Big Data Analytics: Where is it Going and How Can it Be Taught at the Undergraduate Level?

Dr. Frank Lee
Chair, ECE/CS/IT
New York Institute of Technology
Old Westbury, NY 11568

Topics

This talk describes:

- a new course: CSCI-372 Big Data Analytics
- the CS program concentration in
  “Big Data Management and Analytics” and
- our partnership with IBM to use their IBM systems in this concentration.

The Program Goals

- The CS program is designed to allow students to gain theoretical knowledge and apply it to developing an in-depth specialization in one area of concentration.
- It prepares graduates to be creative, inquisitive, analytical, and detail-oriented.

The Undergraduate CS Curriculum

The CS curriculum consists of 57 credits in CS-related courses:

- 36 credits in CS core courses (12 courses),
- 6 credits of CS electives (2 courses),
- One final senior project (3 credits), and
- A 12-credit concentration in either
  - **Network Security** or
  - **Big Data Management and Analytics**

The Undergraduate CS Curriculum

The 12 CS core courses are:

- **CSCI-125 Computer Programming I**
- **CSCI-155 Computer Organization and Architecture**
- **CSCI-185 Computer Programming II**
- **CSCI-235 Elements of Discrete Structures**
- **CSCI-260 Data Structures**
- **CSCI-270 Probability and Statistics for CS**
- **CSCI-312 Theory of Computation**
- **CSCI-318 Programming Language Concepts**
- **CSCI-330 Operating Systems**
- **CSCI-335 Design and Analysis of Algorithm**
- **CSCI-345 Computer Networks**
- **CSCI-380 Introduction to Software Engineering**

The Undergraduate CS Curriculum

- CSCI-455 Senior Project

  all students in the CS program are required to complete a substantial project, which utilizes the full extent of the technical skills and knowledge gained throughout the curriculum.

- CS Program Concentration

  By the end of the first term of their junior year, computer science majors must select a 12 credit concentration in **Network Security** or **Big Data Management and Analytics**

The CS Concentration in Big Data Management and Analytics

To meet the increasing big data challenges and opportunities, the CS department has:

- revised its CS program concentration in **"Big Data Management and Analytics"** and
- created a new course, **CSCI-372: Big Data Analytics**.

The CS Concentration in Big Data Management and Analytics

The students in this concentration must choose 4 courses from following:

- CSCI-365 Information Retrieval
- CSCI-372 Big Data Analytics
- CSCI-401 Database Interfaces and Programming
- CSCI-415 Introduction to Data Mining
- CSCI-405 Distributed Database Systems

The CS Concentration in Big Data Management and Analytics

- This concentration focuses on the management and analysis of big data.
- It provides students with deep analytic skills to design and implement information systems.
- It equips students with both the technical knowledge and analytic acumen necessary to extract meaning from big data.

Our Partnership with IBM

- Our partnership with IBM starts with the "System z Academic Initiative"
- It is prepared to:
  - Assist and enable NYIT to use and teach IBM Enterprise Systems
  - Connect IBM clients with NYIT to hire students learning critical systems skills

Our Partnership with IBM

- Since IBM is a leader in the big data analytics area, the CS faculty decided to address this growing demand by engaging with IBM in their Academic Initiative and by **introducing a course in big data analytics to be included in the concentration.**

CSCI-372: Big Data Analytics

- The new course will embrace the IBM Academic Initiative and will introduce the IBM InfoSphere systems, and in partnership with IBM will provide remote access to the Enterprise System and data for a hands-on-lab experience with big data analytics.

