Programming for Data Science at Scale

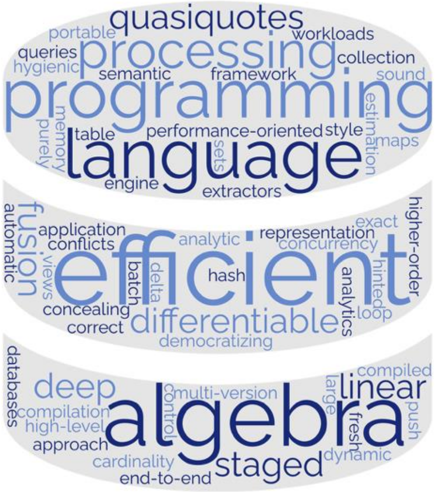
# Introduction to Large-Scale Data Processing

THE UNIVERSITY
*of* EDINBURGH

Amir Shaikhha, Fall 2025

# 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/mf7ce4fxdmg4yq
- Learn: https://www.learn.ed.ac.uk/ultra/courses/_127073_1/outline

# Course Timetable

- Lectures (UPDATED):
  - Thursdays 16:10 – 17:30
  - Weeks 2, 5, 10: Lecture Theatre G.04 - Robson Building
  - Weeks 3, 4, 6, 7, 8, 9: Lecture Theatre G.03 - 50 George Square
- Labs:
  - Mondays 14:00 – 15:30
  - Fridays 15:00 – 16:30
  - Appleton Tower, 6.06

# Course assessment

- 100% coursework → No Exam
- CW1: 70%
- CW2: 30%

# CW1: Group Coursework

- Goal: Learn to design, implement, optimize, evaluate, and document a large-scale data science system
- Group size: 3 students (formed by your own)
- Stage 1: Design, Implement, Optimize, Evaluate
- Stage 2: Write a paper
  - Template will be provided

# CW2: Individual Coursework

- Goal: Learn to assess a large-scale data processing system
- Write a review for the paper and the code
  - Template will be provided

# Coursework Schedule

| | | | |
|---|---|---|---|
| Week 1 (Sep 15) | | Week 7 (Oct 27) | ***CW1*** |
| Week 2 (Sep 22) | | Week 8 (Nov 3) | |
| Week 3 (Sep 29) | | Week 9 (Nov 10) | |
| Week 4 (Oct 6) | ***CW1*** | Week 10 (Nov 17) | ***CW2*** |
| Week 5 (Oct 13) | | Week 11 (Nov 24) | |
| Week 6 (Oct 20) | | | |

# Labs

- Start: Week 3
- End: Week 10
- 3 Lab Sessions
  - Weeks 3, 4, 6
  - Will help you with coursework
  - Not graded
- Rest of the weeks
  - Work on group coursework with your peers
- 2 sessions of 2 hours per week
  - You need to only attend 1 session

