Introduction to DATA7201 Data Analytics at Scale

Dr Gianluca Demartini

DATA7201 Data Analytics at Scale - Week 1

Introductions…

Gianluca Demartini

g.demartini@uq.edu.au

- B.Sc., M.Sc. in Computer Science at U. of Udine, Italy
- Ph.D. in Computer Science at U. of Hannover, Germany
  - on Entity Search
- Worked at the University of Sheffield (UK), eXascale Infolab U. Fribourg (Switzerland), UC Berkeley (on Crowdsourcing), Yahoo! (Spain), L3S Research Center (Germany)
- At the U. of Queensland since Aug 2017
- Tutorials on Entity Search at ECIR 2012 and RuSSIR 2015, on Crowdsourcing at ESWC 2013, ISWC 2013, ICWSM 2016, WebSci 2016, Facebook

www.gianlucademartini.net

Introductions…

- Who you are and what's your background (your undergrad, work experience, etc)?
  - Which semester are you in?
- What do you expect to learn in DATA7201?
  - Technical skills?

- apps.elearning.uq.edu.au/wordcloud/55029

Course Aims

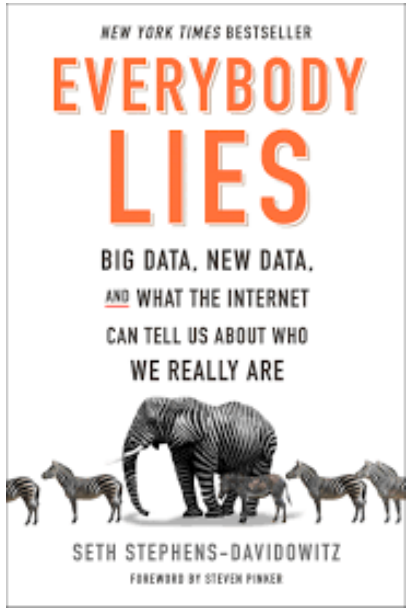
- The aim of this course is to give students knowledge about big data analytics architectures and help them to understand when and how to appropriately use such scalable data processing solutions. The course will help students understand the challenges and opportunities of big data analytics infrastructures. This course aims to:

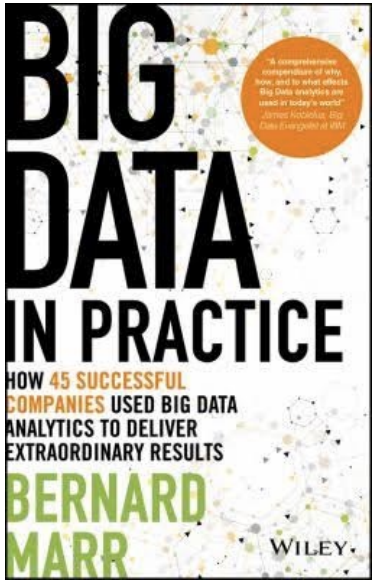
1. Provide an introduction to different big data **computational architectures and algorithms** (e.g., Map/Reduce);
2. Provide an overview of existing **big data analytics products** for volume, velocity, and variety of data;
3. Show how big data analytics is used in industry by means of **use cases**;
4. Provide practical **hands-on experience** through use of cloud-based software for big data processing.

