San José State University
School of Information
INFO 208, Big Data Technologies
Sample Syllabus

Instructor*:*	TBD
Office Location:	Online
Telephone:	TBD
Email*:*	TBD
Office Hours:	TBD
Class Days/Time:	Online
Prerequisites:	Basic knowledge of Linux/UNIX programming. Some familiarity with C, C++, and/or Java. Advantageous: prior programming in Python or R

Course Format

This course is an offering of SJSU's School of Information, which offers all courses completely online. Home computing requirements are posted online for prospective students at http://ischool.sjsu.edu/current-students/technology-support/home-computing-environment. Students must meet those minimum requirements to participate in the activities for this course.

Canvas

In addition to course materials including the syllabus, handouts, notes, assignments, discussion threads, lectures, video presentations that will be provided within the Canvas Learning Management system, the syllabus for this course will be available online on the School of Information Syllabi section of its website: http://ischool.sjsu.edu/current-students/courses/syllabi. The syllabus is found by selecting the current semester and then the course number. Communication with the students, including email messaging, will take place through the Canvas Learning Management system and Blackboard IM.

Course Description

This Big Data Technologies course provides an introduction to the technological ecosystem of Big Data and Hadoop as well as introduces the participants to the goals and work of Data Science.

Course Target Audience

The course is part of the Advanced Certificate in Big Data. It is aimed at IT staff wanting to transition to the role of Data Engineer with a specialization in Big Data and to business users learning Data Science who need a technical and management introduction to the underlying technology to better understand how to implement a Big Data Strategy in a business organization.

The course is intended to move at a fast pace, providing foundational orientation and skills that lead to continuing education or more advanced courses in individual topics. Online lab work, and a small project, are integral to the course, as that is where practical experience in the technologies is gained and strengthened.

Learning Outcomes

Certificate Program Learning Outcomes

Upon successfully completion of the certificate in Big Data curriculum, students will be able to achieve the following Certificate Program Learning Outcomes (CPLO).

CPLO1. Be able to frame Big Data questions, formulate a strategy, and identify applicable technologies and techniques, form a data team, and understand the ethical considerations and risks of Big Data.

CPLO2. Be able to choose, use, and optimize technologies for loading large-scale data from disparate sources into a big data store, for integrating these heterogeneous datasets as well as for big data searching, querying, and analytical processing.

CPLO3. Be able to analyze, display, communicate, and interpret massive amounts of abstract data effectively and efficiently via visual representations.

CPLO4. Be able to choose, use, combine, and evaluate techniques and technologies appropriate for different big data mining tasks, including supervised and unsupervised learning, link analysis, and recommendation systems.

This course supports CPLO1 and CPLO2

Course Learning Outcomes (CLO)

Upon successful completion of this course, participants will be able to:

1. Recognize the importance of the use of Big Data to gain insight into business activities and understand how that insight can be used by companies to achieve competitive differentiation.

2. Develop an in-depth understanding of the open-source Apache/Hadoop ecosystem and its near-term future directions.

3. List the reasons for adopting Hadoop and the Hadoop ecosystem components.

4. Understand the basics of the Hadoop Distributed File System (HDFS), as well as competitive file system (FS) approaches.

5. Compare and evaluate the major Hadoop distributions and their ecosystem components, both their strengths and their limitations.

6. Adapt principles of Data Science to the analysis of data to gain insight into business operations and to solve business problems.

7. Apply ethical practices in everyday business activities and make well-reasoned ethical business and data management decisions.

8. Gain hands-on experience with key components of various Big Data ecosystem components and their roles in building a complete Big Data Solution to common business problems; learning the tools that will enable the participants to continue their big data education after the course.

9. Collaborate and network with other participants, and work under the mentorship of the instructor, to collectively and individually learn to work with Big Data Technology ecosystem systems and components.

Required Texts/Readings

Textbook

White, T. (2015). Hadoop: The definitive guide (4th, revised & updated ed.). Sebastopol, CA: O'Reilly Media. ISBN 978-1-491-90163-2. 727pp. Amazon: $24.25. [4th ed., or later, is important – earlier editions do not have some material.]

Spivey, B., & Echeverria, J. (2015). Hadoop security: Protecting your big data platform. Sebastopol, CA: O'Reilly. ISBN 978-1-695-90098-7. 320pp. Amazon: $43.53.