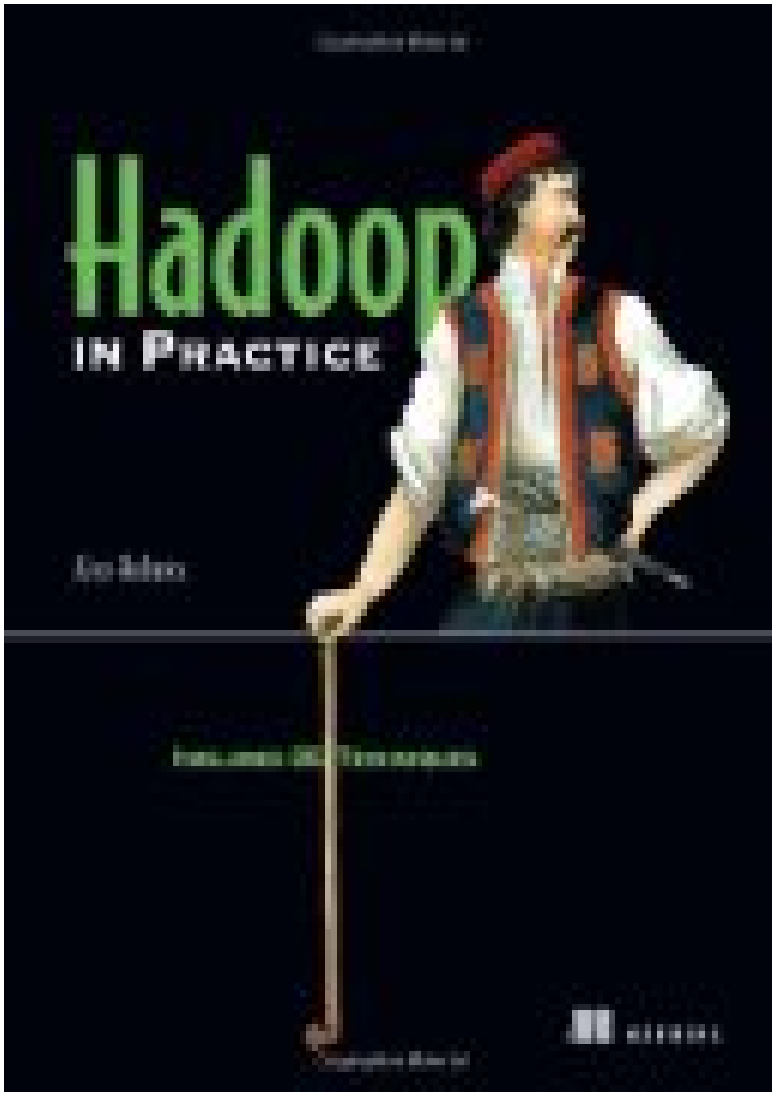
CSCI-372: Big Data Analytics

Course contents:

- the basics of data analysis (e.g. R),
- the tools of big data analytics (e.g. Hadoop, MapReduce, Pig, Hive),
- the analysis of unstructured data using NoSQL and Hadoop/MapReduce,
- IBM InfoSphere systems and application areas including finance, banking, defense, and health.

