

# A BIG DATA ANALYTICS COURSE

**De Liu**

Associate Professor, PhD Coordinator  
3M Fellow in Business Analytics  
Carlson School of Management  
[deliu@umn.edu](mailto:deliu@umn.edu)

Special thanks to Gordon Gao for many discussions and generous sharing of his teaching stack

CARLSON SCHOOL  
OF MANAGEMENT

UNIVERSITY OF MINNESOTA

# My Background

- Joined University of Minnesota in 2014
- Developed two MSBA courses: big data analytics and data management
- Involved in teaching big data modules in two new executive courses
- Hobbies: building my own productivity applications
  - Wrote a full-fledged course portal that serves all my courses with integrated active quiz, photo roster, gradebook, & survey functions
  - Built a customized crawling framework for reusable, distributed web scraping
  - Built dozens of special-purpose websites for online experiments, PhD program management, Faculty/student surveys, Quiz Game etc.
  - Owners of several dozens of github repositories

## Why is it essential for MSBA (and other business programs) to teach Big Data?



**Dan Ariely** ✓

January 6, 2013 · 🌐

 Follow

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

 Like

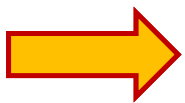
 Comment

 Share

# So What is Big Data?

## People go through stages in thinking about big data

- Stage 1: size of data
- Stage 2: new tools and capabilities
- Stage 3: big opportunities
- Stage 4: new philosophy (of handling and leveraging data)
- ~~– Stage 5: new religion~~



**Use healthcare industry as an example for "big" opportunities**

# Drivers of Big Data in Health Care

- **#1: Health Information digitization is accelerating**
  - Electronic Health/Medical Record data (EHR or EMR) is quickly filling the 'data lake'
  - Optum Lab has collected EHRs of over 30 million patients
  - Records and the volume and detail of patient information is growing rapidly.
  - Plus transaction-level claims data that has been in existence

V olume

# Drivers of Big Data in Health Care

- **#2: 80% of information in healthcare industry is unstructured data**, e.g.
  - Outputs from medical devices
  - Doctor's notes
  - Lab results
  - Medical imaging
  - Medical correspondence
  - Clinical data
  - Patient behavior and sentiment data
  - Genomic data

Variety

# Drivers of Big Data in Health Care

- **#3: Streaming analytics are becoming increasingly important,**  
e.g.
  - Real time claims → fraud detection
  - Patient sensor data → alerts, patient care
  - Real-time medical records → preventive care, reduce re-admission, assist diagnosis

Velocity

## Value of a Big Data Course: anecdotal evidence

- MSBA students told me a big data course was a differentiator for our program.
- CS student who took my course reported that he found a job because of this course. Big data is frequently asked during interviews.
- We start seeing more Carlson Analytics Lab<sup>®</sup> projects involving big data, e.g.
  - An prototype for real-time failure detection leveraging Internet of Things (IOT) and AWS cloud.
  - Using cluster computing to find relevant patents in a large patent database.
  - Analyzing bitcoin blockchain data to detect illicit transactions.



