CLO - Cloud Computing and Big Data: Pre-study

Paul Fremantle, paul@fremantle.org, June 2020

Introduction

The ability to control servers remotely and virtualization technology has led to the creation of cloud computing. In turn, the availability of multiple low-cost servers has dramatically shifted the way in which data is processed. The result is an explosion of tools, technologies and approaches for handling very large datasets using cloud technologies.

In this course, we will study a range of cloud computing and big data technologies. We will primarily look at the Amazon EC2 cloud infrastructure and Apache Spark data processing model, but we will also explore other systems.

The learning objectives of this course are that you should be able to create systems that process large quantities of data in the cloud, including datasets that are too large to satisfactorily manage on any single system.

Remote Teaching and Learning Logistics

Since this course will be held remotely, you will need to use your own system to run the lab exercises. I will be providing a Virtual Machine that can be run on your laptop or desktop computer.

You MUST have a minimum of 8Gb and a modern multi-core processor to run the VM. 16Gb would be better. I will be sending out instructions on how to install the VM and test it at least one week before the class. My preference is that you use **VirtualBox** (https://www.virtualbox.org/) to run the VM although if anyone has an issue with that I can provide a VMWare image as well.

We will be conducting the class via Zoom and Slack (Zoom for video / audio conferencing, screen sharing; and Slack for chat and messaging). The instructions on how to sign into these systems will also be sent during the week before the class.

Pre study

The textbook ***Big Data: Principles and best practices of scalable real-time data systems*** is a strong proponent of an approach called the Lambda Architecture, which was pioneered by Nathan Marz, the author. Although the Lambda Architecture is an important part of this course, we will not be focusing on it as much as the book does! In addition we will not be using all of the projects that the book does. However, it is one of the best currently available books on the subject, and I enjoy the technical level of detail. **Please read at least Chapter 1 of the book.**

The **practicals** will be using the Linux / Unix shell command line under Ubuntu. We will mainly be using **Python 3.8**.

If you have never programmed in Python, please look at the following tutorial:

https://docs.python.org/3.8/tutorial/ (especially sections 1-5)

If you have never used a Linux Command Shell, please read:
http://linuxcommand.org/lc3_learning_the_shell.php

Thanks, Paul