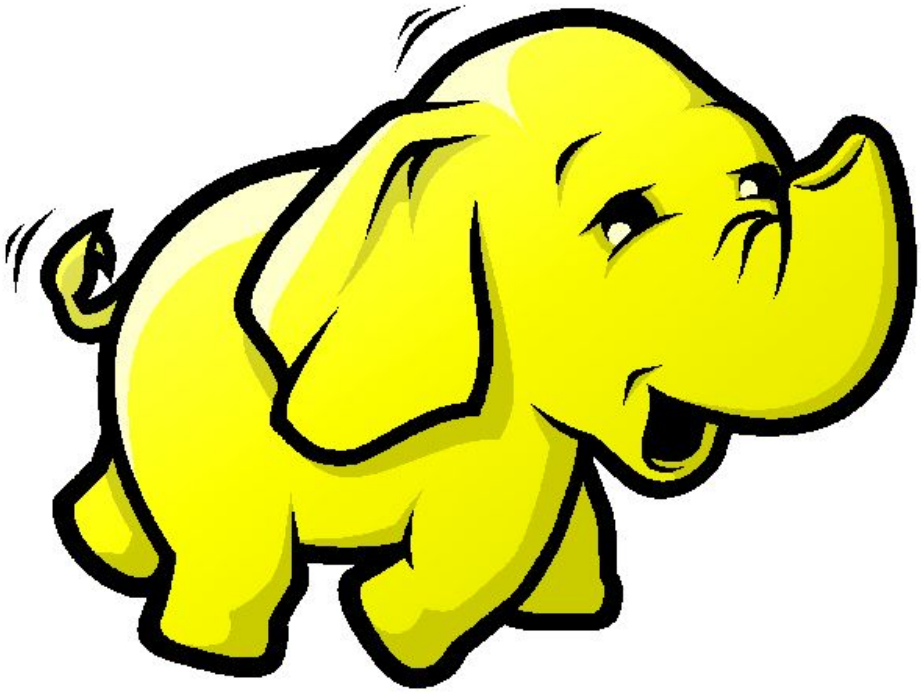
Information Packet

CSSE 490

Introduction to the Hadoop Ecosystem
Fall 2015



Computer Science and Software Engineering
Rose-Hulman Institute of Technology

Computer Science and Software Engineering 490
Introduction to the Hadoop Ecosystem
Fall 2015

Instructor: Sriram Mohan

Office: Moench Hall, Room F226

Phones: (work) 877-8819**;** (cell) 812-219-9658

Email: mohan@rose-hulman.edu

Office Hours: I am usually in my office. Just stop by when you have questions.

Course Prerequisite: CSSE 230 **(**Data Structures & Algorithms) or equivalent

Course Description: CSSE 490 provides a hands-on introduction to the various technologies and tools that make up the Hadoop ecosystem. Hadoop is the data analysis platform of choice for most organizations these days. Here are a few of the topics that will be covered:

1. Overview of the Hadoop ecosystem
2. Internals of MapReduce and the Hadoop Distributed File system (HDFS)
3. Internals of the Yarn distributed operating system
4. MapReduce for data processing.
5. Spark for data processing
6. Use of Hive/Pig for data transformation & analysis at scale
7. Use of Oozie and other workflow engines
8. Data transfer tools such as Sqoop & Flume
9. Real time processing of data with Storm & Kafka

This course will provide an in-depth coverage of analyzing data at scale (processing terabytes and petabytes of information quickly) in both a batch processing setup (using Hadoop) and real time setup (using Storm).

Course Outcomes: Students who complete this course will be able to

1) Apply MapReduce concepts to analyze data at scale.
2) Explain the inner workings of the Yarn distributed operating system.
3) Develop scripts in Pig/Hive to process data in scale using a batch-processing setup.
4) Develop jobs in Spark to process data in scale.
5) Use Data ingestion tools to move data back and forth between Hadoop and data warehouses.

6) Develop an end-end fully automated data analysis system for EDW offloads.
7) Discuss the latest trends and research in the Hadoop ecosystem.
8) Develop a system to process data in real-time using Storm.

