UNIVERSITY OF OXFORD
SOFTWARE ENGINEERING PROGRAMME

Wolfson Building, Parks Road, Oxford OX1 3QD, UK
Tel +44(0)1865 283525 Fax +44(0)1865 283531
info@softeng.ox.ac.uk www.softeng.ox.ac.uk

Part-time postgraduate study in software engineering

Cloud Computing and Big Data, CLO

19th – 23rd September 2016

ASSIGNMENT

The purpose of this assignment is to test the extent to which you have achieved the learning objectives of the course. As such, your answer must be substantially your own original work. Where material has been quoted, reproduced, or co-authored, you should take care to identify the extent of that material, and the source or co-author.

Your answers to the questions on this assignment should be submitted using the Software Engineering Programme website — `www.softeng.ox.ac.uk` — following the submission guidelines. The deadline for submission is 12 noon on Tuesday, 8th November 2016. If you have not already returned a signed assignment acceptance form, you must do so before the deadline, or your work may not be considered. If you cannot meet the deadline and intend to request a later assignment, you must formally withdraw from submission, before the deadline.

We hope to have preliminary results and comments available during the week commencing Monday, 19th December 2016. The final results and comments will be available after the next examiners' meeting, during the week commencing *a date to be confirmed*.

ANY QUERIES OR REQUESTS FOR CLARIFICATION REGARDING THIS ASSIGNMENT OR PROBLEMS INSTALLING SOFTWARE SHOULD, IN THE FIRST INSTANCE, BE DIRECTED TO THE PROGRAMME OFFICE WITHIN THE NEXT TWO WEEKS.

CLO Module Assignment September 2016

Introduction
The assignment is designed to allow you to demonstrate your knowledge of Cloud Computing and Big Data systems, processes and approaches.

You must show a good understanding of Big Data methodologies, including the ability to design Big Data systems in the cloud. You must also show the ability to create applications and systems that can process Big Data.

Assessment objectives
This assignment is being assessed. Like other modules, you will pass or fail dependent on demonstrating certain things. In this case the main criteria for passing is that you understand and can apply cloud and big data concepts, principles and approaches for reasonably complex systems.

You must address issues such as scalability, efficiency and maintainability. You must also address the issue of storing data effectively, as well as processing both real-time and batch data.

Domain
The UK government makes available all the prices paid for UK properties. We will be working with a dataset from the Land Registry known as the Price Paid data. We have data from 1995-2013.

There are three parts to this challenge:
1) creating or using a big data infrastructure in the cloud to process this data
2) calculating summarized statistics about the data
3) designing (but not implementing) a system to handle on-going processing of data

Part 1. Designing a big data analysis system to be deployed in the cloud

In order to successfully pass this assignment you need to process a large amount of data about property prices from the UK collected from 1995-2012.

In order to complete this assignment you need to calculate a number of statistics and summaries from this data. The list is available in part 2.

Please read part 2 carefully before proceeding with part 1.

Your first task is to outline an approach to processing this data in the cloud. The data should be analysed in batch mode.

You should provide a clear outline of your preferred design for solving this:

- Which big data framework or algorithms are you going to choose?
- Which cloud infrastructure?
- How are you going to scale this?
- How are you going to efficiently process the data?
- Which languages are you going to use to process the data?

For each design decision that you make, outline the reasons for this decision and why you took that approach. Please draw an architecture diagram showing the resulting system.

Words: expected 1000, max 1500.

Part 2. Data analytics
The data that is provided for the assignment is price paid data for property in the UK, including the postcode, address and date of each transaction.

The data format is a CSV file, and there are two different versions of the data.
The first is a subset of data, and the second file is the full dataset.
The data is copyrighted by the Land Registry and is made available under the Open Government License.

It is recommended that you use the partial data set to test your system. Once the system is working you **may choose** to then run on the full dataset.