# Preferred Prerequisites

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

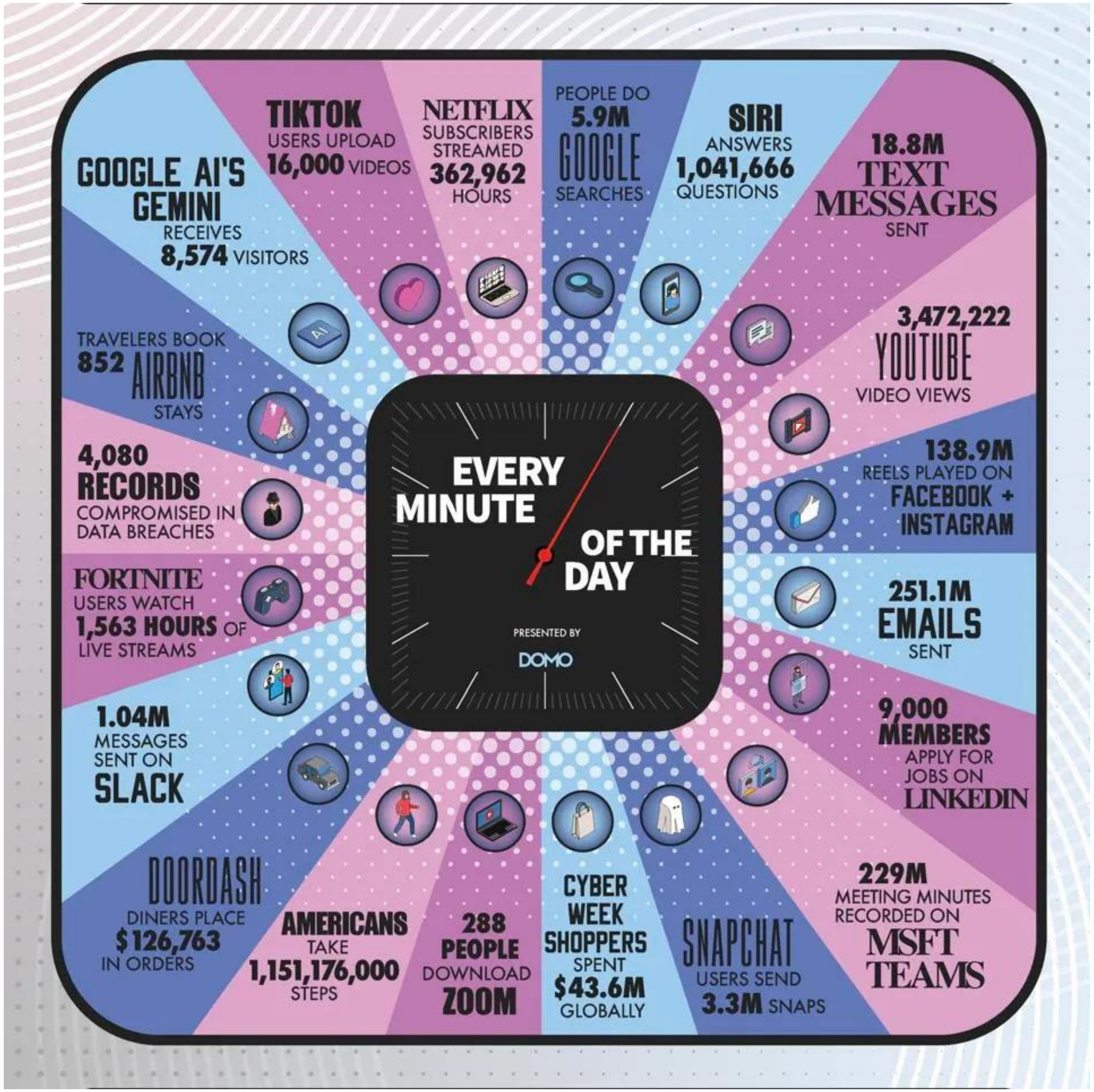
# Acknowledgements

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)

