CS6999 Reading Course Proposal: Efficient Data Warehousing and OLAP over Text Winter 2011

December 1, 2010

This reading course covers data warehousing (DW) and online analytical processing (OLAP). It has two special foci:

- data structures and algorithms enabling OLAP and DW, and
- application of OLAP and DW to text analysis.

It will be offered by Owen Kaser with help from Daniel Lemire and is intended for MCS student Eduardo Gutarra. Its content will be divided into four major modules and there will be a project (with final presentation) that touches on three of these modules. One intended outcome is experience to help the instructor(s) subsequently develop a special-topics or regular graduate course on Data Warehousing/OLAP. While we have previously given special-topics courses that had a significant data warehousing/OLAP component, these courses were given several years ago and the DW/OLAP content was restricted to half a course.

Technology: Open-source or other no-cost software will be used, so that Eduardo can install and develop on his own machines. The plan is to use the Pentaho tools.

Major modules:

DW Data warehousing concepts. Brief idea of a "Warehouse of Words" (WoW). Materialized views (ETL exercise) 4 weeks.

TXT Getting information from text. (NLP basic concepts, named entity extraction, keyword extraction.) Storing this information in a DW. 2 weeks.

OLAP Detailed look at OLAP concepts and operations. Using tools such as Mondrian, JPivot, MDX and SQL. 3 weeks.

DS Data structures and algorithms enabling DW and OLAP. Indexing, esp. papers on bitmap indexing not included in Eduardo's thesis proposal, and comparisons to tree-based indexes. Analytics via Map/Reduce (map/reduce exercise) 3 weeks.

Structure: Class will meet for one hour per week to discuss readings, and for one hour per week to discuss progress on the project. A written summary (2 pages in UNB thesis format) of readings is required at the beginning of each reading-discussion meeting. Typically, two–three papers or book chapters per week will be covered.

The student should expect to put about 180 hours of high-quality effort into this course. Time spent on all aspects of the course should be recorded and will be discussed weekly at the project meeting.

Project: The project is to use the Pentaho tools and GATE to create cubes (and analyses) for a simple WoW. Details can be negotiated, but the anticipated project is

> Re-create a version of a sentence-stats and word co-occurrence cubes from [KKL06] (with rollup of words only by length and part-of-speech), with a jpivot web interface.

There will be a written report on the project (approximately 15 pages, including figures and references, with margins and fonts following UNB thesis regulations), and a final presentation of the project will be given as a half-hour open talk.

Grading scheme:

Class discussions & weekly summaries	20%
Project work (Depth, quality and quantity)	25%
Exercises: Map/Reduce, Kettle ETL	10%
Written report	25%
Presentation	20%

Papers, Books and Other Resources: Although somewhat dated, some of the lecture materials from our Spring 2006 offering of a Web Services and OLAP course (`http://pizza.unbsj.ca/~owen/backup/courses/OLAP-2006/lectures.html`) will be helpful.

"DW" Resources: Kimball has a series of books on data warehousing (most notably [KR02]) that may be a good overall guide to the topic. A reduced summary is his [KRT+08] and the course will (probably) cover his chapters 6–10; other chapters may be referenced. Our 2006 lectures "Introduction", "Data Warehousing", "DW Issues (by examples" (and Daniel's lecture notes for this) can be read for an overall picture.

Although somewhat dated, Johnson's DW chapter[Joh02] is a good overview. Multidimensional design is common to DW and OLAP, and two useful but older papers are [PJ01, C$^+$01].

Two more recent papers that will be examined deal with ETL [TP09] and how dimensional design can be partially automated by analyzing a set of anticipated user queries [RA10].

Materialized views are covered in the lectures "Views and their maintenance in DWs" and "Materialized view selection for DW". These draw from a few papers ([Gup97, Kot02]) that might not need to be read unless more detail is required.

"TXT" Resources: To see one tie between Data Warehousing and Text Analysis, see the papers related to Steven Keith's MCS work ([KKL06, KKL05]) and [MMTA10]. Authorship attribution is surveyed by Stamatotos [Sta09]. Pedersen [Ped07] argues that data warehousing needs to expand its applicability.

The course will cover a 2007 survey of text mining [SAN07], and if more information is required, a book is available for consultation [FS07].

The course will also cover a survey [PBAP08a] that considers incorporating information from text into the DW. One of the (relatively technical) surveyed papers is about "contextualized data warehouses" [PBAP08b]. The course may tackle this paper if time and interest permit.

"OLAP" Resources: Our 2006 lectures # 4 and 5 apply. Two classic papers are by Gray [GBLP96] and Codd [Cod93]. The course also includes two more recent papers that combine OLAP and text [RTTZ08, IT07]. Webb's recent work on the Diamond operator [WKL08] will be covered.

A book by Harinath and Quin [HQ05] on the MDX query language is locally available.

"DS" Resources: The end of our 2006 slides for lecture 4 covers some indexing topics. More bitmapped indexing papers include [OG95, BB10, JL01, DP10, FSV10].

A recent trend is to do large-scale analysis using the Map-Reduce approach [DG04, ABA$^+$09]; the course will finish with these papers.

Project References: There are a number of books that either are already in the library or are being ordered: [Bv09, CBv10, Gor09, HQ05, Pen].

The GATE tools have a website that seems to have much documentation [NLP].

References

[ABA+09] Azza Abouzeid, Kamil Bajda-Pawlikowski, Daniel Abadi, Avi Silberschatz, and Alexander Rasin. HadoopDB: An architectural hybrid of MapReduce and DBMS technologies for analytical workloads. In *VLDB'09*. ACM, August 2009.

[BB10] Ladjel Bellatreche and Kamel Boukhalfa. Yet another algorithms for selecting bitmap join indexes. In *Proceedings, DaWaK'10 (LNCS 6263)*, pages 105–116. Springer-Verlag, 2010.

