.edu at least 8 hours before the upcoming class to arrange potential office hours. Office hours are subject to availability.

**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 insource 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

**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](https://www.nyu.edu/servicelink/KB0018507) 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.

Credit students must use their NYU email to communicate. Non-degree students do not have NYU email addresses. NYU Classes course-mail supports student privacy and FERPA guidelines. All email inquiries will be responded to within 24 hours. I will respond to you using NYU email.

# Structure | Method | Modality

There are 14 session topics in this course. The session topics are organized into three (3) areas of study: 1) History, Abstraction and Theory, 2) Learning Principles, and 3) Instructional Design in Practice.

Active learning experiences and small group projects are key components of the course. Assignments, papers, and exams will be based on course materials (e.g., readings, videos), lectures, and class discussions. Course sessions will be conducted synchronously on NYU

|  |  |
| --- | --- |
| Zoom, which you can access from the course site in [NYU Brightspace.](https://brightspace.nyu.edu/) |  |
|  | |

**Expectations**

Learning Environment

You play an important role in creating and sustaining an intellectually rigorous and inclusive classroom culture. Respectful engagement, diverse thinking, and our lived experiences are central to this course, and enrich our learning community.

Participation

You are integral to the learning experience in this class. Be prepared to actively contribute to class activities, group discussions, and work outside of class.

Assignments and Deadlines

Please submit all assignments to the appropriate section of the course site in [NYU Brightspace.](https://brightspace.nyu.edu/) If you require assistance, please contact me BEFORE the due date.

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.

**Generative AI Use**

**Welcome with Attribution**

You are welcome/ expected to use generative AI tools (e.g. ChatGPT, Dall-e, etc.) in this class as doing so aligns with the course learning goal [insert course learning goal]. You are responsible for the information submitted based on an AI query (for instance, that it does not violate intellectual property laws, or contain misinformation or unethical content). Your use of AI tools must be properly documented and cited. Guidance can be found [here](https://guides.nyu.edu/c.php?g=1307730&p=9624166#s-lg-box-30439868).

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. Attendance will be taken randomly and counted towards your grade. Late arrival, webcam issues/absence or being in a non-learning environment may constitute as a absence.

Refer to the [SPS Policies and Procedures page](https://www.sps.nyu.edu/homepage/student-experience/policies-and-procedures.html) for additional information about attendance.

**Textbooks And Course Materials**

**Required**

* Data Warehouse Essentials (4th edition) o **Editor** – Julio Bolton

o **Publisher** – Larsen and Keller Education (June 28, 2019)o **ISBN-10 : 1641720735** o **ISBN-13 : 978-1641720731**

* PC/Mac Laptop (With Appropriate Administrative Privileges & Technological Capabilities)
* Instructor may also provide session by session content, which will be posted online.
* Assigned **Harvard Business Review Case Studies (Course packet may be created containing the necessary case studies)**.
* Oracle SQL Developer – o [http://www.oracle.com/technetwork/developer-tools/sqldeveloper/overview/index.html](http://www.oracle.com/technetwork/developer-tools/sql-developer/overview/index.html)
* Oracle SQL Developer Data Modeler o  [http://www.oracle.com/technetwork/developertools/datamodeler/overview/index.html](http://www.oracle.com/technetwork/developer-tools/datamodeler/overview/index.html)
* Google Collab Python Notebooks ( o colab.research.google.com
* PyCharm for Education o https://www.jetbrains.com/pycharm-edu/
* puTTy (for windows users) o https://www.putty.org/
* Git Code Repo Account
* AWS Cloud and Redshift Account

# Recommended (Complimentary Resources)

* The Data Warehouse Lifecycle Toolkit (3rd Edition) – Available through Amazon o **Authors** – Kimball, Ross, Thornthwaite, Mundy & Becker o **Publisher** – Wiley, ’2013 o ISBN-13 : 978-1118530801 o ISBN-10 : 1118530802
* Introduction to Data Mining (2nd Edition) o **Authors** - Pang-Ning Tan, Michael Steinbach, Vipin Kumar o **Publisher** - TBS, 2016 ▪ ISBN-10 : 0273769227

▪ ISBN-13 : 978-0273769224

**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 using APA format and must be typed and double-spaced. Grammar, punctuation, and spelling will be considered in grading. Please carefully proof-read your written assignments before submitting them for a grade. I will update the grades on the course site each time a grading session has been completed— typically three (3) days following the completion of an activity.