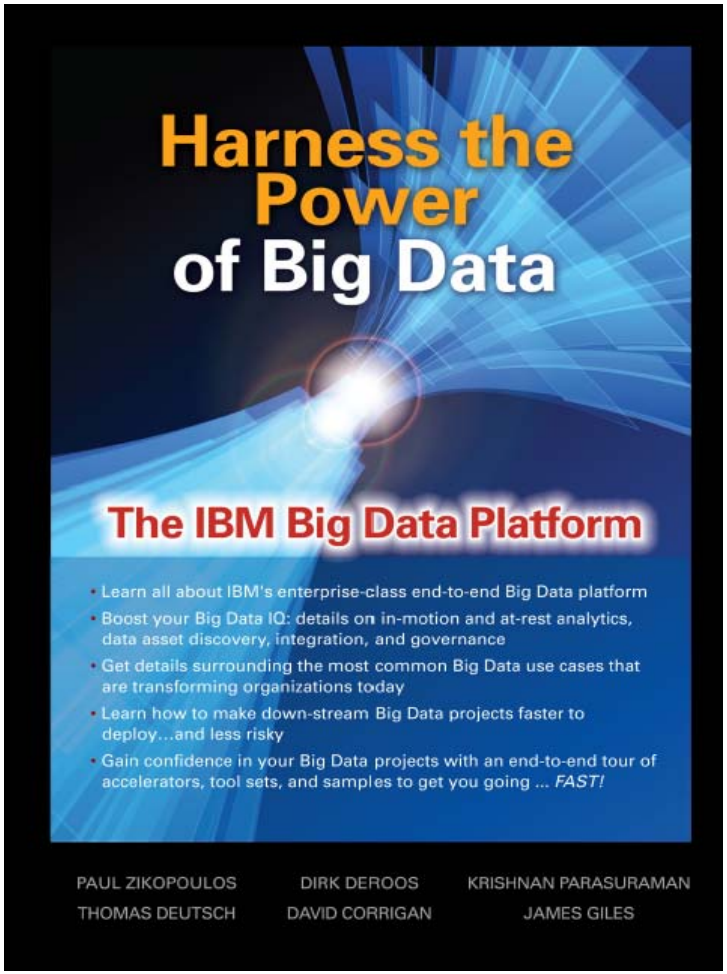
Text book

- To lecture the theoretic fundamentals, the first nine weeks of course work is based on the text book written by Alex Holmes (ISBN: 9781617290237).



Reference

To learn IBM InfoSphere Systems, we study: Paul Zikopoulos, Dirk deRoos, Krishnan Parasuraman, Thomas Deutsch, James Giles, David Corrigan, “Harness the Power of Big Data The IBM Big Data Platform”, McGraw-Hill, 2013



CSCI-372: Big Data Analytics

Its topics:

- Introduction to Big Data Analytics:
  - Big Data overview
  - Relational Databases & Data Mining
  - Cloud & Big Data Architectures
- Basics of Data Analysis
  - Introduction to R
  - Analyzing and exploring data with R
  - Statistics for model building and evaluation
-

CSCI-372: Big Data Analytics

- Introduction to Data Analytics' Tools
  - The Hadoop architecture
  - The Hadoop Distributed File System (HDFS)
  - MapReduce
  - Using the Pig platform to create MapReduce programs
  - Using the Hive data warehouse system to query and analyze large data sets
  - Other related Hadoop technologies

CSCI-372: Big Data Analytics

- Analysis of Unstructured Data
  - Using MapReduce/Hadoop for analyzing unstructured data
  - Using NoSQL
  - Scale up vs. Scale out.

CSCI-372: Big Data Analytics

- Data Lab Exercises using IBM InfoSphere BigInsights
  - What is IBM InfoSphere BigInsights?
  - Downloading BigInsights
  - Installing BigInsights 2.0
  - Setting up a Hadoop cluster on the IBM SmartCloud Enterprise
  - BigInsights Web Console overview

CSCI-372: Big Data Analytics

- Lab Exercises: Stream Computing using IBM InfoSphere Streams
  - What is IBM InfoSphere Streams?
  - Downloading Streams 3.0
  - Installing Streams 3.0
  - Introducing the Streams Studio graphical editing environment
  - Introducing the InfoSphere Streams runtime environment
  - Introducing the data visualization capabilities in the Streams Console
  - Use Cases

Course Work: **Week 1**

- **Reading:** Ch. 1 (Textbook); Ch. 1, 2 (Reference);
- **Topic**: Big Data; Applying Big Data to Business Problems; Data Storage and Analysis; A Brief History of Hadoop; Apache Hadoop and the Hadoop Ecosystem;
- **Exercises**: Downloading and installing Hadoop

Course Work: **Week 2**

- **Reading:** Ch. 2 (Textbook);
- **Topic:** Running Hadoop; Moving data in and out of Hadoop;
- **Exercises:**
  - **#** 1 Using Sqoop to import/export data from/to MySQL

Course Work: **Week 3**

- **Reading:** Ch. 3, 4 (Textbook);
- **Topic:** Data serialization: working with text and beyond; applying MapReduce patterns to big data
- **Exercises:**
  - **#**2 MapReduce with HBase as a data source;
  - #3**:** Integrating Protocol Buffers with MapReduce