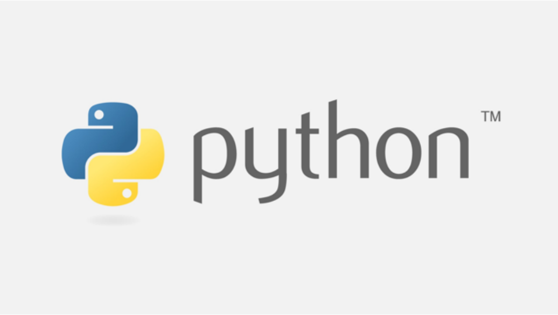
# COURSE OVERVIEW

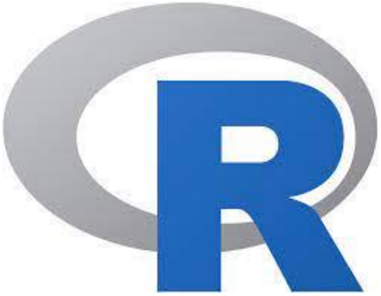
# Internet in 2025

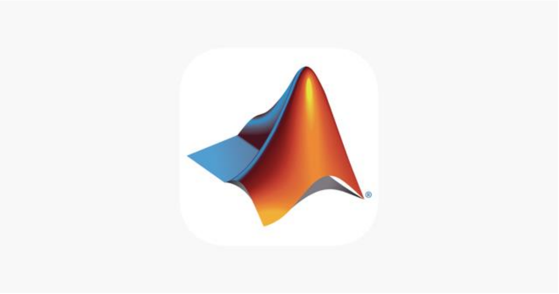

GOOGLE AI'S GEMINI RECEIVES 8,574 VISITORS
TIKTOK USERS UPLOAD 16,000 VIDEOS
NETFLIX SUBSCRIBERS STREAMED 362,962 HOURS
PEOPLE DO 5.9M GOOGLE SEARCHES
SIRI ANSWERS 1,041,666 QUESTIONS
18.8M TEXT MESSAGES SENT
3,472,222 YOUTUBE VIDEO VIEWS
138.9M REELS PLAYED ON FACEBOOK + INSTAGRAM
251.1M EMAILS SENT
9,000 MEMBERS APPLY FOR JOBS ON LINKEDIN
229M MEETING MINUTES RECORDED ON MSFT TEAMS
SNAPCHAT USERS SEND 3.3M SNAPS
CYBER WEEK SHOPPERS SPENT $43.6M GLOBALLY
288 PEOPLE DOWNLOAD ZOOM
AMERICANS TAKE 1,151,176,000 STEPS
DOORDASH DINERS PLACE $126,763 IN ORDERS
1.04M MESSAGES SENT ON SLACK
FORTNITE USERS WATCH 1,563 HOURS OF LIVE STREAMS
4,080 RECORDS COMPROMISED IN DATA BREACHES
TRAVELERS BOOK 852 AIRBNB STAYS
EVERY MINUTE OF THE DAY
PRESENTED BY
DOMO