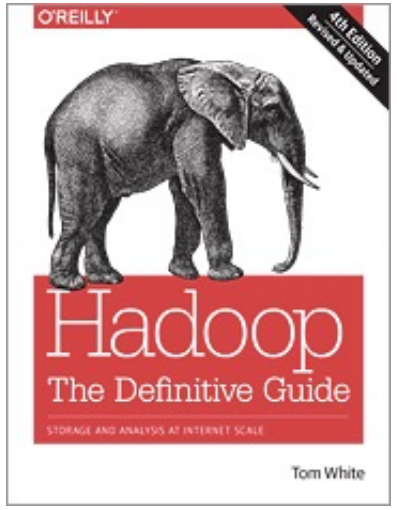
Course Objectives

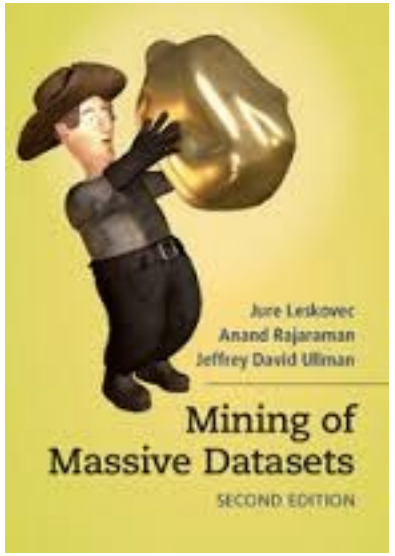
- After successfully completing this course you should be able to:
  - Solve **challenges** and leverage **opportunities** in dealing with Big Data
  - Use Big Data infrastructure solutions for **Volume, Variety, and Velocity** including industry-driven and open-source solutions
  - **Apply data analytics infrastructures** to best support data science practices for non-technical stakeholders (e.g., executives)
  - Compare alternative data analytics infrastructure solutions and **select the most appropriate** one for a certain use case
  - Judge in which situations Big Data analytics solutions are more or less appropriate.
  - Design the **most appropriate Big Data infrastructure solution** given a use case where to deploy Big Data solutions

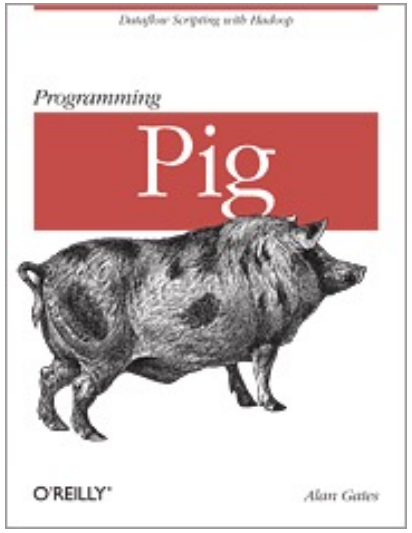
Recommended Material

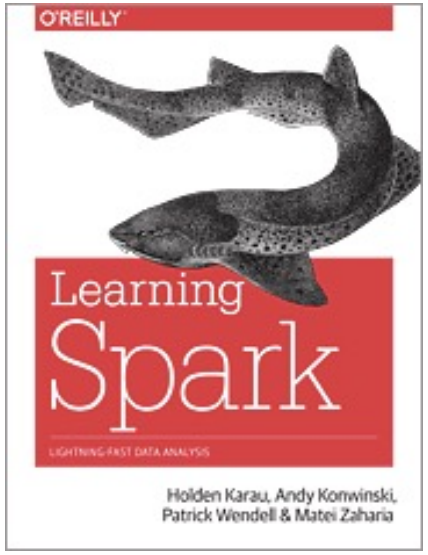


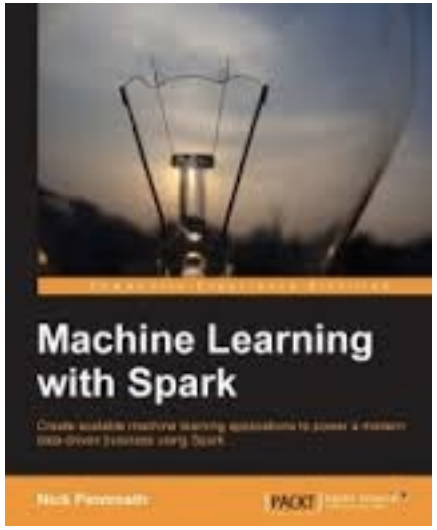












Practical Session Highlights

Week 3: HDFS

Week 4-5: Pig

Week 6-7: Spark, pySpark

Week 8-12: Project drop-in sessions

Practical Session Highlights

- Public cloud solution
  - Based on Amazon AWS – EMR (Spark, HDFS, Pig, etc)
  - DATA7201 cluster
    - Currently 2-4 nodes, 1.5TB HDFS storage,
    - At capacity 40+ nodes, ~60TB HDFS storage, ~2.5TB memory, ~1300 cores

	RAM	**HDFS**
Weeks 1-7	96 GB	3 TB
Weeks 8-10	320 GB	4 TB
Weeks 10-12	2368 GB	5 TB
Weeks 12-14	320 GB	5 TB