# Positioning of the course

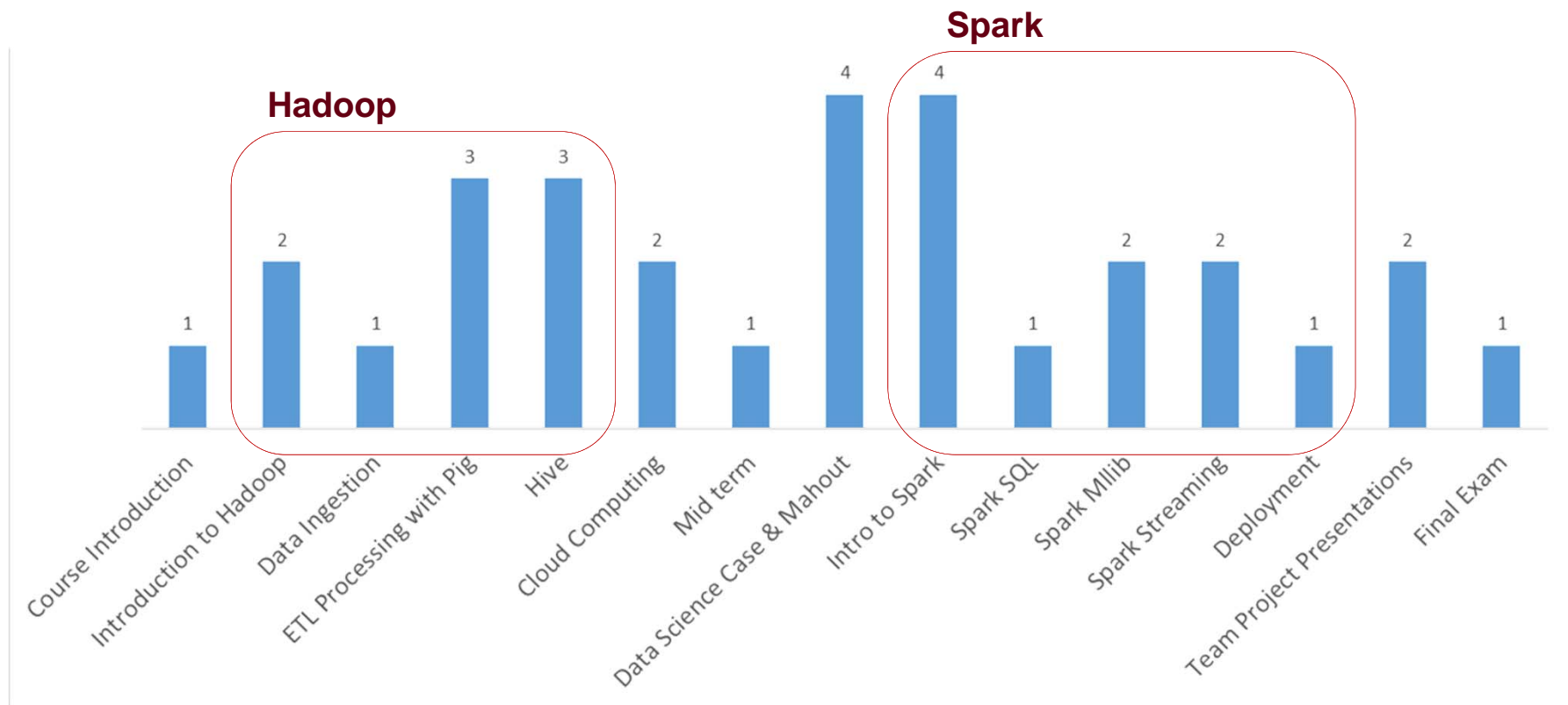
- **Course goals (for executive courses, just Goal 1)**
  - **Goal 1:** develop an understanding of opportunities and challenges associated with big data, from both business and technological perspectives, that allows students to guide businesses in adopting and using big data technologies
  - **Goal 2:** Develop core competences in using a variety of essentially big data tools (e.g. Scoop, MapReduce, Pig, Hive, Spark, and cloud computing) and processes to solve data science problems at scale.
- **What the course is not:**
  - Not a machine learning course (we teach/use scalable machine learning algorithms but assume students have working knowledge of it)
  - Not an big data administrator or developer course – it is oriented towards data analytics leaders, data analysts, and data scientists



**I will focus on the MSBA course first**

# What does the course cover?

- Semester long (15 weeks), meet twice per week (1h15min)



## Some Notes on the Choice of Topics

- The topics are constantly tweaked due to fast-paced innovation
- Did not spend time on
  - Writing java-based MapReduce applications (it is more for developers)
  - Installing and configuring big data environments (it is more for engineers and admins)
- I give a lot of importance to Hive and Spark (and its ML, streaming, and graph modules) – these are workhorses for data analysts.
  - In the future iterations, add more depth on these with goal of better scaling.
- Taught Spark in pySpark but provided Scala versions for references
- Areas for future improvements:
  - Cloud computing is increasingly important (but the space is very fragmented).
  - Streaming, NoSQL (e.g. HBase), and graph analytics are also indispensable advanced components for big data

# Technical Infrastructure for the Course

- Primarily use Cloudera VMs with preinstalled software that run on student laptops
  - Installation notes for windows and Mac users (shared)
  - Laptop RAM: 8GB+, computer classroom with power outlets
- AWS cloud computing
  - Students can get \$75 AWS education grant
  - Tutorials for setting things up (shared)
  - Microsoft Azure & Google cloud may be an alternative
  - Third-party vendors such as DataBricks /Quoble?

# Prerequisites


- Familiarity with SQL and relational database concepts (essential)
  - Working knowledge of common machine learning algorithms (predictive and explorative)
  - Working knowledge of Python for data science (essential)
  - Operating system knowledge, especially Linux command line environment
- } These tend to be problems

# Embrace a "hacker" culture and open-endedness

- Embrace a "hacker" culture




- Yellowdig channel for sharing big data knowledge, news, & questions (3% of course grade)
  - TA can help respond to questions
  - Students can help each other out
  - May switch to Slack after Yellowdig license expires.

