San José State University
College of Science / Computer Science Department
Data BS Management Systems 2, CS 157B-01, Spring, 2017

Course and Contact Information

Instructor: Thanh Tran

Office Location: MacQuarrie Hall 216

Telephone: 924-7227

Email: ducthanh.tran@sjsu.edu

Office Hours: Monday 1:30 – 3:30, please drop me email with time info and subject

Class Days/Time: MoWe 9:00AM - 10:15AM

Classroom: MacQuarrie Hall 422

Prerequisites:

- Programming (CS046A or equivalent),
- Programming languages (at least one of the following: C, C++, Java or Perl),
- Data structures and algorithms (at least CS046B or equivalent, CS146 preferred),
- Database Management Systems (CS157a or equivalent)

Course Format

Faculty Web Page and MYSJSU Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found on my faculty web page at http://sites.google.com/site/kimducthanh and/or on Canvas Leaning Management System course login website at http://sjsu.instructure.com.

Course Description

In the new data-driven era, Big Data technologies have attracted much interest in theory and practice. This advanced data management course will explore novel concepts and cutting-edge technologies in the Big Data area, focusing on the application-related aspects of choosing the right Big Data technologies as well the effective usage, administration and optimization thereof. Since the technology landscape of this field is emerging, diverse and quickly expanding, the main goal of this course is to teach the fundamental concepts, while encouraging students to explore and understand the specific differences among various new solutions in their individual pursuits and experiments. Through technology blogging, tutorials, presentations and a semester-long project, students will discuss and compare existing solutions and also, acquire hands-on experiences with tools they will use to carry out at least three typical Big Data management tasks, including extract-transform-load, cleaning, integration, distributed storage and usage (searching, querying and analytical processing). Since different student teams will compare and experiment with a specific combination of technologies, this course also has a survey character, providing students a broad overview of the Big Data technology landscape. In particular, solutions covered include but are not limited to Sqoop, HCatalog, Talend, FRIL, Amazon S3, maprfs,

GFS, HDFS, Hadoop, HBase, Cassandra, CouchDB, Hive, Neo4J, HiveQL, Pig Latin, Jaql, Impala, Drill/Dremel, ElasticSearch, Solr, Cloudera Search and Blur.

Course Learning Outcomes (CLO)

This course aims to introduce students to the main theories, concepts and solutions for managing Big Data, focusing on the practical aspects of choosing the right Big Data technologies as well the effective usage, administration and optimization thereof.

In particular, the objective is to teach students about the main theories and concepts behind Big Data management as well as technologies widely used in this field:

- Concepts and applications of Big Data management technologies
- Different types of file and database storage solutions for Big Data
- Extracting, transforming, cleaning, integrating and loading Big Data
- Formalisms for querying and analyzing Big Data
- Big Data search solutions

Students will learn how 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.

Required Texts/Readings

Textbook

No textbook is required for the course.

Other Readings

Students may use the following textbooks to augment and reinforce their learning:

- Hadoop: The Definitive Guide, by Tom White
- MongoDB: The Definitive Guide by Kristina Chodorow
- Big Data: Principles and best practices of scalable realtime data systems, by Nathan Marz and James Warren

Course Requirements and Assignments

This course covers both the main theories and widely used technologies in Big Data management. The tentative list of covered topics and technologies is as follows:

- Big Data management use cases, challenges and technologies
- Overview of the Big Data management process
    - Extract-transform-load (ETL)
    - Data cleaning and integration
    - Distributed storage using Big Data file systems / databases
    - Querying (structured query languages), search (NL- and keyword-based) and analytical processing
- Data preprocessing, ETL, integration
    - Sqoop, HCatalog, Talend, FRIL
- Computing paradigms suitable for Big Data scenarios
    - Hadoop, Haloop, Dremel, Dryad

- Distributed file storages
    - Amazon S3, GFS, HDFS, Voldemort
- Big Data / NoSQL databases
    - HBase, Cassandra, MongoDB, CouchDB, Hive, Neo4J
- Querying formalisms
    - HiveQL, Pig, Jaql, Impala, Drill/Dremel, DryadLINQ
- Information retrieval
    - LucidWorks (Solr), ElasticSearch, , Cloudera Search, Blur, Graphinder

In additional, this course features a semester-long project. The project can be conducted individually or collaboratively (in teams with up to 3 members). Besides project examples and recommendations, students are encouraged to propose and pursue their own projects.