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

# Mainstream Languages for Data Scientists







# Mainstream Languages for Data Scientists (cont.)

Pros

✓Rapid Development

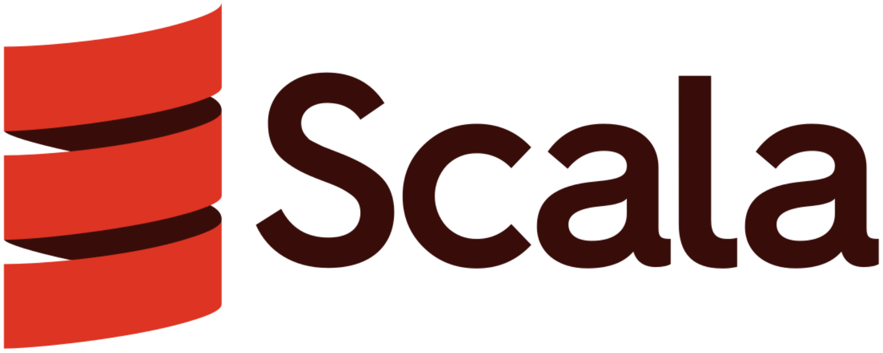
✓Large community

Cons

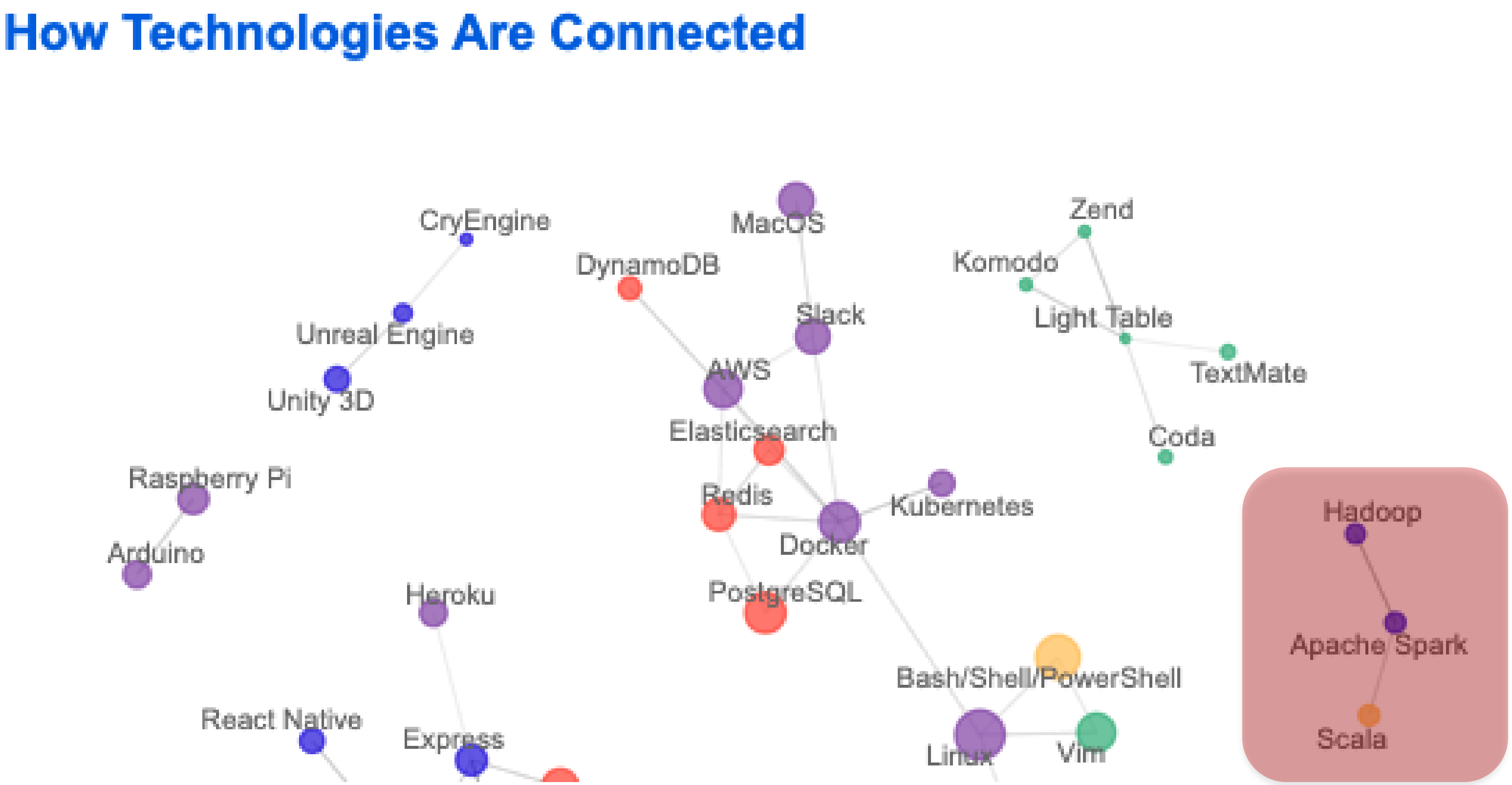
❖What to do with large datasets?