DESCRIPTION PERCENTAGE

Assigned Activities (total of 4 Lab and 1 Group ) 20%

Quizzes & Case Studies (total of 3, Exam Reviews and/or Case studies) 20%

Participation (Attendance is prerequisite to participation) 10%

Midterm Exam 15%

Final Exam 15%

Final Group Project 20%

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_

TOTAL POSSIBLE 100%

See the [“Grades” section of Academic Policies](https://www.sps.nyu.edu/homepage/student-experience/policies-and-procedures.html#Graduate1) 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:** 09/04/2024 - 12/04/2024 | Wednesdays

**Time:** 07:00pm -- 09:35pm

**No Class Date(s):** N/A

**Special Notes:** N/A

Session 1, 09/06/23

Session 1, 09/04/24

**Introduction to Data Warehousing**

* Relationship of Data Mining and Data Warehousing
* What is a 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)**

* **Suggested Reading:**
* Chapter 1 (from both Data Warehouse Lifecycle Toolkit, and Building the Data Warehouse).
* Skim thru glossary (of Data Warehouse Lifecycle Toolkit)
* **Complete Lab Assignment**

Session 2, 09/11/24

**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):**

* **Suggested Reading:** Data Warehouse Essentials (Preface, Introduction chapters)
* **Complete Lab Assignment**

Session 3, 09/18/24

**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):**

* **Suggested Reading:** Data Warehouse Essentials (Dimensional Modeling, STAR Schema Chapters)
* **Complete Lab Assignment**
* **Group Membership for Course Project Due**

Session 4, 09/25/24

**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):**

* **Suggested Reading:** Data Warehouse Essentials (OLAP cubes, MOLAP, ROLAP Chapters)
* **Complete Lab Assignment:**
* **Case Study Review Assignment:**
* Data Warehousing and Multi-dimensional Data Modeling
* [https://store.hbr.org/product/data-warehousing-and-multi-dimensional-datamodelling/A00181](https://store.hbr.org/product/data-warehousing-and-multi-dimensional-data-modelling/A00181)

Session 5, 10/02/24

**The ETL Process**

* The ETL Process
* Extracting the Data into the Staging Area
* The Challenge of Extracting from Disparate Platforms
* Transforming the Data
* Complexity of Data Integration
* Data Transformation Tasks
* Loading the Data
* Aggregating Data
* Goals and Risks of Data Aggregation
* Deciding What to Aggregate
* Design Requirement for Aggregates

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

* **Complete Lab Assignment**
* **Suggested Reading:** Data Warehouse Essentials (Data Aggregation Chapter)
* **Online Midterm Exam\*\***

Session 6, 10/09/24

**Introduction to Data Mining & Techniques**

* Data Mining
* 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
* Predictive Modeling
* Clustering/Segmentation
* Data Mining and Statistics Terminologies
* Supervised vs. Unsupervised
* Tree Induction
* Large Language Models

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

* **Suggested Reading:** Chapter 1,2,3 (**Introduction to Data Mining)**
* **Complete Lab Assignment** • **Case Study Review Assignment:**
* **Data Science and the Art of Persuasion**
* [**https://store.hbr.org/product/data-science-and-the-art-of-persuasion/R1901K**](https://store.hbr.org/product/data-science-and-the-art-of-persuasion/R1901K)

