INFO I416
Applied Cloud Computing for Data Intensive Sciences

Department of Human-Centered Computing
Indiana University School of Informatics and Computing, Indianapolis
Fall 2015

Credit Hours: 3
Prerequisites: INFO-I 123 Data Fluency and INFO-I 308 or equivalent courses.

COURSE DESCRIPTION

This course covers data science concepts, techniques, and tools to support big data analytics, including cloud computing, parallel algorithms, nonrelational databases, and high-level language support. The course applies the MapReduce programming model and virtual-machine utility computing environments to data-driven discovery and scalable data processing for scientific applications.

COURSE TOPICS

This course includes the following topics:

- Data intensive sciences and the data center model
- Clouds with infrastructure, platform, and software as a service
- Virtualization technologies and tools
- Introduction to FutureGrid (or Openstack) as an experimental testbed
- Parallel programming using MapReduce vs. MPI
- MapReduce and data parallel applications using Hadoop
- Iterative MapReduce and data mining algorithms using Twister (expectation maximization, clustering, multi-dimensional scaling, latent Dirichlet allocation, Bayes networks)
- MapReduce on multicore/graphics processing unit (CUDA)
- NoSQL databases (Google BigTable and Hadoop HBase) and parallel query processing
- High level language (Hive and Pig)
- Amazon EC2 and Microsoft Azure and their applications

Required Books

- *The Fourth Paradigm: Data-Intensive Scientific Discovery* (online)
- *The Datacenter as a Computer* (online)
- *Hadoop: The Definitive Guide* (Amazon)

References

- Hadoop Tutorial (online)
- MapReduce Tutorial (online)
- Big Data for Science Workshop at NCSA virtual summer school (VSCSE 2010), July 26-30, 2010

COURSE OBJECTIVES

This course will offer to students programming models and tools of cloud computing to support data intensive science applications. Students will get to know the latest research topics on cloud platforms and have the opportunity to understand some commercial cloud systems through projects using FutureGrid resources.

STUDENT LEARNING OUTCOMES

Upon completion of this course, students will	RBT	PULs	Assessments
1. Explain the main concepts, models, technologies, and services of cloud computing, the reasons for the shift to this model, and its advantages and disadvantages.	2	4	Assignment 0, 3 Final
2. Examine the technical capabilities and commercial benefits of hardware virtualization.	2	4	Assignment 1–2, 3 Final
3. Analyze tradeoffs for data centers in performance, efficiency, cost, scalability, and flexibility.	4	2	Assignment 0 Final
4. Explain the core challenges of cloud computing deployments, including public, private, and community clouds, in terms of privacy, security, and interoperability.	2	4	Assignment 3 Final
5. Create cloud computing infrastructure models.	6	1B	Assignment 4 Presentation
6. Demonstrate and compare the use of cloud storage vendor offerings, such as Amazon S3, Microsoft Azure, OpenStack, and Hadoop distributed file system.	4	2, 1B	Project 3 Assignment 5, 6 Presentation
7. Develop, install, and configure cloud-computing applications under software-as-a-service principles, employing Pig, Hive, and other cloud-computing frameworks and libraries.	6	1B	Assignment 5, 6 Presentation
8. Apply the MapReduce programming model to data analytics in informatics-related domains.	3	1B	Assignment 1–1, 5 Project 1, 2 Presentation
9. Enhance MapReduce performance by redesigning the system architecture (e.g., provisioning and cluster configurations).	6, 5	1B	Assignment 6 Project 3 Presentation

Principles of Undergraduate Learning (PUL):

Learning outcomes are assessed in the following areas:

1A.	Core communication: written, oral and visual skills	
1B.	Core communication: quantitative skills	***Major emphasis***
1C.	Core communication: information resources skills	
2.	Critical thinking	*Some emphasis*
3.	Integration and application of knowledge	
4.	Intellectual depth, breadth, and adaptiveness	*Moderate emphasis*
5.	Understanding society and culture	
6.	Values and ethics	

EXPECTATIONS, GUIDELINES, AND POLICIES

Attendance:

A basic requirement of this course is that you will participate in all class meetings, whether online or face-to-face, and conscientiously complete all required course activities and assignments. Class attendance is required for classroom-based courses. It entails being present and attentive for the entire class period. Attendance shall be taken in every class. If you do not sign the attendance sheet while in class, you shall be marked absent. Signing the attendance sheet for another student is prohibited. The instructor is required to submit to the Registrar a record of student attendance, and action shall be taken if the record conveys a trend of absenteeism.