 **Steffi Jiawen Gu** 02:58 PM CST, 29 Nov MSBA 6330 Harvesti.. ✎ ✕

## Scalable Collaborative Filtering with Apache Spark MLlib



When we learn the tools in MLlib, professor mentioned the collaborative filtering algorithm: alternating least squares (ALS). Here is a post with an example of how to use Spark MLlib for ALS. It also covers some basic comparison between MLlib and Mahout.






Recommendation systems are among the most popular applications of machine learning. The idea is to predict whether a customer would like a certain item: a product, a movie, or a song. Scale is a key concern for recommendation systems, since computational complexity increases with the size of a company's customer base. In this blog post, we ...

<https://databricks.com/blog/2014/07/23/scala...> Pmalink

SPARK MLlib

Love it! -1 Like -3 Not relevant -0 Bookmark Save as New   2

---

 **Sophie Baimin Zheng** 08:10 PM CST, 17 Dec Reply   0 ✕

Hi Steffi,  
Thanks for sharing this post. I think it is nice to learn about how capable of MLlib of spark is via your post or our peers' course presentations. "  
Recently we did an experiment to benchmark ALS implementations in Spark MLlib at scale. The benchmark was conducted on EC2 using m3.2xlarge instances set up by the Spark EC2 script. We ran Spark using out-of-the-box configurations. To help understand state-of-the-art, we also built Mahout from GitHub and tested it. This benchmark is reproducible on EC2 using the scripts at [.github.com/databricks/als-be..](https://github.com/databricks/als-be..)

# Experiential Learning

- Lectures → Labs → Assignments
  - **23 lab notes with solutions** that cover most of the topics (shared)
    - Students need to find time to do labs out of class time
  - **8 assignments** – approximately weekly assignments (shared)
    - Assignments combine conceptual and hands on questions
    - Assignments build on labs so that students have incentives to do labs before assignments.
  - Recorded short videos are helpful – should do more of it

Managed  
through  
umn.github

## MSBA 6330 Homework 7 - Spark I.

### Part I. Short Answers

1. What is Spark and what are the key advantages of Spark (list at least 4)?
2. Explain what is RDD lineage and explain its role in Spark.

### Part II. Hands on.

Our dataset is a .csv file that consists of online auction data. ...

1. Load `auctiondata.csv` ...
2. Report the first 5 elements of `auctionRDD` ...
3. What is the total number of bids? ...

### Part III. Hands on.

We will use the same web log data from Spark Lab 2. Some sample records from the web log are as follows

```
3.94.78.5 - 69827 [15/Sep/2013:23:58:36 +0100] "GET /KBDOC-00033.html HTTP/1.0" 200 14417 "http://www.loudacre.com" "Loudacre Mobile Browser iFruit 1"
3.94.78.5 - 69827 [15/Sep/2013:23:58:36 +0100] "GET /theme.css HTTP/1.0" 200 3576 "http://www.loudacre.com" "Loudacre Mobile Browser iFruit 1"
```

1. Run the following python function, which will be used to parse the web log lines. ...
2. Calculate the average content size (i.e. number of bytes transmitted). ....

# "On the Edge" Team Project Design

- Team project is designed to encourage exploration of new frontiers. Students can choose between (or a combination of):
  - A **special topic track** – develop a tutorial + demo on a special topic
  - A **big-data analytics track** – problem solving on a real-world dataset using scalable tools and processes learned in this course.
  - Provide students sample topics and data sources (shared)
  - **Scalability** is a key criteria for team projects
- Team project flow chart:
  - Project Proposal → Instructor approval & feedback → Presentation (peer evaluation) → Written portfolio (instructor evaluation)
- Team projects feed ideas (& data) into next year's curriculum
- Future design may also include a case competition that emphasizes scalability and performance.