Tentative Menu – Part I

- Week 1: Introduction to Data Analytics at Scale
- Week 2: Supporting infrastructures and Use Cases
- Week 3: Storage Infrastructures for Large Data Volumes
  - Practical: Introduction to Cluster and HDFS
- Week 4: Analytics Queries for Large Data Volumes
  - Practical: PIG
- Week 5: Distributed Data Processing
  - Practical: PIG

Tentative Menu – Part II

- Part II: Other types of data

- Week 6: Processing Large Data Streams
  - Practical: PySpark
- Week 7: Processing Large Graph Data (1)
  - Practical: PySpark
- Week 8: Processing Large Graph Data (2)

Tentative Menu – Part III

- Part III: Scalable Data Analytics Applications

- Week 9: Recommender systems
- Week 10: Opinion Mining & Use Cases
- Week 11: Guest talk on Health Data Analytics
- Week 12: Large Language Models?
  - **Project report due**
- Week 13: Revision & Wrap-up

Assessments

Assessment Task	Due Date	Weighting
Online Quiz Module quizzes (Series of 3)	19 Feb 24 09:00 - 24 May 24 16:00	10%
Project Report Report on Dataset Analytics	20 May 24 16:00	45%
Exam - during Exam Period (Central) Final Exam	Examination Period	45%

- Week 5-6, 8-9, 12-13: Quiz: (60 minutes, online), MCQ, 5+5+5 points, multiple-attempts (max 10 marks, best 2 out of 3, see ECP)
- Week 8-12, Project: analyse a (given) dataset and write a report (1’500 words)
- Final Exam: on system architectures and use cases (120 min, invigilated, closed book)

Big Data

What is Big Data?

- Generated by systems, sensors, devices
- Multiple sources, multiple formats
- Velocity, volume and variety

- “If it fits in memory (32-128GB) it’s not Big Data”, M. Stonebraker

Dimensions of Big Data

- ***Volume*** *(amount of data)*: the large amount of data being generated and stored (normally in the order of TBs or PBs)
- ***Variety*** *(forms of data)*: the range of data types and sources being used, including unstructured data
- ***Velocity*** *(speed of data)*: the rate at which data is collected, shared and analysed - often real time streaming data (e.g., from social media)
- ***Veracity*** *(reliability of data)*: uncertainty in data quality (accuracy, provenance, relevance and consistently)

Data is huge (Volume)

- Facebook
  - processes 750TB/day of data
  - adds 7PB of photo storage / month
- This requires computers (a lot of them)
- Not only internet companies!
  - Banks, package delivery, governments, shops, etc.

Data is fast (Velocity)

- Twitter fire hose
  - In 2011, 1 000 Tweets per second (TPS)
  - In 2014, 20 000 TPS
  - With peaks: 143K TPS
- Services on top
  - DataSift: aggregate, filter and extract insights
- Not only internet companies!
  - Stock exchange, sensors in water network, etc.

Scale-up vs Scale-out

- Scale-up
  - Increasing the power of your computer (i.e, disk, memory, processor)
- Scale-out
  - Use many standard computers and distribute data and computation over them

Facebook Data Center (Sweden)

Facebook Data Center (New Mexico)

Source: facebook.com

Amazon Web Services (AWS)