Only the following are acceptable excuses for absences: death in the immediate family (e.g. mother, father, spouse, child, or sibling), hospitalization or serious illness; jury duty; court ordered summons; religious holiday; university/school coordinated athletic or scholastic activities; an unanticipated event that would cause attendance to result in substantial hardship to one's self or immediate family. Absences must be explained with the submission of appropriate documentation to the satisfaction of the instructor, who will decide whether missed work may be made up. Absences that do not satisfy the above criteria are considered unexcused. To protect your privacy, doctor's excuses should exclude the nature of the condition and focus instead on how the condition impacts your attendance and academic performance.

Missing class reduces your grade through the following grade reduction policy: You are allowed two excused or unexcused absences. Each additional absence, unless excused, results in a 5% reduction in your final course grade. More than four absences result in an F in the course. Missing class may also reduce your grade by eliminating opportunities for class participation. For all absences, the student is responsible for all covered materials and assignments.

Incomplete:

The instructor may assign an Incomplete (I) grade only if at least 75% of the required coursework has been completed at passing quality and holding you to previously

established time limits would result in unjust hardship to you. All unfinished work must be completed by the date set by the instructor. Left unchanged, an Incomplete automatically becomes an F after one year. http://registrar.iupui.edu/incomp.html

Deliverables:

You are responsible for completing each deliverable (e.g., assignment, quiz) by its deadline and submitting it by the specified method. Deadlines are outlined in the syllabus or in supplementary documents accessible through Oncourse. Should you miss a class, you are still responsible for completing the deliverable and for finding out what was covered in class, including any new or modified deliverable. In fairness to the instructor and students who completed their work on time, a grade on a deliverable shall be reduced 10%, if it is submitted late and a further 10% for each 24-hour period it is submitted after the deadline.

WEEKLY SCHEDULE

Class Schedule (Tentative)

Lecture	Topics	Readings	Assignments
1	1. Course introduction 2. Data Intensive Sciences 3. Data Center Model 4. Current Clouds with Infrastructure, Platform and Software as a Service	• The Fourth Paradigm: Data Intensive Discovery • The Data Center as a Computer • Above the Cloud • Distributed System and Cloud Computing (in preparation)	Assignment #0 • Cloud System Stack Correction
2	• Course Projects and Study Groups • Parallel programming/MPI vs. MapReduce/Hadoop • Introduction to FutureGrid • Using FutureGrid	• Hadoop • Overview of FutureGrid • Tutorial on using FutureGrid	Assignment #1, Part 1 • Hadoop word count
3	• Virtualization Technologies and tools • Build your own images to run on a bare cluster • Build your own images to run on academic cloud (e.g. Eucalyptus)	• Xen • KVM	Assignment #1, Part 2 • Hadoop word count on VM • Hadoop word count on Eucalyptus
4	Literature review: MapReduce, Hadoop, DryadLINQ, Twister	• MapReduce • Hadoop • DryadLINQ • Twister	