**Rewrite from scratch ☹**

# Is there any language without this issue?



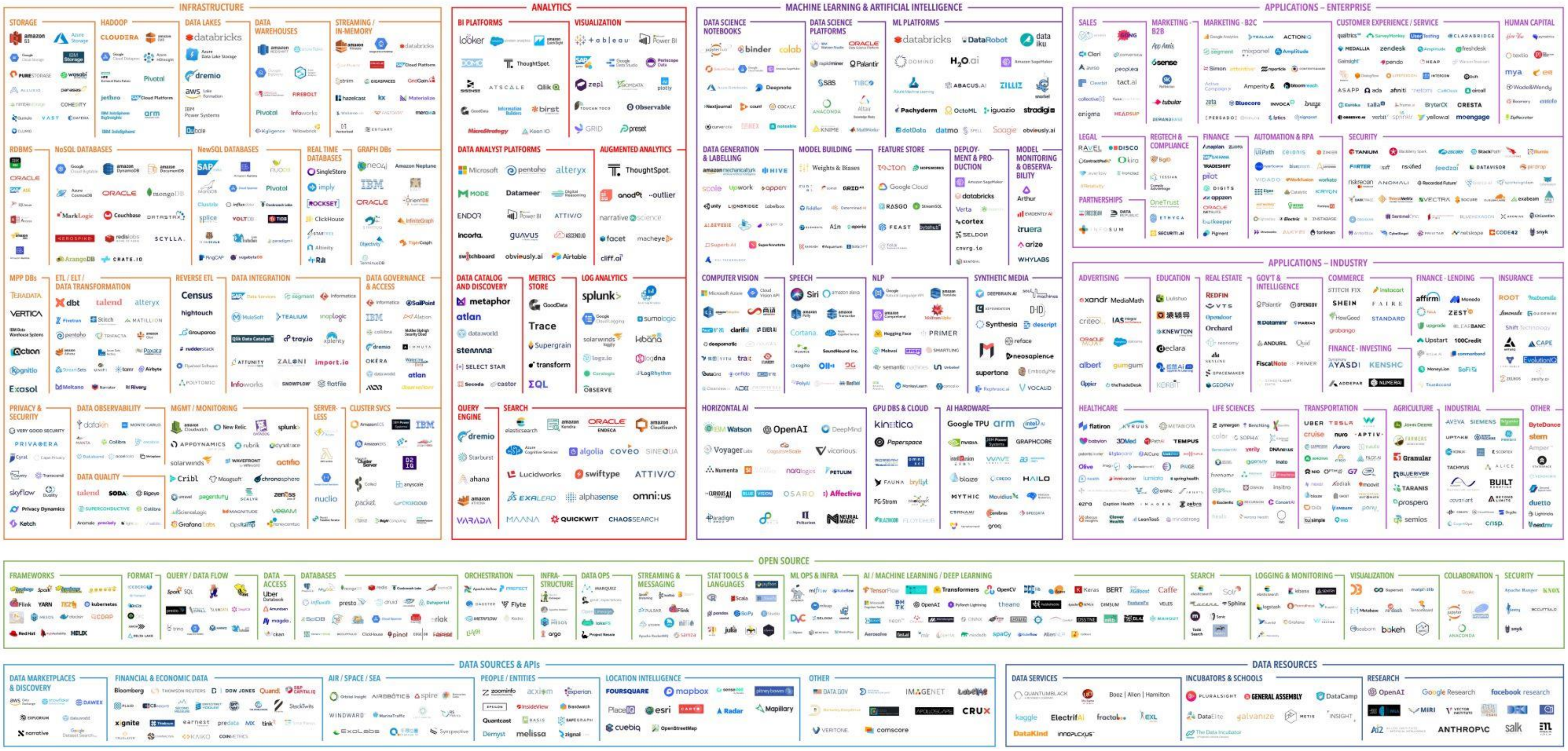
# Why Scala is related to BigData?