Grading:

The learning objectives for this course are listed above. We will be using various mechanisms to measure your progress against these learning objectives. The relative weights for each of these mechanisms are mentioned below:

Mechanism	**Weight**
Participation	10%
Exams	25%
Labs	20%
Research	10%
Project	35%

Please note that these are subject to change and the class will be notified accordingly.

Course Grade Division:

Metrics used to assign final grades are mentioned below. Please note that these are subject to change and the class will be notified accordingly.

90-100	A
85-89	B+
80-84	B
75-79	C+
70-74	C
65-69	D+
60-64	D
0-59	F

Exam Policy:

Exams will be in-class, closed book, and closed notes except for one 8.5 by 11 sheet of paper on which you can put notes on using both sides of the page. No exams will be "dropped". If you have a conflict with a scheduled exam, you should notify me immediately. Giving a makeup exam for an unexcused absence is at the discretion of the instructor. Any requests for re-grading the entire exam must be made in writing by the beginning of the next class period after the exams are returned.

Labs:

Lab exercises for this class will be used to expose you to various tools and concepts in the Hadoop ecosystem that will be pertinent to the final project. Labs unless stated will be done individually.

There will be approximately 10 lab exercises. The labs will comprise of the following:

1) MapReduce & HDFS
2) Serialization
3) Advanced MapReduce – Counters & Joins
4) Hadoop Cluster Installation, Configuration & Maintenance
5) Data analysis using Apache Pig
6) Data analysis using Apache Hive
7) Apache Sqoop, Flume, & other Data Transfer Tools
8) Automation using Oozie
9) Apache Storm – Real time data analysis at scale
10) Vendor led Labs (Avalon/Hortonworks)
11) Spark

Reading Assignment:

Most lectures will have an associated reading assignment. The reading assignment has to be completed before class on the day indicated. Refer to the course calendar for details on reading assignments.

Ethics and Professional Practice:

You are expected to act honestly and professionally in this course at all times, in a manner consistent with the schools honor code.

Class Participation Policy:

There are 40 meeting times during the term. You can potentially receive 10 points towards the class participation portion of your grade for each of those classes in the following fashion:

- If there is a quiz during class, you can earn up to 10 points on it.
- If there is no quiz during class and you attend and make an effort to participate (with a small class there will be lots of discussion), you will earn 10 points.

Late Submissions:

Please note that all deliverables for this class will be due at 11:55 PM, on the day indicated. Late quizzes, and exams will not be accepted. Labs, and Project milestone deliverables will also not be accepted late, with the following exception:

You have four "late day" credits. You may use one of them on any lab, or Project deliverable, which will allow you to submit that assignment up to 24 hours after the due time. Lab's or project assignments, which are more than 24 hours late, will receive a deduction of at least 10% per late day (or not be accepted at all), depending on the circumstances and the degree of lateness.

If you submit something late for which late day credits are allowed, I will assume that you want to use one of your credits unless you tell me otherwise.

Attendance Policy:

Up to 2 unexcused absences allowed. Any additional unexcused absences may result in you receiving a failing grade for the course. You are responsible for making up any missed work.

Laptop Policy:

During class discussion, please do not use your laptops. Laptop use during discussions is distracting to your classmates and also keeps you from focusing on the material. If you typically use your laptop for note taking, please talk to your instructor so he can make an exception.

Collaboration:

You are encouraged to discuss the lab and other parts of the class with other students. Such discussions about ideas are not cheating, whereas the exchange of code or written answers is cheating. However, in such discussions of ideas, you should distinguish between helping and hurting yourself and the other student. In brief, you can help the other student by teaching them, and you can hurt them by giving them answers that they should have worked out for themselves. The same applies to tutoring and getting help from the instructor.

Final Project

Team Composition:

Teams will consist of about 3 members each. We are willing to allow smaller teams, if you have a good reason for wanting to do so. We may consider allowing larger teams, but would require a very strong argument and a very interesting project idea before allowing this. Please note that I will not be assigning the teams.

