Cleveland State University

CIS 612 Modern Database Programming & Big Data Processing (3-0-3) – Fall 2014
Section 50 – Class Nbr. 2670. Tues, Thur – 4:00 – 5:15 PM

Prerequisites: CIS 505 and CIS 530. CIS 611 Preferred.
Instructor: Dr. Sunnie S. Chung
Office Location: BU346  Phone: 216 687 4732  Email: sschung.cis@gmail.com
Webpage: http://grail.csuohio.edu/~sschung
Office Time: Tues, Thurs 1:30 – 3:30 PM and 5:15 – 5:45 PM (or by appointment)
Class Location: BU 0207  Section 50  Tue & Thu 4:00-5:15 PM

Catalog Description: Detailed study of modern database programming and non-relational database systems for big data processing. First, the course presents practical approaches to modern database programming and its applications. The course advances with study of semi-structured/non-structured databases, XML data processing, XPATH, and XQuery. The course continues study of advanced features of modern databases with data warehouse and OLAP, data mining algorithms and applications. The course extends study with architectures and features of modern parallel computing systems for big data processing – Parallel Data Warehouse (PDW) with OLAP and Map Reduce with Hadoop. It continues an exploration of applications and tools of big data processing systems with HIVE, HBase, PigLatin, Sparks, and NoSQL.

Key Concepts: Modern database programming, web data processing, semi-structured/unstructured databases, XML data processing, XPath, XQuery, Parallel Data Warehouse (PDW), OLAP Cube, data mining algorithms, big data processing, Map Reduce, Hadoop, Hive, HBase, PigLatin, Sparks, NoSQL, and Cloud Computing.

Expected Outcomes: Upon successful completion of the course, the student will be able to create modern database applications to process non-traditional data - web data, semi-structured/unstructured data, XML data. The student will be able to design data warehouse, create OLAP Cubes and write analytical OLAP queries. The student further advances skills for modern database applications by obtaining data mining concepts and techniques to create data mining applications with the data warehouse system. The student will be able to have comprehensive understanding of problems of big data processing and approaches for solutions with modern database systems. The student will further extend knowledge with architectures and features of modern parallel computing systems for big data processing – Map Reduce and Hadoop. The student will be exposed to major tools to create Map Reduce applications for big data processing systems. Finally, the student will be able to obtain comprehensive knowledge and analytical skills to build an infrastructure for big data processing and its applications.

List of Required Materials:
Any RDBMS:
Oracle Database 11g or higher - This is available at http://www.oracle.com/us/downloads/index.html
Microsoft SQL Server,
Microsoft Visual Studio 2012 or any higher
Microsoft SQL Server 2012 Business Intelligence – OLAP Server, SQL Server Data Tool
They are available at the Microsoft Academic Alliance program: http://e5.onthehub.com/WebStore/ProductsByMajorVersionList.aspx?ws=31b9929b-c09b-e011-969d-0030487d8897

Text Book:

Supplement Text Book:
   Available at http://pages.cs.wisc.edu/~dbbook/openAccess/thirdEdition/supporting_material.htm

2. Oracle Database 11g PL/SQL Programming (Osborne ORACLE Press Series) by D.C.S. Michael McLaughlin

Official Calendar
Please consult the page http://www.csuohio.edu/enrollmentservices/registrar/calendar/index.html

Final exam: Tues, Dec 10 4:00-6:00 PM.

Term begins Aug 26  Commencement Dec 16
First Weekday Class Aug 26  Fall Incomplete Deadline May 3
Last Day to Add Sep 1  Labor Day (University Holiday) Sep 2
Last Day to Drop Sep 6  Columbus Day (University Holiday) Oct 14
Last Day to Withdraw Nov 1  Veterans Day (Tuesday no classes) Nov 12
Last Day of Classes Dec 6  Thanksgiving Recess Nov 28 – Dec 1
Final Exams Dec 9-14

Grading: The course grade is based on a student's overall performance through the entire Semester. The final grade is distributed among the following components:
1. Exams (Exam 1, 2 & Final) 55% (15% each Exam, 25% Final)
2. Computer Projects 35% (about 4 - 5 lab assignments)
3. Research Topic Presentation: 10%

A  94% +  A: Outstanding (student's performance is genuinely excellent)
A-  90% - 93%
B+  87% - 89%
B  80% - 86%  B: Very Good (student's performance is clearly commendable but not necessarily outstanding)
B-  70% - 79%
C  <70%  C: Good (student's performance meets every course requirement and is acceptable; not distinguished)
D  < 60%  D: Below Average (student's performance fails to meet course objectives and standards)
F  < 60%  F: Failure (student's performance is unacceptable)

Examination Policy: Students are allowed to bring to the tests a summary page (standard letter size) with their own notes. During the exams: (1) the use of books, cell phones, calculators, or any electronic devices is prohibited, and (2) students must not share any materials.

Make-Up Exam Policy: No makeup exams will be given unless notified and agreed to in advance. Requests will be considered only in case of exceptional demonstrated need.

Homework Policy: The students are expected to attend all classes. The students are responsible for collecting the notes, handouts and any other course material distributed during the class period. All assignments must be individually and independently completed and must represent the effort of the student turning in the assignment. Should two or more students turn in substantially the same solution or output, in the judgment of the instructor, the solution will be considered group effort. All involved in group effort
homework will receive a zero grade for that assignment. A student turning in a group effort assignment more than once will automatically receive an “F” grade for the course.