[Bv09] Roland Bouman and Jos van Dongen. *Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL*. Wiley, 2009.

[C+01] Surajit Chaudhuri et al. Database technology for decision support systems. *IEEE Computer*, pages 48–55, December 2001.

[CBv10] Matt Casters, Roland Bouman, and Jos van Dongen. *Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration*. Wiley, 2010.

[Cod93] E. F. Codd. Providing OLAP (on-line analytical processing) to user-analysis: an IT mandate. Technical report, E. F. Codd and Associates, 1993.

[DG04] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In *OSDI 2004*, 2004.

[DP10] François Deliège and Torben Bach Pedersen. Position list word aligned hybrid: optimizing space and performance for compressed bitmaps. In *EDBT '10*, New York, NY, USA, 2010. ACM.

[FS07] Ronen Feldman and James Sanger. *The Text Mining Handbook*. Cambridge University Press, 2007.

[FSV10] F. Fusco, M. P. Stoecklin, and M. Vlachos. NET-FLi: On-the-fly compression, archiving and indexing of streaming network traffic. In *VLDB'10*, San Jose, CA, USA, 2010. VLDB Endowment.

[GBLP96] J. Gray, A. Bosworth, A. Layman, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-total. In *ICDE '96*, pages 152–159, 1996.

[Gor09] Will Gorman. *Pentaho Reporting 3.5 for Java Developers*. Packt Publishing, 2009.

[Gup97] Himanshu Gupta. Selection of views to materialize in a data warehouse. In *ICDT'97*, pages 98–112, 1997.

[HQ05] S. Harinath and S. R. Quinn. *Professional SQL Server Analysis Services 2005 with MDX*. Wrox, 2005.

[IT07] Akihiro Inokuchi and Koichi Takeda. A method of online analytical processing of text data. In *CIKM'07*. ACM, 2007.

[JL01] M. Jurgens and H. J. Lenz. Tree based indexes versus bitmap indexes: A performance study. *International Journal of Cooperative Information Systems*, 10(3):355–376, 2001.

[Joh02] Theodore Johnson. *Handbook of Massive Data Sets*, chapter Data Warehousing, pages 661–710. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[KKL05] Steven Keith, Owen Kaser, and Daniel Lemire. Analyzing large collections of electronic text using OLAP. Technical Report TR-05-001, UNBSJ CSAS, June 2005.

[KKL06] Owen Kaser, Steven Keith, and Daniel Lemire. The LitOLAP project: Data warehousing with literature. In *CaSTA'06*, 2006.

[Kot02] Yannis Kotidis. *Handbook of Massive Data Sets*, chapter Aggregate View Management in Data Warehouses, pages 711–741. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[KR02] Ralph Kimball and Margy Ross. *The data warehouse toolkit: The Complete Guide to Dimensional Modeling*. Wiley, 2nd edition, 2002.

[KRT+08] Ralph Kimball, Margy Ross, Warren Thornthwaite, Joy Mundy, and Bob Becker. *The Data Warehouse Lifecycle Toolkit*. Wiley, 2nd edition, 2008.

[MMTA10] M. J. Martin-Bautista, C. Molina, E. Tejeda, and M. Amparo Vila. Using textual dimensions in data warehousing processes. *Communications in Computer and Information Science*, 81:158–167, 2010.

[NLP] NLP Group, U. Sheffield. GATE, A General Architecture for Text Engineering. `http://gate.ac.uk/`.

[OG95] Patrick O'Neil and Goetz Graefe. Multi-table joins through bitmapped join indices. *SIGMOD Rec.*, 24:8–11, September 1995.

[PBAP08a] Juan Manuel Pérez, Rafael Berlanga, María José Aramburu, and Torben Bach Pedersen. Integrating data warehousing with Web data: A survey. *IEEE Trans. Knowl. Data Eng.*, 20(7):940–955, 2008.

[PBAP08b] Juan Manuel Pérez-Martínez, Rafael Berlange-Llavori, Maria José Aramburu-Cabo, and Torben Bach Pedersen. Contextualizing data warehouses with documents. *Decision Support Systems*, 45:77–94, 2008.

[Ped07] Torben Bach Pedersen. Warehousing the world—a few remaining challenges. In *Proceedings, DOLAP'07*, pages 101–102. ACM, 2007.

[Pen] Pentaho. Pentaho analysis service: Mondrian project. online: `http://mondrian.pentaho.org`. checked 2009-09-19.

[PJ01] Torben Bach Pedersen and Christian S. Jensen. Multidimensional database technology. *IEEE Computer*, pages 40–46, December 2001.

[RA10] Oscar Romero and Alberto Abelló. Automatic validation of requirements to support multidimensional design. *Data and Knowledge Engineering*, 69:917–942, 2010.

[RTTZ08] Franck Ravat, Olivier Teste, Ronan Tournier, and Gilles Zurfluh. Top_Keyword: An aggregation function for textual document OLAP. In *DaWaK'08*, pages 55–64. Springer-Verlag, 2008. LNCS 5182.

[SAN07] Anna Stavrianou, Periklis Andritsos, and Nicolas Nicoloyannis. Overview and semantic issues of text mining. *SIGMOD Record*, 36(23–34), 2007.

[Sta09] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

[TP09] Christian Thompsen and Torben Bach Pedersen. pygrametl: A powerful programming framework for Extract-Transform-Load programming. In *Proceedings, DOLAP'09*, pages 49–56. ACM, 2009.

[WKL08] Hazel Webb, Owen Kaser, and Daniel Lemire. Pruning attribute values from data cubes with diamond dicing. In *International Database Engineering and Applications Symposium (IDEAS'08)*, pages 121–129, 2008.