5	• MapReduce and data parallel applications • Hadoop • Building All-to-All Blast using Hadoop	• mpiBlAST • CloudBAST • CloudBurst • AzureBlast • TwisterBLAST	
6	• Iterative MapReduce and EM algorithms • Twister • Parallel data mining algorithms using Twister	• Pregel • Twister • Twister Kmeans	Project #1 • Hadoop Blast
7	• MapReduce and data parallel applications • DryadLINQ • Dryad • Building Pairwise distance Calculation using DryadLINQ/Dryad	• Dryad: Distributed data-parallel programs from sequential building blocks • Distributed Data-Parallel Computing Using a High-Level Programming Language • All-Pairs: An Abstraction for Data-Intensive Computing on Campus Grids • Cloud Technologies for Bioinformatics Applications	Assignment #2: • Hadoop word count on VM • Hadoop word count on Eucalyptus Assignment #3: • Reading the literature
8	Design your own project – Call for Proposals for term projects • Hadoop • DryadLINQ/Dryad • Twister • Eucalyptus (advanced topic) • Nimbus (advanced topic) • Sector/Sphere (advanced topic) • Virtual Appliances (advanced topic)	• Hadoop • DryadLINQ • Twister • Eucalyptus • Nimbus • Sector/Sphere	Assignment #4 • Project proposal
9	Data mining algorithms • Clustering by Deterministic Annealing (DAC) • Multi Dimensional Scaling (MDS) • Latent Dirichlet Allocation (LDA)	• Paralellized Variational EM for Latent Dirichlet Allocation • Blei's LDA implementation • Variational LDA Latent Dirichlet Allocation • Modern Multidimensional Scaling: Theory and Applications • GTM: The generative topographic mapping • Deterministic Annealing for Clustering , Compression, Classification, Regression, and Related Optimization Problems	Project #2 • Twister K-means

10	MapReduce on Multicore/GPU • Multicore/GPU architecture • Concurrent threading vs. parallel processes programming • Performance Issues	• Multi-Core Programming • Programming Massively Parallel Processors: A Hands-on Approach • Parallel Computer Works!	
11	• Google BigTable • Google Sawzall • Hadoop HBase • Hadoop Hive • Hadoop Pig	The Chubby lock service for loosely-coupled distributed systems • BigTable • Sawzall • HBase • Hive • Pig	Assignment #5 • Poster proposal
12	Midterm Review of term projects		
13	• Amazon EC2 and Microsoft Azure • Discussion of their applications	• Amazon EC2 • Azure	Project #3 • Hadoop/Twister Pairewise distance Calculation using SWG
14	Building Data Intensive Life Sciences applications using Azure, EC2, Eucalyptus, Nimbus and FutureGrid with comparison of Cloud/MapReduce and MPI technologies		Assignment #6 • Final Project report (including code and archives)
15	• Final Exam • Project Presentations		

Assessment:

Homework assignments, including surveys, in-class presentations, and reports		**30%**
Three Projects		**60%**
1. Hadoop	20%	
2. Twister	20%	
3. Dryad/DryadLINQ	20%	
Project extra credit		**5%**
Quizzes		**5%**
Participation		**5%**

Grading Scale:

A+	97 – 100	Outstanding achievement, given at the instructor's discretion
A	93 – 100	Excellent achievement
A–	90 – 92.99	Very good work
B+	87 – 89.99	Good work
B	83 – 86.99	Marginal work
B–	80 – 82.99	Very marginal work
C+	77 – 79.99	Unacceptable work (Core course must be repeated)
C	73 – 76.99	Unacceptable work (Core course must be repeated)
C–	70 – 72.99	Unacceptable work (Elective or core course must be repeated)
D+	67 – 69.99	Unacceptable work (Elective or core course must be repeated)
D	63 – 66.99	Unacceptable work (Elective or core course must be repeated)
D–	60 – 62.99	Unacceptable work (Elective or core course must be repeated)
F	Below 60	Unacceptable work (Elective or core course must be repeated)

CODE OF CONDUCT

All students should aspire to the highest standards of academic integrity. Using another student's work on an assignment, cheating on a test, not quoting or citing references correctly, or any other form of dishonesty or plagiarism shall result in a grade of zero on the item and possibly an F in the course. Incidences of academic misconduct shall be referred to the Department Chair and repeated violations shall result in dismissal from the program.