AWS 2015 revenues: $1.8 billion USD

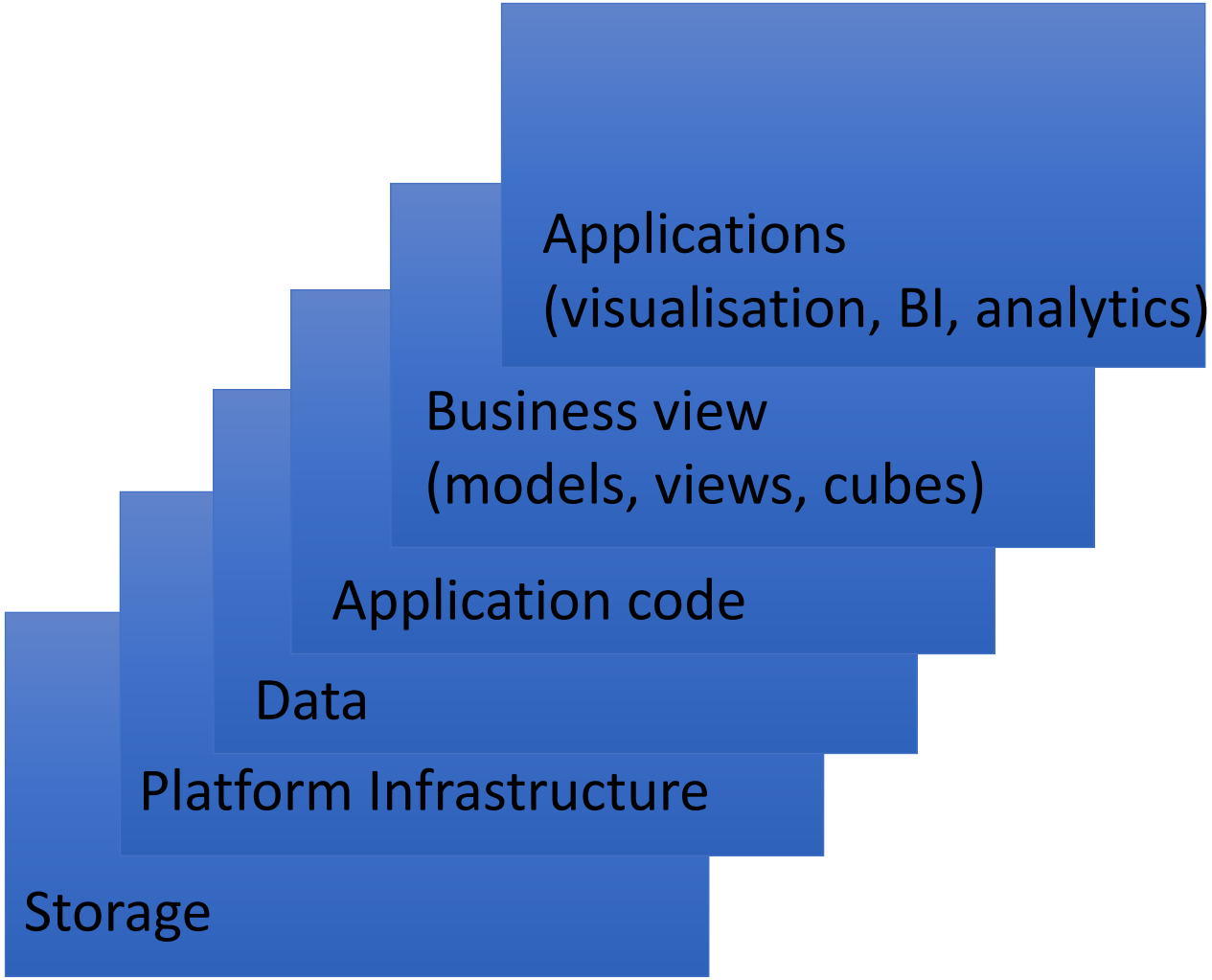
https://aws.amazon.com/about-aws/global-infrastructure/

Machines

- Google has around 900,000 servers (260 million watts == 200K homes)
- Google accounts for roughly 0.013% of the world's energy consumption
- CERN Large Hadron Collider 180MW
- Google's carbon footprint is zero

Data from Aug 2011

Big Data Stack



Big Data at Work: Dispelling the Myths, Uncovering the Opportunities, Thomas H Davenport

Timeline

Week	Date	Lecture	Prac	Assessment
1	21-Feb	Introduction to DATA7201 - Data Analytics at Scale	-	
2	28-Feb	Supporting Infrastructures and Use Cases	-	
3	6-Mar	Storage Infrastructures for Large Data Volumes	Intro to Cluster and HDFS	
4	13-Mar	Analytics Queries for Large Data Volumes	PIG(1)	
5	20-Mar	Distributed Data Processing	PIG (2)	
6	27-Mar	Processing Large Data Streams	PySpark (1)	**Quiz 1 Due (5)**
Semester Break				
7	10-Apr	Processing Large Graph Data (1) + use cases	PySpark (2)	
8	17-Apr	Processing Large Graph Data (2) + use cases	Project support	
9	24-Apr	Recommender Systems	Project support	**Quiz 2 Due (5)**
10	1-May	Opinion Mining + use cases	Project support	
11	8-May	Health Data Analytics (guest speaker)	Project support	
12	15-May	Large Language Models?	Project support	**Report Due (45)**
13	22-May	Course Revision	-	**Quiz 3 Due (5)**