Zikopoulos, P. C., Eaton, C., deRoos, D., Deutsch, T., & Lapis, G. (2012). Understanding big data: Analytics for enterprise class Hadoop and streaming data. New York: McGraw Hill. Available for free download at: https://www.ibm.com/developerworks/vn/library/contest/dw-freebooks/Tim_Hieu_Big_Data/Understanding_BigData.PDF

Zikopoulos, P., deRoos, D., Bienko, C., Buglio, R., & Andrews, M. (2015). Big data beyond the hype: A guide to conversations for today’s data center. New York: McGraw Hill Education. Available for free download at: https://www-01.ibm.com/marketing/iwm/iwm/web/signup.do? source=sw-infomgt&S_PKG=ov28197&dynform=11707

Zhu, W.-D., Gupta, M., Kumar, V., Perepa, S., Sathi, A., & Statchuk, C. (2014). Building big data and analytics solution in the cloud. Poughkeepsie, NY: IBM ITSO (Redpaper).
Available for free download at:
http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/redp5085.html

O'Reilly Media. (2015). Big data now: 2014 Edition: Current perspectives from O'Reilly Media. Sebastopol, CA: O'Reilly Media. Available for free download at:
http://www.oreilly.com/data/free/files/big-data-now-2014-edition.pdf

Other textbooks TBD.

Other Readings

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. Retrieved from http://www.mckinsey.com/~/media/mckinsey/dotcom/insights%20and%20pubs/mgi/research/technology%20and%20innovation/big%20data/mgi_big_data_full_report.ashx

Various Apache Software Foundation (ASF) websites:

Overview: http://hadoop.apache.org and http://hadoop.apache.org/docs/current

MapReduce tutorial: http://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

WordCount 2: http://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Example:_WordCount_v2.0

Spark: http://spark.apache.org

Flume: http://flume.apache.org

Sqoop: http://sqoop.apache.org

Google's Paper on Big Table: http://research.google.com/archive/bigtable.html

Google's Paper on MapReduce: http://research.google.com/archive/mapreduce.html

FTC Report on Big Data: Tool for Inclusion and Exclusion: https://goo.gl/YgPCWv

Hadoop Governance White Paper: http://info.hortonworks.com/rs/549-QAL-086/images/Hadoop-Governance-White-Paper.pdf

Articles on specific topics:

Spark: https://databricks.com/spark & http://ibm.com/spark

Other electronic articles referenced during the course itself.

Library Liaison

Ann Agee, ann.agee@sjsu.edu, Phone: 408-808-2033

Course Requirements and Assignments

SJSU classes are designed such that in order to be successful, it is expected that students will spend a minimum of forty-five hours for each unit of credit (normally three hours per unit per week), including preparing for class, participating in course activities, completing assignments, and so on.

Instructional time may include but is not limited to:
Working on posted modules or lessons prepared by the instructor; discussion forum interactions with the instructor and/or other students; making presentations and getting feedback from the instructor; attending office hours or other synchronous sessions with the instructor.

Student time outside of class
In any seven-day period, a student is expected to be academically engaged through submitting an academic assignment; taking an exam or an interactive tutorial, or computer-assisted instruction; building websites, blogs, databases, social media presentations; attending a study group; contributing to an academic online discussion; writing papers; reading articles; conducting research; engaging in small group work.

More details about student workload can be found in University Policy S12-3 at http://www.sjsu.edu/senate/docs/S12-3.pdf.

This schedule and related dates/readings/assignments is tentative and subject to change with fair notice. Any changes will be announced in due time in class and on the course's web site in the Canvas Learning Management System. The students are obliged to consult the most updated and detailed version of the reading material and syllabus, which will be posted on the course's website.

Mod	Week	CLOs	Topics	Readings	Assignments
0	1	1, 3	Orientation & Introduction to Big Data	http://hadoop.apache.org White, Ch. 1 (pp. 1-17) Spivey & Echeverria, Ch. 1 (pp. 1-19) Manyika et al, pp 1-13 Lecture slides/notes	2
1	2-3	2, 4, 8	Hadoop & HDFS	White, Ch. 2-3 (pp. 19-78) Spivey & Echeverria, pp. 239-42 Lecture slides/notes	3
2	4	2, 4, 8	MapReduce & YARN	White, Ch. 4 (pp. 79-86) Lecture slides/notes	4
3	5	2, 8	Spark	White, Ch. 19 (pp. 549-74) https://databricks.com/spark Lecture slides/notes	5
4	6-7	2, 4, 6, 9	Data Formats & Data Movement	White, Ch. 14-15 (pp. 381-422) White, Ch. 20 (pp. 575-601) http://flume.apache.org http://sqoop.apache.org Zhu et al., Ch. 8 (pp. 77-83). Lecture slides/notes	6 + Start student project
5	8	6	The Role and Work of the Data Scientist	O'Reilly Media (2015), pp. 1-9; Various assigned articles on Internet; Lecture slides/notes	7 = Mid-term Quiz

