McCormick

Northwestern Engineering

Department of Electrical Engineering and Computer Science

MSIA 430 : Introduction to Data Management for Business Intelligence

Spring 2014 Quarter

Instructor:

Goce Trajcevski

Contact:
Office: Tech L360
Telephone: 491-7069
goce@eecs.northwestern.edu**:**

Course Objectives: This course will focus on the different roles of data, and different data-management techniques, in the context of improving the decision-making capabilities in business intelligence. Towards that, the course will aim at introducing the students to complementary "flavors" and tools, in varying degrees of their pure-insight vs. dynamics/reactivity. After a brief positioning of the role of data in the "process-centric world" of business tools (**Part I**) – addressing the Workflow Management models – the course will proceed with the following main categories of topics.

Part II:
We will study the Dimensional Fact Model (DFM) that provides the conceptual layer of a DW and then discuss a number of logical models that are used to represent a multidimensional data structures: ROLAP and MOLAP. Next, we will discuss the steps involved in populating a DW: ETL—Extract, Transform and Load. This part of the course will have homework and a project.

Part III:
While warehouses provide a varying degree of "views" on the data, the problem is that they are static – in the sense that one does need to "explicitly pull" the data of interest. In this part of the course, we will discuss in a tad-deeper manner the concept of reactive behavior and database triggers. In addition to the project, it is possible that another homework may be given from this part of the course.

Part IV:
This last "traditional" part of the course will take the reactive behavior to the "next-level". Namely, we will detach from the DBMS as the main carrier, and venture into the broader realm of Complex Events Processing. After defining and presenting examples of the necessary formalisms, we will introduce tools for CEP and define a project.

Part V:
The last part of the course will have a "potpourri" nature. We will go over topics such as: NoDB/NoSQL paradigm; In-memory Databases; etc… (TBD). We will try to "tune" the manner certain popular themes are perceived (e.g., are cloud-based/Hadoop-based system solutions always the right/best choice?). For this part of the course, there may be presentations "fusing" contemporary research papers and some industry white-papers/reports.

Required texts:
Not sure whether it would serve justice to the course and the intended educational purpose if any particular books are claimed as "required". Hence, instead, we can say that the "most-recommended" reference textbooks are:

The Data Warehouse Life Cycle Toolkit, 2nd Edition, by R. Kimball et. al, Wiley, 2008
Event Processing in Action, by O. Etzion and P. Niblett, Manning Publications, 2011

Other Recommended text and materials:

Data Warehouse Design: Modern Principles and Methodologies, by M. Golfarelli and S. Rizzi, McGraw-Hill, May 2009.
Multidimensional Databases and Data Warehousing, by C.S. Jensen, T.B. Pedersen and C. Thomsen, Morgan & Claypool, 2010.
Note: Handouts will be passed/posted for reading.

Tentative Outline of Topics

Part 1: Introduction and Motivation; Workflows	Business Process Modelling;
	Workflows and Workflow Management Systems;

Part 2: Data Warehousing	Introduction: DW and Event Processing for Business Intelligence. DW –requirements, basic architecture and life-cycle. Conceptual Modeling of DW: The DFM: facts, measures dimensions and cubes Events and Aggregation : additive, non-additive, aggregations with hierarchies Advanced Concepts: slowly-changing dimensions and dynamic hierarchies. Logical Modeling of DW: ROLAP versus MOLAP Star schemas and snowflake schemas View materialization and greedy algorithm for their selection ETL—Extract, Load and Transform: Immediate and delayed extraction, computing deltas Loading dimension, fact tables and populating materialized views Data Cleansing Real-time Business Intelligence with DW: Detecting changes with sentinels: a new data mining type In memory implementations for DW
Part 3: Databases and Reactive Behavior	Triggers Basic definitions Semantic Dimensions Relationship among “components” Good vs. Bad vs. Ugly Triggers and external world
Part 4: Complex Events Processing	Introduction: Event driven behavior and computing Business value of event processing Introductory Case-Study Event-Based Programming: Main concepts and architecture Processing Networks Defining Events: Attributes; Payloads; Relationships Producers vs. Consumers of Events: Hardware/Software/Humans Interfacing Event Processing Networks and Contexts Agents/Channels/States

	Application/Spatial/Temporal Context Filtering, Transformations and Patterns Detection: Basic Patterns Dimensional Patterns Policies
Part 5:	Potpourri topics (see “Course Objectives” above)

I. Workload and Evaluations

Your grades will be based on:

- **Programming Projects (~54%):** three projects, from Part II, III and IV. Projects will be done in teams
- **Homework Assignments (~6-10%)**— one or two HWs
- **Midterm (~27%):** will take place in week 7 of the course.
- **Presentations** (~10%): These will also be done in teams; each team will have 25min.s to present a mini-survey on a particular topic (list of topics will be posted during week#4). More detailed discussions in class.

***Please note*:** the distribution given above is approximate and may be subject to some very minor changes. However, the firm-policy will be announced during the last week of classes (Week #10 of the Spring quarter).

Awareness, Academic Responsibilities and Closing Remarks:

be advised that it is each student’s *individual responsibility* to keep him/herself up-to-date with the announcements *made in class*, *distributed via email*, or *otherwise posted*. If a particular homework or project is designated as an individual assignment: you are encouraged to discuss certain high-level aspects and/or design approaches with your classmates – however, most of the work needs to be done independently. Similarly for teams assignment – cross-teams discussion regarding high-level aspects is OK – but most of the work needs to be done within a particular team. The policies for cheating are well-defined and there will be no exceptions made for any excuse whatsoever – if caught cheating (both in terms of borrowing someone else’s code, as well as allowing someone to borrow your code), you will automatically fail the class and face a possible expulsion from the University. In addition, notwithstanding our willingness to be understanding for the students’ commitments and time-constraints, please do not attempt to obtain an incomplete grade for the course, based solely on your poor performance – it is against the University regulations.

Lastly, please note that a substantial part of your grade is based on the projects. Hence, you really need to keep yourself up-to-date with the material lectured and start working on the projects as early as possible. You should not allow yourself to fall behind with the topics, as the new ones will be building incrementally upon the older ones, and it will be very hard to catch up.