I535: Management, Access, and Use of Big and Complex Data
2017

Instructor: Dr. Inna Kouper inkouper@indiana.edu. Office hours: Friday 10:45-11:45am, Info East Room 256A

Associate Instructor(s): Isuru Suriarachchi isuriara@umail.iu.edu. Office hours: Wednesday 4 - 5 pm for residential students (Info East Room 256), by email for online students.

Course Summary and Goals

The course aims to develop a working knowledge of important topics affecting management and use big and complex data. Students in this course will be introduced to capabilities and benefits of big data, key components of big data projects, and major steps in data analysis and visualization.

The following concepts are covered in the course:

- Big data in science and business
- Data pipelines
- Complexity in software systems
- Modeling data storage in noSQL stores
- Data replication
- Distributed computing
- Data coding and cleaning
- Data provenance
- Data trustworthiness
- Economies of data sharing

This course is a “flipped course” in that a student views recorded lectures then has opportunity to reinforce their learning through discussions, assignments, and hands on activities. It is expected that a student will put in 6-7 hours a week every week into the course which includes time spent in readings, reflections, and engaging with instructional content.

Prerequisites

Students are expected to have undergraduate level expertise in computational thinking, but not a strong programming background. For students with weak backgrounds in databases and file systems, the following tutorials on MySQL and the Linux File System will be useful:
- MySQL Database Tutorial 1 - Introduction to Databases, Bucky Roberts, 2012
- MySQL Database Tutorial 2 - Getting a MySQL Server
- MySQL Database Tutorial 3 - Creating a Database
- MySQL Database Tutorial 4 - SHOW and SELECT
- MySQL Database Tutorial 5 - Basic Rules for SQL Statements
- MySQL Database Tutorial 6 - Getting Multiple Columns
Course Policies

This course follows Indiana University’s academic calendar for Fall 2017 and ends with submission of final grades Friday, December 15 2017.

Each unit has a small number of activities associated with it. Students are expected to carry out all activities within each unit.

Students are expected to maintain more than one copy of their projects and assignments in case of equipment failures (disk crashes, laptop malfunctions, and so on). **Equipment failure is not a sufficient reason for deadline extension.**

Students are expected to work throughout the week and do the readings and watch videos before the discussion date of each lesson. Online discussions open on Monday and continue through Sunday of the unit week. Participation in discussions is an obligatory part of the course.

Students should bring a laptop to weekly discussions.

Course Evaluation

Competency in the course is evaluated on a student’s engagement with and mastery of the content. This assessment is done through:

- **Class participation and interactions** (20% of the final grade). Evaluated on a 3-point scale through engagement in offline and online discussions and student-to-student interactions.
- **Assignments** (30% of the final grade, due Friday night of the unit week). Evaluated on a 3-point scale per question plus extra credit for relating class content to one’s own experience.
- **Peer review** (10% of the final grade, 3 peer reviews in Unit 1, 2, and 6). Evaluated on a 3-point scale and allows students to evaluate their peers and learn about assignment quality through comparison with their peers.
- **Projects** (40% of the final grade) are larger efforts that provide hands on experience with tools and data. 100 points for each project (A and B).

**Knowledge checks** are ungraded online activities that allow you to quiz yourself and make sure you have a good understanding of the corresponding unit.
Integrity of a College Student

Students are expected to conduct themselves in a manner befitting their status as a student of a respected and distinguished institution of higher education. In college courses, we are continually engaged with other people’s ideas: we read them in texts, hear them in lecture, discuss them in class, and incorporate them into our own writing. As a result, it is very important that we give credit where it is due. Plagiarism is using others’ ideas and words without clearly acknowledging the source of that information.

See http://wts.indiana.edu/pamphlets/plagiarism.shtml (Links to an external site.)Links to an external site. for help in addressing plagiarism in your own work.

Articles and data posted to Canvas for this course are for the convenience of students only and should not be shared outside the course.

Accommodations for Students with Disabilities

Every attempt will be made to accommodate qualified students with disabilities (e.g. mental health, learning, chronic health, physical hearing, vision neurological, etc.). You must have established your eligibility for support services through the appropriate office that services students with disabilities. Note that services are confidential, may take time to put into place and are not retroactive; captions and alternate media for print materials may take three or more weeks to get produced. Please contact your campus office as soon as possible if accommodations are needed. Find your office at: http://ada.iu.edu/students/index.shtml (Links to an external site.)Links to an external site.

Weekly Schedule

8/25/2017 Unit 0: Course Introduction and Overview

Syllabus overview, course requirements

Assignment: Review MySQL and Linux tutorials (see above)

9/1/2017 Unit 1: Introduction to Big Data in Business and Science

Key topics: big data, data science

Learning Objectives: Gain perspective of how society thought and talked about big data as it first entered our lexicon. Taken together, the readings highlight how differently industries and scientific disciplines see data challenges.

Readings:
“Data, data everywhere: A special report on managing information” by The Economist, February 25, 2010


9/8/2017 Unit 2: Data Processing Pipelines

Key Topics: Data is rarely show up ready to use. It undergoes numerous steps, some of which are end products in and of themselves. What is a data pipeline? What is data pipeline connection to big data? How is data pipeline used in science? Why does software complexity hurt when data gets big? How does business view a data pipeline? Role of cloud computing in the business view of data pipelines.

Learning Objectives:

- Gain understanding of the concept of a data pipeline, its connection to the lifecycle of data, and its use in science and business.
- Acquire basic understanding of the complexity of a software algorithm and its relationship to the number of steps in the algorithm.
- Apply the data pipeline concept to an example from students’ own experience. Student has opportunity to research a popular form of data pipeline in business today called the Amazon Web Service (AWS) Data Pipeline.