Session 7, 10/16/24

**Data Mining : Regression**

* Simple Linear Regression
* Assumptions
* Multivariate- Regression
* Logistic Regression
* Impact on Data Analytics and Datawarehouses today **Assignments (due one week from today):**
* **Suggested Reading:** Appendix D (**Introduction to Data Mining)**
* **Complete Lab Assignment**
* **Group Assignment 1 Due**

Session 8, 10/23/24

**Data Mining: Classification**

* Naïve Bayes Classifier
* Hunts Algorithm
* Decision Trees: Tree Induction
* How it relates to Data Warehouses and Data Analytics **Assignments (due one week from today):**
* **Suggested Reading:** Chapter 4,5 (**Introduction to Data Mining)**
* **Complete Lab Assignment** • **Case Study Review Assignment:**
* Miroglio Fashion (A)
* <https://store.hbr.org/product/miroglio-fashion-a/519053>

Session 9, 10/30/24

**Data Mining : Clustering**

* Distance Matrix, Proximity Matrix
* K-Means, K-Medoids
* DBSCAN
* How it relates to Data Warehouses and Data Analytics **Assignments (due one week from today):**
* **Suggested Reading:** Chapter 8 (**Introduction to Data Mining)**
* **Complete Lab Assignment**

Session 10, 11/06/24

**Data Mining: Text Mining**

* Term Frequency Matrix
* Natural Language Processing
* Stopwords
* Stemming
* Document Similarities
* Topic Modeling
* LDA
* Text Classification
* Text Clustering **Assignments:**
* **Complete Lab Assignment**

Session 11, 11/13/24

**Data Mining: Anomaly Detection (Subject to Change)**

* Model based Approaches : Supervised and Unsupervised • Graphical Approaches : Boxplots, Convex Hull Method
* Statistical Approaches:
* Distance based Approaches: K-NN
* Density based Approaches: LOF Approach
* Clustering Approaches
* How it relates to Data Warehouses and Data Analytics

**Assignments:**

* **Suggested Reading:** Chapter 10 (**Introduction to Data Mining)**
* **Complete Lab Assignment** • **Case Study Review Assignment:**
* Data Science at Target
* <https://store.hbr.org/product/data-science-at-target/118016>

Session 12, 11/20/24

**Data Warehousing and Data Mining: Current Trends and Special Topics:**

* SaaS
* ChatGPT (LLMs)
* Snowkflake (Cloud based Data Warehouses)
* AWS Redshift

Session 13, 11/27/24

Special Topic or Guest Speaker.

Session 14, 12/04/24

* Online Final Exam
* Group Course Project Due and Presented



**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*)](https://www.sps.nyu.edu/homepage/about-us/idbea/about-idbea.html).

**New York University School of Professional Studies Policies**

1. Policies - You are responsible for reading, understanding, and complying with [University Policies and Guidelines,](http://www.nyu.edu/about/policies-guidelines-compliance.html) [NYU SPS Policies and Procedures,](http://sps.nyu.edu/academics/academic-policies-and-procedures.html) and [Student Affairs and Reporting.](https://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/student-services.html)
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.](https://www.nyu.edu/students/communities-and-groups/student-accessibility.html) If you are interested in applying for academic accommodations, contact the [Moses Center](https://www.nyu.edu/students/communities-and-groups/student-accessibility/academic.html) 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](https://www.nyu.edu/students/communities-and-groups/student-accessibility.html) 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.](https://www.nyu.edu/students/health-and-wellness/wellness-exchange.html) 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.

1. 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.](https://www.sps.nyu.edu/homepage/student-experience/resources-and-services.html)

1. 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](https://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/university-calendar-policy-on-religious-holidays.html) for the complete policy.

1. 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,](https://www.nyu.edu/servicelink/KB0018471) 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](https://www.sps.nyu.edu/homepage/student-experience/policies-and-procedures.html) page.

1. 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.