**Late Assignment:** All lab assignments are due at the beginning of class on the date specified. Laboratory Assignments handed in after the class has begun will be accepted with a 25% grade penalty for up to a week and then not accepted at all. All laboratory assignments must be completed. *Failure to do so will lower your course grade one additional letter grade.*

**Student Conduct:** Students are expected to do their own work. Academic misconduct, student misconduct, cheating and plagiarism will not be tolerated. Violations will be subject to disciplinary action as specified in the CSU Student Conduct Code. A copy can be obtained on the web page at: [http://www.csuohio.edu/studentlife/StudentCodeOfConduct.pdf](http://www.csuohio.edu/studentlife/StudentCodeOfConduct.pdf) or by contacting Valerie Hinton Hannah, Judicial Affairs Officer in the Department of Student Life (MC 106 email v.hintonhannah@csuohio.edu). For more information consult the following web page CSU Judicial Affairs available at [http://www.csuohio.edu/studentlife/jaffairs/faq.html](http://www.csuohio.edu/studentlife/jaffairs/faq.html)

**Course Schedule:** The schedule of topics and their order of coverage is given below. The schedule and topics to be covered may vary depending upon the progress made.

<table>
<thead>
<tr>
<th>Week of</th>
<th>Topic</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>Review: Database Foundations</td>
<td>Elmasri. Chp. 8,10,11</td>
</tr>
<tr>
<td></td>
<td>DBMS Architecture, SQL Query Processing</td>
<td>Lecture Notes</td>
</tr>
<tr>
<td></td>
<td>Complex Queries, Views</td>
<td></td>
</tr>
<tr>
<td>2, 3, 4</td>
<td>Database Programming:</td>
<td>Elmasri, Lecture Notes</td>
</tr>
<tr>
<td></td>
<td>Embedded SQL, Dynamic SQL, PHP</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stored Procedure,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>User Defined Function (UDF), User Defined Type (UDT),</td>
<td></td>
</tr>
<tr>
<td></td>
<td>User Defined Aggregate (UDA),</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Table Function,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Database Triggers</td>
<td></td>
</tr>
<tr>
<td>4, 5, 6</td>
<td>Modern Databases: Enhanced Data Models for Advanced Applications</td>
<td>Elmasri 12, 19, 20, 21</td>
</tr>
<tr>
<td></td>
<td>Semi Structured and Unstructured Databases:</td>
<td>Listed papers. Lecture Notes</td>
</tr>
<tr>
<td></td>
<td>XML Data Processing:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- XML Schema, Syntax/Semantics, Protocol</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- XPath</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- XQuery</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Introduction to Information Retrieval and Web Search</td>
<td></td>
</tr>
<tr>
<td>7, 8</td>
<td>Review of Query Processing Techniques and Operators:</td>
<td>Ramakrishnan – 11, 12</td>
</tr>
<tr>
<td></td>
<td>Join, Projection, Selection, Group By, Aggregations</td>
<td>Lecture Notes. Selected Papers</td>
</tr>
<tr>
<td></td>
<td>Review of Query Optimization Techniques</td>
<td></td>
</tr>
<tr>
<td>8, 9</td>
<td>Data Warehouse and OLAP:</td>
<td>J. Han Chap 1,2,3. Listed Papers, Lecture Notes</td>
</tr>
<tr>
<td></td>
<td>- Decision Support Technology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- On Line Analytical Processing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Star Schema</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- OLAP (Aggregation) Operators:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data Cube, Roll Up, Drill Down. OLAP Window Functions</td>
<td></td>
</tr>
</tbody>
</table>
• An Overview of Data Warehousing and OLAP Technology by Surajit Chaudhuri (Microsoft) and Umeshwar Dayal (HP Labs), in the proceedings of IEEE 1995
• Data Cube: A Relational Aggregation Operator Generalizing Group By, Cross Tab, and SubTotals by Jim Gray (Microsoft), et al, in the proceedings of IEEE 1996

9, 10, 11  Data Mining
   - Main Concepts
   - Support and Confidence
   - APRIORI Algorithm and Optimized Algorithms
   - Frequent Pattern Tree Algorithm
   - Association Rules
   - Decision Tree Algorithm
   - Lift

Advanced Topics for Data Mining Applications

12, 13, 14  Big Data Processing and Parallel Computing:

Google Map Reduce Tutorial
Apache Hadoop File System
Parallel Data Warehouse with OLAP Query Processing
Pig Latin on Apathe Hadoop by Yahoo and Apathe
Data Warehouse HIVE with Hadoop by Facebook
HBase
SPARKS
MongoDB
SQL vs NoSQL
Map Reduce Join Algorithms
Extended PDW for MR/Hadoop : Oracle, Teradata

14, 15, 16  Presentation of Significant Database Industry Research Papers on Big Data Processing: List of Selected Papers will be given in class.

Technical Topics: Select one and prepare a 25 min talk on the subject.

1. Semistructured/Unstructured Data Processing using Structured Data Model
   a. FaceBook
   b. EBay
2. Data Warehousing and Analytics Infrastructure at Facebook
3. Parallel Computing for Big Data Processing:
   • Google Map Reduce with Apache Hadoop
   • Parallel Data Warehouse (PDW)
   • Columnar Databases : SAP Hana
4. MapReduce: Simplified Data Processing on Large Clusters by Google
5. Lammal, Ralf, Google's MapReduce Programming Model Revisited.
6. Open Source Apache Hadoop
7. Pig Latin, Hbase, Hive
8. Map Reduce Join Algorithms,
9. Data Partition Techniques
10. SQL vs NoSQL
11. Processing MR/Hadoop with PDW : Oracle, Teradata
Works
13. Cloud Computing

....

More New Topics to come here.

NOTE: The instructor reserves the right to retain, for pedagogical reasons, either the original or a copy of your work submitted either individually or as a group project for this class. Students' names will be deleted from any retained items.