Programming for Data Science at Scale

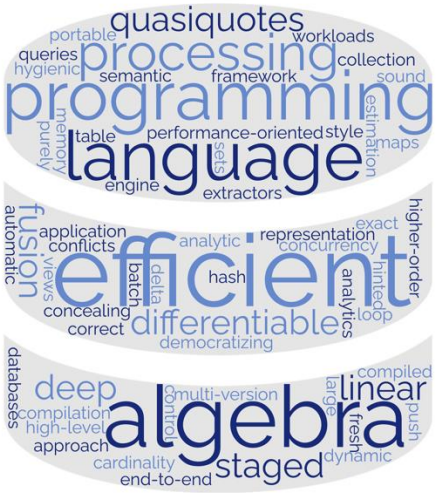
Introduction to Large-Scale Data Processing

THE UNIVERSITY *of* EDINBURGH

Amir Shaikhha, Fall 2024

Lecturer

- Amir Shaikhha
  - Reader
  - https://amirsh.github.io
  - Interests
    - Programming Languages
    - Database Systems
    - Compilers
    - Domain-Specific Languages



Essentials

- Webpage: http://course.inf.ed.ac.uk/pdss
- Piazza: https://piazza.com/class/m0mg39x2zlt3i7
- Learn: https://www.learn.ed.ac.uk/ultra/courses/_120834_1/outline

Course Timetable

- Lectures:
  - Tuesdays 11:10 – 12:30
  - Drill Hall, Forresthill
- Labs:
  - ~~Wednesdays 16:00 – 17:30~~
  - Tuesdays 14:00 – 15:30
  - Appleton Tower, 6.06

Course assessment

- 100% coursework → No Exam
- CW1: 25%
- CW2: 50%
- 2 x Quiz: 25%

Coursework Schedule

Week 1 (Sep 16)		Week 7 (Oct 28)	
Week 2 (Sep 23)		Week 8 (Nov 4)	***CW2***
Week 3 (Sep 30)		Week 9 (Nov 11)	
Week 4 (Oct 7)	***CW1***	Week 10 (Nov 18)	
Week 5 (Oct 14)		Week 11 (Nov 25)	***Quiz2***
Week 6 (Oct 21)	***Quiz1***		

Labs

- Will help you with coursework
- 1 session of 2 hours

- Start: Week 3
- End: Week 10

- ~~Time: Wednesday 16:00 – 17:30~~
- Time: Tuesday 14:00 – 15:30
- Location: Appleton Tower, 6.06

Preferred Prerequisites

- Programming Languages
  - Strong programming skills
    - Java
    - Scala
    - C++
    - Python

