INSTRUCTOR APPROACH

- Curriculum designed to teach parallel programming concepts using Hadoop tools
- Distributed computing system that has become
- Widely adopted by a variety of industries
- Low cost Hadoop solution

- Early homework labs completing using Hadoop clusters
- Hadoop class
- Course fee, fully configured Hadoop virtual machine (VM)

- Lectures on ‘Core Hadoop’, which is comprised of a
  - Distributed storage component: HDFS (Hadoop Distributed File System)
  - Distributed compute component: MapReduce

- Lectures on the Hadoop ecosystem, including
  - Hive, a SQL-like language for storing datasets
  - Pig, a data flow language for transforming datasets

- Both of these languages facilitate programming tasks, particularly for users who are less comfortable programming in languages supported at the MapReduce level (e.g., Java, C++, Python, etc.).

- Introduces Hadoop programming, from C++, computing advertising, computing disciplines, and programming languages.

- Programming challenge

- Computer Science and Engineering students typically learn C, C++, Java, or Python, all of which are supported by Hadoop’s compute component, MapReduce. However, students in other disciplines may program in higher-level languages such as SQL or R, for example.

- Fortimately, Hadoop provides several programming language options to bridge the programming gap and make Hadoop available to a wider user base.

- Another challenge in creating a Hadoop course is the continuously and rapidly changing Hadoop landscape. For example, the diagrams below depict a major shift that occurred in the past year where enterprises started adopting YARN/MapReduce 2 in place of MapReduce 1. In order to keep pace with the technology, this course must be updated frequently.

- The next major course update includes the addition of Spark, a language that introduces the concept of resilient distributed datasets (RDDs), which can be persisted in memory to allow for highly performant and efficient distributed data processing (1 2).

THE PROJECT

In the second half of the course, students form project teams to develop self-designed analytics projects using the tools learned in the first half of the course.

- Project Requirements
  - Form a team with maximum four team members
  - Complete a project proposal describing the analytic to be developed
  - Who will benefit from the analytic?
  - What insight(s) are expected to be produced by the analytic?
  - How will the goodness of the analytic be assessed?
  - What are the runtime use cases directly or indirectly related to the analytic?
  - Identify and obtain at least two large data sources – this is challenging due to a number of issues:
    - Data owner is not willing to release data
    - Privacy concerns
    - Data requires owner’s time to perform anonymization
    - Difficulty with transfer of data
    - Not enough storage to store the data source
    - Difficulty with data format
  - Utilize at least two Hadoop compute technologies, e.g. MapReduce, Hive, Pig
  - Identify a platform on which to run the analytic, for example
    - University Hadoop cluster
    - Amazon Elastic MapReduce (EMR)
    - Self-installed Hadoop on Elastic Compute Cloud (EC2)

- Pig
- Spark
- SQL
- Spark Streaming

STUDENT ANALYTICS PROJECTS

- Business, Finance, Legal Projects
  - News-Based Stock Performance Prediction
  - Predictions for Success of Startup Investments
  - U.S. Patent Office Analytic
  - Predicting Local Real Estate Value
  - Social Media-Based Stock Performance Prediction
  - Yelp Business Data Analysis to Recommend Business Location

- Health and Safety
  - Breast Cancer Staging and Prediction
  - Healthcare Relationship to Employment in the United States
  - Public Safety and Mapping Analytic

- Sports
  - Prediction of NBA Game Pace
  - Soccer: Dynamically Ranking Football Players

- Entertainment
  - Effect of Title, Posting Time, Comments and Community on Popularity of Social Media Posts
  - Rating Prediction on NetFlix Dataset
  - Song Recommendation
  - Song Popularity Based on User Listening Patterns
  - Book Recommendation

- Travel and Transportation
  - Detecting Movement Paths and Patterns Using Wireless Access Point Logs
  - Urban Traffic Monitoring
  - Impact of Weather on Road Travel
  - Creating a Smarter Transportation Plan for New York City

- Miscellaneous
  - Drought Prediction
  - Earthquake Data Analytic
  - Resume Matching for Job Seekers
  - Understanding Hierarchies in Technical Conferences
  - Twitter Feed Sentiment Analysis

REFERENCES


ACKNOWLEDGEMENTS

Thanks to Amazon for supporting student projects through educational grants. Thanks also to Pennsylvania State University’s Chancellors’ team for providing access to the Chancellors’ data used in student projects. Finally, thanks to the domain experts for supporting the development of student analytics projects.