Mod	Week	CLOs	Topics	Readings	Assignments
6	9-10	2, 4, 5, 8	Programming for Big Data	White, Ch. 16-17 (pp. 423-518) Lecture slides/notes	8
7	11	3, 5	The Hadoop Ecosystem	Lecture slides/notes	9

8	12-13	7	Data Governance & Data Security	Spivey & Echeverria, Ch. 2-3 (pp. 23-48)
				Hadoop Governance White Paper;
				FTC Report on Big Data;
				Lecture slides/notes
					10 = Research Paper

9	14	2, 8	(If time allows) Advanced & Additional Topics	TBD	TBD
10	15	9	Project presentations		11 = Project presentation in Canvas
	16	All	Final Exam (comprehensive)		12 = Final Exam

Grading Policy

Assignments and Grading Policy

Assignments: Each student is required to complete the following

Assignments	*CLOs*	*Possible Points*
Group discussion participation in Canvas	9	100 pts (generally 10 points max per occasion, but bonus points will also be available). (Note 1).
2. Download, install, and learn to use two (2) of Hadoop distributions (Note 2); Manage pseudo-cluster with Ambari	2	Installation of two different distributions required – running and documented (150 pts). Bonus point potential for additional, documented installations (i.e., other distributions and/or VM environments)
3. Simple HDFS commands; compile and run simplified WordCount program in Java	4, 8	50 pts
4. Compile, run, and monitor WordCount 2 under YARN	5, 8	50 pts
5. Simple program running under Spark	5, 8	50 pts

Assignments	*CLOs*	*Possible Points*
6. Exploration of different data formats, including HBase, JSON, etc.	4-6	50 pts
7. Mid-term Quiz	1-4	100 pts
8. Programming with Pig & Hive	2, 4, 5, 8	50 pts
9. SQL and SQL-like programming: HiveQL, Cloudera's Impala, & IBM's Big SQL	3, 5, 8	50 pts
10.Write paper on Security & Ethics related to Big Data	7	5+ page (double spaced) Research Paper on one aspect of Security and/or Ethics related to Big Data technologies and their use in a business environment (100 pts)
11.Data Project	9	Emphasis here is on the planning and documentation of a data science project (150 pts). Bonus points can be earned for more substantial completion. (Notes 4 & 5)
12. Final Comprehensive Exam	All	200 pts
Total Available Points		**1100**

Grading: There are a possible 1100 points for 10 discussion topics, 7 hands-on exercises, one research paper, one project, and one quiz and a final comprehensive exam. The following grading scale will be used:

Points Earned	**Range**	**Grade**
976 +	97-100%	A
850 – 975	94-96%	A minus
775 – 849	91-93%	B plus
748 – 774	88-90%	B
723 – 747	85-87%	B minus
697 – 722	82-84%	C plus

Points Earned	Range	Grade
672 – 696	79-81%	C
646 – 671	76-78%	C minus
621 – 645	73-75%	D plus
595 – 620	70-72%	D
570 – 594	67-69%	D minus
0 – 569	Below 67%	F

Notes and Appendices

1. Learning is always best done in a collaborative and peer team environment. Thus extra credit will be available to those participants who take the time and provide extra effort in helping others.

2. The Hadoop distributions that we will work with in this course are: Cloudera CDH 5.5, Hortonworks HDP 2.3, and IBM BigInsights 4.1 – or later, as available from each at the time of the course, running as VMware, Virtual Box, Docker Images, or in the Cloud. The participants need to become familiar with at least two of these (recommended, for variety, are Cloudera CDH and IBM BigInsights). Note that Hortonworks and IBM both use the Open Data Platform (ODP) approach.

3. As a number of lab projects require documentation as proof of completion and results achieved, documentation would be uploaded to Canvas in MS Word file format (.doc / .docx), including screen shots taken with HotShots or other screen capture software and annotated as appropriate. The description provided should enable another participant (and the instructor) to follow the steps in sequence. The uploaded MS Word files will be shared and available to other participants in the course.

4. The project can be done as a single participant or paired with another participant. This should be declared up-front. The project will be done in a series of stages, the first of which will be a two-page project proposal.

