Designing and Implementing a Data Warehouse

MET CS 689

BLENDED FORMAT - Fall 2019

This course surveys state-of-the art technologies in DW and Big Data, and provides students with the engineering skills required to evaluate, implement, and scale a modern data warehouse using commercially available and open source software. It describes logical, physical and semantical foundation of modern DW infrastructure. Students will create a cube using OLAP and implement decision support benchmarks on Hadoop/Spark vs Vertica database. Students will do 6 two-week-long assignments and one final project.

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. Students will do 6 two-week-long assignments and one final project.

INSTRUCTOR

Mary E. Letourneau, Lecturer
maryleto@bu.edu
I am your instructor, Mary E. Letourneau. I have worked in the computer industry for over 30 years, starting with chip design and including consulting, programming, teaching, and for the last 12 years databases. I am currently employed as the Director of Information Systems. I earned my M.S. in Computer Information Systems from BU MET in 2015, and have been facilitating and/or teaching part-time for Boston University almost every semester since.
Office hours: by appointment

PREREQUISITES

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

MATERIALS

Required Books:

Kimball, Ralph and Ross, Margy. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3rd Edition. Indianapolis, IN: John Wiley & Sons, 2013. ISBN-13: 978-1-118-53080-1

Krishnan, Krish. *Data Warehousing in the Age of Big Data* , 1st ed., Krish Krishnan. Waltham, MA: Morgan Kaufmann, 2013. ISBN: 978-0-12-405891-0.

Optional:

McKinney, Wes. *Python for Data Analysis*. Second Edition. Sebastopol, CA: O'Reilly Media, 2013. ISBN-13: 978-1-491-95766-0.

COURSEWARE

Python https://docs.python.org/2/tutorial/
Python Pandas library: http://pandas.pydata.org/pandas-docs/stable/tutorials.html
Vertica: https://my.vertica.com/docs/5.1.6/HTML/index.htm - 8871.htm
Analytical functions in Vertica https://my.vertica.com/docs/5.1.6/HTML/index.htm - 10955.htm
Microsoft OLAP: https://technet.microsoft.com/en-us/library/ms170208(v=sql.100).aspx
Hadoop: http://hadoop.apache.org/docs/r2.7.3/hadoop-mapreduce-client/hadoop-mapreduce-clientcore/MapReduceTutorial.html

CLASS RESOURCES

This course will provide students with the following resources:

- 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 MEETINGS, LECTURES AND ASSIGNMENTS

Week	Description	Due / On
1: Sep 3 – Sep 9	Lecture 01: Introduction	Sep 3
	Reading: Module 1	Sep 10
	Reading: Kimball/Ross Chapter 1	Sep 10
	Optional reading: McKinney Chapter 1	Sep 10
	Install tools	Sep 10
2: Sep 10 – Sep 16	Lecture 02: Analytic Functions	Sep 10
	Assignment 1	Sep 17
	Term Project submission – Project description and plan	Sep 17
	Quiz 1	Sep 17
3: Sep 17 – Sep 23	Lecture 03: Dimensional Data Modeling	Sep 17
	Reading: Module 2	Sep 24
	Reading: Kimball/Ross Ch 2, 18	Sep 24
	Reading: Krishnan Ch 6, 7	Sep 24
4: Sep 24 – Sep 30	Lecture 04: Time, Bitemporality, Slowly-Changing Dimensions	Sep 24
	Assignment 2	Oct 1
	Quiz 2	Oct 1
5: Oct 1 – Oct 7	Lecture 05: Extract and Transform	Oct 1
	Reading: Module 3	Oct 8
	Reading: Kimball/Ross Ch 19, 20	Oct 8
6: Oct 8 – Oct 21	Lecture 06: Load and Verification	Oct 8
(No class Oct 15)	Assignment 3	Oct 22
	Quiz 3	Oct 22
7: Oct 22 – Oct 28	Lecture 07: Reporting	Oct 22
	Reading: Module 4	Oct 29
	Reading: Krishnan Ch 12 & 13	Oct 29
8: Oct 29 – Nov 4	Lecture 08: Forwarding Data to Further Stores and Uses	Oct 29
	Assignment 4	Nov 5
	Quiz 4	Nov 5
9: Nov 5 – Nov 11	Lecture 09: Big Data Approaches to Modelling	Nov 5
	Reading: Module 5	Nov 12
	Reading: Krishnan Ch 2, 3, 4 & 11	Nov 12
10: Nov 12 – Nov 18	Lecture 10: Dealing with Velocity, Volume, Variability	Nov 12
	Assignment 5	Nov 19
	Quiz 5	Nov 19
11: Nov 19 – Nov 25	Lecture 11: Alternative Storage for Big Data	Nov 19
	Reading: Module 6	Nov 26
	Reading: Krishnan Ch 8, 9	Nov 26
12: Nov 26 – Dec 2	Lecture 12: Performance Analysis and Tuning for Data Warehousing and Big Data	Nov 26
	Assignment 6	Dec 3
	Quiz 6	Dec 3
13: Dec 3 – Dec 9	Lecture 13: Course Wrap-Up and Final Exam Preparation	Dec 3
	Term Project	Dec 10
14: Dec 10 – Dec 16		
Dec 17	Final Exam	

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 %
Quizzes	30 %
Final Project	10 %
Final Exam	30 %
Total	**100 %**

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.	

Participation:

Participation includes asking questions, offering insights, sharing experiences, etc. relevant to the material being discussed. As such, participation implies attendance to lectures. Still 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 your grade, if notice is provided in advance.

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.

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 6 AM 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.