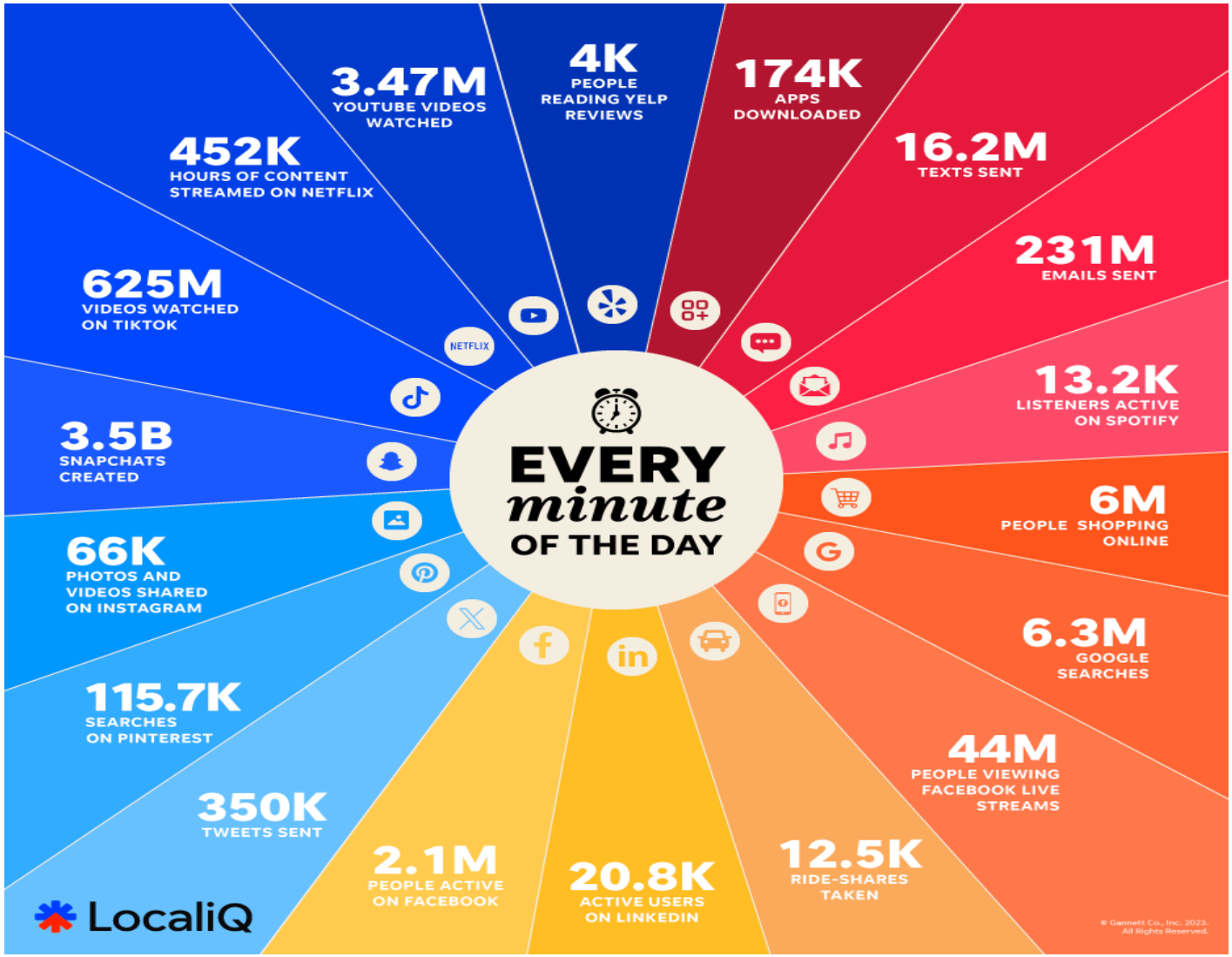
Acknowledgements

The lecture slides draw on notes by several folks to which I am grateful, in particular:

- P. Bhatotia (formerly Univ. of Edinburgh, now TUM)
- M. Odersky (EPFL)
- C. Koch (EPFL)
- H. Miller (CMU)
- M. Zaharia (Berkeley & DataBricks)
- The many researchers whose work I will mention in the slides (I will give pointers to their research papers)

COURSE OVERVIEW

Internet in 2024



https://localiq.com/blog/what-happens-in-an-internet-minute//

Mainstream Languages for Data Scientists

Mainstream Languages for Data Scientists (cont.)

