### Programming for Data Science at Scale

# Introduction to Large-Scale Data Processing



Amir Shaikhha, Fall 2025

## Lecturer

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





### **Essentials**

- Webpage: http://course.inf.ed.ac.uk/pdss
- Piazza: <a href="https://piazza.com/class/mf7ce4fxdmg4yq">https://piazza.com/class/mf7ce4fxdmg4yq</a>
- 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 largescale 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)	
Week 2 (Sep 22)		Week 8 (Nov 3)	CW1
Week 3 (Sep 29)		Week 9 (Nov 10)	
Week 4 (Oct 6)		Week 10 (Nov 17)	CW2
Week 5 (Oct 13)	CW1	Week 11 (Nov 24)	CVVZ
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

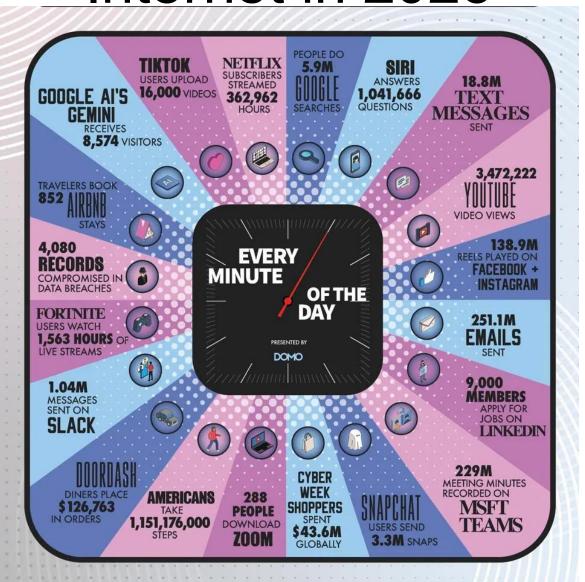
# Acknowledgements

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



# Mainstream Languages for Data Scientists







# Mainstream Languages for Data Scientists (cont.)

#### Pros

- ✓ Rapid Development
- ✓ Large community

#### Cons

What to do with large datasets?

