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: <u>http://course.inf.ed.ac.uk/pdss</u>
- Piazza: <u>https://piazza.com/class/m0mg39x2zlt3i7</u>
- 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 \rightarrow 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)	
Week 3 (Sep 30)		Week 9 (Nov 11)	CW2
Week 4 (Oct 7)	CW/1	Week 10 (Nov 18)	
Week 5 (Oct 14)	CWI	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?



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			APPLICATIONS - ENTERPRISE			
STORAGE DATA LAKES DATA STREAMING /	BI PLATFORMS VISUALIZATION	DATA SCIENCE ML PLATFORMS	SALES MARKETING - B2C CUSTOMER EXPERIENCE / SERVICE HUMAN CAPITAL			
Anne CLOUDERA States States WAREHOUSES	looker 🚌 mar 🕢 🛲 🥸 + a b le a u 🖬 Rover Bi	NOTEBOOKS PLATFORMS				
A terrer and a terrer and a terrer a terre a terrer a terrer a terre a terrer a terrer a terrer a terr		Night Shinder colob	Ca Clari @comence ApAntis @isegment mixponel @Ampilitude @HEDALLA zendesk @Ampinuk @freshdesk Otectio @Be-			
Contraction Control Co	T. ThoughtSpot.	Chanced States and Annual Southing Palantin COMING H2O.CI O Annual Southing	A anto peopleg Simon minimum Zingarise @ commune			
C NURESTORIAGE @ WORKER Physical Physical Officeration Physical Construction Constr	S OT C Czepl Varman		P Canter tact.al Q Adve tact.al Atvantation (Ad			
A ALLER DATASAS AND ALLER	sander ATSCALE GIR O	Abretterielaring social	ASAPP Q ads afiniti metors carous Q eleval			
a motive management of the second sec	G contras Mitimutias & birst Provens roco O Observable	Nextjournal De court @ COCALC	eniona HEADUP 2012 (Busecore HNOCA [®] bruge (Courses talla [®] Armes BryterOX CRESTA			
Carele VAST Corres Sparses		Graverten EREX Constanting Automatic Constanting Constanting Statements	angen angen angen (Persano) Derve Spice General Commen vebti Spinkr y jelowal moengage Spicewar			
Citication International Control of the Control of	Marodinategy Akaonio 2 0000 70 1000	and the second dated set. Succession	LEGAL REGIECH & FINANCE AUTOMATION & RPA SECURITY			
RDBMS NoSQL DATABASES NewSQL DATABASES REAL TIME GRAPH DBs	DATA ANALYST PLATFORMS AUGMENTED ANALYTICS	DATA GENERATION MODEL BUILDING FEATURE STORE DEPLOY MODEL	RAVEL BEDISCO COMPLIANCE Anaplan zuora DiPath Celonis @Itwee CTANUM @IndurySon Carcola Stackfor > 2000			
Conference SAP SAP	Microsoft @pentaho alteryx	amazon mechnicistura di HIVE 11 Weights & Biases TOCTON @ HORMONIA	Consistence: Uting Wildo Bacconer Barron Ause Forter sit risided (redzo' & Datavisor Spiratop			
CRACLE TRM III		Anacon Separate BUTY	ATTIGENTY Creek Dig TESSAR PROT VIDADO PROStation Hotelato (Skepcin ANOMALI & Rounded Rame Contract Contraction Co			
Carried ORACLE ImongeDB Areas Protes Oracle Oracle	MODE Datameer 😅 🖓 anod®t -outlier	Stole opwork support at twatter of cooge coose of the operation of the ope				
Marking Couthage announced and announced		KOwey Lighternoe labebax Orideter Coverned in RASOO StreetSO. Verta Annual Internation	LICENSE DATA			
	intervence intervence intervence	ALETTER & # 2000 Summer Ale seens @ FEAST BEERS Scortex	butkeeper Dimmer Allenke in Dathake Orienter Allenke in Dathake			
Toricash SCYLLA.	Incorta. GUAVUS CASERUID @facet macheye>	SS SELDON	Californital Provent Provent Control Control Sciences Sciences Sciences Control Control Sciences			
Arange B - CRAYE.10 PropCAP @ reparterion	switchboard obvieusly.al 🚧 Airtable cliff.gi	A subsection of the second of				
The bringed		Summer Contractor	APPLICATIONS - INDUSTRY			
MPP DBs ETL/ELT/ DATA INTEGRATION DATA GOVERNANCE DATA TRANSFORMATION & & ACCESS	AND DISCOVERY STORE	COMPUTER VISION SPEECH NLP SYNTHETIC MEDIA	ADVERTISINGEDUCATION _ REAL ESTATE _ GOV'T & COMMERCE FINANCE - LENDING INSURANCE			
TERADATA Middle talend alterux Census 20 tous troom (2 Segment (2 Information (2 Inform	M metaphor / splunk>	Microsoft Assee 2 Vision Art Si Camacon Sico	exandr MedaMath Busishoo REDFIN ORionic Overant SHEIN LALUE affirm Monedo ROOT Instromula			
VERTICA	atlan Asimpionis		Q 建築研究 Opendoor 学行mine Carbon Cortain Cortai			
antine and a state of the second state of the	Trace	Contanta darifai #BBRA Contanta de ma	Criteou, IASantana SKNEWTON Orchard Orchard Goodborgo Userski al. LASBANC Shift Tochnoogy			
terence game () periodo og teosota to or () and our contraction () of traylo splenty () formio () and the set	Supergrain solarwinds Hobono	O deepomotic @ market	Constant Con			
Cotion de man Brown No Pavela Preserver	Sternina Slogz.lo Slogdna	V RE VITU trat On Sendeund inc. O Mobiel Direct Sendetund				
Rognitio & more war & tam @ Avenue O Participation of the second states tates of the second states of the second s	[+] SELECT STAR • transform	Consider ground to the semantic reduces Un overeal supertone Control (Merceller Street)				
Excosol Miletano @senter NiRivery APourtounto Inforvorios securptor Statile ACC diservation	Secoda Castor ZQL OBSERVE	S Chances A DE - Connection Project S				
PRIVACY & DATA OBSERVABILITY MGMT / MONITORING SERVER CLUSTER SVCS	QUERY SEARCH	HORIZONTAL AI GPU DBS & CLOUD AI HARDWARE	HEALTHCARE LIFE SCIENCES TRANSPORTATION AGRICULTURE , INDUSTRIAL OTHER			
SECURITY & distance memory cance and an and the second and the second se		kiostico Google TPU grm (ma)				
	Enotes visition and and another and another and another and another and another anothe	WEM Watson @ OpenAI O DeepMind				
PRIVAGERA MARTA COMBRA (FINAMICS Orubrik Originations)		O Voyagerus opposisonie Vvicorious O Paperspace O Pa	atematical and a second and as			
Solarwinds www.man actrico	Starburst					
Same Strangered DATA QUALITY >Cribl & Maggatt Chronosphere	Anana Lucidworks Swiftype ATTIV/O		Disan almoscor (mith) samplem freename almost files and free files			
skyflow Duty talend SODA: @ Boove During peordulty	A Excercision illi alphasense omnicus	-CURRUSAL CIER VIEW OSARO :) Affectiva	At an and a state of the state			
O Phacy Dynamics @ suresconductive @ cellers	Cuer and the second of the sec	PG-Stom configure	ezha Gusten Heaten III x x o z x Zzebra Stadet Official Constru			
Katch Avenue peckety ways / with Grafana Lates Opticarity Conversion Care Separation		4Paradigm Printeen Mindel Palatone FLOYDHIES	Other Environment All Learning Institution Description Description <t< td=""></t<>			