Pros

✓Rapid Development

✓Large community

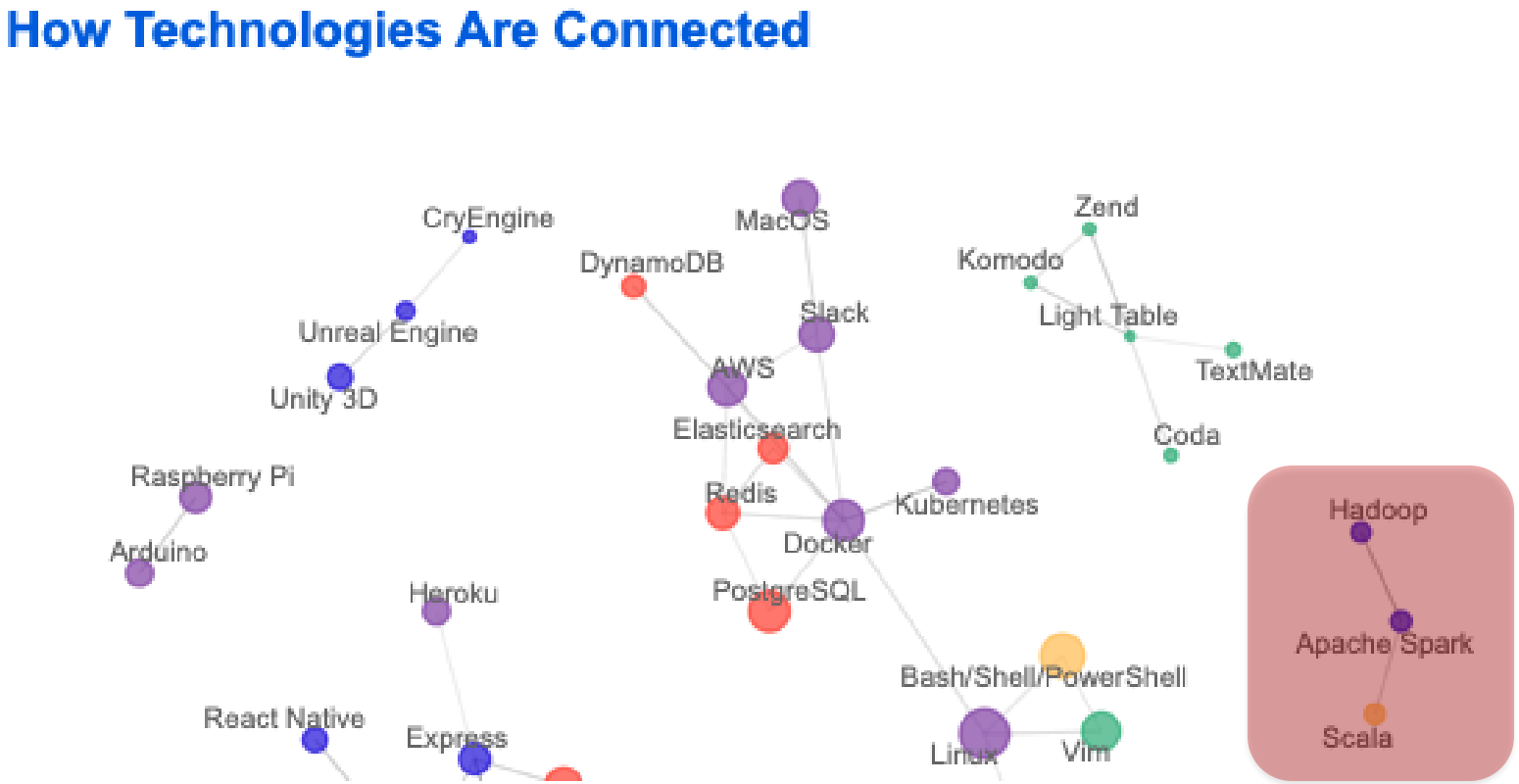
Cons

❖What to do with large datasets?