## Team Project Sample Topics

- **Real Time Twitter Analytics Using Apache Storm**
- **Recommender System Using Outbrain Data (Kaggle)**
- **Spark Tutorials for Machine Learning**
- **Harry Potter Books (stats.ox.ac.uk)**
- **Real time visualization of New York City Traffic Data**
- **Twitter Data and Company Stock Price**
- **Meetup.com streaming data**
- **Best Buy Mobile User Behavior (Kaggle)**
- **Analyzing Reddit comments**
- **Stackoverflow Questions & Answers (Kaggle)**

# TEACH BIG DATA TO EXECUTIVE AUDIENCE

## Executive Course Backgrounds

- A 3-day open-enrollment executive program of "Leading Business Analytics"
  - Marketed as "demystify descriptive, predictive and prescriptive business analytics, giving executives concrete direction to drive business change and growth using analytics"
  - Students from multiple industries (agriculture, chemical, manufacture) and functional areas
- A tailored 3-day program for a big health analytics company that sends its directors for training in machine learning
  - Client wants lots of machine learning and a use-case based program.
  - Most students are IT directors

## Course Designs and Lessons

- More centered around understanding the technologies and their values
  - Followed the technology evolution
  - Designed with some hands-on using online notebooks backed by a AWS cluster.
- Students were not into the "hands-on" part; many do not consider themselves as "doers."
- Lack relatable real-world examples
- More centered around values and opportunities, high-level understanding, & how to map business problems into big data solutions
  - No hands-on
  - but has a case discussion and many use cases (though time constrained)
- There were more engaged discussions with students (related to their own organizations)

# Content Organization

- Start by calling out a myth, "My Data is not big"
- Use that to discuss different perspectives on big data
  - moving from big challenges to big opportunities, and to some fundamental shifts in our philosophy about data.
- Discuss opportunities in the health care industry
  - that involve using big data to create values (along the lines of volume, variety, and velocity).
- Brief introduction to Hadoop MapReduce and ecosystems
  - what challenges do they address, what are some of the advantages (e.g. cost, reliability, and scalability), system architecture.
- Use a case to discuss how a business problem could be formulated and solved using big data

## Case Study – Cloudera Movies

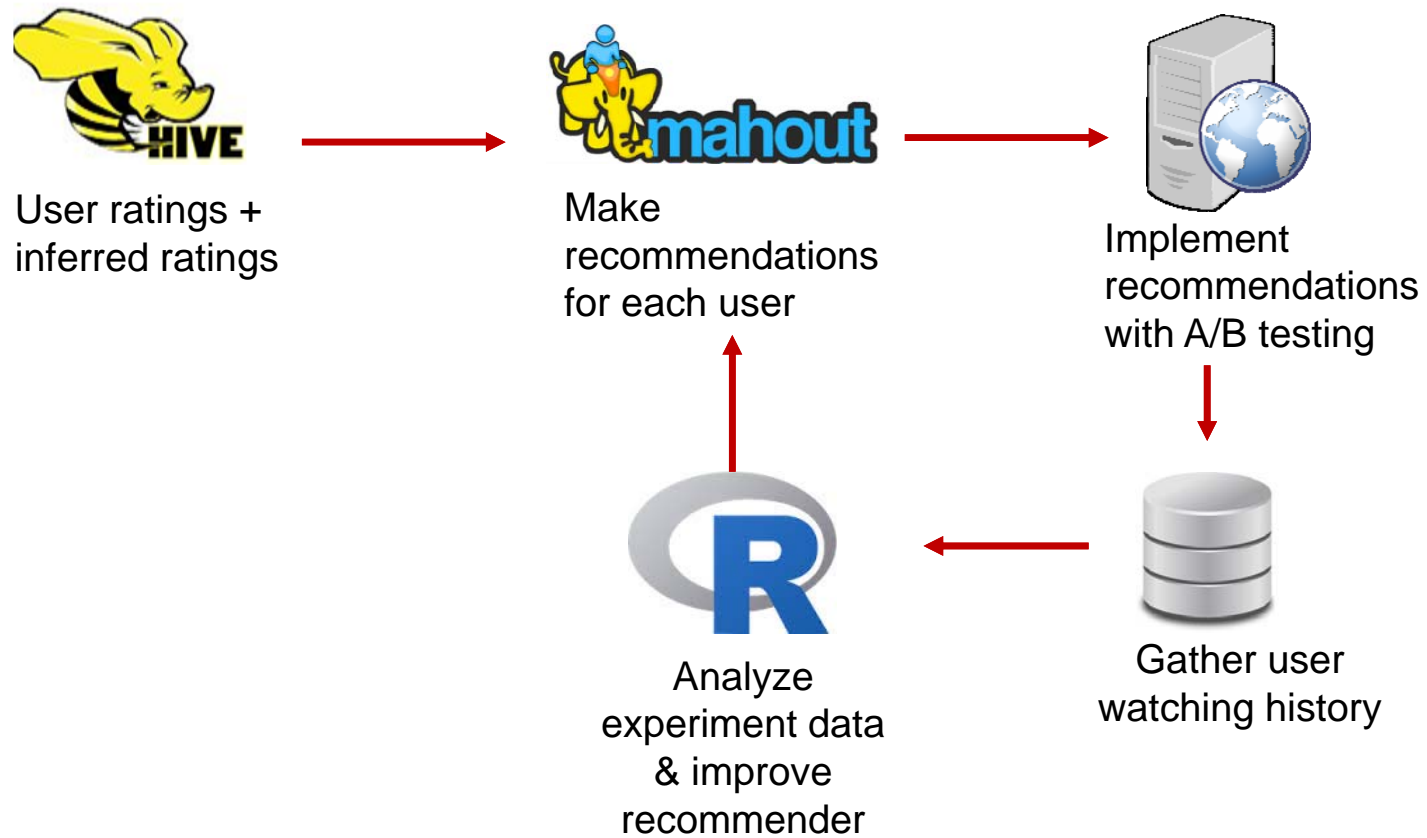
- **Situation:** Cloudera Movies (CM) is a successful online movie streaming service
  - **14 million** customers pay a monthly subscription fee
- **Complication:** Unfortunately, our success has started to wane
  - Revenues dropped last quarter by 11%
  - Projections show revenues decreasing this quarter by 17%
- **Key Question:** Can we turn the ship around (with big data analytics)?