Project Ideas:

Students are recommended to identify, explore and analyze datasets that they find interesting. There are several publicly available datasets that can be used. I have included links to a few here:

1. Open Government Initiative - http://www.data.gov/open-gov/
2. World Bank Data Catalog - http://datacatalog.worldbank.org/
3. Wikimedia Statistics - http://dumps.wikimedia.org/other/pagecounts-raw/
4. Public Datasets on AWS - http://aws.amazon.com/public-data-sets/

5. Common Crawl - http://commoncrawl.org/
6. National Weather Dataset - ftp://ftp.ncdc.noaa.gov/pub/data/noaa/
7. Internet Census - http://internetcensus2012.bitbucket.org/paper.html
8. Million Songs - http://labrosa.ee.columbia.edu/millionsong/
9. Stack Exchange Dump - https://blog.stackexchange.com/2014/01/stack-exchange-cc-data-now-hosted-by-the-internet-archive
10. Machine Learning Data Sets - http://archive.ics.uci.edu/ml/datasets.html
11. Open Street Map Data - http://wiki.openstreetmap.org/wiki/Planet.osm
12. MLB Data - http://www.retrosheet.org/game.htm

Once a team has identified an interesting data set, the team is expected to identify interesting questions that can be answered using the data set and then perform analysis to answer said questions. Teams can also choose to use Hadoop to extract information from large data sets and make them easily searchable. The main objective of this project is that the data analysis/information extraction be done on a Hadoop cluster.

Milestones:

At each milestone, each team is required to submit the required documents. All team members are required to participate in generating and writing the milestone documents.

Milestone 0: Project and Team Identification (Due Week 2 Day 4)

Please notify the instructor of the project title, the data set to be used for exploration, and a rough description (two paragraphs) of the goals and the team of students who will work together to perform the identified project.

Milestone 1: Project proposal (Due Week 3 Day 4)

During this phase, the team must develop a problem statement for their project. The teams need to identify a detailed list of features (analysis to be supported, other utilities that will be supported – search, ability to repeat the analysis of just a random day's data, ability to export the results of the analysis to a data warehouse etc.). Please ensure that your features are clearly explained. The problem description must identify the programming language/framework/tools in the Hadoop ecosystem that the team plans to utilize during the implementation of the project. Your initial problem description will be graded and you may receive some suggestions. Note that an updated version of this document should be turned in at the end of Week 9

Milestone 2: Cluster Inspection & General Readiness (Due Week 4 Day 4)

Students will demonstrate the general readiness of their cluster. All tools identified by the team in the previous week will be installed and configured on the cluster. The dataset will be

hosted in a manner that is conducive for easy access for the analysis workflow. This means that the dataset will either be available on a public FTP server or organized on a student server for easy access by the team.

Milestone 3: Architecture & Workflow (Due Week 5 Day 2)

Each team will develop two architectures/workflows that will indicate how the team expects to meet the features identified in Milestone 1. The team will include an analysis of the two suggested approaches and identify weaknesses and strengths of each approach. The team will use the above analysis to identify an optimal choice based on current information. Please note that the identified workflow is a living document and I fully expect this to change as the team proceeds with the project.

Intermediate Status Reports:

Each team has complete freedom in deciding their internal organization and workload arrangement. Each team is expected to meet with the instructor on a weekly basis, to discuss the project progress. To ensure that the teams are making adequate progress, each team will demonstrate the current status of the project at the end of each week. These status reports will begin Week 6 and the team should do the following during the meeting:

a) Demonstrate the progress made during the previous week.
b) Identify[1] and confirm the user scenario that will be demonstrated during next week's meeting.
c) Updated workflow/architecture diagram.

There are no other documentation requirements for the various intermediate status reports.

Milestone 4: Updated Problem Statement (Due Week 9 Day 4):

Your team should turn in a revised version of your problem statement that incorporates the changes we suggested in response to your initial problem statement. We will compare your initial and final problem statements and will also use your final problem statement in evaluating your project.