How Technologies Are Connected

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

## MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021



INFRASTRUCTURE

STORAGE · HADOOP · DATA LAKES · DATA WAREHOUSES · STREAMING / IN-MEMORY · RDBMS · NoSQL DATABASES · NewSQL DATABASES · REAL TIME DATABASES · GRAPH DBs · MPP DBs · ETL / ELT / DATA TRANSFORMATION · REVERSE ETL · DATA INTEGRATION · DATA GOVERNANCE & ACCESS · PRIVACY & SECURITY · DATA OBSERVABILITY · DATA QUALITY · MGMT / MONITORING · SERVERLESS · CLUSTER SVCS

ANALYTICS

BI PLATFORMS · VISUALIZATION · DATA ANALYST PLATFORMS · AUGMENTED ANALYTICS · DATA CATALOG AND DISCOVERY · METRICS STORE · LOG ANALYTICS · QUERY ENGINE · SEARCH

MACHINE LEARNING & ARTIFICIAL INTELLIGENCE

DATA SCIENCE NOTEBOOKS · DATA SCIENCE PLATFORMS · ML PLATFORMS · DATA GENERATION & LABELLING · MODEL BUILDING · FEATURE STORE · DEPLOYMENT & PRODUCTION · MODEL MONITORING & OBSERVABILITY · COMPUTER VISION · SPEECH · NLP · SYNTHETIC MEDIA · HORIZONTAL AI · GPU DBS & CLOUD · AI HARDWARE

APPLICATIONS – ENTERPRISE

SALES · MARKETING – B2B · MARKETING – B2C · CUSTOMER EXPERIENCE / SERVICE · HUMAN CAPITAL · LEGAL · PARTNERSHIPS · REGTECH & COMPLIANCE · FINANCE · AUTOMATION & RPA · SECURITY

APPLICATIONS – INDUSTRY

ADVERTISING · EDUCATION · REAL ESTATE · GOV'T & INTELLIGENCE · COMMERCE · FINANCE - LENDING · INSURANCE · FINANCE - INVESTING · HEALTHCARE · LIFE SCIENCES · TRANSPORTATION · AGRICULTURE · INDUSTRIAL · OTHER

OPEN SOURCE

FRAMEWORKS · FORMAT · QUERY / DATA FLOW · DATA ACCESS · DATABASES · ORCHESTRATION · INFRASTRUCTURE · DATA OPS · STREAMING & MESSAGING · STAT TOOLS & LANGUAGES · ML OPS & INFRA · AI / MACHINE LEARNING / DEEP LEARNING · SEARCH · LOGGING & MONITORING · VISUALIZATION · COLLABORATION · SECURITY

DATA SOURCES & APIs

DATA MARKETPLACES & DISCOVERY · FINANCIAL & ECONOMIC DATA · AIR / SPACE / SEA · PEOPLE / ENTITIES · LOCATION INTELLIGENCE · OTHER