All students are responsible for reading, understanding, and applying the *Code of Student Rights, Responsibilities and Conduct* and in particular the section on academic misconduct. Refer to *The Code* > *Responsibilities* > *Academic Misconduct* at http://www.indiana.edu/~code/. All students must also successfully complete the Indiana University Department of Education "How to Recognize Plagiarism" Tutorial and Test. https://www.indiana.edu/~istd You must document the difference between your writing and that of others. Use quotation marks in addition to a citation, page number, and reference whenever writing someone else's words (e.g., following the *Publication Manual of the American Psychological Association*). To detect plagiarism instructors apply a range of methods, including Turnitin.com. http://www.ulib.iupui.edu/libinfo/turnitin

Academic Misconduct:

1. **Cheating:** Cheating is considered to be an attempt to use or provide unauthorized assistance, materials, information, or study aids in any form and in any academic exercise or environment.
   a. A student must not use external assistance on any "in-class" or "take-home" examination, unless the instructor specifically has authorized external assistance. This prohibition includes, but is not limited to, the use of tutors, books, notes, calculators, computers, and wireless communication devices.
   b. A student must not use another person as a substitute in the taking of an examination or quiz, nor allow other persons to conduct research or to prepare

work, without advanced authorization from the instructor to whom the work is being submitted.

c. A student must not use materials from a commercial term paper company, files of papers prepared by other persons, or submit documents found on the Internet.

d. A student must not collaborate with other persons on a particular project and submit a copy of a written report that is represented explicitly or implicitly as the student's individual work.

e. A student must not use any unauthorized assistance in a laboratory, at a computer terminal, or on fieldwork.

f. A student must not steal examinations or other course materials, including but not limited to, physical copies and photographic or electronic images.

g. A student must not submit substantial portions of the same academic work for credit or honors more than once without permission of the instructor or program to whom the work is being submitted.

h. A student must not, without authorization, alter a grade or score in any way, nor alter answers on a returned exam or assignment for credit.

2. **Fabrication:** A student must not falsify or invent any information or data in an academic exercise including, but not limited to, records or reports, laboratory results, and citation to the sources of information.

3. **Plagiarism:** Plagiarism is defined as presenting someone else's work, including the work of other students, as one's own. Any ideas or materials taken from another source for either written or oral use must be fully acknowledged, unless the information is common knowledge. What is considered "common knowledge" may differ from course to course.

   a. A student must not adopt or reproduce ideas, opinions, theories, formulas, graphics, or pictures of another person without acknowledgment.

   b. A student must give credit to the originality of others and acknowledge indebtedness whenever:

      1. directly quoting another person's actual words, whether oral or written;
      2. using another person's ideas, opinions, or theories;
      3. paraphrasing the words, ideas, opinions, or theories of others, whether oral or written;
      4. borrowing facts, statistics, or illustrative material; or
      5. offering materials assembled or collected by others in the form of projects or collections without acknowledgment

4. **Interference:** A student must not steal, change, destroy, or impede another student's work, nor should the student unjustly attempt, through a bribe, a promise of favors or threats, to affect any student's grade or the evaluation of academic performance. Impeding another student's work includes, but is not limited to, the theft, defacement, or mutilation of resources so as to deprive others of the information they contain.

5. **Violation of Course Rules:** A student must not violate course rules established by a

department, the course syllabus, verbal or written instructions, or the course materials that are rationally related to the content of the course or to the enhancement of the learning process in the course.

6. **Facilitating Academic Dishonesty:** A student must not intentionally or knowingly help or attempt to help another student to commit an act of academic misconduct, nor allow another student to use his or her work or resources to commit an act of misconduct.