Milestone 5: Project delivery: Demo, and Presentation (Due Week 10 Day 4)

Each team will spend the rest of the quarter completing their project. Each team will demonstrate their final project and present it to the class during 10th week. The final presentation should comprise of the following:

a) Problem solved by the analysis
b) Solution (Architecture)

[1] Please identify this before the meeting.

c) Technical issues faced
d) Pro's and Con's of Design
e) Demo

Milestone 6: Project Reflection and Team Evaluation (Due Week 10 Day 4)

Each team member will produce a one-page report that reflects on his or her experience with the project. Topics to reflect on include:

a. General Lessons Learned
b. Pro's and Con's of Design
c. Pro's and Con's of Process

Each team member must independently complete a form evaluating his or her contribution to the team, and the contributions of the other team members. This form will be available Monday of 10th Week.

Project Grading

Grade will be determined based on several aspects. The quality of the project, the milestone reports and the final demonstration will determine the major portion of the project grade. Creativity and extensibility in design is highly appreciated. Other issues, such as teamwork will also be taken into consideration.

We will not necessarily assign the same grade to each member of a team. We reserve the right to adjust individual grades up or down based on peer evaluations and our observations about your teamwork. A team member can hurt his or her team either by being a slacker or by running roughshod over other team members. Thus, taking over a project and doing all the work might actually hurt your grade.

Research

There are numerous tools that make up the Hadoop ecosystem. In addition to these tools, several open source projects have been developed to improve the usability and utility of the Hadoop ecosystem. A Hadoop developer needs the ability to identify the right tool, learn it quickly and leverage the tool to fit the requirements for a project.

The purpose of this research is to provide a realistic experience of the above process to the student. This research will be a team effort. The team for this will be the same team that is on the final project. Each team will agree on a tool that they will choose for their research. Teams will not be allowed to research the same tool and tools will be assigned to teams on a strictly first come first served basis. Some potential tools/concepts for students to research include:

1. Machine Learning with Apache Mahout

2. CRUD operations with Apache HBase
3. Securing a Hadoop Cluster using Kerberos
4. Integration search with Hadoop using SOLR and Elastic Search
5. Developing a YARN Application
6. Data processing pipelines using Falcon
7. Real time data processing with STORM & KAFKA
8. Hadoop Security with Apache Ranger
9. Integrating your favorite NoSQL database with Hadoop and performing computations with Hadoop.

Please note that the above list of research topics is not exhaustive and teams can choose other tools and concepts and explore them with instructor approval.

Each team is required to research and produce a lab (similar to the ones used in class) on their topic. The teams will be graded both on the content of the lab as well as the quality of the lab. The lab should address at the very least the following components

1) Introduction to the tool.
2) Installation of the tool.
3) Configuration of the tool.
4) Simple activities to introduce relevant ideas and concepts.
5) A mini-project that demonstrates common use of the tool.
6) A follow on project that interested students completing the lab can work on to demonstrate their mastery of the tool.
7) Lists of issues and unique things to keep an eye on while using this tool.
8) Links to relevant documentation about the tool.
9) Performance tuning if relevant.

Each member of the team is required to work on a substantial portion of the lab. The instructor might call upon any member of the team to explain different aspects of the research performed by the team. Note that individual grades for the research track may be different from the team's grade.

Milestones:

Research Milestone 0: Topic Identification (Due Week 3 Day 4)

Please notify the instructor of the title, and a rough description (2 paragraphs) of the goals of the research to be performed by the team.

Research Milestone 1: Cluster Inspection & General Readiness (Due Week 4 Day 4)

Students will demonstrate the general readiness of their cluster. All tools identified by the team in the previous week will be installed and configured on the cluster.

Intermediate Status Reports:

Each team has complete freedom in deciding their internal organization and workload arrangement. Each team is expected to meet with the instructor on a weekly basis, to discuss the progress of the research. To ensure that the teams are making adequate progress, each team will demonstrate the current status of the project at the end of each week. These status reports will begin Week 6 and the team should do the following during the meeting:

a) Demonstrate the progress made during the previous week.
b) Identify[2] the user scenario that will be demonstrated during next week's meeting.
c) Review of the current state of the lab.

Research Milestone 2: Final Lab Delivery

The team will provide a copy of the final lab to the instructor. The instructor reserves the right to have the lab reviewed by another team and their feedback will be taken into account to determine the final grade.

[2] Please identify this before the meeting.