The data has a first line header with column names. There is further information about the data here:
https://www.gov.uk/guidance/about-the-price-paid-data

Partial Data (~100Mb compressed)
https://s3-eu-west-1.amazonaws.com/oxclo/landreg-pp-1995-2012-sampled.zip

Complete Data (~1Gb compressed)
https://s3-eu-west-1.amazonaws.com/oxclo/landreg-pp-1995-2012-full.zip

In addition, you will also need a mapping of postcodes to longitude/latitude to answer some of the questions. This is also available in a CSV file with a header line.

Postcode locations (90Mb uncompressed)
https://s3-eu-west-1.amazonaws.com/oxclo/ukpostcodes.csv

For **either** the partial dataset **or** the complete PP dataset, calculate:

1. How many records are there in the dataset?
2. Some postcodes no longer exist (they get deprecated). How many records have a valid postcode as defined by the postcode location dataset? From now on, only use data with valid postcodes.
3. How many records have a valid date, postcode and price paid?
4. Find the minimum, maximum, and average price paid overall for each year.
5. Calculate a median price paid for each year. You may estimate this.
6. We define the postcode prefix as the first digits of the postcode up to the space.
   For each postcode prefix, calculate the difference between the minimum and maximum price paid for each year, and then show the 10 records with the largest difference for the year 2008.
7. Identify the 10 postcode prefixes that had the biggest increase in average price paid from 1995 to 2012.
8. Which month has on average the most transactions across all available years?
9. Using a k-means algorithm, create 20 geographic clusters of transactions for the year 2010, and plot these onto a map of the UK. Only use the number of transactions not the prices for this.
10. Assume that the centre of London is marked by the postcode W1J 7NT. Taking the year 2010, for each transaction identify the distance between the property and the centre of London. Use any form of regression analysis you like to identify what correlation there was in 2010 between prices paid and distance from London.

You must show the code that you wrote to calculate each of these measures and some of the log output from each interaction. *You must demonstrate that you are doing these calculations in a way that utilizes either multiple computers or multiple threads.*

No word limits, but please post a *summary* of your logs as well as your code in an appendix and explain what each part is.

You may run this on a cloud service or locally. You may choose to only process the partial data set. No further marks will be given for processing the complete data set, but you may wish to do so if you can afford the cloud costs or you have enough time on your local machine(s).

Part 3. Extension of this system to provide continuous analytics
Your client would like to turn this into a production system providing continuous analytics of this data, including comparison of each statistic with the previous years data. Assume that we get a daily feed of new events. The system needs to be able to calculate live data on a daily basis. In addition, your client would like to be able to analyse cross-correlations with other datasets. For example, correlating this data against the results of school and hospital ratings, crime statistics, rental prices, and other geographic data.

The client would like this system to run 100% in the cloud.

Do not implement this system. However, please design such a system. Provide an updated architecture diagram, as well as a description of the system.

Please address:

- How the data is ingested into the system and fed around?
- How current data is compared with historical data?
- How you would add new datasets and queries to the system?
- How a live dashboard can be built that displays data from the analytics to your client?
- How the system is deployed, maintained and monitored?
- How the system is scaled?
- How the system is secured?

Expected 1000 words. Max 1500.

Final thoughts

- You are not expected to completely implement a production system! A real-life solution is out-of-scope. *We will not install or test any code you write.*
- Clearly document any assumptions you make.
- You (and the examiner) must be confident that the there are no major flaws in the design and that it is implementable.
- Properly formatted references/a bibliography will be appreciated.

Derivative works
You must not directly copy any other work. If you do use any source materials, you must make sure that you make clear the source and the extent of any derivative material, and reference it clearly.

Overall Assessment Criteria
Assessment will be according to the following criteria:

- Have you understood the principles and design characteristics of a big-data architecture?
- Can you implement and design simple analytics to run big-data jobs? Can you design a system that scales in the cloud? Are you able to define and design continuous feed big-data analytics systems?
- Have you addressed high-availability, reliability and failover in your design?

- Do you understand the challenges, emerging work and tradeoffs between different approaches? In particular, can you articulate clearly why different big data and cloud technologies are better or worse for certain tasks?

Paul Fremantle, September 2016