5. The main requirement of the project is that it exercises concepts covered in the class. Don't worry too much about covering a lot of the concepts things. It is also not critical that the project complete, since projects in Data Science require a lot of time in understanding the date and wrangling it to achieve your needs. It is more important for you to work on a project that you are excited about (and can excite the rest of us about, since we want you to share what you learn), or one that would be particularly useful. For example, in the latter case you might build a significant component of a project you are working on professionally (but, in that case, it must be one that you can share with us, of course).

Note that "All students have the right, within a reasonable time, to know their academic scores, to review their grade-dependent work, and to be provided with explanations for the determination of their course grades." See University Policy F13-1 at http://www.sjsu.edu/senate/docs/F13-1.pdf for more details.

University Policies

General Expectations, Rights and Responsibilities of the Student

As members of the academic community, students accept both the rights and responsibilities incumbent upon all members of the institution. Students are encouraged to familiarize themselves with SJSU's policies and practices pertaining to the procedures to follow if and when questions or concerns about a class arises. To learn important campus information, view University Policy S90–5 at http://www.sjsu.edu/senate/docs/S90-5.pdf and SJSU current semester's Policies and Procedures, at http://info.sjsu.edu/static/catalog/policies.html. In general, it is recommended that students begin by seeking clarification or discussing concerns with their instructor. If such conversation is not possible, or if it does not address the issue, it is recommended that the student contact the Department Chair as the next step.

Dropping and Adding

Students are responsible for understanding the policies and procedures about add/drop, grade forgiveness, etc. Add/drop deadlines can be found on the current academic year calendars document on the Academic Calendars webpage at http://www.sjsu.edu/provost/services/academic_calendars/. The Late Drop Policy is available at http://www.sjsu.edu/aars/policies/latedrops/policy/**.** Students should be aware of the current deadlines and penalties for dropping classes.

Information about the latest changes and news is available at the Advising Hub at http://www.sjsu.edu/advising/.

Consent for Recording of Class and Public Sharing of Instructor Material

University Policy S12-7, http://www.sjsu.edu/senate/docs/S12-7.pdf, requires students to obtain instructor's permission to record the course and the following items to be included in the syllabus:

- "Common courtesy and professional behavior dictate that you notify someone when you are recording him/her. You must obtain the instructor's permission to make audio or video recordings in this class. Such permission allows the recordings to be used for your private, study purposes only. The recordings are the intellectual property of the instructor; you have not been given any rights to reproduce or distribute the material."
    - It is suggested that the greensheet include the instructor's process for granting permission, whether in writing or orally and whether for the whole semester or on a class by class basis.
    - In classes where active participation of students or guests may be on the recording, permission of those students or guests should be obtained as well.
- "Course material developed by the instructor is the intellectual property of the instructor and cannot be shared publicly without his/her approval. You may not publicly share or upload instructor generated material for this course such as exam questions, lecture notes, or homework solutions without instructor consent."

Academic integrity

Your commitment, as a student, to learning is evidenced by your enrollment at San Jose State University. The University Academic Integrity Policy S07-2 at http://www.sjsu.edu/senate/docs/S07-2.pdf requires you to be honest in all your academic course work. Faculty members are required to report all infractions to the office of Student Conduct and Ethical Development. The Student Conduct and Ethical Development website is available at http://www.sjsu.edu/studentconduct/.

Campus Policy in Compliance with the American Disabilities Act

If you need course adaptations or accommodations because of a disability, or if you need to make special arrangements in case the building must be evacuated, please make an appointment with me as soon as possible, or see me during office hours. Presidential Directive 97-03 at http://www.sjsu.edu/president/docs/directives/PD_1997-03.pdf requires that students with disabilities requesting accommodations must register with the Accessible Education Center (AEC) at http://www.sjsu.edu/aec to establish a record of their disability.

Accommodation to Students' Religious Holidays

San José State University shall provide accommodation on any graded class work or activities for students wishing to observe religious holidays when such observances require students to be absent from class. It is the responsibility of the student to inform the instructor, in writing, about such holidays before the add deadline at the start of each semester. If such holidays occur before the add deadline, the student must notify the instructor, in writing, at least three days before the date that he/she will be absent. It is the responsibility of the instructor to make every reasonable effort to honor the student request without penalty, and of the student to make up the work missed. See University Policy S14-7 at http://www.sjsu.edu/senate/docs/S14-7.pdf.