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

UNIVERSITY OF OXFORD

Cloud Computing and Big Data, CLO

23rd – 27th November 2015

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 to:

Software Engineering Programme
Department of Computer Science
Wolfson Building
Parks Road
Oxford OX1 3QD

Alternatively, you may submit using the Software Engineering Programme website — `www.softeng.ox.ac.uk` — following the submission guidelines. The deadline for submission is 12 noon on Tuesday, 12th January 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. We hope to have preliminary results and comments available during the week commencing Monday, 22nd February 2016. The final results and comments will be available after the next examiners' meeting, during the week commencing Tuesday, 26th April 2016.

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

CLO Module Assignment November 2015

Introduction

The assignment is designed to allow you to demonstrate a 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 – what is the problem!?

Under a Freedom of Information request, we have access to several gigabytes of data relating to taxi journeys in New York City during 2013. The assignment will be about handling this data as well as designing a system to handle real-time taxi data.

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 real-time 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 very large amount of data about trips and fares for New York Taxis collected in 2013.

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?
- How can you monitor and maintain this system.

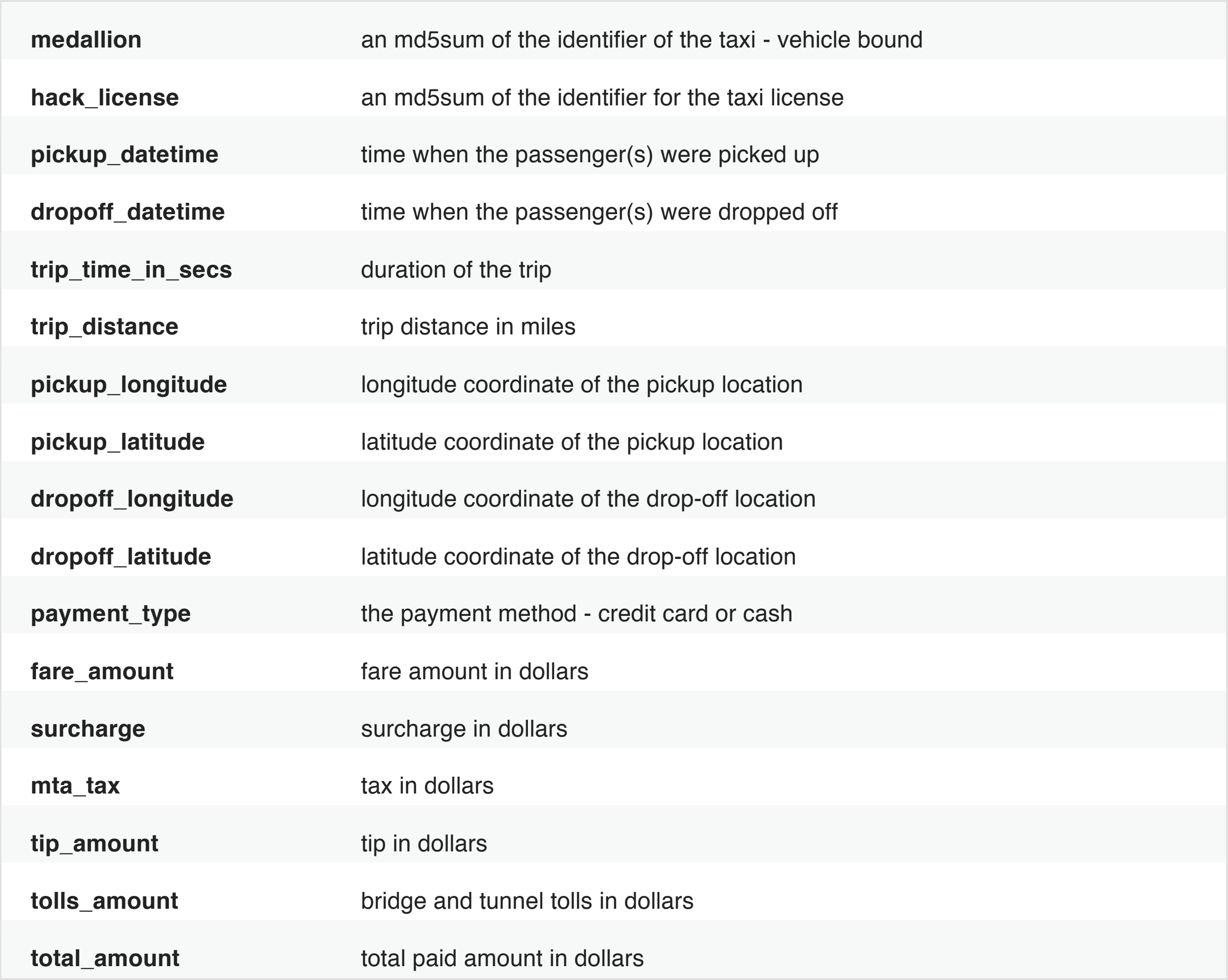
For each design decision that you make, outline the reasons for this decision and why you took that approach. Please draw an architecture diagram showing the resulting system.