	OFEN SOURCE			
FRAMEWORKS FORMAT QUERY / DATA FLOW DATA DATABASES	ORCHESTRATION INFRA DATA OPS STREAMING & STAT TOOLS & ML OPS & INFRA AI / MACHINE LEARNING / DEEP LEARNING	SEARCH LOGGING & MONITORING	VISUALIZATION COLLABORATIO	IN J SECURITY
ACCESS Standing and Standing S	Re readvolue Prozence Structure Caffe	Automatik Sol/" Ekizens Ekizens)) OO Sayerset matp(\Sbb	Apache Hanger KNOX
Caribost Oradost Oradost Oradost Oradost Oradost Oradost Oradost Oradost Oradost	Seattree W Flyte States States and States an	Lucana. #Sphinx Logitash Orandess Yant	Information Start Supervised	Run MONTAN
🔝 🖗 Halos 🐠 door GCDAP 🔍 👘 👘 👘 👘 👘 👘 Galaxie 👔 🖓 magdo. ZBoOB 😥 💁 💁 🛶 🎰 crick		Contract Vision		
Andrea house HELX	Lifer Lange Lange Antonio Status 2 Market Stat	tant 📶 📌	Osesborn bakeh	Ø snyk
		a state and a state of the stat	· · · · · · · · · · · · · · · · · · ·	

	DATA	SOURCES & APIs		 	DATA RESOURCES	
A A DISCOVERY BIOMOGENE D DOW JONES Querd Company A	MIR / SPACE / SEA	PEOPLE / ENTITIES		DATA SERVICES	INCUBATORS & SCHOOLS	RESEARCH
Concerner Concer		Versioner Versioner	Pacelle @esri A Radar Ampliary	kaggle ElectrifAl fractales) Exc.	PLURALSIGHT GENERAL ASSEMBLY DataCamp A DataElite galvanize The Data Decement	

Version 3.0 - November 2021

© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)

mattturck.com/data2021

FIRSTMARK

Mainstream Big Data models

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

 MapReduce family: simpler, more constrained

АРАСНЕ



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



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?