C S 71A: DATA ANALYTICS & MANAGEMENT

Foothill College Course Outline of Record

Heading	Value
Units:	4.5
Hours:	4 lecture, 2 laboratory per week (72 total per quarter)
Advisory:	MATH 10, C S 31A, C S 21A or 21B.
Degree & Credit Status:	Degree-Applicable Credit Course
Foothill GE:	Non-GE
Transferable:	CSU
Grade Type:	Letter Grade (Request for Pass/No Pass)
Repeatability:	Not Repeatable

Description

Introduction of Big Data ecosystems, tool infrastructure and industrial applications. Overview of the evolution, characteristics and significance of Big Data and the analytics process model. Hands-on exploration of Big Data solutions for specific industries. Concept topics include data management such as acquiring, cleansing and normalizing Big Data; application to log analytics, fraud detection, social media patterns, call centers, etc.; review of traditional SQL based Relational Database Management and issues with scaling when datasets are too big; methodology of NoSQL; big data technology infrastructures, such as the Hadoop framework and ecosystem components including Hadoop Distributed File Systems (HDFS), Hbase, MapReduce, Oozie, Pig and functionality used in Big Data; survey of tools in analytics and data visualization (DVT); survey of deployment patterns used in various industries.

Course Objectives

The student will be able to:
A. Harness "big data" solutions and tools and solve business problems.
B. Understand the MapReduce Programming Model.
C. Understand the Hadoop ecosystem and how they function within the stack.
D. Use a Hadoop File System (HDFS).
E. Understand how to integrate Hadoop within the production environment.
F. Set up and use MySQL database.
G. Set up a Hadoop cluster via platforms such as Cloudera Quickstart VM.
H. Use NoSQL database solutions such as DataStax.

Course Content

A. Big Data Concepts & Characteristics
1. Introduction to Data at Scale
2. Define "big" data
3. The impact of "big" data
4. Adoption and use cases
5. Major players in the ecosystem
B. Describe the evolution of how we store data and why
1. Structured, Unstructured and Semi-Structured data
2. Reasons to store unstructured data
C. Big Data & Analytics
1. Analytics & Analytical Model Requirements
2. Data Collection, Sampling, Preprocessing
3. Relational Databases & SQL
4. NoSQL Data Model & Solutions
5. Data Management Practice: MySQL (Server/Personal or AWS), HBase and/or Cassandra
D. Technology Infrastructure
1. MapReduce Programming Model
2. Basics of Hadoop Framework
3. Hadoop Distributed File System (HDFS) for large data sets
4. Hadoop Solutions including types of analyses powered by Hadoop and industry use cases for Hadoop
5. Hadoop ecosystem composed of Hive, Pig, Impala, HBase, Flume, Sqoop, Kafka, and Oozie
6. Spark clustering framework
E. Data Visualization
1. ETL (Extract, Transform, Load) Processing
2. Explore how data visualization helps in the analysis and understanding of complex data
3. Overview of Data visualization tools
4. Follow good design practices for visualization and various approaches for different data types

Lab Content

A. Implementation and report exploring a specific case study of a Big Data Industry Application.
1. Explore applications such as Banking, Healthcare, Energy, Retail, Media, Manufacturing, Insurance, Telecommunications.
2. Understand differences between specific case studies such as a Banking Application specific case study can be Fraud Detection Or Trade Surveillance or Customer Relationship Management.
B. Building a project to manage large data sets.
1. Learn how to download datasets.
2. How to load data from relational databases and other sources.
C. Develop in an ecosystem that handles large data.
1. Work with Hadoop ecosystem.
2. Use language syntax and data formats supported by related tools.
3. Manage data and export to other systems (for example, manage data in HDFS and export it for use with other systems).
D. Perform analysis tasks on large datasets.
1. Improve productivity for typical analysis tasks using Hadoop ecosystem tools.
2. Design and execute queries on data stored in HDFS.
3. Join diverse datasets to gain business insight.
4. Analyze structured, semi-structured, and unstructured data.
5. Compare techniques on how to store and query data for better performance.
6. Determine which tool is the best choice for a given task.

Special Facilities and/or Equipment

A. Access to a computer laboratory with Python and Java interpreters.
B. Website or course management system with an assignment posting component (through which all lab assignments are to be submitted) and a forum component (where students can discuss course material and receive help from the instructor). This applies to all sections, including on-campus (i.e., face-to-face) offerings.
C. When taught via Foothill Global Access on the Internet, the college will provide a fully functional and maintained course management system through which the instructor and students can interact.

D. When taught via Foothill Global Access on the Internet, students must have currently existing email accounts and ongoing access to computers with internet capabilities.

Method(s) of Evaluation

A. Tests and quizzes
B. Written laboratory assignments which include source code, sample runs and documentation
C. Final examination

Method(s) of Instruction

A. Lectures which include overview of the big data ecosystem, tool infrastructure, and industrial applications.
B. Online labs (for all sections, including those meeting face-to-face/on campus) consisting of:
1. A programming assignment webpage located on a college-hosted course management system or other department-approved Internet environment. Here, the students will review the specification of each programming assignment, submit their completed lab work and get feedback from the instructor.
2. A discussion webpage located on a college-hosted course management system or other department-approved Internet environment. Here, students can request assistance from the instructor and interact publicly with other class members.
C. Detailed review of programming assignments which includes model solutions and specific comments on the student submissions.
D. In person or online discussion which engages students and instructor in an ongoing dialog pertaining to all aspects of designing, implementing and analyzing programs.
E. When course is taught fully online:
1. Instructor-authored lecture materials, handouts, syllabus, assignments, tests, and other relevant course material will be delivered through a college-hosted course management system or other department-approved Internet environment.
2. Additional instructional guidelines for this course are listed in the attached addendum of CS department online practices.

Representative Text(s) and Other Materials

Baesens, Bart. Analytics in a Big Data World: The Essential Guide to Data Science and its Applications. Wiley and SAS Business Series. John Wiley & Sons, 2014.

Types and/or Examples of Required Reading, Writing, and Outside of Class Assignments

A. Reading

1. Textbook assigned reading averaging 30 pages per week

2. Reading the supplied handouts and modules averaging 10 pages per week

3. Reading online resources as directed by instructor though links pertinent to programming

4. Reading library and reference material directed by instructor through course handouts

B. Writing

1. Writing technical prose documentation that supports and describes the programs that are submitted for grades

Discipline(s)

Computer Science