Words: expected 1000-1250, max 1500.

Part 2. Data analytics

The data that is provided for the assignment is taxi data from New York City in 2013. The data includes the fares, distance, time, number of passengers and other interesting data.

The data format is a CSV file, and there are two different files.
The first is a subset of data, and the second file is the full dataset.
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 the following format:



medallion	an md5sum of the identifier of the taxi - vehicle bound
hack_license	an md5sum of the identifier for the taxi license
pickup_datetime	time when the passenger(s) were picked up
dropoff_datetime	time when the passenger(s) were dropped off
trip_time_in_secs	duration of the trip
trip_distance	trip distance in miles
pickup_longitude	longitude coordinate of the pickup location
pickup_latitude	latitude coordinate of the pickup location
dropoff_longitude	longitude coordinate of the drop-off location
dropoff_latitude	latitude coordinate of the drop-off location
payment_type	the payment method - credit card or cash
fare_amount	fare amount in dollars
surcharge	surcharge in dollars
mta_tax	tax in dollars
tip_amount	tip in dollars
tolls_amount	bridge and tunnel tolls in dollars
total_amount	total paid amount in dollars

Partial Data (126Mb compressed)

HTTP URL:
https://s3-eu-west-1.amazonaws.com/nyctaxidata-oxclo/partial/sorted_data.csv.gz
S3 URL:
s3n://nyctaxidata-oxclo/partial/sorted_data.csv.gz

Complete Data (11Gb compressed)

HTTP URL:
https://s3-eu-west-1.amazonaws.com/nyctaxidata-oxclo/complete/sorted_data_complete.csv.gz
S3 URL:
s3n://nyctaxidata-oxclo/complete/sorted_data_complete.csv.gz

You need to create big data analysis programs to calculate the following:

- Taking 30 minute periods throughout the day (from midnight to midnight):
    - What is the average speed of taxis during each period?
    - Which is the period where drivers in total earn the most money in terms of fares?
    - Which is the period where drivers in total earn the most in tips?
    - *Where a trip crosses a boundary (where the drop off is in a different period to the pickup), assign that trip to the period which it is in more. If the trip exactly straddles two periods then assign it to the earlier period. If a trip crosses more than two boundaries, assign it to the period where the midpoint of the journey happened.*
- Which is the vehicle (medallion) that earned the most revenue (fares+tips) in each day, month, and in the whole year?
    - Provide a chart of the highest takings by day.
    - Provide the actual data for 5th January, January and the year.
- Who is hardest working driver (license) of 2013? We define hardest working as the most distance driven in fares.
- Indicate which data set you used to produce this data.

You must show the code that you wrote to calculate each of these measures and some of the log output from each interaction.

How long did the calculations take and on how many concurrent processors or threads did you run the jobs?

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 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 real time.

The NYC Taxi regulator would like real-time analytics of this data, including comparison of each statistic with the previous years data. Of course some statistics (such as the hardest working driver of the year) are not important to be calculated in real-time. However, the system needs to be able to calculate live data, such as who drive the most distance in the last 1 hour, 4 hours, 24 hours (all rolling).

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 a live dashboard can be built that displays data from the analytics to the NYC taxi regulators.
- How the system is deployed and maintained.

Expected 1000-1250 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 *real-time* 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, November 2015