CISC-525 Big Data Architectures

Course Description

Government, academia and industry have in recent years spent a great deal of time, effort, and money dealing with increases in the volume, variety, and velocity of collected data. Collection methods, storage facilities, search capabilities, and analytical tools have all needed to adapt to the masses of data now available. Traditional storage and computing methods were found to be insufficient to the task, and new tools were needed. Google paved the way for a new paradigm in Big Data, with two seminal white papers describing the Google File System, a distributed file system for massive storage, and MapReduce, a distributed programming framework designed to work on data stored in the distributed file system. Since those papers were published, various open-source and proprietary systems have been developed that implement the concepts of distributed, or cloud, computing designed to efficiently operate on Big Data.
This class will introduce the students to the concepts of Big Data, and describe the architectures, tools, and techniques that exist to work with data at scale. Practical exercises in using distributed file systems and the MapReduce programming framework will provide the students with skills applicable to developers and data scientists in any facet of industry. Students will learn about the ecosystem of software that supports Big Data, including NoSQL databases such as HBase and MongoDB, SQL-like languages such as Hive and Pig, distributed indexing tools such as Solr and ElasticSearch, and other tools designed to make Big Data accessible to the analyst, such as Sqoop, Oozie, Hue, and others.

Course Objectives

- Develop an understanding of the characteristics of Big Data.
- Be able to describe the architecture of a distributed file system and MapReduce framework.
- Demonstrate the ability to use a distributed file system and the MapReduce framework to store, access, and manipulate data.
- Understand the concepts of NoSQL databases, including identifying the different types of NoSQL databases and the use cases each type addresses.

- Use a NoSQL database to store, query, and analyze data.
- Understand the ecosystem tools and their uses, including the SQL-like languages, distributed indexes, and utilities.
- Complete a practical project that uses Big Data tools to analyze a chosen data set and reach a conclusion.

Prerequisites

A Bachelor of Science degree in Computer Information Systems, Computer Science, or related field.

Textbooks

- Hadoop Application Architectures, 1st Edition (ISBN 1491900083) (Required)
- Hadoop: The Definitive Guide, 3e, White, 2012 (Optional)
- HBase: The Definitive Guide, George, 2011 (Optional)

Additional Requirements:
Students must have access to sufficient computing resources to complete the assignments and projects for this class. This can include a laptop or desktop computer with at least 8GB of system RAM and 20GB of free disk space, or an equivalent virtual environment such as Amazon Web Services or similar.
All assignments and projects for the course can be completed using the free MapR Sandbox virtual environment (https://www.mapr.com/products/mapr-sandbox-hadoop), which contains a fully functional distribution of Hadoop and the MapReduce framework as well as many other ecosystem components, and can be run in the free VMWare Player or Virtual Box software.

Course Schedule:

Week	Topic	Assignment
1	**Introduction** • What is Big Data? • Architecture of a Big Data system. • Introduction to the Hadoop File System **Reading** • Google File System paper • Textbook Chapter 1	**Assignment 1:** Compare and contrast Hadoop distributions. **Discussion Questions**
2	**The MapReduce Framework** • The advantages of distributed programming • Data locality • Performance and communication costs **Reading** • Google MapReduce framework paper • Textbook Chapter 2	**Discussion Questions**
3	**MapReduce Programming** • Creating a MapReduce Job • The Map function • The Reduce function	**Assignment 2:** Simple MapReduce program

	• Streaming MapReduce **Reading** • Textbook Chapter 3	**Discussion Questions**
4	**Introduction to NoSQL** • The 3 classes of NoSQL Databases • Strengths and weaknesses **Reading** • Textbook Chapter 4	**Discussion Questions**
5	**HBase – a NoSQL Columnar Store** • Schema Design • Storage methodology • Operations • MapReduce and HBase **Reading** • Textbook Chapter 5	**Assignment 3:** Using HBase **Discussion Questions**
6	**SQL-Like Languages – Hive, Pig, Drill** • Strengths and weaknesses • Comparison of functionality **Reading** • Textbook Chapter 6	**Discussion Questions**
7	**Hive – An In-depth view** • Data Definition Language • Data Manipulation Language • Metastore and Catalog **Project Assignment**	**Assignment 4:** Using Hive **Discussion Questions**

	Create a presentation outlining the project concept, including a description of the data, ETL mechanism, analytic logic, and desired outcome. **Reading** • Textbook Chapter 7	
8	**Midterms**	**Midterm Exam**
9	**Big Data Tools and Use Cases** • Discussion of commercial tools for Big Data. • Discussion of analytic use cases. • Demonstration of a Big Data analytic framework. **Reading** • Textbook Chapter 8	**Discussion Questions**
10	**Distributed Indexing – Solr and ElasticSearch** • Features and Use Cases • Comparison of approaches • Integration with Big Data systems **Reading** Textbook Chapter 9	**Assignment 5:** A Solr Search **Discussion Questions**
11	**Ecosystem Tools – Sqoop, Oozie, Hue, and More** • Data import/export with Sqoop • Workflow management with Oozie • User experience with Hue **Reading** • Textbook Chapter 10 **Project Progress Review**	**Discussion Questions**

12	**Big Data Analysis** • Tools and techniques • Visualization **Reading** • Textbook Chapter 11	**Discussion Questions**
13	**Project Turn-in** See Term Project section below for project deliverables	**Term Project Due**
14	**Finals**	**Final Exam**
15	**Wrap Up** • Return final exam • Return Term Project • Final discussions	**None**

Term Project:

This course's term project is intended to allow the students to experiment with Big Data tools and techniques using a data set of their own choosing. Students will select a data set, design a methodology to import the data into a distributed file system, use one or more tools to analyze the data, and produce a report.