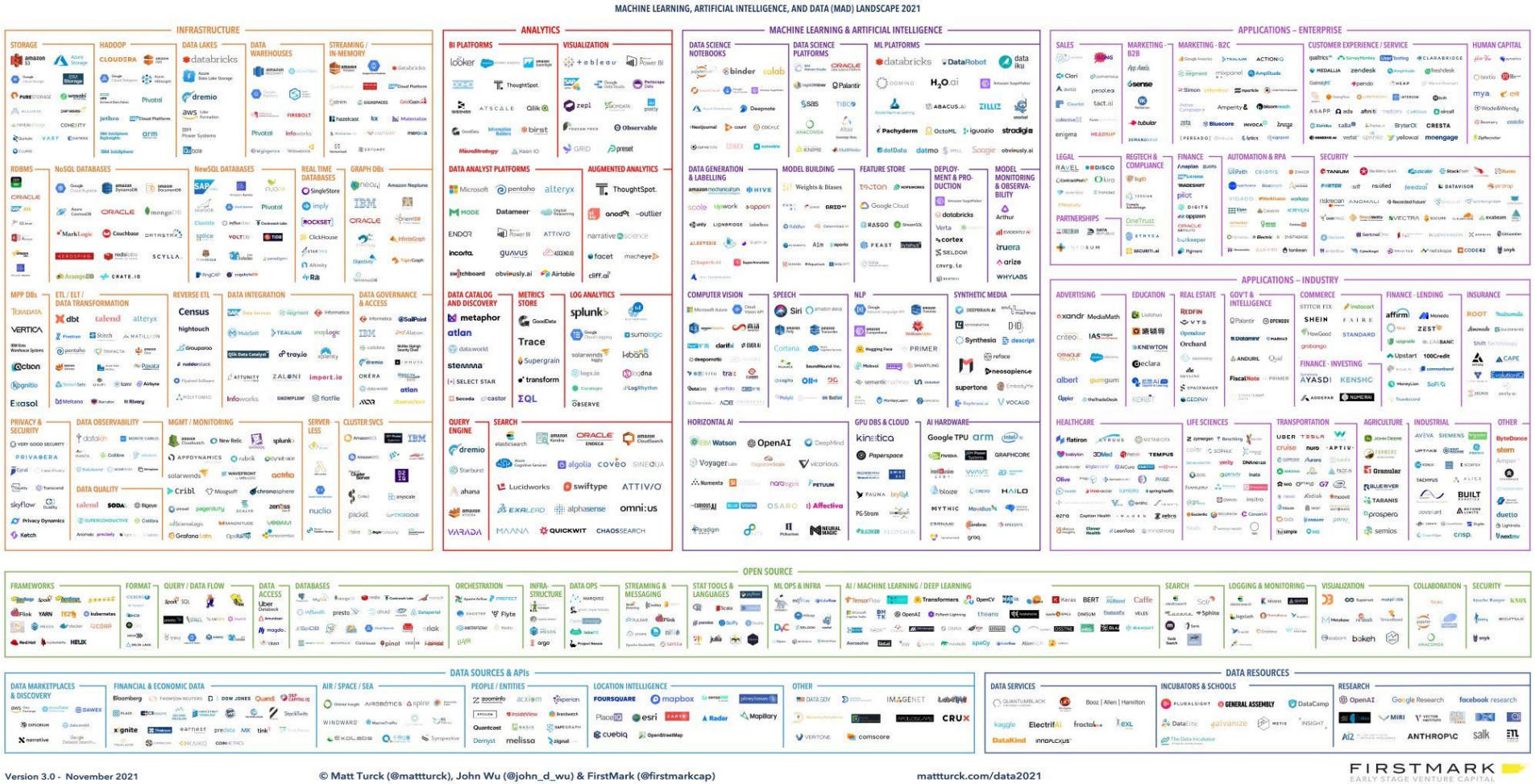
Rewrite from scratch ☹

Is there any language without this issue?

Why Scala is related to BigData?



https://insights.stackoverflow.com/survey/2019


MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021
INFRASTRUCTURE
ANALYTICS
MACHINE LEARNING & ARTIFICIAL INTELLIGENCE
APPLICATIONS – ENTERPRISE
APPLICATIONS – INDUSTRY
OPEN SOURCE
DATA SOURCES & APIs
DATA RESOURCES
Version 3.0 - November 2021
© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)
mattturck.com/data2021
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Mainstream Big Data models

How to store, manage and process Big Data by harnessing large clusters of commodity nodes