# Explore Possible Solutions

- Can we stop the revenue decline?
  - Decrease subscription cost
    - discard: price is not the problem
  - Social media integration
    - discard: may violate privacy laws
  - **Improve customer experience with better movie recommendations**
    - We'll pursue this one!

- What kind of data is needed? How to address data sparseness (supplement it using twitter data).
- What tools, algorithms, and processes should one go through to arrive at the recommendations?

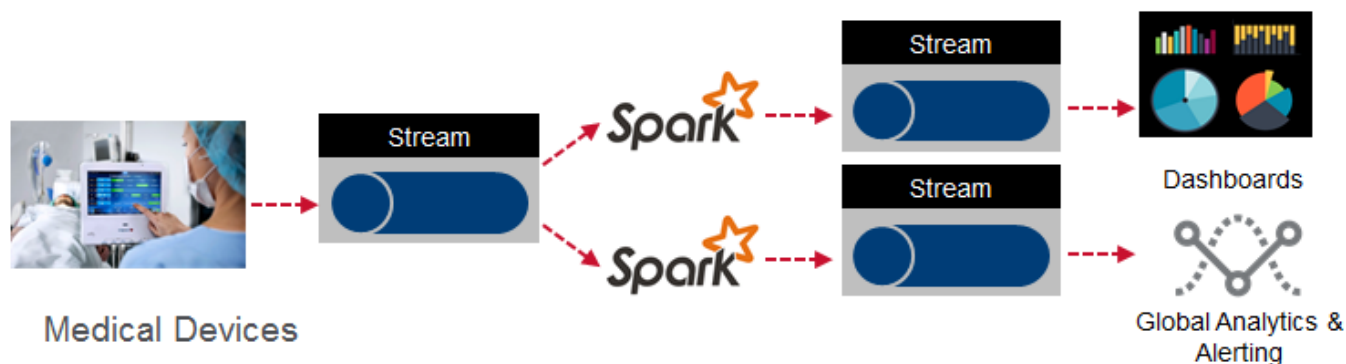
# Build & Test Recommendations





# Spark for Big Data ML

- Why is Spark a significant improvement over Hadoop?
- Lead a discussion on what the companies should use, Spark or Hadoop MapReduce?
- Discuss how Spark integrates machine learning, streaming, and other analytics workloads and increase productivity.
- Examples on how Spark machine learning and streaming have been used to drive value.



## Some further thoughts

- For execs, the priority should be high-level understanding of technology/algorithms and their use cases, rather than how to do it.
  - One could start with use cases to sets the stage for big data applications.
  - Cases can be a good way for execs to learn:
    - problem formulation, translation into data analytics problems, data needs, type of tools and algorithms most suitable, processes, architecture, challenges and risks.
    - The industry of use cases matter – variety, realism, and relatability.
  - Shortage of good big data cases: Gordon Gao (Maryland) uses a Harvard case "Kyruus: big data's search for the Killer app", but it is not ideal
- Gordon and I are starting a platform for b-school faculty who teach big data to share resources and best practices (still a work-in-progress). Ideas are welcome.

# Thank you!

## Google Drive Folder

 De Liu  
deliu@umn.edu