Designing and Implementing a Data Warehouse

MET CS 689 B1 for Fall 2018
Charles River Campus at MCS room B19
Math and Computer Science Building, 111 Cummington Mall
Tuesdays, 18:00 – 20:45

Mary E. Letourneau
maryleto@bu.edu
Office hours: by appointment

COURSE DESCRIPTION

This course provides the student with the ability to analyze, design, and implement a data warehouse. The student will gain important foundational skills in applying database analytical functions and implementing extract-transform-load processes. From this point, we cover the modeling and implementation techniques for dimensional data warehouses, star/snowflake schemas, OLAP, and data lakes. The course also introduces Big Data concepts and technologies, including entity resolution in unstructured data and one or more massive-parallelism platforms.

PREREQUISITES

> MET CS 579 or MET CS 669
> MET CS 521 or MET CS 520

BOOKS

The Data Warehouse Toolkit, 3rd ed., Kimball and Ross. ISBN: 9781118530801. Indianapolis: Wiley, 2013.
Data Warehousing in the Age of Big Data, 1st ed., Krish Krishnan. ISBNs: 9780124058910 (paperback), 9780124059207 (eBook). Waltham, MA: Morgan Kaufmann, 2013.
Python for Data Analytics, 1st ed., McKinney. ISBN: 9781449319793. Sebastopol, CA: O'Reilly, 2012.

COURSEWARE

Courseware will be *Blackboard*.

CLASS MEETINGS, LECTURES & ASSIGNMENTS

Week	Description	Due / On
1: 9/4/2018 – 9/10/2018	Lecture 01: Introduction	9/4/2018
	Reading: Module 1	9/10/2018
	Reading: McKinney Chapter 1	9/10/2018
	Reading: Kimball/Ross Chapter 1	9/10/2018

	Reading: Krishnan Chapter 6	9/10/2018
2: 9/11/2018 – 9/17/2018	Lecture 02: Analytic Functions	9/11/2018
	Lab 1A: Software and Appliance Installations	9/17/2018
	Lab 1B: Analytical/Windowed Functions	9/17/2018
	Quiz 1	9/17/2018
3: 9/18/2018 – 9/24/2018	Lecture 03: Extract and Transform	9/18/2018
	Reading: Module 2	9/24/2018
	Reading: McKinney Chapters 6 & 7	9/24/2018
	Reading: Kimball/Ross Chapters 19 & 10	9/24/2018
	Reading: Krishnan Chapter 7	9/24/2018
4: 9/25/2018 – 10/1/2018	Lecture 04: Load and Verification	9/25/2018
	ETL/ELT Workshop	9/25/2018
	Lab 2A: Python Familiarization	10/1/2018
	Lab 2B: ETL with Python	10/1/2018
	Quiz 2	10/1/2018
5: 10/2/2018 – 10/8/2018	Lecture 05: Dimensional Data Modeling	10/2/2018
	Reading: Module 3	10/15/2018
	Reading: Kimball/Ross Chapters 2 & 18	10/15/2018
	Reading: Krishnan Chapter 11	10/15/2018
6: 10/9/2018 – 10/15/2018	No class – Columbus Day	
7: 10/16/2018 – 10/22/2018	Lecture 06: Time, Bitemporality, Slowly-Changing Dimensions	10/16/2018
	Lab 3: Dimensional data modeling	10/22/2018
	Quiz 3	10/22/2018
8: 10/23/2018 – 10/29/2018	Lecture 07: Big Data Approaches to Modeling	10/23/2018
	Reading: Module 4	10/29/2018
	Reading: Krishnan Chapters 12, 13	10/29/2018
9: 10/30/2018 – 11/5/2018	Lecture 08: Reporting	10/30/2018
	Lab 4: Business Reporting with Data Warehouses	11/5/2018
	Quiz 4	11/5/2018
10: 11/6/2018 – 11/12/2018	Lecture 9: Forwarding Data to Further Stores and Uses	11/6/2018
	Reading: Module 5	11/12/2018
	Reading: Krishnan Chapters 2, 3, 4 & 9	11/12/2018
11: 11/13/2018 – 11/19/2018	Lecture 10: Dealing with Velocity, Volume, Variability	11/13/2018
	Lab 5: Business Reporting with Data Warehouses	11/19/2018
	Quiz 5	11/19/2018
12: 11/20/2018 – 11/26/2018	Lecture 11: Alternative Storage for	11/20/2018

