**CLO - Cloud and Big Data Module**

***Pre-study***

*Paul Fremantle,* [*paul@fremantle.org*](mailto:paul@fremantle.org)*, November 2015*

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 Hadoop 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.

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.

**Please read at least Chapter 1 of the book.**

The book does not focus on the Cloud aspects of this course. If you have not used Amazon EC2 before, then I suggest you might wish to sign up for a free account and try this first exercise:

<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EC2_GetStarted.html>

However, this will be covered in the Labs if you don’t get a chance to do it.

The **practicals** will be using the Linux / Unix shell command line, as well as Python 2.7 and some Java.

If you have never used a Linux Command Shell, please read:

<http://linuxcommand.org/lc3_learning_the_shell.php>

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

<https://docs.python.org/2/tutorial/> (especially sections 1-5)

Finally, this will be the very first running of this course, which means that there will inevitably be issues with both the slides and the lab exercises. However, I have often found that the students on the first run of a course can often learn even more because they have to work harder and dig a little deeper to understand the material and make things work!

Thanks

Paul