Readings and Resources:

- Primer on MapReduce / Hadoop http://readwrite.com/2013/05/23/hadoop-what-it-is-and-how-it-works
- Data Pipeline samples https://github.com/awslabs/data-pipeline-samples
UK Data Archive: the Research Data Lifecycle, [http://www.data-archive.ac.uk/create-manage/life-cycle](http://www.data-archive.ac.uk/create-manage/life-cycle) (Links to an external site.)

9/15/2017 Unit 3: Software Systems Design

**Key topics:** distributed systems, emergent behavior, tradeoffs in software system design

**Learning objectives:** Understand the general concepts of software systems used during design of the large software systems.


9/22/2017 Unit 4: Complexity in Software Systems

**Key topics:** complexity, layering, abstraction, modularity, hierarchy

**Learning objectives:** Examine the sources of that complexity, and design principles that are used to overcome the complexity.


9/29/2017 Unit 5: NoSQL data stores

**Key topics:** document, key-value, graph, and column stores

**Learning objectives:** This unit introduces large-scale data stores and the concept of data services. It gives an overview of noSQL data stores using a 2012 talk by Martin Fowler at the GoTo Aarhus Conference 2012.


**Project A:** Twitter Dataset Analysis and Modeling: Cleaning Twitter data, storing it to a noSQL data store, and using tools to map the user profile locations to a geospatial map.

10/13/2017 Unit 6: Comparison of Data Models through Example

**Key topics:** schema, data model

**Learning objectives:** While efficient location and retrieval of data from a distributed noSQL data stores is important, equally important is the flexibility that noSQL stores give you in storing data objects. Relational databases provide structured and
normalized tables for rapid and precise querying. NoSQL stores support less structured data but also less rich querying. This unit walks you through a comparison of data models on different storage systems (relational, document store, key-value pair, and column store data model) using real ecological data from a science research project at Indiana University.

Readings:

- Cassandra SET and MAP data types [http://www.Tutorialspoint.Com/cassandra/cassandra_cql_collections.htm](http://www.Tutorialspoint.Com/cassandra/cassandra_cql_collections.htm) (Links to an external site.)

**10/20/2017 Unit 7: Distributed File Systems**

*Introduction*

**Key topics:** transparencies, session semantics, fault tolerance, naming

**Learning objectives:** Gain understanding of distributed systems concepts that underlie today’s NoSQL stores, including such key concepts as transparencies, session semantics, fault tolerance, and naming.


**10/27/2017 Unit 8: Caching in Distributed Applications**

**Key topics:** caching, locality of reference, cache replacement strategy, cache coherency

**Learning objectives:** Caching of data is a key concept to efficient performance of a large distributed application. We turn back to the Levy and Silberschatz paper to see what the authors have to say about caching.


**11/3/2017 Unit 9: Fault Tolerance in Distributed Applications**

**Keywords:** stateful and stateless servers, idempotence, transactions
Learning objectives: When distributed systems span multiple locations or computers, the incidence of failure increases substantially. In this lesson we turn back to the Levy and Silberschatz paper to learn about fault tolerance.


Project B: Students have a choice between two options for this project. Option 1 – create a database of books and learn how to analyze and visualize their text. Option 2 – select a topic related to the concepts studied in this course and write a research report on it.

11/10/2017 Unit 10: Consistency in Distributed NoSQL Stores

Key topics: eventual consistency, CAP Theorem, Quorum for voting

Learning objectives: Understand how consistency of reads and writes becomes a paramount issue when storage in a data store is distributed across multiple storage devices.


11/17/2017 Unit 11: Data Cleaning and Coding

Key topics: data cleaning, missing data, quantitative coding. Tagging data, categorizing data, coding data, feature extraction

Learning objectives: Learn data cleaning through real use cases from environmental science and social science. Gains basic knowledge about coding data – purpose of coding and methodology for coding. Coding is at its most basic level the tagging or categorization of data on important features of the data so that themes emerge.


12/01/2017 Unit 12: Data Provenance

Key topics: Data provenance, causality graph, Open Provenance Model

Learning objectives: As data sharing increasingly moves from a friendly exchange between two parties that know each other to a transaction on an open data sharing market, the need also grows for data to carry with it sufficient information that the recipient can use to establish whether or not they trust and can use the data. Data provenance lies at the heart of the descriptive data needed to discern data
trustworthiness. This lesson will introduce data provenance and give you a sense of what provenance data is and how it is represented.


### 12/08/2017 Unit 13: Overcoming Social and Technical Barriers to Data Sharing

**Key topics:** data trustworthiness, data sharing, economies of data sharing

**Learning objectives:** Gain appreciation for the social and economic barriers to sharing data.

**Readings:**

- L. Borgman, “If Data Sharing is the Answer, What is the Question? (Links to an external site.)” Available at [https://ercim-news.ercim.eu/en100/special/if-data-sharing-is-the-answer-what-is-the-question](https://ercim-news.ercim.eu/en100/special/if-data-sharing-is-the-answer-what-is-the-question)

### 12/15/2017 Unit 14: Routing in NoSQL Data Stores

**Key topics:** routing, distributed hash tables, Chord, peer-to-peer, local versus global knowledge

**Learning objectives:** When data are stored across multiple computers in a single NoSQL data store, and the data store can be accessed any of the multiple servers that support the NoSQL data store, how does the NoSQL data store ensure that a request for a data object is routed to the right location where the data are stored? This lesson discusses ways of keeping track of the information needed to route requests to the right server.