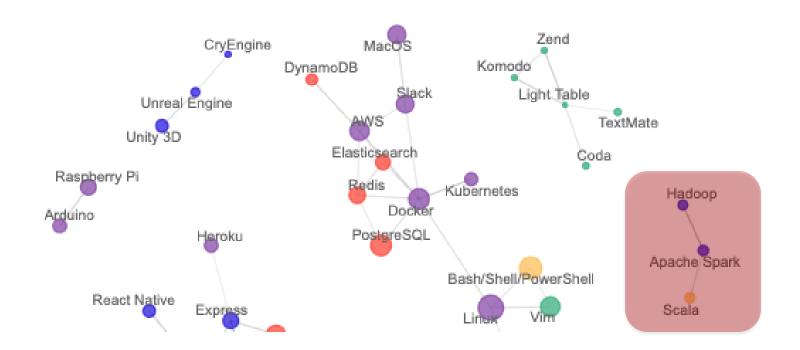
Rewrite from scratch 😊

# Is there any language without this issue?



# Why Scala is related to BigData?

#### **How Technologies Are Connected**



#### MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021 INFRASTRUCTURE MACHINE LEARNING & ARTIFICIAL INTELLIGENCE APPLICATIONS - ENTERPRISE DATA LAKES -VISUALIZATION -DATA SCIENCE -MARKETING - MARKETING - B2C ΗΠΜΑΝ CAPITA ML PLATFORMS -SALES -Harthe Gara CLOUDERA = awase databricks +ableau Power Bl sdatabricks DataRobot O data um ACTIONIG Strange (2) O Warning Sbinder cold Stoph .... Coople Option H.O.ai spidminer QPalantin Owner Other Street A auso peopled O ... (S)= mya 🕜 es dremio @ ATSCALE Qlik@ Zepl Microsta plothy MABACUS AI ZILLIZ A Appellances Despricts @Wade8Wendy aws Life @Beerry CEE zefa 👰 Blue COHESTY BryterOX CRESTA Provenu roco Observable \*birst Nextjournal Shoount @ COCALC € Pachyderm OctoML 1-iguazio stradigia Divoted Information enigma HEADSU Gesevente IIIEX nateab Dubole SRID Apreset AKNIME . MARINA Onumb LEGAL -FINANCE -COMPLIANC AUGMENTED ANALYTICS DATA GENERATION MODEL BUILDING -RAVEL ...DISCO Anaplan zuos Otenname Okira 8190 MONITORIN TRIM 186 mneoul Am & OBSERVAmazoe desame Microsoft pentaho alteryx ThoughtSpot. Weights & Risses TOUTON A HOPSHORKS DUCTION C) SingleStre amazon mechanicalturis di HIVE BILITY piot CRACLE MEI 517 Ameron Se (a) imply Google Cloud 8cole Upwork sapper Come GRIDA Comple DIGITS. 222 ASE Arthur III Bee Access KEND Datameer = 049201 anod<sup>o</sup>t -outlier databricks ORACLE - Oriental PARTMERCHIPS ez oppzen ROCKSET «Винку привиюсе забаба RASGO ( Street SQ. Verta 🐞 ORACLE MENUIS all evidently CONTROL 3 DATA ENDO2 ATTIVO narrative ascience D Ame splice ClickHouse 9 cortex Aim caporle FEAST MINE truera + INTOSUM gunvus @xxxxxxx S SELDON 134 150 redsions SCYLLA facet macheve Tiger Cough ^ arize O Altinity envrg. to 68 PROGRE & supetieted obvieusly.ai 🔊 Airtoble cliff.ai ArangaDB - CRATE IO de Rati WHYLABS APPLICATIONS - INDUSTRY MPP DRs ETL/FET/ -REVERSE ETL DATA GOVERNANCE -DATA CATALOG COMPUTER VISION -SYNTHETIC MEDIA -ADVERTISING -FOLICATION - REAL ESTATE COMMERCE -FINANCE-LENDING - INSURANCE -DATA INTEGRATION -METRICS LOG ANALYTICS -SPEECH -NIP -GOVT8 -STITCH FIX Finstone Microsoft Azore Charles Siri O arrescon alexy (II) Comple AMASSA. CO DECPORAIN AL SOUT L TERADATA splunk> **B** Luisho affirm Moned ROOT Pale xandr MediaMati Mdbt talend atteryx M metaphor SHEIN FAIRE GoodData CYTE QPalantir @ OPENDON Marrowson D-ID (i) ---- (ii) \* Ow ZEST® VERTICA hightouch IBM aware. amazen Garqueriana ◎ 塘辆导 Amenade Source M MuleSoft >TEALIUM Inoplogic atlan Google Classifungsing HowGood STANDARD criteo IAS 1994 🔾 Synthesia 📴 descrip Trace Collibra Searly God clarify 4 BBH KNEWTON Orchard adata.world PRIMER W Hugging Face solarwinds Hoana GRACIE S (ii) reface & ANDURIL Child A ACAPE Cdremio I HHUTA Supergrain FINANCE - INVESTING stemma Geclara Payata Desire A Commonton (i) logdna Vaminitat O ▼ Evolution® ZALONI import.io OKERA Wheeling e' transform AYASDI KENSHO el SELECT STAR progito Olles DG albert gumgum O MINIO Access to the Witness & Address supertone @ Embodyl S SPACEMAKE KERBÎT Secoda @castor EQL 11733 MonkeyCasers PROPERSOLA V VOCALID Fygeol OBSERVE CHANGE ADE CARREST PRIVACY 8 HORIZONTAL AI OUERY GPU DRS & CLOUD AI HARDWARE HEALTHCARE LIFE SCIENCES TRANSPORTATION AGRICULTURE - INDUSTRIAL ENGINE amazon ORACLE amazon Cloudieuro Google TPU arm intell UBER TESLA W O America TRM anazoe O New Relic kin≣tica AVRUUS (SMITABIOD ADHN Depre nuro -APTIV dremin UPTAKE @IZZIes ... O Amountes 10° -PRIVAGERA APPOYNAMICS O rubrik Capital Services | algolia COVEO SINEQUI ESCORTEX A and the second second 🏐 Starbu rad@rim VVAVE 23 === O second Co. A PACE 0 Chipler Server 50 BORRION PRO PAGE TACHYUS A ALICE A. Numenta Towny @ Transcard Lucidworks swiftype ATTIV/O nordingles PETULM AND OFFICE G7 (2) Cribl Thougsoft 0 A ahana blaze CORECO HAILO W FAUNA brottet A BUILT Rh amerale w neural Kodiak Smoowin Mowey insitro 32 TADANIS skyflow O & EXALERO | alphasense omni:us MYTHIC Movidus N 4 THE Desare @ over arrayan covariant Assesse nuclio amazon PG-Strom packet ath инколи Zzebra prospero duetto Cipio Wasser core CTINAMI Gerebras - TELLENIA after course (Scotteres - Supple II MAGIC NGNFAN MAANA \*QUICKWIT CHAOSSEARCH PRINCE STORES Clover di Leoniosi il ministrong tusmoe Ovo a semios **€** Ketch Anomalo precisely \$1000 miles (Grafang Lobs OpeRamp Conscious) Common Crisp. STREAMING & -MLOPS & INFRA -6 CV 23 Ser. OO Superset matphill Uber Arme S. training (Courts) 3 © ■Scala ■ nantes & Flyte Millerout DM @ OpenAI Or De Series Manatese Pash Terroritore Lineston Orandes Y. SCHLERE GERM infotor \_\_ Beers MCCLITTAL THE PERSON LINES VICTOR per HIGHE (Spender Carry Chican @ 7:00 #ScOB 🥳 🙇 🐧 www. O mer 🙆 nille MISON MINES 💯 juli 🌦 🌚 Ossborn bokeh 0 # snyk Task pain T grop L population

