

Course Introduction

Big Data Engineering in the Cloud

March 2019



© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

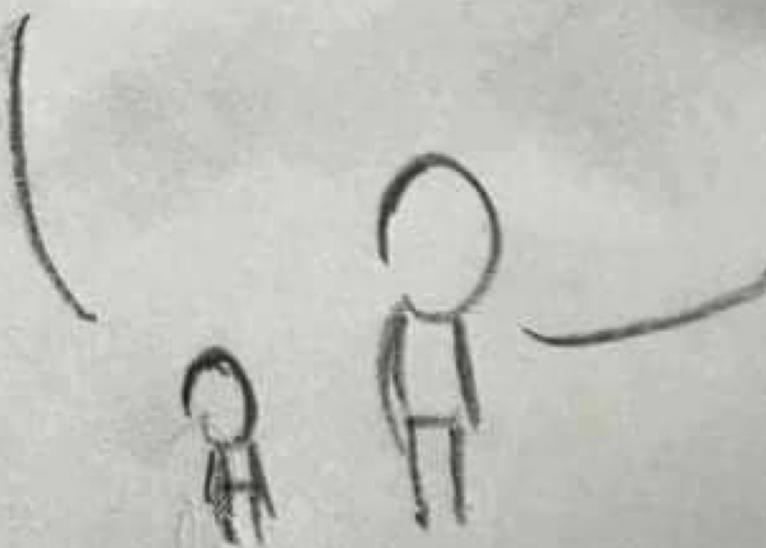
Introduction

- Aims
- Pre-requisites
- Contents
- Objectives
- Resources
- Rules of Engagement
- Introductions



© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

DADDY, WHAT ARE
CLOUDS MADE OF?



LINUX SERVERS,
MOSTLY



© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

Big Data learning objectives

- Principles
- Theoretical background and origins
- Practical experience of modern big data processing systems technologies
- Architecture and design
- Wider context



© Paul Fremantle 2015. This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

Pre-requisites

Covered by the Pre-Study Guide

- **Command line tooling and Unix commands**
- **Some Python programming and text editors**



© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

Format

- A mixture of lectures and practical labs
- Lectures aim to provide the wider context and background
 - Independent of specific technologies
- Labs are based on specific technologies
 - Designed to demonstrate the principles



© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

Lab model

- Local Virtual Machine
 - Ubuntu
 - Pre-installed big data software
 - E.g. Apache Spark, Cassandra, Python
- Amazon Web Services
 - Virtual machines in the cloud



Contents

- Big Data motivation and overview
- Using Python for Data Analysis
- Map Reduce and Directed Acyclic Graphs
- Apache Spark
- Spark and SQL
- Theory of scaling
- Running Spark on Amazon
- Introduction to NoSQL databases
- Introduction to Machine Learning



© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

Practicals

- Python Data Analysis
- Spark, SparkSQL
- Spark on Amazon
- Cassandra and NoSQL
- Machine Learning libraries
- Visualisation



© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

Improve your CV?



Leverage the NoSQL boom



© Paul Fremantle 2
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

Beyond the scope of this course

- Detailed Data Science techniques
- Understanding **all** of Hadoop, Spark, HDFS, Machine Learning



© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

Rules of Engagement

- ***Ask questions as we go along***
 - We will “park” any that are better answered later
 - Don’t wait till the end to ask or raise concerns
 - If you don’t ask we can’t help you



There ~~might~~ will be bugs!



- Please help out:
 - Please create new issues on the Github repository
 - <https://github.com/julieweeds/BigData/issues/>



© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

Julie Weeds



© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

Simon Wibberley



© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

You?



© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

Approximate Schedule

- Monday
 - Introductions
 - Overview and Motivation
 - Data Analysis with Python and Pandas
 - Map Reduce
 - Apache Spark
- Tuesday
 - SQL
 - Theoretical background on scaling systems
 - Scaling Spark on AWS
 - Visualisation
- Wednesday
 - Introduction to Machine Learning
 - Realtime systems
 - Architecting big data systems
 - Completion of labs



Let's get started



© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>