Prerequisites: IST 604

Instructor: Dr. Sunnie S. Chung

Office Location: BU346  Phone: 216 687 4732  Email: sschung.cis@gmail.com

Catalog Description: Detailed study of modern relational database systems and non-relational database systems for Enterprise Business Analytics. First, the course studies core components of relational database systems with the query processing concept and execution techniques, query optimization strategies of modern relational databases. Then the course advances with the study of semi-structured databases and non-structured data processing. The course continues with the study of the advanced features of modern Business Intelligence systems with Data Warehouse and OLAP, Data Mining algorithms and applications. The course continues with the study of the modern parallel computing systems for big data processing – Parallel Data Warehouse (PDW) with OLAP and Map Reduce with Hadoop. It continues with an exploration of big data processing systems and tools such as HIVE, HBase, PigLatin, Mongo DB, and VoltDB as SQL vs NoSQL. Finally, the course will explore current database industry research papers on big data processing and cloud computing.

Key Concepts: Query processing and execution techniques, analysis of query processing cost, query optimization strategies and performance, query rewrites, modern database systems, big data processing, semi-structured/unstructured databases, Parallel Data Warehouse (PDW), OLAP cube, data mining algorithms, Map Reduce, Hadoop, Hive, Hbase, PigLatin, NoSQL, and Cloud Computing.

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

List of Required Materials:
Microsoft SQL Server 2012,
Microsoft Visual Studio 2012 or any higher
Microsoft SQL Server Business Intelligence Data Analytic Tool 2012 – OLAP Server and SQL Server Data Tool
They are available at the Microsoft Academic Alliance program:
Adventure Works 2012 Data Warehouse Database for SQL Server 2012 – Will be directed in class to get this
Applied Analytics Using SAS Enterprise Miner – available in Lab 17A and will be directed in class to get this

**Text:**


2. List of Selected Database Industry Research Papers on Big Data Processing and Business Analytics will be given in class

**Supplement Text:**


**Official Calendar**

Please consult the page [http://www.csuohio.edu/enrollmentservices/registrar/calendar/index.html](http://www.csuohio.edu/enrollmentservices/registrar/calendar/index.html)

*Final exam: Wed, Dec 10  6:00-10:00 PM.*

<table>
<thead>
<tr>
<th>First Saturday Class begins</th>
<th>Aug 23</th>
<th>Veteran Day</th>
<th>Nov 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Weekday Class</td>
<td>Aug 25</td>
<td>Thanksgiving Recess</td>
<td>Nov 27 - 30</td>
</tr>
<tr>
<td>Last Day to Add</td>
<td>Aug 31</td>
<td>Last Day to Withdraw</td>
<td>Oct 31</td>
</tr>
<tr>
<td>Last Day to Drop</td>
<td>Aug 29</td>
<td>Last Day of Classes</td>
<td>Dec 5</td>
</tr>
<tr>
<td>Labor Day (University Holiday)</td>
<td>Sept 1</td>
<td>Final Exams</td>
<td>Dec 8 - 13</td>
</tr>
<tr>
<td>Columbus Day (University Holiday)</td>
<td>Oct 13</td>
<td>Commencement</td>
<td>Dec 14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fall Incomplete Deadline</td>
<td>May 1, 2015</td>
</tr>
</tbody>
</table>

**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 (Midterm & Final) 50% (20% Midterm, 30% Final)
2. Computer Labs 20% (about 3-4 Assignments)
3. 1 Project (15%)
4. Research Topic Presentations (2): 15%

<table>
<thead>
<tr>
<th>Grade</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>94% +</td>
</tr>
<tr>
<td>A-</td>
<td>90% - 93%</td>
</tr>
<tr>
<td>B+</td>
<td>87% - 89%</td>
</tr>
<tr>
<td>B</td>
<td>80% - 86%</td>
</tr>
<tr>
<td>B-</td>
<td>70% - 79%</td>
</tr>
<tr>
<td>C</td>
<td>&lt;70%</td>
</tr>
<tr>
<td>D</td>
<td>&lt;60%</td>
</tr>
<tr>
<td>F</td>
<td>&lt;50%</td>
</tr>
</tbody>
</table>

A: Outstanding (student's performance is genuinely excellent)
A-: Outstanding (20% Midterm, 30% Final)
B+: Good (student's performance is clearly commendable but not necessarily outstanding)
B: Good (student's performance meets every course requirement and is acceptable; not distinguished)
B-: Good (student's performance fails to meet course objectives and standards)
C: Acceptable (student's performance is unacceptable)
D: Below Average (student's performance is unacceptable)
F: Failure (student's performance is unacceptable)

Grading may vary depending on the overall class average.
Examination Policy: Students are allowed to bring to the final exam 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 or extra assignments 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 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

Course Schedule: The schedule of topics and their order of coverage is given below. Every effort will be made to follow the schedule, but topics covered may vary depending upon the progress made.

