Big Data Analytics

Information Systems Area
PGP Term V, 2015-16

Instructor
Prof. Srikumar Krishnamoorthy
Wing 4D, Extension 4834
E-mail: srikumark@iimahd.ernet.in

Course Objective

IDC estimates that the market for big data including hardware, software and services is around $18Bn in 2015 and it is expected to grow at a CAGR of 17%. Despite significant market potential, there is a dearth of analytical talent – data scientists (akin to the Wall Street quants of the 1990s), analysts and managers – to leverage the economic value of big data.

Big data analysis is likely to fuel the next wave of growth in productivity, innovation, and competition in the market place. The organizations ability to unlock the potential of big data and lead in the market place will be largely determined by its ability to tackle major hurdles in effectively managing big data – identifying the business use case, hiring, nurturing and retaining the right analytical talent for conducting big data analytics, and embracing data driven culture for making business decisions.

This course aims to help the participants to build a solid foundation on big data. It will enable the participants to learn, design and build big data analytic solutions to solve business problems and improve decision making. The course will also help participants to understand various issues, challenges and best practices in implementing big data analytic solutions in organizations.

Session Plan

Session	Date	Topic	Case / Reading
1	08 Sep	Introduction to big data	*Case:* Volkswagon group: Driving big business with big data (Ivey) *Read:* Big data: The management revolution
2	09 Sep	Data management and big data	*Read:* CAP Twelve Years Later: How the “Rules” have changed
3-4	14 Sep 15 Sep	Distributed file system & Data modeling for big data	*Case:* Data modeling and management in the big data era (IIMA)

			Read: The hadoop distributed file system
5-6	21 Sep 22 Sep	Map reduce paradigm	*Read:* 1. Bigtable: A distributed storage system for structured data 2. MapReduce: Simplified data processing on large clusters
7	23 Sep	Hadoop: Big data ecosystem	*Hands-on:* Hadoop fundamentals
8	28 Sep	Hive: Data Warehousing & Analytics for Big Data	*Hands-on:* Hadoop
9-10	29 Sep 30 Sep	Recommender systems for large scale e-commerce personalization	*Case:* Netflix: Designing the Netflix Prize (A) (HBS) *Read:* Recommender systems in E-commerce
11	05 Oct	Design and build recommender system	*Read:* Collaborative filtering with temporal dynamics *Hands-on:* Hadoop
12-13	06 Oct 07 Oct	Mining unstructured text documents	*Read:* Multi-label classification: An overview
14	12 Oct	Unstructured text mining: Design and build social tag prediction system	*Read:* Predicting tags for stackoverflow posts *Hands-on:* R / Hadoop
15-16	13 Oct 26 Oct	Mining social media data	*Read:* Mining social media: A brief introduction
17	27 Oct	Design and build social network based prediction model	*Read:* Predicting purchase behaviors from social media *Hands-on:* R / Hadoop
18	28 Oct	Big data applications in Finance	
19	02 Nov	Building and managing data driven organization	*Read:* 1. Data Scientist: The Sexiest Job of the 21st Century 2. The parable of Google flu: Traps in big data analytics

20	03 Nov	Course wrap-up and summary	

Pedagogy
This course will have a mix of lectures, cases, and hands-on sessions.

Preparation
Each student needs to spend about 80 hours for class preparation (cases and readings), group assignment and term exam.

Evaluation
The course grade will be based on the following components and weights:

Class participation	20%
Group assignment (3) and presentation	50%
Term exam	30%

Acknowledgement

This course is supported by AWS Education Program.

Key References

1. McAfee, Andrew and Brynjolfsson, Erik and Davenport, Thomas H and Patil, DJ and Barton, Dominic, *Big data: The management revolution,* Harvard Business Review, 90(10), 2012
2. Shvachko, Konstantin and Kuang, Hairong and Radia, Sanjay and Chansler, Robert, *The hadoop distributed file system,* IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), 2010
3. Chang, Fay and Dean, Jeffrey and Ghemawat, Sanjay and Hsieh, Wilson C and Wallach, Deborah A and Burrows, Mike and Chandra, Tushar and Fikes, Andrew and Gruber, Robert E, *Bigtable: A distributed storage system for structured data*, ACM Transactions on Computer Systems (TOCS), 26(2), 2008
4. Dean, Jeffrey and Ghemawat, Sanjay, *MapReduce: Simplified data processing on large clusters*, Communications of the ACM, 51(1), 2008
5. Davenport, Thomas H and Patil, DJ, *Data Scientist: The Sexiest Job of the 21st Century*, Harvard Business Review, 90, 2012