Project

This course is application-oriented, putting a strong emphasize on a semester-long project. The project shall be conducted collaboratively (in teams with 3 members) – or individually in exceptional cases. Besides project samples and recommendations, students are encouraged to propose and pursue their own projects.

Give a large amount of data, the task is to choose, apply and optimize Big Data technologies for one particular usage scenario. While student teams are encouraged to choose an ambitious scenario (with an end user application such as “the next Facebook” in mind), they are advised to devise a contingency plan. That is, every team might have an ambitious plan A (that they might want to pursue even beyond the scope of this course) and a plan B that precisely focuses on the core course requirements.

There are two main types of requirements:

- The first is task coverage: in their projects, students shall go through at least three main tasks typically involved in the Big Data processing pipeline, including ETL, cleaning, integration, distributed file / database storage and usage (searching, querying, and analytical processing).
- The second is technology coverage: for every such task, students shall work with a specialized solution such that in the end, the project showcases the usage of an interesting combination of Big Data technologies and tools.

The overall project quality is measured with respect to these requirements. That is, students should prepare for these questions:

- Does the project involve at least three main tasks?
- Does the project uses a different technology for every task?
- What is the quality of the task execution / outcomes?

In particular, every project shall cover at least 2 and focus on one of the following tasks:

- Extract, transform, load: students will identify one or several large datasets they want to use, extract them from existing sources / databases, transform them to the format needed and load them into a distributed file storage (DFS)
- Distributed file storage: students will perform basic data management operations provided by the DFS platform of choice

- Big data preprocessing and transformation: students will perform data cleansing and data integration operations (in parallel on DFS) to improve data consistency / quality
- Big Data management & querying: using a cutting / bleeding edge database solution for Big Data management, students will go through the steps of data loading, querying, and query optimization
- Searching: using a cutting / bleeding edge search engine for Big Data, users will query information via intuitive interfaces (keywords / NL)
- Analytical processing: students will use the advanced querying and data processing capabilities of the Big Data platform to process analytical workloads.

Note that in order to satisfy the course requirements, there is no need to build a fancy UI. In particular, there might be not enough time to develop a sophisticated Big Data application. However, teams are encouraged to do so and can get bonus credits, depending on the project outcomes.
Pointers to datasets to be considered for the project are:

- Social Computing Data Repository at ASU
- List of Free Public Datasets Available on the Web - FlipKarma Blog
- Publicly available large data sets for database research – Lemire's Blog
- Linked Data, CKAN
- NYC Open Data, see also datasets published by Chicago or Palo Alto
- Amazon Public Datasets

Ideas for projects:

- 5 Big Data Projects That Could Impact Your Life
- 3 big ideas for big data in the public sector
- Using OpenData to help drivers navigate NYC

The progress of the projects will be documented in a technology blog and discussed in class on a regular basis. In particular, the blog will contain a

- Project proposal (1-2 pages): title, team, idea, datasets to be used, main tasks to be addressed, technologies that might be relevant, project management (milestones, work allocation: who does what?)
- Mid-term report (4-5 pages): title, team, idea, datasets, solution architecture, component descriptions, work / experiments / evaluations done so far and check-point assessment: does the team proceed as planned? does every team member proceed as planned? what is the relative contribution of team member in percentage?
- Final-report (8-10 pages): same structure as mid-term report + main achievements + lessons learned

There will be several slots for teams to present and discuss their progress in class. Also, there will be a final project presentation.

The breakdown of the project score (60) is as follow:

- Project management, team communication, pacing (5)
- Project progress communication & discussion (5)
- Presentation (5)
- Project blog (5)
- Overall project quality (40)
    - Focused task (20 max)
    - Each additional task (10 max)

For more precise assessment, student's individual contributions to the project shall be documented in the report and correspondingly, reflected in the project presentation. Also, every student will be asked to provide an assessment of their own and other team members' contributions (check-point / final assessment).

Technology Survey Presentation / Blog

Every student shall prepare either a presentation or a written report (technology blog) for a particular technology. These presentations and reports are of the following two types:

- Technology overview: the goal of this is to provide an overview of the (1) motivation, (2) key concepts, (3) applications, (3) strengths and weaknesses, (5) to discuss its role in the landscape of Big Data technologies and (6) to engage the class in discussion.
- Tutorial: the goal here is to provide detailed instructions so that fellow students can follow and learn how to use the technology for a particular project task. Fellow students should be able to successfully complete the tutorial and acquire hands-on experiences from this.