<table>
<thead>
<tr>
<th>Week of</th>
<th>Topic</th>
<th>Reading</th>
</tr>
</thead>
</table>
| 1, 2    | Review: Architecture of Modern RDBMS  
Overview of SQL Query Processing.  
The relational model of data. Attributes and Atomic domains, Key and  
Referential Integrity rules  
Relational Algebra  
Listed papers. Lecture Notes |
| 3, 4    | Advanced Topics in View  
Functional Dependencies and Normalization for Relational Databases.  
Redundancy and Abnormal behavior. Normal Forms, 1NF, 2NF, 3NF, BCNF.  
Inference Rules. Lossless-Join decompositions. Decompositions that  
Preserve Dependencies. | Lecture Notes  
Selected Papers |
| 5, 6    | File Structure, Disk Storage  
Disk Access Fundamental  
Index: Primary, Secondary, Clustered, Multi-Level Index  
B/B+ Tree  
External Sorting  
External Hashing Techniques  
Access Path | Lecture Notes,  
Elmasri – 17, 18  
Ramakrishnan – Chap 6, 12, 13 |
Evaluation of Relational Operators:
Projection, Join Types, Group-By, Aggregation
Query Processing Concepts
Query Execution Steps
SQL Query Processing Cost

Query Optimization Concept and Techniques
Advanced Query Optimization techniques for Complex Queries with:
Join Types
Correlated Subquery with/without Aggregation

- *Design Note for Correlated Subqueries and Exists Predicate* in DB2 IBM (DN-2208-02)

Query Rewrite Optimization

Data Warehouse and OLAP - Lecture Notes:
- Decision Support Technology
- On Line Analytical Processing
- Star Schema
- OLAP Aggregation Operators:
  Data Cube, Roll Up, Drill Down

- *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

Data Mining for Business Analytics
- Concept & Architecture
- Support and Confidence
- Association Rules
- Data Mining Algorithms:
  - Associate Rule Mining
  - APRIORI Algorithm and Optimization
  - Frequent Pattern Tree
  - Decision Tree
  Integrating association rule mining with relational database systems: Alternatives and implications by S. Sarawagi, et al (IBM Almaden Research Center) in the proceedings of SIGMOD'98
  - Lift and Other Correlation Measures

Enhanced Data Models for Advanced Applications: Semi-structured Databases and XML

- Semi-structured Databases
- XML Schema, Syntax, Semantics

Introducing to Information Retrieval and Web Data Processing
Data Models for Unstructured Data Processing

- Bigtable: A Distributed Storage System for Structured Data, by Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Google, Inc. in
Big Data Processing

Introduction of Google Map Reduce
Apache Hadoop File System
Pig Latin on Apache Hadoop by Yahoo and Apache
HIVE with Hadoop by Facebook
HBase by Facebook
Parallel Data Warehouse with OLAP Query Processing
Extended PDW with Map Reduce and Hadoop: Oracle, Teradata
Extended PDW with Columnar Data Processing: Teradata, Microsoft
Columnar Databases: SAP HANA Databases
ACID
SQL vs NoSQL
MongoDB
VoltDB

- MapReduce: Simplified Data Processing on Large Clusters by Jeffrey Dean (Google) and Sanjay Ghemawat (Google) *in the proceedings of OSDI 2004*
- Apathy Hadoop *in White Papers by Apache, Yahoo*
Data Warehousing and Analytics Infrastructure at Facebook by Ashish Thusoo, et al. (Facebook) *in the proceedings of SIGMOD 2010*

Presentation of Significant Database Industry Research Papers on Big Data Processing: List of Selected Papers for each topics will be given in class.

**Technical Topics:** Select one and prepare a 30 min talk on the subject.

1. Semistructured/Unstructured Data Processing using Structured Data Model
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) with Map Reduce and Hadoop
   - Columnar Databases: SAP HANA
4. MapReduce: Simplified Data Processing on Large Clusters by Google
6. *Open Source* Apache Hadoop
7. Pig Latin, Hbase, Hive
8. Map Reduce Join Algorithms,
9. Data Partition Techniques
10. Information Retrieval: How Google Search Engine Works
11. Cloud Computing

More New Topic Selections to come here.