COURSE CALENDAR DESCRIPTION

Storing, managing, and processing datasets are foundational to both computer science and data science. The enormous size of today’s data sets and the specific requirements of modern applications necessitated the growth of a new generation of data management systems, where the emphasis is put on distributed and fault-tolerant processing. New programming paradigms have evolved, an abundance of information platforms offering data management and analysis solutions appeared and a number of novel methods and tools have been developed. This course introduces the fundamentals of big data storage, retrieval, and processing systems. As these fundamentals are introduced, exemplary technologies are used to illustrate how big data systems can leverage very large data sets that become available through multiple sources and are characterized by diverse levels of volume (terabytes; billion records), velocity (batch; real-time; streaming) and variety (structured; semi-structured; unstructured). The course aims to provide students with both theoretical knowledge and practical experience of the field by covering recent research on big data systems and their basic properties. Students consider both small and large datasets because both are equally important and justify different trade-offs. Topics include: software frameworks for distributed storage and processing of very large data sets, MapReduce programming model, querying of structured data sets, column stores, key-value stores, document stores, graph databases, distributed stream processing frameworks. Prerequisites: Cumulative GPA of 4.50 or better over all major EECS courses (without second digit "5"), LE/EECS 3421 3.00, LE/EECS 3101 3.00

INSTRUCTOR(S)

<table>
<thead>
<tr>
<th>Name</th>
<th>Section / Format / Term</th>
<th>Contact Email</th>
<th>Contact Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Papangelis, Emmanouil</td>
<td>Sec. A / LECT / F</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ADDITIONAL INFORMATION

COURSE WEBSITE

https://www.eecs.yorku.ca/~papaggel/courses/eeecs4415/

TOPICS

* Data-driven organizations
* Data ingestion
* Data quality
* Data storage (data lakes, RDBMS, columnar DBMS, NoSQL, HDFS, Key-Value stores, object storage)
* Data definition (CAP theorem, schema-on-read, schema-on-write)
* Big data analytics architectures
* Batch processing
* Interactive query processing
* Data stream processing
* Unified processing engines
• Tools/systems for big data analytics (examples: OpenRefine, Apache Hadoop/MapReduce, Google BigTable/BigQuery, Twitter Storm/Huron, Apache Spark)

**TEXTBOOKS**

• There is no single text for this course. The course will rely mainly on the following suggested textbooks:
  
  **Note:** The aforementioned book is freely available online.
  
  
  
  **Note:** The aforementioned book is freely available online.

In addition, a number of recent research papers in the area of big data systems will be distributed every week.

**LECTURE 1. INTRODUCTION**

Introduction, administrivia.

*Readings:*
  - Mining of Massive Datasets, 2nd Edition (chapter 1)
  - Introduction to Data Mining (chapter 1)

**LECTURE 2. DATA-DRIVEN ORGANIZATIONS**

Data-driven organizations, DDO solutions reference model.

*Readings:*

**LECTURE 3. DATA INGESTION AND DATA QUALITY**

Data ingestion, ETL, data quality, data quality reference model, record linkage, entity resolution, string similarity, data quality scaling issues.

*Readings:*

**LECTURE 4. COMPUTING PLATFORMS AND STORAGE SYSTEMS**

Computing platforms, single-node computing, parallel computing, cluster computing, grid computing, data storage, data warehouse model, data lakes, data storage systems, relational DBMS, columnar DBMS, NoSQL, HDFS, Key-Value stores, object storage, software defined storage, CAP theorem, moving large data, data definition, schema-on-read, schema-on-write, big data analytics architectures, lambda architecture, kappa architecture.

*Readings:*
- Goetz Graefe (1993). Query Evaluation Techniques for Large Databases (ACM survey)

**LECTURE 5. PROCESSING SYSTEMS - BATCH PROCESSING**

Batch processing, Hadoop MapReduce.

*Readings:*

**LECTURE 6. PROCESSING SYSTEMS - STRUCTURED DATA (DREMEL/BIGQUERY)**

Structured data processing, Interactive query processing, Google Dremel/BigQuery.

*Readings:*

**LECTURE 7. PROCESSING SYSTEMS - STREAMING DATA (TWITTER STORM/HERON)**

Data stream processing, Twitter/Apache Storm, Twitter Heron.

*Readings:*
- Toshniwal, Ankit, et al. Storm@Twitter. SIGMOD, 2014.

**LECTURE 8. PROCESSING SYSTEMS - UNIFIED ENGINE (APACHE SPARK)**

Unified processing engines (Spark), Resilient Distributed Dataset (RDDs).

*Readings:*

**LECTURE 9. SERVING DATA**
Analytics reporting, business intelligence (BI) tools, OLAP cube, cuboids, ROLAP, MOLAP, in-application/real-time analytics, Serving at-Scale.

Readings:

LECTURE 10. COURSE REVIEW
Comprehensive course review.

ACADEMIC INTEGRITY LINKS
- Senate Policy on Academic Honesty - http://secretariat-policies.info.yorku.ca/policies/academic-honesty-senate-policy-on/
- Academic Integrity - http://lassonde.yorku.ca/academic-integrity

STUDENT LINKS
- Student Rights and Responsibilities - http://oscr.students.uit.yorku.ca/student-conduct
- Religious Observance - https://w2prod.sis.yorku.ca/Apps/WebObjects/cdm.woa/wa/regobs
- Student Accessibility Services (SAS) - https://accessibility.students.yorku.ca/
- York University’s Policies on Gender/LGBTQ*/Positive Space - http://rights.info.yorku.ca/lgbtq/

LAND ACKNOWLEDGEMENT
- We acknowledge our presence on the traditional territory of many Indigenous Nations. The area known as Tkaronto has been care taken by the Anishinabek Nation, the Haudenosaunee Confederacy, the Huron-Wendat, and the Métis. It is now home to many Indigenous Peoples. We acknowledge the current treaty holders, the Mississaugas of the New Credit First Nation. This territory is subject of the Dish With One Spoon Wampum Belt Covenant, an agreement to peaceably share and care for the Great Lakes region.
- The Indigenous Framework for York University: A Guide to Action can be found here: http://indigenous.info.yorku.ca/
- Meaning of a land acknowledgement: http://healthydebate.ca/opinions/indigenous-land-acknowledgements

Many courses utilize Moodle, York University's course website system. If your course is using Moodle, click here to access it.
Moodle @ York University