	Big Data	
	Reading: Module 6	11/26/2018
	Reading: Krishnan Chapter 8	11/26/2018
13: 11/27/2018 – 12/3/2018	Lecture 12: Performance Analysis and Tuning for Data Warehousing and Big Data	11/27/2018
	Lab 6: Big Data Workshop	12/3/2018
	Quiz 6	12/3/2018
14: 12/4/2018 – 12/10/2018	Lecture 13: Course Wrap-Up and Final Exam Preparation	12/4/2018
	Term Project	12/10/2018
12/17/2018 – 12/21/2018	Final Exam	TBD

*On 10/9/2018 BU will substitute Tuesday classes for Monday, due to the Columbus Day holiday. This lecture may need to be moved to another location or day, depending on student availability.

CLASS RESOURCES

This course will provide students with the following resources:

- Virtual Machines for Labs and Experimentation
- Access to Software with Free or Academic Licenses
- Access to Microsoft Azure data warehousing functionality
- Access to Hadoop cluster computing resources
- Large-scale datasets suitable for warehousing

Recommended minimum system requirements:

- Intel-based
- i5 Core or equivalent
- 12 GB RAM
- 100 GB free disk space (if external, USB 3 or faster)

CLASS POLICIES

Attendance & Absences –

Students are expected to attend all classes or notify the instructor for an excuse with good reason three hours before class. After two unexcused absences the student forfeits all class participation credit.

Assignment Completion & Late Work –

All assignments will be submitted through Blackboard, and all quizzes and examinations will be administered through Blackboard. Students may receive a 36-hour extension without penalty, on a single assignment or assessment, by notifying the instructor 36 hours before that assignment or assessment is due, giving reason. Other extensions will be granted at the instructor's discretion based on student circumstances. No access to take a quiz/assessment will be allowed 5 days after its original due date. The instructor will apply late penalties at his or her discretion,

up to and including forfeiture of grade on any assignment. The instructor may apply additional penalties for repeated seeking of extensions or other late submission of work.

Academic Conduct Code –

WRITE IT, OR CITE IT!
Please review the Policy on Academic Conduct:
http://www.bu.edu/met/metropolitan_college_people/student/resources/
conduct/code.htm

Neither the University, nor I, nor your classmates can tolerate plagiarism or other academic misconduct in any formal submission for this class. Please show appropriate respect for all – and for yourself – by expressing your own mastery of the material in your own words, diagrams, programming, etc. You must include references for everything you copy or quote. When you make such inclusions, mark and attribute them clearly and in appropriate academic style. You may not submit any other student's work as your own, nor may you provide anyone else, in class or outside, with your own work on this class. Contact your instructor with any questions.

Grading Criteria

Overview:
Grades of coursework will be applied to the final course grade with the following weights:

Component		Weight
Lab Assignments		30%
Labs 1A, 1B, 2A, 2B	2.5% each	
Labs 3 - 6	5% each	
Term Project		10%
Participation / Online Discussion (5% each)		10%
Quizzes		15%
Final Exam		35%

Participation / Online Discussion:
Participation includes asking questions, offering insights, sharing experiences, etc. relevant to the material being discussed. As such, participation implies attendance to lectures. But it is understood that Life happens. Let the instructor know as soon as possible if you cannot attend class. Up to two classes can be missed without impacting the participation part of your grade, if notice is provided in advance.

Every two weeks will have a new discussion topic in the Blackboard to discuss. These topics rarely have a "correct" answer, and can be approached from many perspectives. An "A" grade in this portion of the grade requires substantive content relative to the topic posted on five different days during the two weeks. The purpose of requiring that posts appear on different days is to encourage you to post early in the period and then go back later to read and respond to other students' posts. ("I agree" or repeating someone else's post is not substantive; neither is one

short sentence considered substantive.) Please be respectful in your posts. Feel free to debate and disagree, but do it with extreme sensitivity.

Term Project:
While this one-semester course provides a solid foundation in data warehouses and big data, it is not exhaustive. The term project is intended to be an opportunity for you to further explore a topic from this course that is of interest to you. You will spend the first few weeks reviewing the topics and selecting one. The remaining weeks will be spent researching materials not already part of the curriculum and experimenting. The project submission will be a short report describing your research and sharing your findings, along with any successful code, design, project, etc. created during the experimentation.