In order to make the class more discussion-oriented, the presenter is encouraged to ask questions and to spur discussion. For this, try to avoid tough questions. As a rule of thumb, prepare questions that are more open to facilitate interesting exchange and class discussion.

While presentations are preferred, the number of available time slots will be not sufficient to accommodate all students. Thus, slots will be assigned on a first-come and first-served basis. Students without an assigned slot shall prepare a report that can serve the same purpose as the presentation.

For both the report and the presentation, the main goal is to provide an analysis of the key concepts and to cover the aspects as outlined above. In addition, students are encouraged to communicate their critical opinion, point out flaws, limitations, or unconventional applications of the studied technology. In particular, bonus credits will be available for one particular type of presentation / report:

- Critical analysis and comparison of technologies: this type requires information that goes beyond what are explicitly available. Students are expected to identify the criteria used for the comparison, critically analyze the technologies, derive conclusions and if necessary, run experiments to verify them.

The breakdown of the score (20) is as follow:

- Management / pacing (5): all reports / presentations require instructor's guidance and final approval; the highest score can be achieved when the result is delivered as requested by the instructor and requires the least amount of guidance
- Quality of the report / presentation (12)
- Discussion and involvement (3): students are expected to actively contribute to the technology discussion and complete all tutorials presented in class

Final Examination or Evaluation

There will be one midterm exam and one (non-comprehensive) final exam. The date of the midterm exam is subject to change. The final exam date is firm and cannot be changed. No make-up exams will be offered.

Grading Information

Determination of Grades

The tentative components of the final grade are:

- Project 60%
- Technology survey 20%
- Final Exam 20%

A more fine-grained breakdown of these components are given above.

Late progress will result in lower score (see scores for pacing etc.).

Each assignment, project, and exam will be scored (given points) but not assigned a letter grade.
Final individual class letter grades will be assigned based on the class curve.
Your final class grade can be adjusted up or down depending on your level and quality of class / project participation.

Classroom Protocol

Attendance is not required. Late attendance or any behavior (such as usage of cell phone) that leads to distraction should be avoided.

University Policies

Per University Policy S16-9, university-wide policy information relevant to all courses, such as academic integrity, accommodations, etc. will be available on Office of Graduate and Undergraduate Programs' Syllabus Information web page at http://www.sjsu.edu/gup/syllabusinfo/

Course Schedule

This schedule is subject to change with fair notice. Please check your Canvas (https://sjsu.instructure.com) for announcements.

Course Schedule

Week	Date	Topics, Readings, Assignments, Deadlines
1-2	02/01/17	Big Data Management – Introduction • Big Data: why, what, use cases • Big Data Management: RDBMS/OLTP, DWH/OLAP, the big picture, technology, architecture, projects • Course Organization: what, why, how, project, technology presentation
2-5	02/08/17	Machines and Parallel Computing Models for Big Data Computing • Parallel Computing Models: Master-Worker, Divide-Conquer, Pipelining...

Week	Date	Topics, Readings, Assignments, Deadlines

		• Parallel Computing Machines: multi-cores, multi-nodes, clusters/clouds/grids
		• Parallel Programming Models / Execution Environments
		  ○ Basic MapReduce (Hadoop)
		  ○ Iterations (Haloop, Spark)
		  ○ Complex programs (Dryad)
		  ○ Graph computations (Pregel, Graphlab)

6	03/06/17	Extract-Transform-Load • What? • Why? • Details: extraction, transform, load, architectures, tools
7	03/13/17	Distributed File System (DFS) • DFS main concepts, NFS • HDFS
7-9	03/15/17	High-Level Languages • Pig • Hive • JAQL

9-13	03/27/17	Introduction to DBMS for Big Data
		• History of DBMS
		• Scalability
		• Availability
		• Consistency
		• Overview of Big Data DBMS solutions
		  ○ Hadoop-based
		  ○ Key-value: big hash table with uninterpreted key-value, in memory, column-based key value
		  ○ Document-based
		  ○ Graph-based
		  ○ NewSQL

13-14	04/24/17	Key-value Stores • Big hash table with uninterpreted key-value: Dynamo, Voldemort • In-memory big hash table: Memcached, Redis • Column-based: BigTable, Hbase, Cassandra
14-15	05/01/17	Document-based Stores

Week	Date	Topics, Readings, Assignments, Deadlines
15	05/03/17	Graph-based Stores
16	05/08/17	NewSQL Solutions
16-17	05/10/17	Exam Preparation
Final	05/22/2016	7:15-9:30