Course Work: **Week 4**

- **Reading:** Ch. 5, 6 (Textbook);
- **Topic:** Streamlining HDFS for big data; Diagnosing and tuning performance problems
- **Exercises:**
  - #4 Compression with HDFS, MapReduce, Pig, and Hive,
  - **#**5 Using stack dumps to discover unoptimized user code

Course Work: **Week 5**

- **Reading:** Ch. 7 (Textbook)
- **Topic:** Utilizing data structures and algorithms
- **Exercises:**
  - #6 Calculate PageRank over a web graph

Course Work: **Week 6**

- **Reading:** Ch. 8 (Textbook)
- **Topic:** Integrating R and Hadoop for statistics and more
- **Exercises:**
  - **#7** Calculate the cumulative moving average for stocks using RHadoop.

Course Work: **Week 7**

- **Reading:** Ch. 9 (Textbook)
- **Topic:** Predictive analytics with Mahout;
- **Exercises:**
  - #8 Using Mahout to train and test a spam classifier

Course Work: **Week 8**

- **Mid-term Exam:** Ch. 1 -9 (Textbook); Ch. 1 -2 (Reference)

Course Work: **Week 9**

- **Reading:** Ch. 10, 11 (Textbook)
- **Topic:** Hacking with Hive; Programming pipelines with Pig
- **Exercises:**
  - **#**9 Tuning Hive joins,
  - #10 Combining data in Pig,
  - #11 Pig optimizations

Course Work: **Week 10**

- **Reading:** Ch. 3, 4 (Reference)
- **Topic:** The IBM PureData Systems: A Big Data Platform for High-Performance Deep Analytics
- **Exercises:** none

Course Work: **Week 11**

- **Reading:** Ch. 5 (Reference)
- **Topic:** IBM's Enterprise Hadoop: InfoSphere BigInsights
- **Exercises:** Installation of IBM InfoSphere BigInsights 2.0

  Case study No. 1

Course Work: **Week 12**

- **Reading:** Ch. 6 (Reference)
- **Topic:** Real-Time Analytical Processing with IBM InfoSphere Streams
- **Exercises:** Installation of IBM InfoSphere Streams 3.0

  Case study No. 2

Course Work: **Week 13**

- **Reading:** Ch. 7 (Reference)
- **Topic:** Unlocking Big Data: Data Exploration and Discovery
- **Exercises**: Case study No. 3

Course Work: **Week 14**

- **Reading:** Ch. 8, 9 (Reference)
- **Topic:** Text Analysis: The IBM Big Data Analytic Accelerators
- **Exercises:** Text Analysis: The IBM Big Data Analytic Accelerators

Learning Outcomes

At the completion of this course, the students will be able to:

1. **Describe** the characteristics of the Big Data model vs. the Relational Database model.
2. **Use** the programming language R for Big Data analysis.
3. **Use** the MapReduce algorithm as a data mining tool.
4. **Distinguish** between the Map step and Reduce step of the MapReduce algorithm.
5. **Analyze** data with Hadoop.
6. **Describe** the key attributes and differences of the NoSQL database model vs. Relational database model.

Assessment

- LOs 1, 4, 6 will be assessed using Essay, and homework questions to assess the student's understanding of the Big Data Model, the MapReduce algorithm and the NoSQL database model.
- LOs 2, 3, 5 will be assessed with programming projects: these projects will include data analysis, and data mining using R, Hadoop, and MapReduce.

Assessment

- LOs 1, 3, 4, and 5 will be assessed through exam questions. These exam questions will assess the student's ability to use the MapReduce algorithm, and basic statistical data analysis using R and Hadoop.

Conclusion

- The new course and concentration meet the demands of industry.
- The new course and concentration meet the Accreditation Board for Engineering and Technology (ABET) criteria which were the primary challenges for the CS faculty.
- The new concentration will embrace the IBM z Enterprise Academic Initiative and will introduce the IBM systems.

Conclusion

- The partnership with IBM will, in addition to providing remote access to the IBM systems, provide access to data for a hands-on lab experience with big data analytics.
- The **“Big Data Management and Analytics”** concentration and lab experience will give our students an important advantage in the data engineering marketplace.