OTHER POLICIES

1. **Right to revise:** The instructor reserves the right to make changes to this syllabus as necessary and, in such an event, will notify students of the changes immediately.
2. **IUPUI course policies:** A number of campus policies governing IUPUI courses may be found at the following link: http://registrar.iupui.edu/course_policies.html
3. **Classroom civility**: To maintain an effective and inclusive learning environment, it is important to be an attentive and respectful participant in lectures, discussions, group work, and other classroom exercises. Thus, unnecessary disruptions should be avoided, such as ringing cell phones engagement in private conversations and other unrelated activities. Cell phones, media players, or any noisy devices should be turned off during a class. Texting, surfing the Internet, and posting to Facebook or Twitter during class are generally not permitted. Laptop use may be permitted if it is used for taking notes or conducting class activities. Students should check with the instructor about permissible devices in class. IUPUI nurtures and promotes "a campus climate that seeks, values, and cultivates diversity in all of its forms and that provides conditions necessary for all campus community members to feel welcomed, supported, included, and valued" (IUPUI Strategic Initiative 9). IUPUI prohibits "discrimination against anyone for reasons of race, color, religion, national origin, sex, sexual orientation, marital status, age, disability, or [veteran] status" (Office of Equal Opportunity). Profanity or derogatory comments about the instructor, fellow students, invited speakers or other classroom visitors, or any members of the campus community shall not be tolerated. A violation of this rule shall result in a warning and, if the offense continues, possible disciplinary action.
4. **Bringing children to class:** To ensure an effective learning environment, children are not permitted to attend class with their parents, guardians, or childcare providers.
5. **Email:** Indiana University uses your IU email account as an official means of communication, and students should check it daily for pertinent information. Although you may have your IU email forwarded to an outside email account, please email faculty and staff from your IU email account.
6. **Disabilities Policy:** In compliance with the Americans with Disabilities Act (ADA), all qualified students enrolled in this course are entitled to reasonable accommodations. Please notify the instructor during the first week of class of accommodations needed for the course. Students requiring accommodations because of a disability must register with Adaptive Educational Services (AES) and complete the appropriate AES-issued

before receiving accommodations. The AES office is located at UC 100, Taylor Hall (Email: aes@iupui.edu, Tel. 317 274-3241). Visit http://aes.iupui.edu for more information.

7. **Administrative Withdrawal:** A basic requirement of this course is that students participate in all class discussions and conscientiously complete all required course activities and/or assignments. If a student is unable to attend, participate in, or complete an assignment on time, it is the student's responsibility to inform the instructor. If a student misses more than half of the required activities within the first 25% of the course without contacting the instructor, the student may be administratively withdrawn from this course. Administrative withdrawal may have academic, financial, and financial aid implications. Administrative withdrawal will take place after the full refund period, and a student who has been administratively withdrawn from a course is ineligible for a tuition refund. Contact the instructor with questions concerning administrative withdrawal.
8. **Emergency Preparedness:** Safety on campus is everyone's responsibility. Know what to do in an emergency so that you can protect yourself and others. For specific information, visit the emergency management website. http://protect.iu.edu/emergency

MISSION STATEMENT

The Mission of IUPUI is to provide for its constituents excellence in

- Teaching and Learning;
- Research, Scholarship, and Creative Activity; and
- Civic Engagement.

With each of these core activities characterized by

- Collaboration within and across disciplines and with the community;
- A commitment to ensuring diversity; and
- Pursuit of best practices.

IUPUI's mission is derived from and aligned with the principal components—Communities of Learning, Responsibilities of Excellence, Accountability and Best Practices—of Indiana University's Strategic Directions Charter.

STATEMENT OF VALUES

IUPUI values the commitment of students to learning; of faculty to the highest standards of teaching, scholarship, and service; and of staff to the highest standards of service. IUPUI recognizes students as partners in learning. IUPUI values the opportunities afforded by its location in Indiana's capital city and is committed to serving the needs of its community. Thus, IUPUI students, faculty, and staff are involved in the community, both to provide educational programs and patient care and to apply learning to community needs through service. As a leader in fostering collaborative relationships, IUPUI values collegiality, cooperation, creativity, innovation, and entrepreneurship as well as honesty, integrity, and support for open inquiry and dissemination of findings. IUPUI is committed to the personal and professional development of its students, faculty, and staff and to continuous improvement of its programs and services.