Version 3.0 - November 2021 © Matt Turck (@mattturck), John Wu (@john\_d\_wu) & FirstMark (@firstmarkcap)

O OSSESSED ANDROTICS ASPIRE # 127

€ExoLabs 0,1808 & Sympetive

THOMSON REUTERS D | DOW JONES Quand Cantal

Main Charles Street Street Street Street

THE SHARE CONCERCS

DATA MARKETPLACES

aws on Securior Spawn

В сочония — — фактор

X namative Goods

**DATA SOURCES & APIS** 

PEOPLE / ENTITIES -

Demyst melissa

Z zoominfo acxiem texperior

Quanticast Salasia Salasia

LOCATION INTELLIGENCE

FOURSQUARE ( mapbox ( mapbox

Place esri

mattturck.com/data2021

VERTONE. COMSCOR

DATA SERVICES

O GUNNTUMBLACK

DataKind inona cyus

Booz | Allen | Hemilton

kaggle ElectrifAl froctobes XEXL

OpenAI Google Research facebook research

FIRSTMARK

MIRI V VICTOR

ANTHROPIC Salk

- DATA RESOURCES

O PLURALSIGHT O GENERAL ASSEMBLY DataCan

A DataElite galvanize P Heris

RESEARCH

INCUBATORS & SCHOOLS —

C The Data Incubator

# 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



Pregel

Google Microsoft

# The Hadoop Ecosystem

Other Hadoop Storm, Flink **Giraph, Hama** Pig, Hive libraries (Streaming) (Graph) (Query) **Hadoop MapReduce** Distributed Batch Processing Framework **YARN** Cluster resource management **HDFS** Hadoop Distributed File System

# Spark Software Stack

Other **Spark MLlib** Spark SQL **GraphX Spark Streaming** (Machine (SQL) (Graph processing) Learning) (Streaming) 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



## **QUESTIONS?**