Note on Lab assignments:
Labs will be graded using the following rubric:

	Letter Grade	Qualities Demonstrated by the Lab Submission	Grade Assigned
Answers and Methodology **Measures the correctness and completeness of the answers and methodology used for lab steps**	A+ ➔ 100	The answers, and answer justifications where required, are entirely complete and correct for all steps. The methodologies used to derive the answers are entirely applicable to the given problems, and are implemented correctly, for all steps. There are absolutely no technical or other errors present.	
	A ➔ 96	One insignificant technical or other error is present, but otherwise the answers, and answer justifications where required, are entirely complete and correct for all steps. Excluding the insignificant error, the methodologies used to derive the answers are entirely applicable to the given problems, and are implemented correctly, for all steps.	
	A- ➔ 92	One or two technical or other errors are present, but otherwise the answers, and answer justifications where required, are entirely complete and correct for all steps. Excluding the one or two errors, the methodologies used to derive the answers are entirely applicable to the given problems, and are implemented correctly, for all steps.	
	B+ ➔ 88	The answers, and answer justifications where required, are complete and correct for most steps. Likewise, the methodologies used to derive the answers are applicable to the given problems, and are implemented correctly, for most steps.	
	B ➔ 85	The answers are correct or almost correct for most steps. Some answer justifications may be missing or incorrect, but most are present and correct where required. The methodologies used to derive the answers are applicable and implemented correctly for most steps.	
	B- ➔ 82	The answers, and answer justifications where required, are complete and correct for about ¾ of the steps. Likewise, the methodologies used to derive the answers are applicable to the given problems, and are implemented correctly, for about ¾ of the steps.	
	C+ ➔ 78	The answers are correct or almost correct for about ¾ of the steps. Some answer justifications may be missing or incorrect. The methodologies used to derive the answers are applicable to the given problems, and are implemented correctly, for about ¾ of the steps.	
	C ➔ 75	The answers for about half of the steps are either missing or incorrect. Likewise, the methodologies used for about half of the steps are either inapplicable to the given problem, or are implemented incorrectly. Some answer justifications are missing or incorrect where required.	
	C- ➔ 72	The answers for most of the steps are either missing or incorrect. Likewise, the methodologies used for most of the steps are either inapplicable to the given problem, or are implemented incorrectly. Some answer justifications are missing or incorrect where required.	
	D ➔ 67	The answers for almost all of the steps are either missing or incorrect. Likewise, the methodologies used for almost all of the steps are either inapplicable to the given problem, or are implemented incorrectly. Some answer justifications are missing or incorrect where required.	
	F ➔ 0	The answers for virtually all of the steps are either missing or incorrect. Likewise, the methodologies used for virtually all of the steps are either inapplicable to the given problem, or are implemented incorrectly. Some or all answer justifications are missing or incorrect where required.	

Simple completion of specified laboratory tasks will earn an A grade (96 out of 100). To earn a higher grade, the student must demonstrate mastery of the task with additional work, for example:

- Variations in lab tasks that add substantially to the skills demonstrated
- Correct and insightful explanations of tasks in the lab
- Wrap-up essay effectively tying together the tasks of the lab, identifying goals and achievements in the lab work

Submission of work:

All labs and the term project will be submitted through the Assignments links in the Blackboard. The quizzes will be done online in the Assessments section of Blackboard. All work should be submitted by midnight of the day it is due. If an assignment or quiz will be submitted late let the instructor know as soon as possible, but at least by noon of the due date. Up to two assignments can be submitted late without penalty only if pre-approved by the instructor. Otherwise, there will be a 5-point penalty for each day an assignment or quiz is late. Quizzes can be up to three days late before a "0" grade is posted. Assignments can be up to five days late before a "0" grade is posted. Quiz and assignment grades cannot be released until either all students have submitted or the late period has expired.

CONCLUSION

Ask questions early and often. I check my email frequently throughout the day, including weekends. I do not have an office on-campus, but I can arrange to meet on-campus before or after class, or online any other day of the week through the Blackboard Live Office feature.

This syllabus is subject to change. Announcements of changes will be made as early as possible.