- MapReduce family: simpler, more constrained

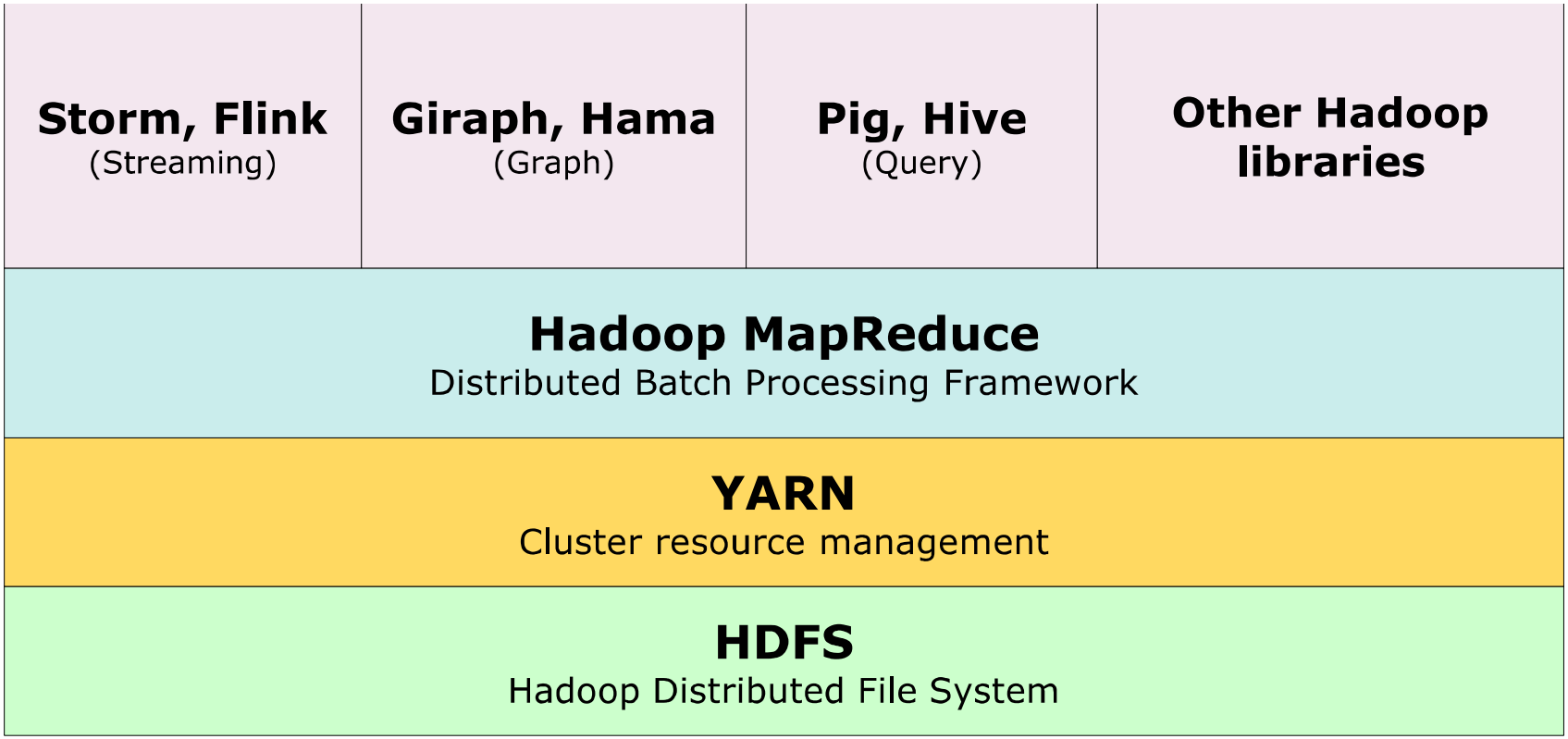
HadoopDB

- Dataflow family: enables more complex processing & data, optimization opportunities