DATA RESOURCES

DATA SERVICES · INCUBATORS & SCHOOLS · RESEARCH

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

hadoop MapReduce

HIVE

hadoop

HadoopDB

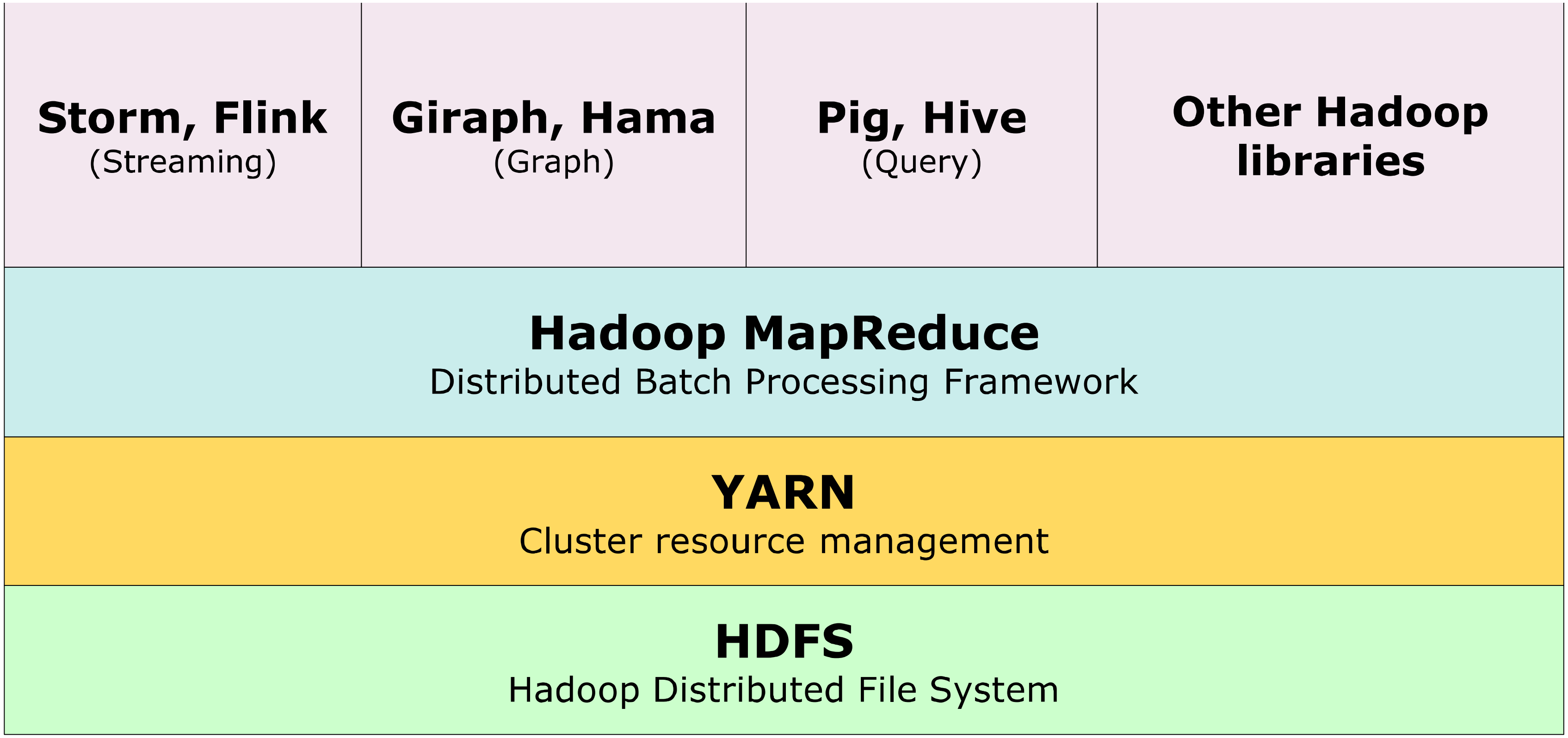
- Dataflow family: enables more complex processing & data, optimization opportunities