- **Data Selection:** Students will choose a source of data to use. Any source and type of data is acceptable as long as there is sufficient data to produce a meaningful result. Students will be provided with examples and suggestions of data sources along with the term project assignment.
- **Data Ingest:** Students will design a process to acquire the data and deliver it to the distributed file system. This can include creating and using a NoSQL database such as HBase, or using Hive, Pig, Sqoop, or other tool to store the data.
- **Data Analytics:** Students can use MapReduce, Hive, Pig, Drill, or other tools to design a data analytic to examine the data. The output of the analytic will serve as the basis for a report or visualization about the data.
- **Deliverables:** The following items will be turned in for grading:
    - **Initial Presentation:** A short (~5 slides) presentation, with notes, describing the data set that will be used and notional processes for ingesting and analyzing the data.
    - **Final Report:** A longer presentation (15-20) slides, with accompanying notes, giving details of the entire process, from data selection, to ingest, to analysis, to the final output of the process.

- **Timeline:** The initial presentation will be due at the end of Week 9. A progress review and question and answer period will take place during the weekly live session in Week 11. The final presentation will be due at the end of Week 13.

Course Guidelines:

The following guidelines will ensure a smooth and productive educational experience for everyone.

- Harrisburg University's **Moodle** software will be the platform for course materials, assignments, tests, and course discussion.
- The class will meet **once per week** in an online live meeting.
- **Attending live sessions** and actively participating in the discussions is highly recommended, as that is where the basic concepts of each unit are explained, assignments are discussed, and your questions answered.
- Weekly assignments are expected to take approximately **3 to 4 hours** to complete. This is separate from the work on the term project.
- **You are responsible for all the readings**, even if the material is not explicitly covered in class. You should read the class materials prior to class and be prepared to discuss and ask questions about the readings and assignments.
- You should also **re-read the material** after class as not every topic will be covered during class time. Many passages in the text may need to be read several times to gain clarity. Also, taking notes on the material you are reading and reflecting on the reading and these notes will help you better understand the issues, concepts and techniques that are being presented.
- All work must be completed and turned in on or before the due date. **Late work will result in lowered score. No late work will be accepted after one week beyond the due date.** Note that a computer's failure is not an excuse (it represents poor planning on your part).
- Your work should be properly referenced and adhere to standards of both academic integrity and proper form. Generally, the APA style should be used. (See http://www.apa.org/).
- All class-related electronic mail must be done using **Harrisburg's electronic mail service** and the student's assigned Harrisburg University ID.
- All assignments are to be completed **individually**, unless otherwise specified.

HU Core Competencies:

At the conclusion of this course a student will have met the following core competencies that reflect HU's mission:

- **Critical Thinking and Problem Solving** skills are demonstrated by the student’s ability to:
  Identify and clarify the problem**,** Gather information, Evaluate the evidence, Consider alternative solutions, Choose and implement the best alternative.
- **Communications** skills are demonstrated by the student’s ability to:
  Express ideas and facts to others effectively in a variety of formats, particularly written, oral, and visual formats, Communicate effectively by making use of information resources and technology.
- **Teamwork and Collaboration** - The students will be working with others to increase involvement in learning and by sharing one's own ideas and responding to others' reactions to sharpen thinking and deepen understanding.
- Information Technology - The students will be making effective use of the information resources and technology.
- **Competency Assessment:** The term project in this class will be assessed to evaluate your level of proficiency in the HU core competencies directly connected to that assignment. (http://www.harrisburgu.net/academics/core-competencies.php) This competency assessment will not impact your grade in this course, but can be used as a gauge for you to self-evaluate your progress in developing your skill level in specified core competencies attached to the assignment. This additional evaluation can be a point of discussion between you and your academic advisor.

Statement on Academic Integrity:

According to the University's Student Handbook: Academic integrity is the pursuit of scholarly activity free from fraud and deception, and is the educational objective of this institution. Academic dishonesty includes, but is not limited to cheating, plagiarism, fabrication of information or citations, facilitating acts of academic dishonesty by others, unauthorized possession of examinations, submitting work of another person, or work previously used without informing the instructor, or tampering with the academic work of other students. Any violation of academic integrity will be thoroughly investigated, and where warranted, punitive action will be taken. Students should be aware that standards for documentation and intellectual contribution may depend on the course content and method of teaching, and should consult the instructor for guidance in this area.

Honor Code **- We as members of Harrisburg University community pledge not to cheat, plagiarize, steal, or lie in matters related to academic work. As a Community of Learners, we honor and uphold the *HU Honor Code*.**

Last modified: Thursday, 7 Jan 2016, 4:05 PM