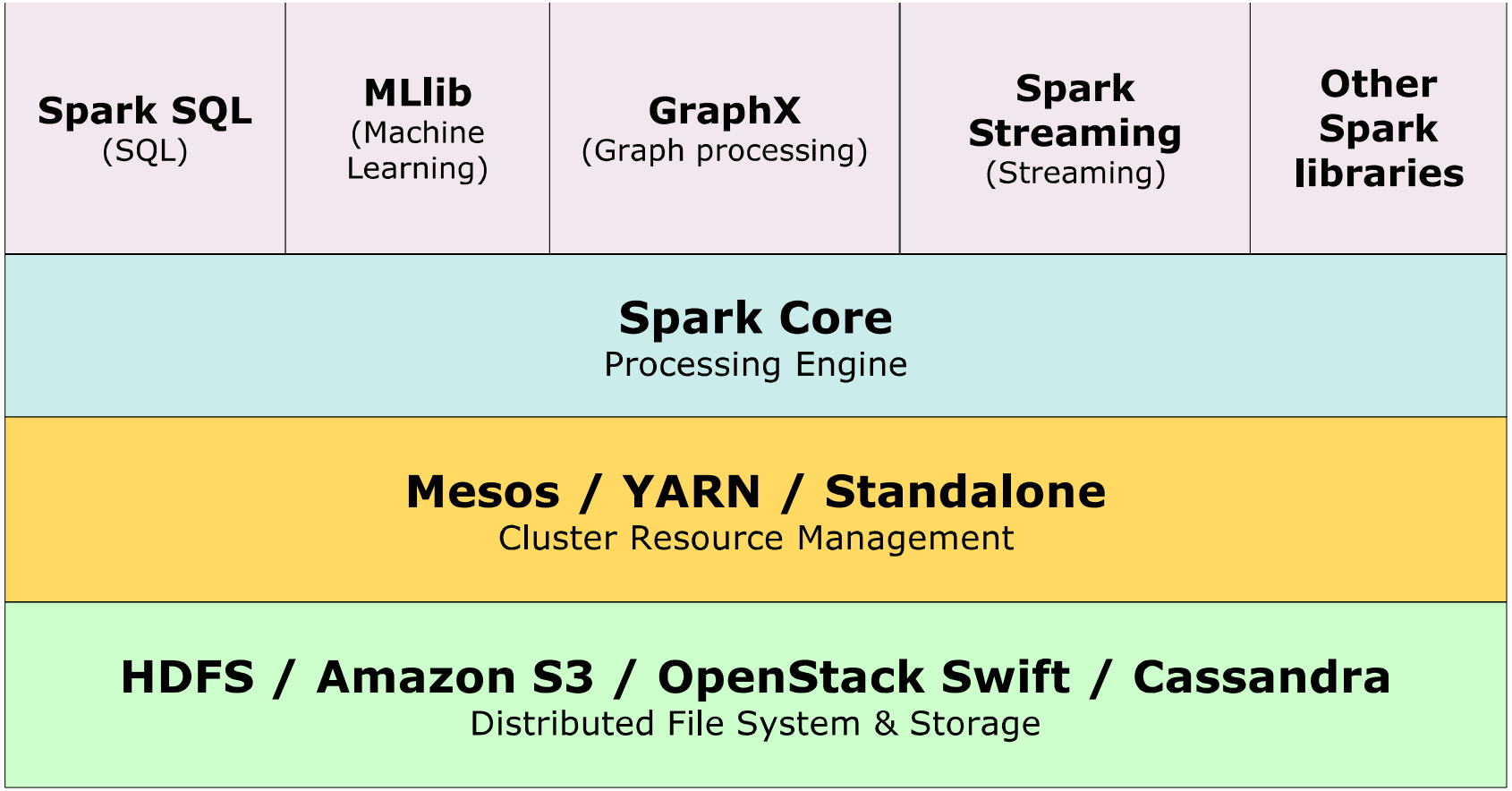
Google Pregel

Microsoft Dryad

The Hadoop Ecosystem



Spark Software Stack



Syllabus

- Data-Parallel Programming, Functional Collections
- Distributed Data-Parallel Programming
- Distributed Key-Value Processing
- Distributed Query Processing
- Optimizing Distributed Data Processing
- Distributed Graph Processing
- Efficient Non-Distributed Processing
- Distributed Tensor Processing

Guest Lecture

- Week 10
- Dr. Manos Karpathiotakis

QUESTIONS?