APACHE Spark™

Google Pregel

Microsoft Dryad

# The Hadoop Ecosystem



**Storm, Flink**
(Streaming)

**Giraph, Hama**
(Graph)

**Pig, Hive**
(Query)

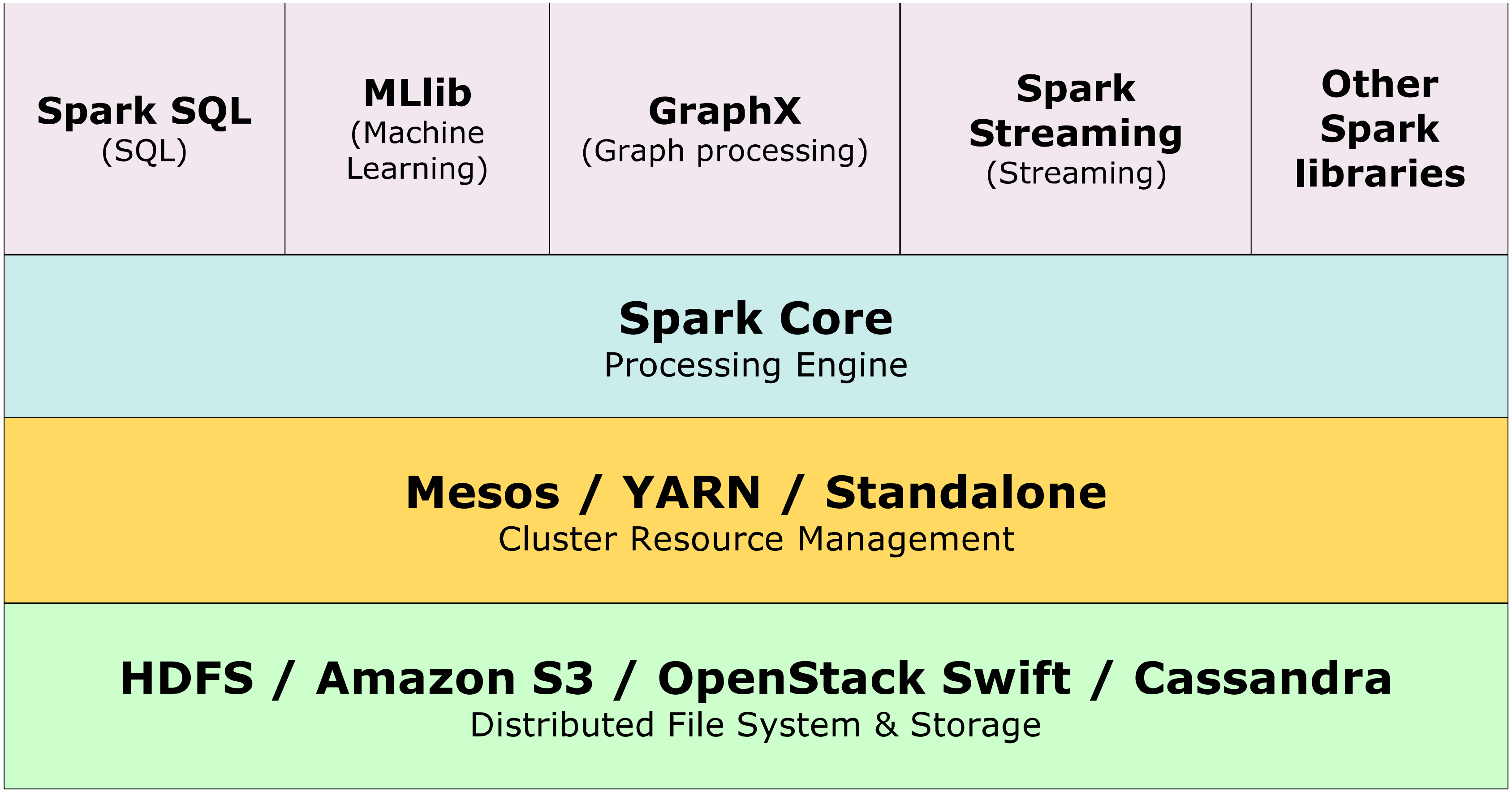
**Other Hadoop libraries**

**Hadoop MapReduce**
Distributed Batch Processing Framework

**YARN**
Cluster resource management

**HDFS**
Hadoop Distributed File System

# Spark Software Stack



**Spark SQL**
(SQL)

**MLlib**
(Machine Learning)

**GraphX**
(Graph processing)

**Spark Streaming**
(Streaming)

**Other Spark libraries**

**Spark Core**
Processing Engine

**Mesos / YARN / Standalone**
Cluster Resource Management

**HDFS / Amazon S3 / OpenStack Swift / Cassandra**
Distributed File System & Storage

# Syllabus

- Data-Parallel Programming
- Functional Collections
- Distributed Data-Parallel Programming
- Distributed Key-Value Processing
- Optimizing Distributed Data Processing
- Distributed Query Processing
- Distributed Graph Processing

# Guest Lecture

- Week 10
- Dr. Manos Karpathiotakis

Meta

# QUESTIONS?